Iteratively improving Hi-C experiments one step at a time

Authors:

Rosela Golloshi, Jacob Sanders, and Rachel Patton McCord

Affiliation:

Department of Biochemistry & Cellular and Molecular Biology, M407 Walters Life Sciences Building, 1414 Cumberland Ave., University of Tennessee, Knoxville, TN 37996

Corresponding author contact: rmccord@utk.edu

Abstract: (250 words)

The 3D organization of eukaryotic chromosomes affects key processes such as gene expression, DNA replication, cell division, and response to DNA damage. The genome-wide chromosome conformation capture (Hi-C) approach can characterize the landscape of 3D genome organization by measuring interaction frequencies between all genomic regions. Hi-C protocol improvements and rapid advances in DNA sequencing power have made Hi-C useful to study diverse biological systems, not only to elucidate the role of 3D genome structure in proper cellular function, but also to characterize genomic rearrangements, assemble new genomes, and consider chromatin interactions as potential biomarkers for diseases. Yet, the Hi-C protocol is still complex and subject to variations at numerous steps that can affect the resulting data. Thus, there is still a need for better understanding and control of factors that contribute to Hi-C experiment success and data quality. Here, we evaluate recently proposed Hi-C protocol modifications as well as often overlooked variables in sample preparation and examine their effects on Hi-C data quality. We examine artifacts that can occur during Hi-C library preparation, including microhomology-based artificial template copying and chimera formation that can add noise to the downstream data. Exploring the mechanisms underlying Hi-C artifacts pinpoints steps that should be further optimized in the future. To improve the utility of Hi-C in characterizing the 3D genome of specialized populations of cells or small samples of primary tissue, we identify steps prone to DNA loss which should be considered to adapt Hi-C to lower cell numbers.

Keywords: Hi-C; chromosome conformation capture; 3D genome structure; high throughput sequencing

Highlights:

- **Variability in Hi-C libraries can arise from early steps of cell preparation**
- **Hi-C 2.0 changes also benefit libraries generated by 6-cutter enzymes**
- **Artificial molecule fusions can arise during end repair and PCR, increasing noise**
- **Common causes of Hi-C DNA loss identified for future optimization**

1. Introduction

The billions of DNA bases in the eukaryotic genome must be efficiently packed to fit inside the micron-sized nucleus and to perform necessary cellular functions including gene expression as well as DNA replication and repair [1-3]. Abnormal 3D packaging of chromatin may disrupt cellular homeostasis and lead to diseases [4, 5].

Many techniques have been devised to understand how chromatin is packed into the nucleus and how that packaging affects genome function. Light and electron microscopy led to observations of densely packed heterochromatin and open euchromatin as well as the small scale wrapping of DNA around nucleosomes [6, 7]. Fluorescence microscopy approaches such as fluorescence *in situ* hybridization (FISH) [8] and *in vivo* techniques such as bacterial operator arrays [9] and CRISPR imaging [10] have allowed visualization and tracking of the nuclear arrangements of specific genome loci. In recent years, the development of the chromosome conformation capture (3C) technique [11], and variations such as 4C [12, 13], 5C [14], and Hi-C [15], have accelerated the progress of the genome organization field and sparked the interest of researchers with numerous biological problems and across disciplines from genetics to physics, mathematics, and computer science.

As high throughput sequencing becomes increasingly powerful and inexpensive, many researchers have adopted Hi-C as a method to capture a snapshot of the 3D nuclear organization genome-wide. The ability to represent 3D genome organization in terms of matrices of contact frequencies allows the characterization of structures across a variety of length scales, from the positioning of whole chromosome territories to the folding of small-scale enhancer-promoter loops [3, 16, 17]. Even when more focused resolution is desired, many techniques begin with an approach similar to Hi-C and then isolate a subpopulation of interactions of interest. This is the case with approaches such as Capture-C [18, 19], Hi-ChIP [20], and T2C [21]. Further, it has recently become clear that Hi-C is a very useful tool not only for identifying 3D genome folding patterns, but also for *de novo* whole genome sequence assembly [22-24] and translocation detection [25]. As basic researchers from many fields, commercial ventures (such as Phase Genomics and Dovetail Genomics), and clinical researchers adopt Hi-C based techniques, it is more important than ever to optimize the methodological details of the Hi-C protocol and understand how they affect the resulting data and interpretation.

Several excellent discussions of Hi-C theory and optimizations of Hi-C protocols have been published in recent years [17, 26-28], and numerous other insights about key steps of the Hi-C protocol have been discovered while researchers optimize this technique in various systems. Yet, Hi-C is still a technique that often requires troubleshooting and can sometimes unexpectedly result in low quality datasets even when the best current protocols are followed. Rather than a comprehensive guide or update to every step of the Hi-C approach, here, we will describe our observations about certain key variables in the Hi-C protocol and the resulting data. We will discuss the effect of certain protocol modifications on Hi-C data quality and utility, propose explanations for previously unreported artifacts that can arise in Hi-C libraries, describe considerations for applying Hi-C to smaller populations of cells, and discuss key areas of uncertainty that will be ripe candidates for future careful optimization. To increase the clarity of our references to sequenced Hi-C interaction molecule types, we have included a terminology definition table (Table 1).

2. Early steps in the Hi-C protocol with downstream implications

2.1 Initial sample preparation: an overlooked variable

The key initial steps of Hi-C involve the covalent crosslinking of chromatin regions interacting in 3D space, followed by restriction enzyme digestion, marking digested ends with biotin, and proximity ligation of fragments from interacting regions. As we will discuss below, digestion, biotin fill-in, and ligation conditions certainly can have a large impact on Hi-C library quality, but another potential source of variation in Hi-C is the handling of the starting material before restriction enzyme digestion takes place. For cells in culture, variables such as exactly how the formaldehyde is added, how adherent cells are removed from culture dish, and how nuclei are isolated and permeabilized can have downstream implications for Hi-C library quality (Figure 1A). Even if optimized kits such as the Arima Hi-C approach (https://arimagenomics.com/products/) are employed, these early steps still need to be evaluated for new cell types and experimental conditions.

Higher reproducibility and quality of Hi-C datasets occurs when digestion, biotinylation, and ligation steps occur within a permeabilized nucleus that is not lysed into the solution. This type of approach has been variably called the "in situ"[17] or "in nucleus"[29] Hi-C protocol. Despite minor variations in exact protocols, the basic idea remains the same: Avoiding high concentrations of SDS, nuclei lysis, and dilution to large volumes during digestion and ligation increases Hi-C reproducibility and decreases the capture of random background interactions [17, 29]. The generally accepted explanation for this improvement has been that intact nuclei constrain the movement and random collisions of crosslinked complexes, but other factors could be at work as well. Other work has suggested that high concentrations of SDS and subsequent Triton sequestration of SDS are more likely to lead to aggregates of material that reduce digestion, fill-in, and ligation efficiency [30].

The effect of nuclei lysis on the quality of Hi-C libraries supports the idea that the initial state of the biological material going into a Hi-C experiment can affect the likelihood of generating a high-quality Hi-C library. Here, we will discuss aspects of the initial Hi-C sample preparation that may affect whether nuclei stay intact, whether random chromatin damage is generated, and whether chromatin is efficiently digested and recovered from nuclei. Unfortunately, even when a theoretically consistent protocol is used, there can be variation in what actually happens to a given cell type during the initial steps of Hi-C. Such variations may lead to downstream effects such as a low ratio of cis to trans interactions (which may reflect high background noise) and a high percentage of dangling ends (sequenced fragments that do not contain chimeric interactions; see Table 1), reducing the fraction of valid interaction pairs among sequenced reads. In general, for the areas described below, we advocate a thorough documentation of the methods used and their effects on the appearance of nuclei and chromatin. By better linking such metadata to downstream Hi-C library results, the field can gain a better understanding of how to standardize or optimize these steps of the protocol.

2.1.1 Formaldehyde Crosslinking

The effects of variations in formaldehyde concentration and crosslinking time on Hi-C results have been considered previously [31, 32]. It has also been noted that the presence of serum in the cell culture media can inhibit effective crosslinking [33]. Other work has suggested that the preparation of fresh formaldehyde immediately before each experiment, to prevent uncontrolled formaldehyde polymerization, is helpful to increase reproducibility [30]. In addition to these previous observations, we have noticed that changes in cell morphology can occur depending on how formaldehyde is added (Figure 1B). When concentrated formaldehyde (37%) is added directly to adherent cells, an instantaneous and irreversible shrinkage of the cells nearest to the site of formaldehyde addition is

observed (Figure 1B). This effect of adding concentrated formaldehyde directly into the sample is even more noticeable for isolated nuclei immobilized on a poly-D-lysine coated dish (Figure 1B). (Note that isolated nuclei are often used for Hi-C experiments on cells with thick cell walls or from complex tissues [11, 34]). Evident, though less dramatic, alterations in nucleus appearance were also observed when formaldehyde pre-diluted to 1% in water was added to nuclei already incubated in water. This result is a cautionary note for experiments in which cells need to be in an unusual buffer related to the treatment of interest. The best preservation of pre-existing nuclear appearance was achieved by replacing cell culture media with a pre-mixed 1% solution of formaldehyde in either serum-free cell culture media, or other buffered solutions (PBS, HBSS, HEPES). There is no clear evidence at this time about how changes in nuclear appearance relate to the chromosomal interactions detected by Hi-C, but changes in appearance after localized addition of highly concentrated formaldehyde suggest that certain cells are experiencing a much higher effective concentration of formaldehyde than others. This would likely lead to heterogeneity of crosslinked complex size across the cell population and to some cells having crosslinked networks impenetrable to digestion enzymes [35]. To avoid this, we recommend that formaldehyde be pre-diluted in an appropriate buffer before addition to the cells destined for Hi-C.

2.1.2 Cell Harvest

The harvest of crosslinked cells is a common step to many genomic experiments (including ChIP-Seq and others), but there are still considerations in this step that are often unexplored. We have observed that different cells types vary greatly in how easily they can be recovered by manual scraping from culture dishes after crosslinking. Differentiated myotubes can require trypsinization during scraping after crosslinking to aid release from the dish [36] while cells like HepG2 detach very easily (unpublished observations). Consequently, different cellular preparations may yield variable results: aggregates of cells that are later more difficult to permeabilize, physically damaged cells, or loss of cellular material. One option for increasing the reproducibility and recovery of material is to gently enzymatically detach cells from the plate using, for example, Accutase, before crosslinking and then to perform crosslinking steps in a single cell suspension. Our results indicate that this approach can lead to better estimation of cell numbers going into the experiment and fewer cellular aggregates that may be hard to permeabilize. Visualizing DNA recovered from cells that were harvested with Accutase before crosslinking was easier than cells crosslinked and scraped, but this in part reflects the fact that DNA from these cells was more fragmented, while DNA from scraped cells was so intact that it became sticky and floated out of the gel (Figure 1C). Also, since cellular attachments to different substrates can affect nuclear structure [37], cells may not retain their native 3D genome conformation after being detached. So, crosslinking adherent cells in their native state is often preferable, but we recommend evaluating crosslinked plates under the microscope during scraping to document the amount of scraping needed to retrieve the cells and the appearance of the cells after harvest.

2.1.3 Nuclei Isolation / Permeabilization

Many Hi-C protocols recommend that nuclei be released by dounce homogenization before restriction digestion begins. But, like cell scraping, the results of this technique vary with douncing speed and cell type. Indeed, we have observed that different cell types can look quite different after the same nuclei isolation steps. While GM12878 nuclei are cleanly isolated after incubation in a detergent-containing buffer and 15 strokes of the dounce homogenizer, mouse melanoma cells still retain a large amount of debris around their nuclei after the same treatment (Figure 1D). If similar variation in cell lysis across cell

types occurs during early Hi-C steps, this could result in nuclei that are too inaccessible or excessively ruptured going into the later steps of the Hi-C protocol. Relatedly, changes in buffer conditions can cause dramatic effects on isolated nuclei, again, depending on cell type. We have observed that upon addition of 1 drop of pure water to isolated nuclei from GM12878 cells, the nuclei can rupture (Figure 1E), while a HeLa epithelial-type cancer cell nucleus is minimally affected by the same treatment. While it would be unusual for a researcher to place cells or nuclei into pure water, other buffer components more likely to be overlooked (like the addition of EDTA [38]) could cause a similar effect. In practice, resulting variations are often noted in the visible properties of Hi-C nuclei before the digestion step: some nuclei settle quickly or are more clumpy, granular, or sticky. But such differences are rarely reported or considered. So that the field can develop a more complete understanding of what upstream steps lead to these differences, and how they affect the quality of the resulting Hi-C library, we recommend documenting microscopic images of cells and nuclei to link their appearance before and after crosslinking, and before digestion to their downstream Hi-C success.

2.1.4 DNA damage affects Hi-C library quality

It is possible that variations in the early physical harvesting steps described above may impact the initial integrity of chromatin. In both long read Sanger sequencing of isolated Hi-C library molecules and high throughput sequencing results, we have observed that different replicates and cell types can have varying proportions of breaks in the DNA at sites that do not correspond to restriction enzyme sequences (discussed in detail as “random breaks” in Imakaev et al. [39]). While the portion of these random breaks that end up forming ligated Hi-C interaction products likely result from off target restriction enzyme activity, many of these non-canonical breaks correspond to unligated but biotinylated molecules that end up being uninformative reads (“dangling ends”) at the end of the Hi-C experiment, reducing the number of interaction pairs recovered. In particular, these non-canonical dangling ends also seem to be recalcitrant to attempts to remove biotin at fragment ends (Figure 1F) and may therefore arise from biotin incorporation at nicks in the middle of fragments. These may arise from factors in the experiment other than off-target restriction enzyme activity. For example, if the nuclei rupture during cell harvesting or nuclei isolation, this may expose the chromatin to DNA damaging agents or physically damage the DNA. Notably, the example in Figure 1F of extensive random breaks in the final Hi-C library was the end product of the “Accutase” sample shown in Figure 1C, in which the DNA appeared to be a bit more fragmented after initial harvesting. However, a conventional 0.8% gel is not the best approach for visualizing such large DNA fragments. In the future, pulsed-field gel electrophoresis (PFGE) [40] could greatly aid the investigation of what initial conditions lead to the recovery of the most intact DNA.

2.1.5 Future optimization of cell preparation needed

Poorly controlled variations in the steps described above can have a negative impact on the downstream quality of the Hi-C library. Insufficient cell recovery during cell harvest steps can lead to low amounts of DNA and low library complexity. Incomplete nucleus isolation or permeabilization may affect the efficiency of downstream steps such as digestion, biotin fill-in, and ligation, in which enzymes must access the chromatin. As previously described, DNA recovery and digestion efficiency can be checked on a gel before subsequent steps, though this check would likely be more efficient using PFGE [26, 33]. With data from a variety of Hi-C experiments, we have observed that when samples from otherwise similar experiments have varying degrees of DNA digestion, those with less well-digested DNA have a

smaller proportion of valid pair Hi-C interaction reads, higher fractions of dangling ends, and higher levels of background ligation noise at the end of the experiment (Appendix A).

All the variables discussed above apply in particular to Hi-C being performed on cells in culture. Controlling the consistent initial preparation of crosslinked material before digestion is even more challenging when whole primary tissues are the source of biological material. Numerous upstream approaches to prepare diverse biological materials have been developed [22, 41-43], but many of these require similar downstream steps discussed above. To increase the success of the complex preparations of diverse biological materials for future Hi-C experiments, future work should explore protocols that reduce variability in ambiguous steps such as scraping and douncing. There is also a great need for more informative quality control tests of initial Hi-C steps that will allow researchers to assay whether a given preparation of material is suitable (good chromatin accessibility and material recovery with low DNA damage and nuclei lysis) for the expensive downstream steps of digestion, biotin fill-in, and ligation, not to mention library preparation and sequencing.

Currently, Hi-C data gathered by the 4D Nucleome Consortium [44] is archived along with a variety of metadata about the experiment (https://data.4dnucleome.org/). We encourage researchers to provide such details about both intermediate phases and final library statistics about Hi-C experiments, including cell appearance at various stages described above as well as DNA integrity and digestion efficiency information. To correlate these early checkpoints with final Hi-C results, we encourage all researchers to report statistics informative about the quality of libraries including the percentage of dangling ends, the cis/trans interaction ratio, the percent of reads that are PCR duplicates, and the percentage of reads that represent mid-range distance (25 kb – 10 Mb) interactions (poor quality libraries are often enriched for both extremely proximal and extremely distant interactions).

2.2 Utility of Hi-C 2.0 Biotin Fill-in conditions for HindIII libraries

Recently, Belaghzal et al. have developed an optimized in nucleus Hi-C protocol called Hi-C 2.0 which uses the DpnII enzyme and is designed for single aliquots of 5 million cells [26]. Some changes in the protocol from previously published approaches [17, 33] might seem to be minor or just specific to the use of biotin-dATP and DpnII. Here, we have compared an in nucleus version of the Belton et al. Hi-C protocol (referred to as Hi-C 1.0)[33] with the modifications of the Hi-C 2.0 method. We find that the technical variations in Hi-C 2.0 notably improve HindIII Hi-C library quality in a direct comparison using aliquots of 5 million mouse melanoma cells (B16-F1) identically prepared in initial steps (see Appendix B for protocols used). In particular, the modified Hi-C 2.0 protocol involves performing biotin-dCTP fill-in at 23°C for 4 hours instead of 37°C for 2 hours. When applied to a HindIII Hi-C library preparation, these modifications resulted in a nearly 10% increase in valid interaction pairs due to a decrease in unligated dangling ends (Figure 2A). Notably, the decrease in dangling end molecules corresponded to a reduction of the bias toward inward facing reads among valid pairs. This indicates that a fraction of the excess inward facing interaction pairs are actually undigested dangling ends, which are then also reduced when dangling ends are removed.

Dangling ends are reduced even further when DpnII with biotin-dATP is used instead of HindIII in the Hi-C 2.0 protocol (Figure 2A). However, an increase in interchromosomal ("% trans") interactions is observed, which could indicate higher background noise in this condition. This increase in the fraction of interchromosomal interactions when the cells were derived from the same initial preparation steps shows that nuclear lysis (which would be the same in each of these aliquots) is not the only factor that

can contribute to higher levels of potential background ligation. Overall, all of these approaches lead to reproducible final Hi-C libraries according to concordance scores calculated by GenomeDISCO [45] and visual comparisons of the resulting heatmaps at 40 kb resolution (Figure 2B and C). TAD boundary locations and strengths at 40 kb resolution are also overall similar according to an insulation score profile (Figure 2D) [43]. Thus, the major gain in these particular optimizations will lie primarily in the fraction of useful reads gained from a certain number of cells, allowing the generation of high resolution heatmaps from smaller numbers of total sequenced reads.

3. Approaches and Artifacts during Hi-C library preparation for sequencing

After the "3C-like" phase of the Hi-C experiment is complete (digestion, biotin fill-in, and ligation), the Hi-C DNA follows some steps that are common to nearly all high throughput sequencing experiments but with variations and factors unique to Hi-C that require special consideration. Even with a high quality library at the beginning of this stage, variations in library preparation can matter, and problems with low quality libraries become exacerbated during preparations for sequencing.

3.1 Reproducible NEBNext library preparation of captured interactions

Like some previously published protocols [17], we find it useful to perform streptavidin bead capture of biotinylated interactions immediately after size selection and before any end repair. We have adapted the NEBNext® Ultra™ II DNA Library Prep protocol for use on beads, resulting in lower reaction volumes, fewer steps, and reduced time than some previous library preparation approaches [33] (Appendix B). In our hands, this protocol is easier and more reproducible specifically in cases of indexed adaptor ligation than adapting the Illumina protocol to Hi-C, possibly due to the divergent design of NEB adaptors. We find that substantial dilution of adaptors (at least 1:15) is necessary to avoid excess adaptors in the final Hi-C library. In the end, the final Hi-C library statistics are very similar when the same ligated Hi-C material is used for an Illumina library preparation with streptavidin bead capture immediately before adaptor ligation as compared to an NEBNext library preparation (Figure 2E) NEBNext library preparation in fact resulted in a slightly higher percentage of sequencing reads for which both paired reads align to the genome.

3.2 Disadvantages of over-sonication to short DNA fragment length

Given that ligated interacting DNA molecules are often multiple kilobases in length, fragmentation of DNA to smaller sizes for sequencing on the Illumina platform is essential. However, some evidence suggests that it is better to leave some larger DNA fragments rather than trying to achieve average sizes of 200 bp, as has sometimes been recommended [26, 33]. Smaller fragments reduce the overall mappability of the Hi-C library after sequencing, in part because shorter sequences are less likely to be uniquely mappable, particularly in repetitive regions (Figure 3A). For short fragments, it is more likely that one of the two interacting fragments will be too short to be aligned uniquely to the genome before a ligation junction is encountered (Figure 3A). This phenomenon results in a higher proportion of "single side mapped" reads. Also, evidence further discussed below suggests that shorter molecules are more likely to contain end-repair related artifacts.

3.3 Dangers of chimera formation and end repair based artifacts

In all high throughput genome-wide sequencing approaches, the complexity of DNA sequences in the same tube raises the possibility that small regions of homology between molecules could hybridize and form chimeras during PCR amplification. However, in techniques other than Hi-C, chimeric reads can easily be excluded due to their impossible genomic positions (paired reads from different chromosomes, reads from the same DNA strand, etc.). In Hi-C, by contrast, informative paired sequence reads should come from all orientations and locations in the genome, making these possible artifacts harder to detect or exclude.

We have found evidence of one particular type of artifact that we propose occurs due to complementary base pairing during end repair. In the paired end sequencing results of some Hi-C experiments, we found an unusually low rate of mapping for Read 2 (81% mappability for Read 1 vs. 45% for Read 2). Strangely, we discovered 10-15 bases of the 5' end of the library molecule had been copied artificially onto the 3' end, resulting in 10-20 identical bases at the beginning of paired end Read 1 and Read 2. Thus, Read 2 would not map unless the first bases were excluded (Figure 3B and C). Further investigation revealed that molecules with this artifact had 4-6 bp of microhomology within the mappable portions of Read 1 and Read 2 (Figure 3A and D). This indicates that the original ligated Hi-C fragments contained a complementary sequence between the 5' and 3' ends that existed before the copying artifact occurred (Appendix C). This artifact is most common in libraries that have undergone extensive biotin removal steps and sonication to small fragment sizes (Figure 3E), conditions that would tend to expose single stranded overhangs. We proposed the following explanation for this phenomenon, based on the evidence we have collected: during the room temperature end repair step, short complementary single stranded 5' and 3' ends of the Hi-C molecule could anneal to each other, and then the T4 DNA polymerase would be able to elongate the DNA from this "priming", copying the 5' sequence (Read 1 when sequenced) onto the 3' end of the molecule (which becomes Read 2) (Figure 3F). In libraries in which this phenomenon occurs, a percentage of Read 2 sequences will not align to the genome, but instead contain identical sequences at the beginning of Read 1 and Read 2 that span a smooth distribution of lengths between 7-25 base pairs. In some libraries, identical sequences between Read 1 and 2 of exactly 13 bp are observed (Figure 3C and Appendix C), but this can be explained by the sequencing of adaptor sequence at both ends of the molecule, which can occur in adaptor dimers or if more than one adaptor is ligated to the same library molecule end. In cases not explained by adaptor dimers, no enriched sequence motifs are noted in the copied or complementary DNA pieces, but these sequences are AT-rich (Appendix C), which would correspond to greater fragility [46] and likelihood of ssDNA ends.

Stable base pairing resulting from a few complementary bases may sound unbelievable, but, interestingly, a very similar mechanism is known to happen *in vivo* during certain types of DNA repair [47]. If there are 3' single stranded overhangs on damaged DNA ends, microhomologies of 1-4 bp can anneal and undergo fill-in synthesis by DNA polymerase [48, 49]. Microhomology annealing can be a small bias in non-homologous end joining repair or microhomology mediated repair pathways.

This artifact is minor in most Hi-C libraries but suggests that care must be taken not to over-sonicate the DNA or to overdo the biotin removal recessing of DNA ends. Future library preparation protocols could also consider variations to the end repair steps to avoid the possibility of low temperature priming and extension from small identical regions. Notably, we have not observed this artifact in Hi-C libraries prepared with tagmentation for adaptor incorporation [42] (Figure 3C) or in 5C libraries, in which no random fragmentation is performed and adaptors are added by PCR [50] (Appendix C). On the other

hand, consistent with the proposal that this artifact occurs during the end repair step common to many techniques other than Hi-C, we have observed the very same end copying in sequencing data from a ChIP-Seq library [51] (Appendix C).

Fortunately, the direct copying of sequence from Read 1 onto Read 2 is easily detectable, and the resulting sequence either does not map or is discarded as two reads within the same fragment. But, chimeras that are more difficult to detect could form in a similar way either during the end repair or final PCR steps. As previously documented [33], concatenations of Hi-C library molecules cause an upward shift in overall Hi-C library molecular weight after too many rounds of PCR amplification. But, such chimeras may form before this shift is visibly detectable. Indeed, in libraries with more noisy random interchromosomal ("trans") ligations, a higher proportion of these interchromosomal interaction reads lack the canonical Hi-C junction sequence (Figure 3G). Such random interaction noise without true ligation junctions could arise from PCR chimeras formed in an attempt to amplify the few real interactions that exist in a poor library. This observation again indicates that increases in random background signal may not always stem from actual random ligation in solution due to nuclei lysis but also can be caused by other factors.

4. Considerations in sequencing and mapping

4.1 How much can be gained by deeper sequencing?: Issues of library complexity

The potential number of pairwise genomic interactions in a mammalian genome is on the order of 10^{11} – 10^{12}, so large numbers (hundreds of millions to a billion) of paired sequence reads are often needed to thoroughly sample this interaction space at high resolution. However, deeper sequencing will not lead to better Hi-C maps for all samples. The input number of cells and the efficiency of digestion and ligation limit the number of possible captured interactions. If digestion with a 6-cutter enzyme and interaction ligation were perfectly efficient, 5 million cells could indeed lead to a library of 10^{12} captured interactions. However, inefficiencies and losses occur at many steps, so that the interaction complexity in practice is often much more limited. Thus, while a highly complex library can provide better interaction maps with additional sequencing, the interaction map resulting from a low complexity library will barely change even when the number of reads is increased to 1 billion (Figure 4A). A high proportion (more than a few percent) of PCR duplicate molecules in a Hi-C library can also signal a lack of complexity. But, the increased prevalence of optical duplicates on patterned HiSeq 3000 and 4000 flowcells [52] adds uncertainty to this potential measure of complexity.

4.2 How finely can the interactions be binned?

4.2.1 Structural information derived from read location within restriction fragments

Interactions detected by sequencing Hi-C products are often mapped to the midpoint of each represented restriction enzyme, based on the idea that the crosslinked interaction could have occurred anywhere across the given fragment. We have found, however, that sometimes one end of the fragment is favored for an interaction over the other end, and that this bias can contain biologically significant information about the position of the specific interaction. For example, for interactions of between long fragments that are within 40 kb of each other along the linear genome, more ligations are observed

between the fragment ends that are closer together (Figure 4B). This shows that the natural decay of interaction frequency along the chromatin polymer is revealed in a restriction fragment end preference for sequence reads. We therefore argue that it is valid to bin sequence reads to constant bin sizes, taking into account exactly where each read lands rather than only which restriction fragment it belongs to. Employing this approach, however, absolutely requires that resulting interaction maps be carefully corrected for artifacts by a fragment-position naïve method such as iterative correction [39].

4.2.2 Comparing restriction fragment size distributions to a chosen bin size

When choosing a high-resolution bin size for Hi-C data generated with a given restriction enzyme, it is also important to consider the restriction fragment size distribution to limit the number of bins containing no fragment ends. Publications often quote the average or median restriction fragment size, but we note that this can be a misleading value. HindIII and DpnII fragments follow a non-normal distribution (Figure 4C) in the human genome, so the average or median size does not imply that most fragments are close to this size. In fact, while the average HindIII fragment size in the mouse genome is 3.2 kb, nearly 30% of fragments are less than 1 kb in size, so there are numerous interactions that can be detected at a higher resolution than the average fragment size. Conversely, the median DpnII fragment size in the mouse genome is 260 bp, but nearly 30% of DpnII fragments are larger than 500 bp, so using a bin size of 200-300 bp to match the median fragment size will result in numerous bins with no information.

5. Considerations for applying Hi-C to lower input cell numbers

Interpreting Hi-C data always requires the consideration that the data represents the average structure across a population of cells [27]. Single cell Hi-C approaches have made it possible to capture interactions that occur in individual cells [53], but these datasets are inherently limited in their information, since a maximum of 4 interactions (2 ends of each fragment on 2 alleles) can be captured from each genomic location in any given cell. Therefore, there is a need for continued development of protocols effective for small populations of cells, particularly as Hi-C is applied to more primary tissues or to capture genome structure during rare cellular events. We have shown earlier that results between replicate batches of 5 million cells are highly reproducible (Figure 2), but many applications will require lower cell numbers than this. We have noted several experimental considerations that will be important as Hi-C is adapted to lower numbers of cells.

5.1 Steps prone to substantial DNA loss

With small numbers of cells, recovery of all possible DNA is important at every step. We have found unexpected losses of DNA at certain steps of the Hi-C protocol, which would be unacceptable if working with low cell numbers.

5.1.1 Ethanol precipitation

Ethanol precipitation performed in large volumes as recommended by many Hi-C protocols can be inefficient. Indeed, we find that substantial amounts of Hi-C DNA may still be in the ethanol supernatant after the precipitation of ligated Hi-C DNA molecules. In two independent experiments, we have performed another round of DNA precipitation on the ethanol supernatant from the first DNA

purification attempt and have recovered microgram quantities of additional DNA (10-40% of the original amount recovered) (Figure 5A). This indicates that in the first round of ethanol precipitation, substantial quantities of DNA were lost. Indeed, as 4C experiments are adapted to lower cell numbers, smaller volume DNA precipitations and minimizing tube transfers have been recommended [54].

5.1.2 Reversing crosslinks

Many Hi-C protocols recommend reversing formaldehyde crosslinks at 65 degrees C overnight. We have observed that this lengthy high temperature step can result in DNA degradation, presumably from trace amounts of DNase enzymes. Extreme caution must be taken to avoid any DNase contamination before this step, and minimizing the length of this incubation may help reduce the likelihood of DNA degradation.

5.1.3 AMPure purification

AMPure XP beads are a powerful and convenient way to purify a given size range of sequencing libraries and to remove adaptor and primer dimers. However, this step can also result in substantial DNA loss in the desired size range as well. We have observed a 50% loss of Hi-C library material after AMPure purification was performed to remove excess adaptor dimers (Figure 5B). Troubleshooting guidelines provided with AMPure beads should be considered to avoid loss of material. Further, since any purification step will involve some DNA loss, protocols should be optimized to minimize the number of purification steps required. For example, increasing the dilution of adaptors in the adaptor ligation step can minimize the need for purification to remove adaptor dimers.

5.2 Difficulties with assessing quality for Hi-C with low cell numbers

Previously published Hi-C protocols often recommend checking the quality of a ligated Hi-C library after by performing a targeted PCR on one or a few neighboring interaction products [26, 33]. However, we find that this control is less likely to be a reliable indicator of Hi-C library success when smaller input cell numbers are used. Though interaction patterns are reproducible as cell numbers are decreased from 20 million to 5 million (Figure 2B), the likelihood of detecting with PCR any given single interaction out of a small sample of the Hi-C library decreases with lower cell numbers. We have observed cases in which the quality control PCR of one interaction amplified no detectable product, but the final Hi-C library and sequenced results were high quality (Figure 5C).

Indeed, even with larger numbers of cells, a PCR digest control can be unreliable. In a recent publication of Hi-C in barley, for example, two variations of Hi-C were used. Effective NheI digestion of the single interaction PCR product band suggested that one protocol worked much better than the other, but NheI digestion of the complex final library reveals that both experiments are similarly high quality [22]. In general, different quality control metrics may be needed to evaluate Hi-C experiments with low cell numbers. Experiments adapting 4C to low cell numbers have found, for example, that short range interaction profiles from low cell numbers can be reproducible even while weak long-range interactions are harder to detect reproducibly [54].

6. Conclusions

The complex nature of Hi-C experiments means that there are many steps, from cell harvesting to sequence mapping, which can affect the quality of the downstream genome interaction maps. Protocol improvements are continuing to increase the power, information, and resolution that can be gleaned from Hi-C experiments. Adjustments to the protocol such as those suggested in Hi-C 2.0, and ideas suggested here, such as avoiding over-sonication, can notably increase the recovery of informative interaction pairs. Artifacts and noise in the Hi-C data can arise from variables in the early cell preparation steps or as a result of the complex mixtures subjected to end repair and PCR steps. Some of these sources of artifacts occassionally lead to Hi-C experiments with lower numbers of valid interacting pairs or higher background noise. Fortunately, as we have shown here, many features of Hi-C maps are reproducible despite protocol variations. As more scientists apply Hi-C to study 3D genome structures in diverse systems, careful attention to and documentation of the details of successful and problematic Hi-C experiments will lead to continued insights into how to optimize this important approach.

Acknowledgements

We thank Yang Xu for his help in calculating Hi-C dataset reproducibility. We thank Job Dekker, Johan Gibcus, Jon Belton, Houda Belaghzal, Natalia Naumova, Noam Kaplan, and Bryan Lajoie for helpful discussions about concepts discussed here. Comparative Hi-C protocol experiments were in part supported by a Ralph E. Powe Junior Faculty Enhancement Award from Oak Ridge Associated Universities.

Figure Legends

Figure 1. Variable factors in early steps of the Hi-C protocol. A) Hi-C experiments for cells in culture begin with decisions about initial steps such as formaldehyde crosslinking, cell harvest, and cell lysis and permeabilization. B) Different approaches to adding formaldehyde may affect the state of the nucleus. In each case, the final concentration of formaldehyde is 1%. (First Panel Top) MDA-MB-231 cells directly under the site of 37% formaldehyde addition show a notable shape change. (First Panel Bottom) Cell morphology changes are less apparent when formaldehyde is pre-diluted to 1% in 1xHBSS. (Second Panel Top) GM12878 nuclei immobilized on a poly-D-lysine coverslip show a dramatic shape change when directly under the site of 37% formaldehyde addition. (Second panel middle) Nuclear shape change is still observed when 1% formaldehyde is pre-diluted in water (Second panel bottom) Nuclear structure is preserved when formaldehyde is pre-diluted in 0.25x PBS. C) The appearance of DNA varies for HepG2 cells harvested by either scraping after crosslinking or enzymatic removal followed by crosslinking. In each case, crosslinks were reversed in harvested cells for 1 hour at 65^{o}C with proteinase K and then DNA was visualized on a 0.8% agarose gel. “Scraped” DNA presented as sticky threads and partly floated out of the well before entering the gel. D) B16-F1 mouse melanoma cells retain extensive cytoplasmic debris in the same nucleus purification protocol that effectively purifies GM12878 nuclei. E) GM12878 nuclei can lyse when 1x PBS buffer is replaced with water, an example to show that unusual buffer conditions could contribute to higher or lower rates of intact nuclei entering the downstream Hi-C protocol. F) Hi-C protocols performed simultaneously on two different cell types can result in differing proportions of informative interaction pairs (Valid Pairs). While unligated HindIII fragments (HindIII dangling ends) are consistent between the cell types, one experiment suffered from more biotin

incorporation at nicks resulting in non-canonical dangling ends. These non-canonical ends are not are recalcitrant to additional biotin removal from DNA ends with the exonuclease activity of T4 DNA polymerase.

Figure 2. Hi-C 2.0 [26] protocol changes benefit HindIII Hi-C libraries as well as DpnII libraries. A) Hi-C library statistics for aliquots of 5 million B16-F1 mouse melanoma cells prepared with: Hi-C 1.0 with HindIII (top), Hi-C 2.0 with HindIII (middle), and Hi-C 2.0 with DpnII (bottom). Left: proportions of reads that are valid interaction pairs as compared to unmapped or single fragment reads. Middle: proportions of mapping orientations of valid interaction pairs. Hi-C 2.0 gives less inward bias, indicating a reduction of biases arising from undigested fragments. Right: Fraction of intrachromosomal (cis) and interchromosomal (trans) interaction pairs. B) Concordance between interaction maps resulting from different Hi-C protocol approaches as calculated by GenomeDISCO. C) Similarity of 40 kb binned and iteratively corrected [39] heatmaps from the 3 parallel experiments for a 7.5 Mb region of chr19. D) Insulation score profiles (sliding window size of 500 kb, heatmap resolution of 40 kb) for a section of chromosome 19 for each Hi-C protocol. Overall TAD boundaries (major dips in the insulation score) are conserved in position and strength between the Hi-C protocols at this resolution.

Figure 3. Artifacts arising from Hi-C sequencing library preparation. A) Fragment size distributions of the final Hi-C library (ligated DNA insert with ligated adaptors) from an Agilent Bioanalyzer trace. The shorter fragments with a narrower distribution have a higher percentage of single side mapped reads. This is likely attributable to the higher chance of encountering a ligation junction in the first few base pairs of the shorter fragment (blue line, bottom panel), resulting in one unmappable interaction partner. B) Sample reads demonstrating the artifact in which a piece of the 5' end of a Hi-C interaction molecule is artificially attached to the 3' end. As a result, paired end Read 1 and 2 are identical for the first 10-15 bp and Read 2 is unmappable without 5' trimming. C) A Hi-C experiment on GM12878 nuclei showed an extremely high amount of copying from the 5' to 3' end of interaction library molecules, resulting in a large % of the total read pairs with 10-20 bases identical between Read 1 and Read 2. This corresponds to a only 45% alignment rate of Read 2. This effect is observed to a lesser, but notable extent in other experimental data, such as published HFF1 metaphase Hi-C [31]. Other published datasets, such as pro B cell Hi-C data [3] show only a peak at 13 bp, which corresponds to the amount of identity between adaptor dimers. A Hi-C library prepared by tagmentation [42] (SRR3105477) shows none of this copying artifact. D) Most library molecules that experience this copying artifact have 4-6 bp of complementary sequence between the two ends of the molecule. The histogram shown represents the length of identical sequence between Read 1 and Read 2 that maps on both sides and is immediately adjacent to unmappable copied sequence on Read 2 (blue box from B). E) Reads containing these artifacts (blue) tend to arise from shorter DNA molecules than non-artifact reads (red). F) Model of explanation of how short complementary regions ("microhomologies", blue boxes) could anneal and then elongate during end repair, copying the 5' end of the top strand and appending it to the 3' end of the top strand. During sequencing, this chimeric 3' end is sequenced as Read 2 as shown. G) In a Hi-C library with large numbers of noisy interchromosomal read pairs ("High % trans"), a high percentage of these reads contained no Hi-C junction sequence, suggesting that they may arise from something other than ordinary Hi-C ligation. When the number of interchromosomal pairs is low, on the other hand, most of these "trans" reads contain a canonical Hi-C junction sequence.

Figure 4. Considerations in Hi-C data alignment and binning. A) 10 kb binned and corrected heatmaps in a 4 Mb region of chr22 with varying input read numbers for poor complexity and good complexity

replicates of a GM12878 Hi-C experiment. The poor complexity Hi-C library (left) gives no additional interaction resolution or detail even when input reads are increased to 1 billion (upper triangle) vs. 335 million (lower triangle). The high complexity library (right) on the other hand gives much more detailed information at 863 million reads (upper triangle), while only 27 million reads (lower triangle) are needed to give the amount of detail captured by 355 million reads of the poor complexity library. B) Three scenarios of proximal interactions are considered for their effect on interacting strand/side bias: “Frag1 long”: With increasing upstream restriction fragment length, positive strand interactions (“top”, green line) are increasingly favored compared to negative strand interactions (“bottom”, purple line). “Frag2 long”: with increasing downstream fragment length, the opposite pattern holds. “Either/both long”: if either fragment can be equally long, no strand bias is seen, only a preference for inward facing reads (closest on the linear genome) vs. outward facing reads (farthest apart on the linear genome). C) Distributions of HindIII (left) and DpnII (right) restriction fragment lengths in the mouse genome (mm9).

Figure 5. A) After reversal of crosslinks, three different GM12878 Hi-C libraries (A, B, and C) were purified and ethanol precipitated as recommended [33], yielding 14-22 micrograms of DNA. The ethanol supernatant removed from the DNA pellet was then re-centrifuged at the same speed as in the first purification, and another 2 – 6 micrograms (10-40% more) of DNA was recovered. B) Bioanalyzer traces before (black) and after (red) 1.5x Ampure XP purification are shown, scaled according to their 35 bp standard peak. Size-selection of Hi-C library DNA with 1.5x Ampure XP beads successfully removes the ~80 bp excess adaptor peak (black), but reduces the overall concentration of DNA by nearly half. C) For the HindIII Hi-C 2.0 experiment on 5 million B16-F1 cells, no PCR product was observed for a neighboring interaction primer pair in the GAPDH gene region (primers described in [33]) (left). However, the final Hi-C library could be amplified with 10 cycles of PCR (middle), and digested well with NheI, as expected for a high quality Hi-C library (right). This final library produced good Hi-C interaction data (see Figure 2) despite the failed quality control neighboring interaction PCR. Note: some lanes in gels have been excised for clarity, but all lanes shown in contiguous images were run on the same gel.

References

[1] J. Dekker, L. Mirny, The 3D Genome as Moderator of Chromosomal Communication, Cell 164(6) (2016) 1110-21.
[2] V. Dileep, F. Ay, J. Sima, D.L. Vera, W.S. Noble, D.M. Gilbert, Topologically associating domains and their long-range contacts are established during early G1 coincident with the establishment of the replication-timing program, Genome Res 25(8) (2015) 1104-13.
[3] Y. Zhang, R.P. McCord, Y.J. Ho, B.R. Lajoie, D.G. Hildebrand, A.C. Simon, M.S. Becker, F.W. Alt, J. Dekker, Spatial organization of the mouse genome and its role in recurrent chromosomal translocations, Cell 148(5) (2012) 908-21.
[4] D.G. Lupianez, K. Kraft, V. Heinrich, P. Krawitz, F. Brancati, E. Klopocki, D. Horn, H. Kayserili, J.M. Opitz, R. Laxova, F. Santos-Simarro, B. Gilbert-Dussardier, L. Wittler, M. Borschiwer, S.A. Haas, M. Osterwalder, M. Franke, B. Timmermann, J. Hecht, M. Spielmann, A. Visel, S. Mundlos, Disruptions of topological chromatin domains cause pathogenic rewiring of gene-enhancer interactions, Cell 161(5) (2015) 1012-25.
[5] R.P. McCord, A. Nazario-Toole, H. Zhang, P.S. Chines, Y. Zhan, M.R. Erdos, F.S. Collins, J. Dekker, K. Cao, Correlated alterations in genome organization, histone methylation, and DNA-lamin A/C interactions in Hutchinson-Gilford progeria syndrome, Genome Res 23(2) (2013) 260-9.

[6] E. Heitz, Das Heterochromatin der Moose, Jahrb Wiss Botanik 69 (1928) 762–818.
[7] A.L. Olins, D.E. Olins, Spheroid Chromatin Units (v Bodies), Science 183(4122) (1974) 330-332.
[8] S. Shachar, T.C. Voss, G. Pegoraro, N. Sciascia, T. Misteli, A high-throughput imaging-based mapping platform for the systematic identification of gene positioning factors, Cell 162(4) (2015) 911-923.
[9] C.C. Robinett, A. Straight, G. Li, C. Willhelm, G. Sudlow, A. Murray, A.S. Belmont, In vivo localization of DNA sequences and visualization of large-scale chromatin organization using lac operator/repressor recognition, The Journal of Cell Biology 135(6) (1996) 1685-1700.
[10] H. Ma, A. Naseri, P. Reyes-Gutierrez, S.A. Wolfe, S. Zhang, T. Pederson, Multicolor CRISPR labeling of chromosomal loci in human cells, Proceedings of the National Academy of Sciences 112(10) (2015) 3002-3007.
[11] J. Dekker, K. Rippe, M. Dekker, N. Kleckner, Capturing chromosome conformation, Science 295(5558) (2002) 1306-11.
[12] M. Simonis, P. Klous, E. Splinter, Y. Moshkin, R. Willemsen, E. de Wit, B. van Steensel, W. de Laat, Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture-on-chip (4C), Nat Genet 38(11) (2006) 1348-1354.
[13] Z. Zhao, G. Tavoosidana, M. Sjolinder, A. Gondor, P. Mariano, S. Wang, C. Kanduri, M. Lezcano, K. Singh Sandhu, U. Singh, V. Pant, V. Tiwari, S. Kurukuti, R. Ohlsson, Circular chromosome conformation capture (4C) uncovers extensive networks of epigenetically regulated intra- and interchromosomal interactions, Nat Genet 38(11) (2006) 1341-1347.
[14] J. Dostie, T.A. Richmond, R.A. Arnaout, R.R. Selzer, W.L. Lee, T.A. Honan, E.D. Rubio, A. Krumm, J. Lamb, C. Nusbaum, R.D. Green, J. Dekker, Chromosome Conformation Capture Carbon Copy (5C): A massively parallel solution for mapping interactions between genomic elements, Genome Research 16(10) (2006) 1299-1309.
[15] E. Lieberman-Aiden, N.L. van Berkum, L. Williams, M. Imakaev, T. Ragoczy, A. Telling, I. Amit, B.R. Lajoie, P.J. Sabo, M.O. Dorschner, R. Sandstrom, B. Bernstein, M.A. Bender, M. Groudine, A. Gnirke, J. Stamatoyannopoulos, L.A. Mirny, E.S. Lander, J. Dekker, Comprehensive mapping of long-range interactions reveals folding principles of the human genome, Science 326(5950) (2009) 289-93.
[16] J.H. Gibcus, J. Dekker, The hierarchy of the 3D genome, Mol Cell 49(5) (2013) 773-82.
[17] S.S. Rao, M.H. Huntley, N.C. Durand, E.K. Stamenova, I.D. Bochkov, J.T. Robinson, A.L. Sanborn, I. Machol, A.D. Omer, E.S. Lander, E.L. Aiden, A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping, Cell 159(7) (2014) 1665-80.
[18] J.R. Hughes, N. Roberts, S. McGowan, D. Hay, E. Giannoulatou, M. Lynch, M. De Gobbi, S. Taylor, R. Gibbons, D.R. Higgs, Analysis of hundreds of cis-regulatory landscapes at high resolution in a single, high-throughput experiment, 46 (2014) 205.
[19] R. Jager, G. Migliorini, M. Henrion, R. Kandaswamy, H.E. Speedy, A. Heindl, N. Whiffin, M.J. Carnicer, L. Broome, N. Dryden, T. Nagano, S. Schoenfelder, M. Enge, Y. Yuan, J. Taipale, P. Fraser, O. Fletcher, R.S. Houlston, Capture Hi-C identifies the chromatin interactome of colorectal cancer risk loci, Nat Commun 6 (2015) 6178.
[20] M.R. Mumbach, A.J. Rubin, R.A. Flynn, C. Dai, P.A. Khavari, W.J. Greenleaf, H.Y. Chang, HiChIP: efficient and sensitive analysis of protein-directed genome architecture, Nat Methods 13(11) (2016) 919-922.
[21] P. Kolovos, H.J. van de Werken, N. Kepper, J. Zuin, R.W. Brouwer, C.E. Kockx, K.S. Wendt, I.W.F. van, F. Grosveld, T.A. Knoch, Targeted Chromatin Capture (T2C): a novel high resolution high throughput method to detect genomic interactions and regulatory elements, Epigenetics Chromatin 7 (2014) 10.
[22] M. Mascher, H. Gundlach, A. Himmelbach, S. Beier, S.O. Twardziok, T. Wicker, V. Radchuk, C. Dockter, P.E. Hedley, J. Russell, M. Bayer, L. Ramsay, H. Liu, G. Haberer, X.Q. Zhang, Q. Zhang, R.A. Barrero, L. Li, S. Taudien, M. Groth, M. Felder, A. Hastie, H. Simkova, H. Stankova, J. Vrana, S. Chan, M. Munoz-Amatriain, R. Ounit, S. Wanamaker, D. Bolser, C. Colmsee, T. Schmutzer, L. Aliyeva-Schnorr, S.

Grasso, J. Tanskanen, A. Chailyan, D. Sampath, D. Heavens, L. Clissold, S. Cao, B. Chapman, F. Dai, Y. Han, H. Li, X. Li, C. Lin, J.K. McCooke, C. Tan, P. Wang, S. Wang, S. Yin, G. Zhou, J.A. Poland, M.I. Bellgard, L. Borisjuk, A. Houben, J. Dolezel, S. Ayling, S. Lonardi, P. Kersey, P. Langridge, G.J. Muehlbauer, M.D. Clark, M. Caccamo, A.H. Schulman, K.F.X. Mayer, M. Platzer, T.J. Close, U. Scholz, M. Hansson, G. Zhang, I. Braumann, M. Spannagl, C. Li, R. Waugh, N. Stein, A chromosome conformation capture ordered sequence of the barley genome, Nature 544(7651) (2017) 427-433.
[23] N. Kaplan, J. Dekker, High-throughput genome scaffolding from in vivo DNA interaction frequency, 31 (2013) 1143.
[24] J.N. Burton, A. Adey, R.P. Patwardhan, R. Qiu, J.O. Kitzman, J. Shendure, Chromosome-scale scaffolding of de novo genome assemblies based on chromatin interactions, 31 (2013) 1119.
[25] L. Harewood, K. Kishore, M.D. Eldridge, S. Wingett, D. Pearson, S. Schoenfelder, V.P. Collins, P. Fraser, Hi-C as a tool for precise detection and characterisation of chromosomal rearrangements and copy number variation in human tumours, Genome Biol 18(1) (2017) 125.
[26] H. Belaghzal, J. Dekker, J.H. Gibcus, Hi-C 2.0: An optimized Hi-C procedure for high-resolution genome-wide mapping of chromosome conformation, Methods (2017).
[27] B.R. Lajoie, J. Dekker, N. Kaplan, The Hitchhiker's guide to Hi-C analysis: practical guidelines, Methods 72 (2015) 65-75.
[28] A.D. Schmitt, M. Hu, B. Ren, Genome-wide mapping and analysis of chromosome architecture, Nat Rev Mol Cell Biol 17(12) (2016) 743-755.
[29] T. Nagano, C. Varnai, S. Schoenfelder, B.M. Javierre, S.W. Wingett, P. Fraser, Comparison of Hi-C results using in-solution versus in-nucleus ligation, Genome Biol 16 (2015) 175.
[30] A. Gondor, C. Rougier, R. Ohlsson, High-resolution circular chromosome conformation capture assay, Nat Protoc 3(2) (2008) 303-13.
[31] N. Naumova, M. Imakaev, G. Fudenberg, Y. Zhan, B.R. Lajoie, L.A. Mirny, J. Dekker, Organization of the mitotic chromosome, Science 342(6161) (2013) 948-53.
[32] A.S. Belmont, Large-scale Chromatin Organization: The Good, the Surprising, and the Still Perplexing, Current opinion in cell biology 0 (2014) 69-78.
[33] J.M. Belton, R.P. McCord, J.H. Gibcus, N. Naumova, Y. Zhan, J. Dekker, Hi-C: a comprehensive technique to capture the conformation of genomes, Methods 58(3) (2012) 268-76.
[34] P. Dong, X. Tu, P.Y. Chu, P. Lu, N. Zhu, D. Grierson, B. Du, P. Li, S. Zhong, 3D Chromatin Architecture of Large Plant Genomes Determined by Local A/B Compartments, Mol Plant 10(12) (2017) 1497-1509.
[35] E.A. Hoffman, B.L. Frey, L.M. Smith, D.T. Auble, Formaldehyde crosslinking: a tool for the study of chromatin complexes, J Biol Chem 290(44) (2015) 26404-11.
[36] R.P. McCord, V.W. Zhou, T. Yuh, M.L. Bulyk, Distant cis-regulatory elements in human skeletal muscle differentiation, Genomics 98(6) (2011) 401-411.
[37] E. Makhija, D.S. Jokhun, G.V. Shivashankar, Nuclear deformability and telomere dynamics are regulated by cell geometric constraints, Proceedings of the National Academy of Sciences of the United States of America 113(1) (2016) E32-E40.
[38] H. Takata, T. Hanafusa, T. Mori, M. Shimura, Y. Iida, K. Ishikawa, K. Yoshikawa, Y. Yoshikawa, K. Maeshima, Chromatin compaction protects genomic DNA from radiation damage, PLoS One 8(10) (2013) e75622.
[39] M. Imakaev, G. Fudenberg, R.P. McCord, N. Naumova, A. Goloborodko, B.R. Lajoie, J. Dekker, L.A. Mirny, Iterative correction of Hi-C data reveals hallmarks of chromosome organization, Nat Meth 9(10) (2012) 999-1003.
[40] H.E. Bryant, DNA double-strand break damage and repair assessed by pulsed-field gel electrophoresis, Methods Mol Biol 920 (2012) 315-21.

[41] W. Schwarzer, N. Abdennur, A. Goloborodko, A. Pekowska, G. Fudenberg, Y. Loe-Mie, N.A. Fonseca, W. Huber, C. Haering, L. Mirny, F. Spitz, Two independent modes of chromatin organization revealed by cohesin removal, Nature advance online publication (2017).
[42] I. Gabdank, S. Ramakrishnan, A.M. Villeneuve, A.Z. Fire, A streamlined tethered chromosome conformation capture protocol, BMC Genomics 17 (2016) 274.
[43] E. Crane, Q. Bian, R.P. McCord, B.R. Lajoie, B.S. Wheeler, E.J. Ralston, S. Uzawa, J. Dekker, B.J. Meyer, Condensin-driven remodelling of X chromosome topology during dosage compensation, Nature 523(7559) (2015) 240-4.
[44] J. Dekker, A.S. Belmont, M. Guttman, V.O. Leshyk, J.T. Lis, S. Lomvardas, L.A. Mirny, C.C. O'Shea, P.J. Park, B. Ren, J.C.R. Politz, J. Shendure, S. Zhong, D.N. Network, The 4D nucleome project, Nature 549(7671) (2017) 219-226.
[45] O. Ursu, N. Boley, M. Taranova, Y.X.R. Wang, G.G. Yardimci, W.S. Noble, A. Kundaje, GenomeDISCO: A concordance score for chromosome conformation capture experiments using random walks on contact map graphs, bioRxiv (2017).
[46] M. Manghi, N. Destainville, Physics of base-pairing dynamics in DNA, Phys Rep 631 (2016) 1-41.
[47] D. Ottaviani, M. LeCain, D. Sheer, The role of microhomology in genomic structural variation, Trends in Genetics 30(3) (2014) 85-94.
[48] M.R. Lieber, The Mechanism of Double-Strand DNA Break Repair by the Nonhomologous DNA End Joining Pathway, Annual review of biochemistry 79 (2010) 181-211.
[49] E. Crespan, T. Czabany, G. Maga, U. Hübscher, Microhomology-mediated DNA strand annealing and elongation by human DNA polymerases λ and β on normal and repetitive DNA sequences, Nucleic Acids Research 40(12) (2012) 5577-5590.
[50] E.M. Smith, B.R. Lajoie, G. Jain, J. Dekker, Invariant TAD Boundaries Constrain Cell-Type-Specific Looping Interactions between Promoters and Distal Elements around the CFTR Locus, Am J Hum Genet 98(1) (2016) 185-201.
[51] V.B. Teif, J.P. Mallm, T. Sharma, D.B. Mark Welch, K. Rippe, R. Eils, J. Langowski, A.L. Olins, D.E. Olins, Nucleosome repositioning during differentiation of a human myeloid leukemia cell line, Nucleus 8(2) (2017) 188-204.
[52] S. Wingett, Illumina Patterned Flow Cells Generate Duplicated Sequences, QCfail.com https://sequencing.qcfail.com/articles/illumina-patterned-flow-cells-generate-duplicated-sequences/ (2017).
[53] T. Nagano, S.W. Wingett, P. Fraser, Capturing Three-Dimensional Genome Organization in Individual Cells by Single-Cell Hi-C, Methods Mol Biol 1654 (2017) 79-97.
[54] J.O. Davies, J.M. Telenius, S.J. McGowan, N.A. Roberts, S. Taylor, D.R. Higgs, J.R. Hughes, Multiplexed analysis of chromosome conformation at vastly improved sensitivity, Nat Methods 13(1) (2016) 74-80.

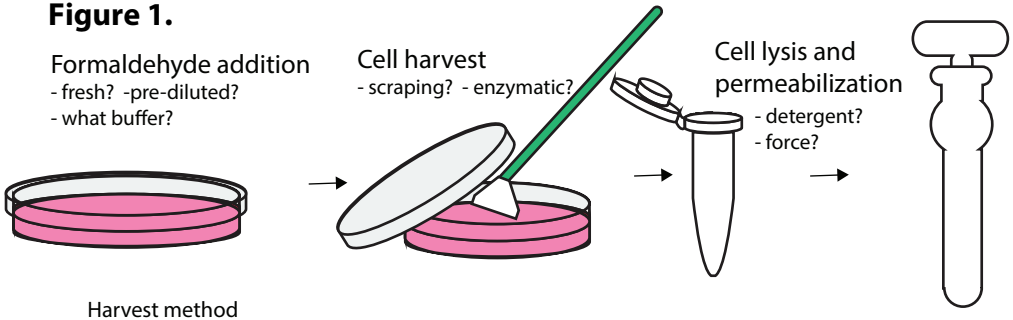
Figure 1.

A

Formaldehyde addition
- fresh? -pre-diluted?
- what buffer?

Cell harvest
- scraping? - enzymatic?

Cell lysis and permeabilization
- detergent?
- force?

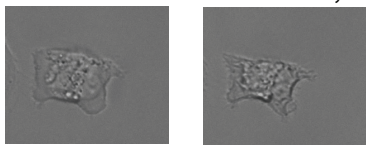


B

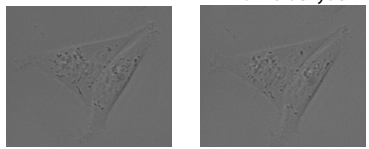
Adherent MDA-MB-231 cells

Pre-crosslinking | Crosslinking

1xHBSS | 37% Formaldehyde added directly



1xHBSS | Pre - diluted 1% Formaldehyde



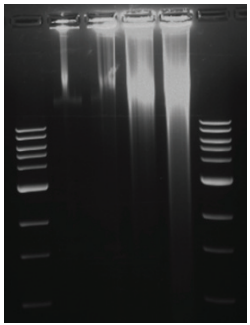
C

Harvest method

Scrape Accutase

10 kb

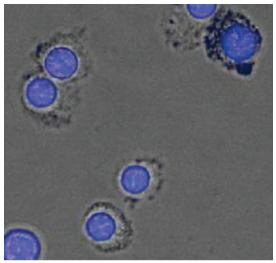
5 kb



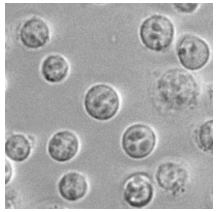
D

0.06% NP-40 lysis buffer + 15 dounce strokes for:

B16-F1

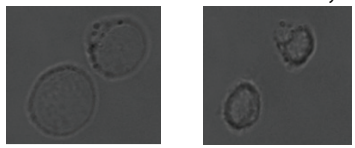


GM12878

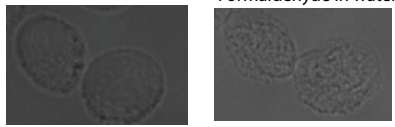


Immobilized GM12878 nuclei

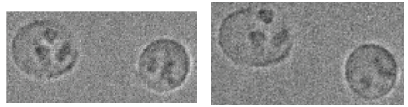
water | 37% Formaldehyde added directly



water | Pre - diluted 1% Formaldehyde in water



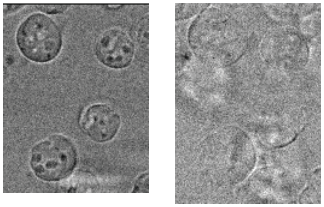
0.25x PBS | Pre - diluted 1% Formaldehyde in 0.25x PBS



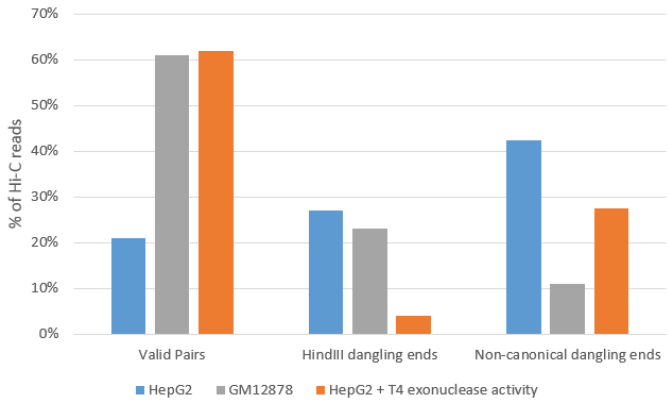
E

GM12878 nuclei

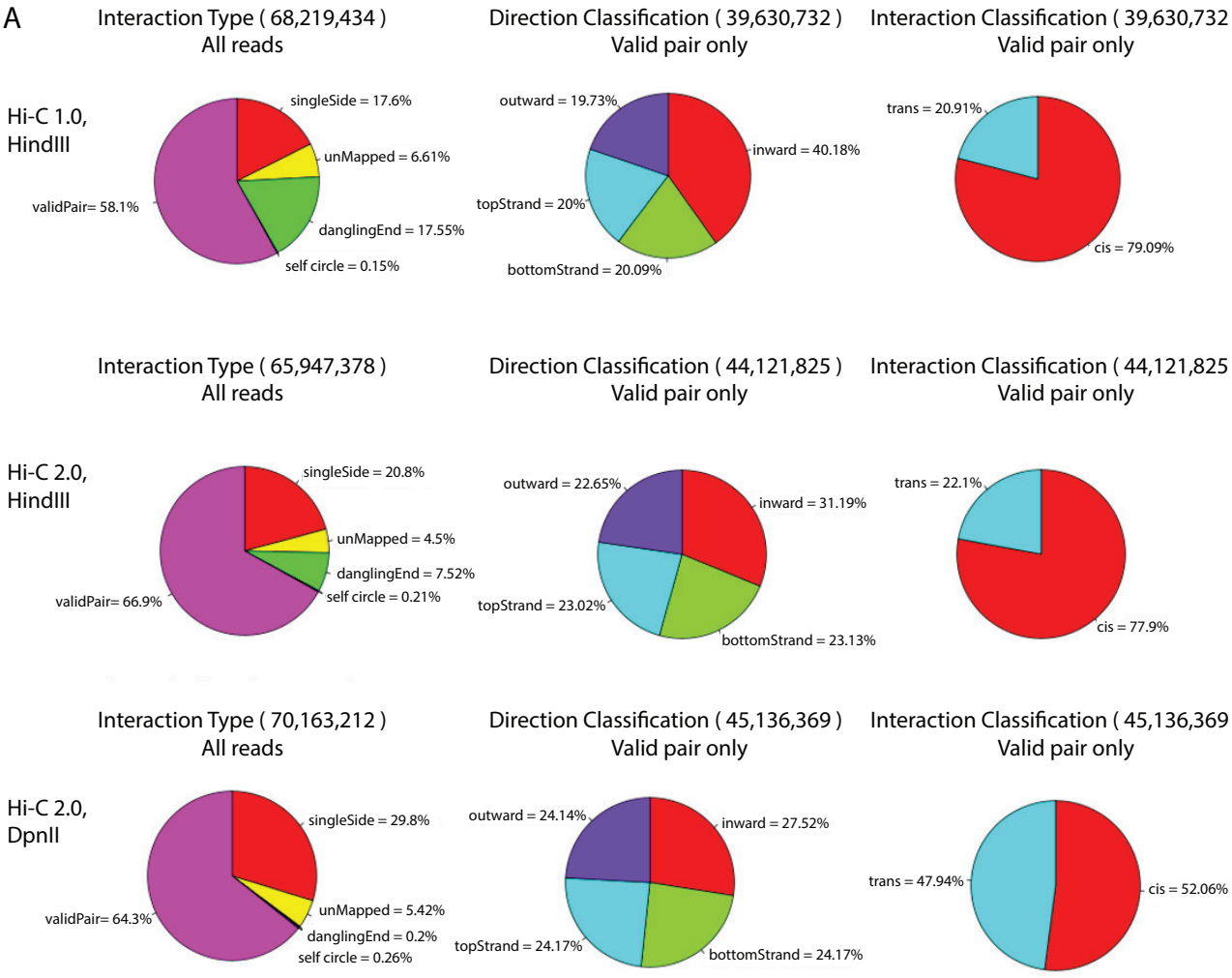
1x PBS | water



F



A

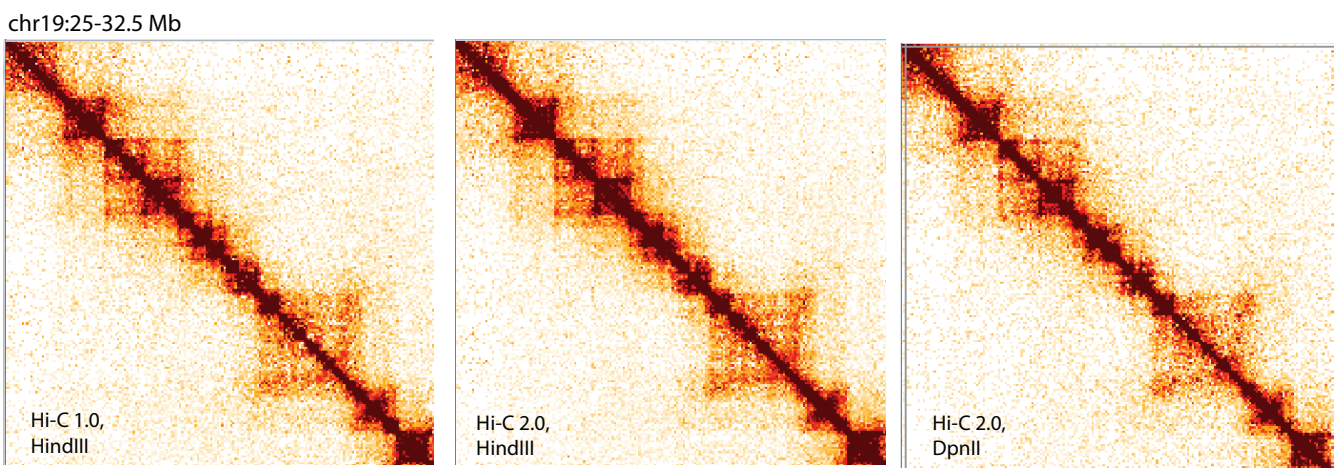


B

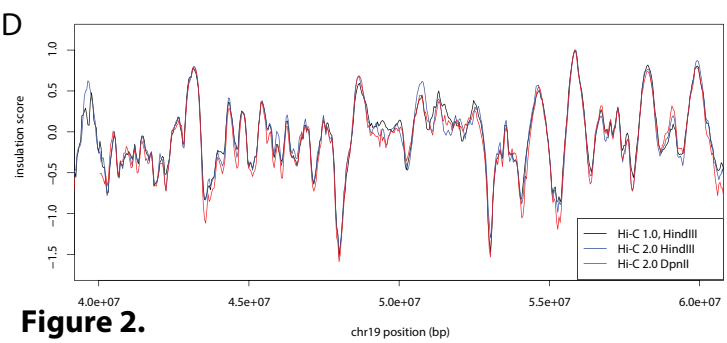
Hi-C concordance scores (GenomeDISCO):

2.5 Mb, whole genome	1	2	3	4
Hi-C 1.0 , HindIII, Illumina prep (1)	1	0.949	0.955	0.819
Hi-C 1.0, HindIII, NEB prep (2)	0.949	1	0.964	0.825
Hi-C 2.0, HindIII, NEB prep (3)	0.955	0.964	1	0.819
Hi-C 2.0, DpnII, NEB prep (4)	0.819	0.825	0.819	1

C



D



E

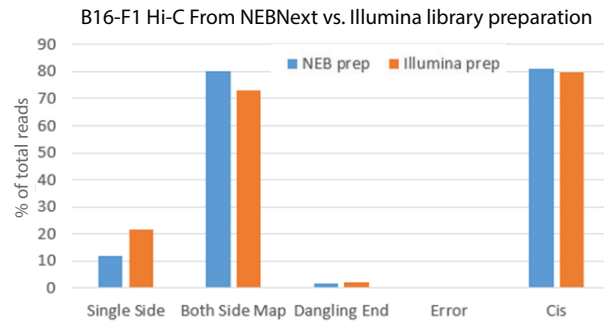


Figure 2.

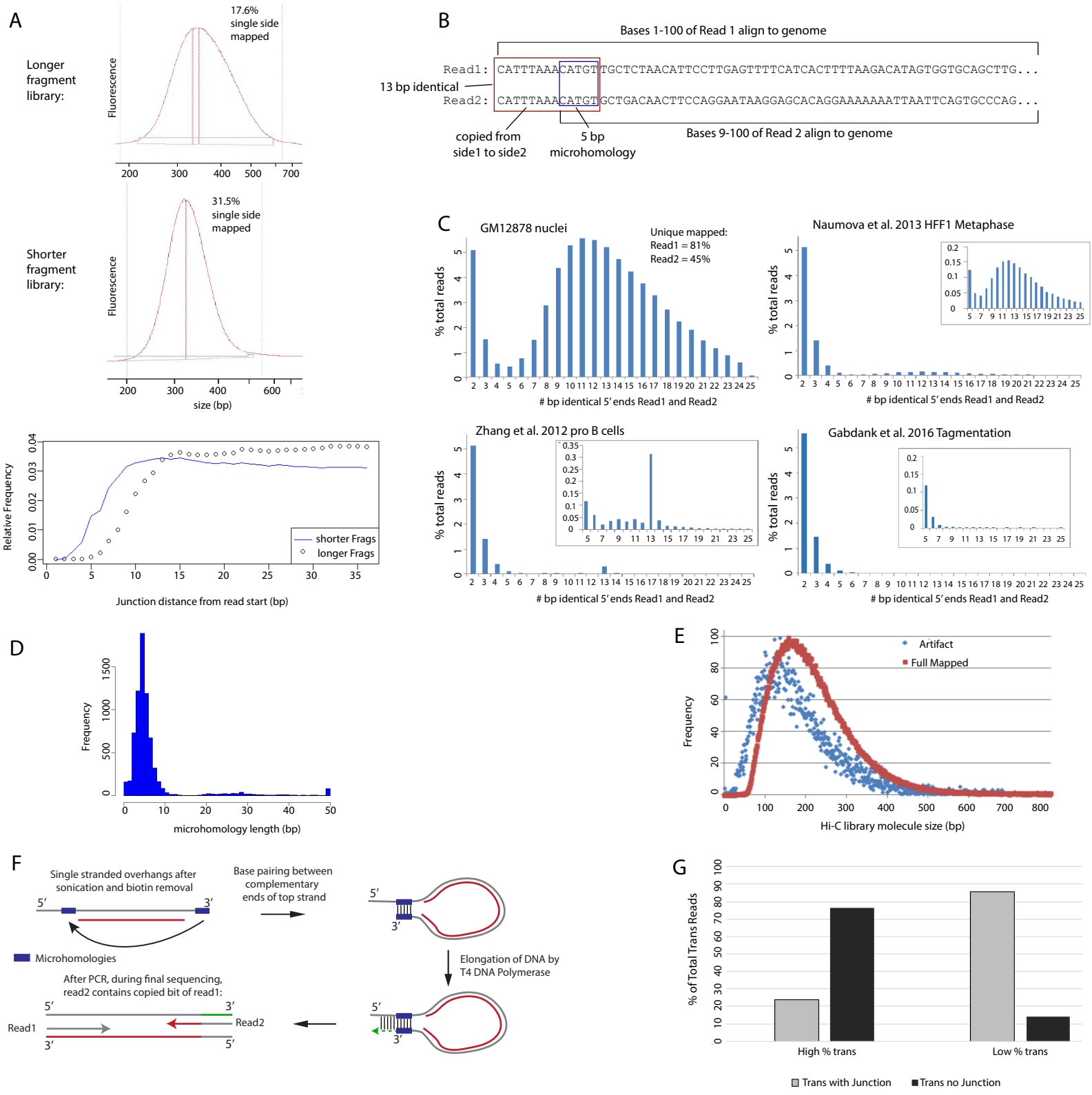
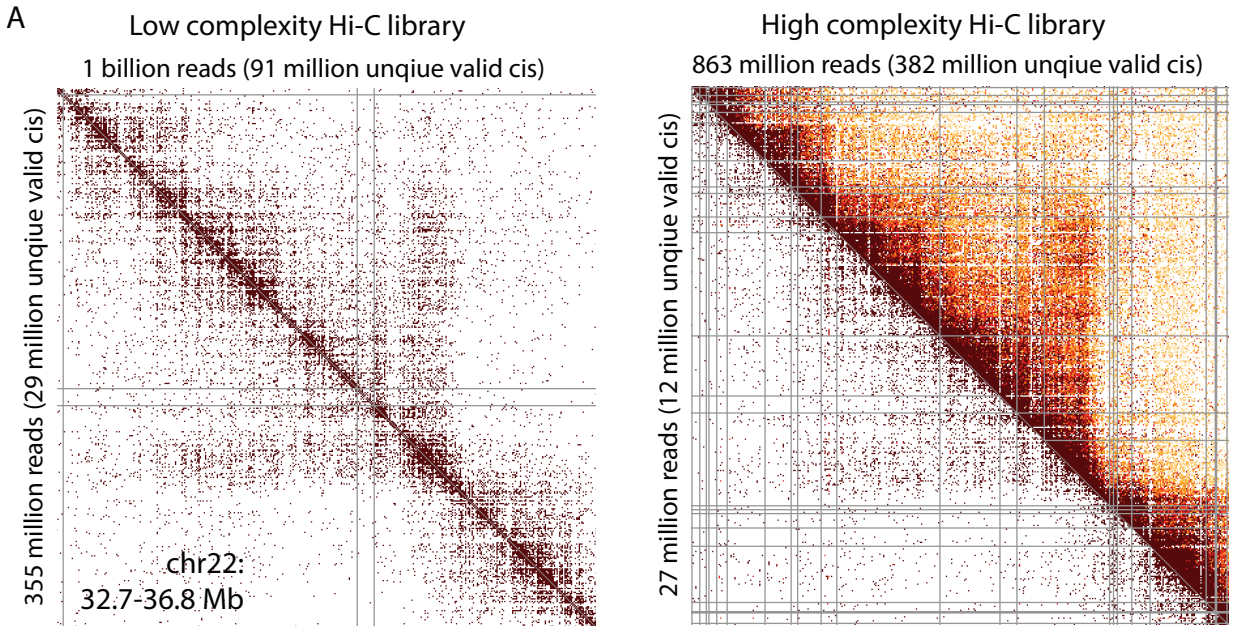
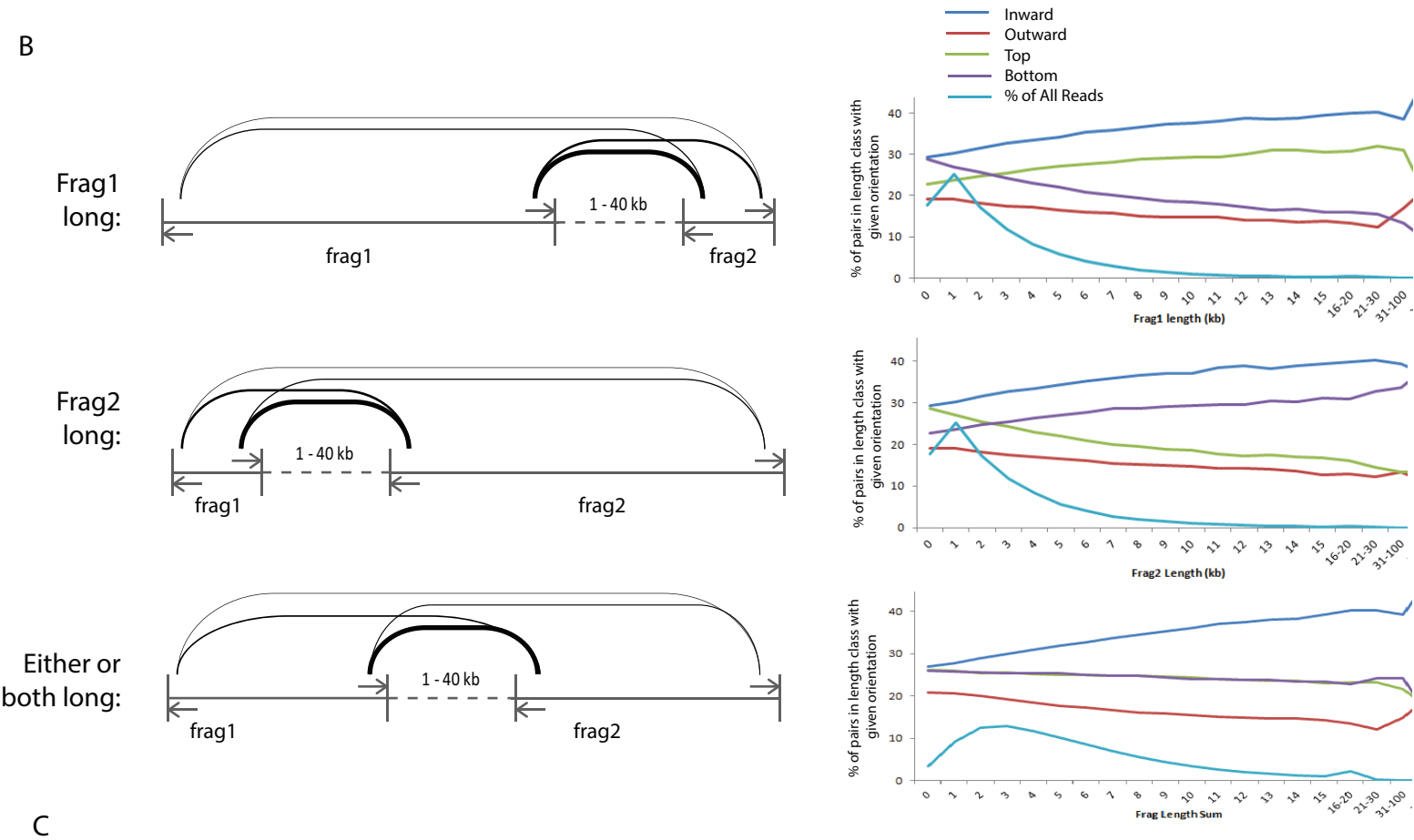
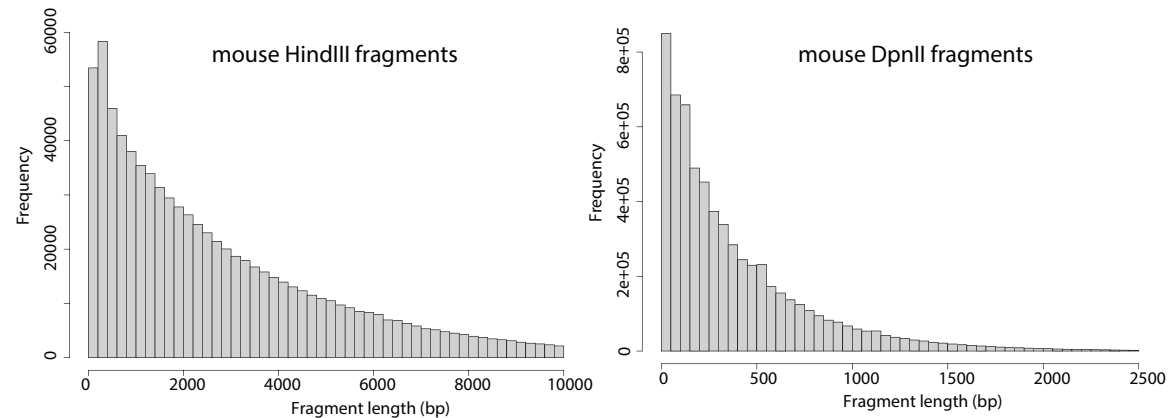

A
Longer fragment library:
17.6% single side mapped
Fluorescence
200 300 400 500 700
Shorter fragment library:
31.5% single side mapped
Fluorescence
200 300 400 600
size (bp)
Relative Frequency
0.00 0.01 0.02 0.03 0.04
shorter Frags
longer Frags
0 5 10 15 20 25 30 35
Junction distance from read start (bp)
B
Bases 1-100 of Read 1 align to genome
Read1: CATTTAAACATGTTGCTCTAACATTCCTTGAGTTTTCATCACTTTTAAGACATAGTGGTGCAGCTTG...
13 bp identical
Read2: CATTTAAACATGTGCTGACAACTTCCAGGAATAAGGAGCACAGGAAAAAAATTAATTCAGTGCCCAG...
copied from side1 to side2
5 bp microhomology
Bases 9-100 of Read 2 align to genome
C
GM12878 nuclei
Unique mapped:
Read1 = 81%
Read2 = 45%
% total reads
# bp identical 5′ ends Read1 and Read2
Naumova et al. 2013 HFF1 Metaphase
% total reads
# bp identical 5′ ends Read1 and Read2
Zhang et al. 2012 pro B cells
% total reads
# bp identical 5′ ends Read1 and Read2
Gabdank et al. 2016 Tagmentation
% total reads
# bp identical 5′ ends Read1 and Read2
D
Frequency
0 500 1000 1500
0 10 20 30 40 50
microhomology length (bp)
E
Artifact
Full Mapped
Frequency
0 20 40 60 80 100
0 100 200 300 400 500 600 700 800
Hi-C library molecule size (bp)
F
Single stranded overhangs after sonication and biotin removal
5′ 3′
Base pairing between complementary ends of top strand
5′ 3′
Microhomologies
Elongation of DNA by T4 DNA Polymerase
After PCR, during final sequencing, read2 contains copied bit of read1:
5′ 3′
Read1
Read2
3′ 5′
5′ 3′
G
% of Total Trans Reads
0 10 20 30 40 50 60 70 80 90 100
High % trans
Low % trans
Trans with Junction
Trans no Junction

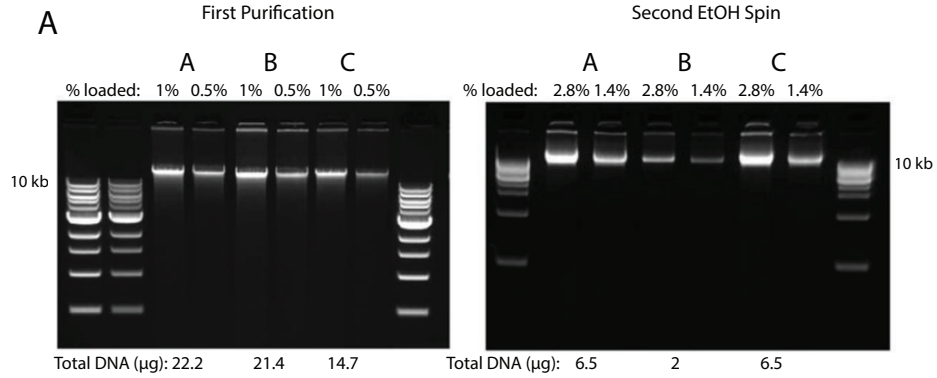
Figure 3.

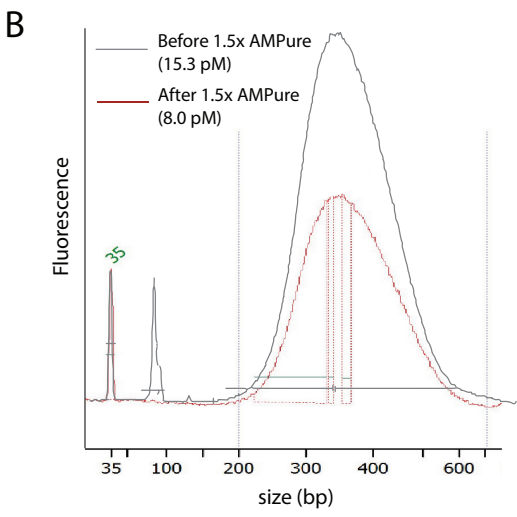
Figure 4.


A
Low complexity Hi-C library
1 billion reads (91 million unqiue valid cis)
355 million reads (29 million unqiue valid cis)
chr22:
32.7-36.8 Mb
High complexity Hi-C library
863 million reads (382 million unqiue valid cis)
27 million reads (12 million unqiue valid cis)


B
Frag1 long:
1 - 40 kb
frag1
frag2
Frag2 long:
1 - 40 kb
frag1
frag2
Either or both long:
1 - 40 kb
frag1
frag2
Inward
Outward
Top
Bottom
% of All Reads
% of pairs in length class with given orientation
Frag1 length (kb)
Frag2 Length (kb)
Frag Length Sum


C
mouse HindIII fragments
Frequency
Fragment length (bp)
mouse DpnII fragments
Frequency
Fragment length (bp)


A
First Purification
Second EtOH Spin
A B C
% loaded: 1% 0.5% 1% 0.5% 1% 0.5%
10 kb
Total DNA (µg): 22.2 21.4 14.7
A B C
% loaded: 2.8% 1.4% 2.8% 1.4% 2.8% 1.4%
10 kb
Total DNA (µg): 6.5 2 6.5


B
Before 1.5x AMPure
(15.3 pM)
After 1.5x AMPure
(8.0 pM)
Fluorescence
35
35 100 200 300 400 600
size (bp)

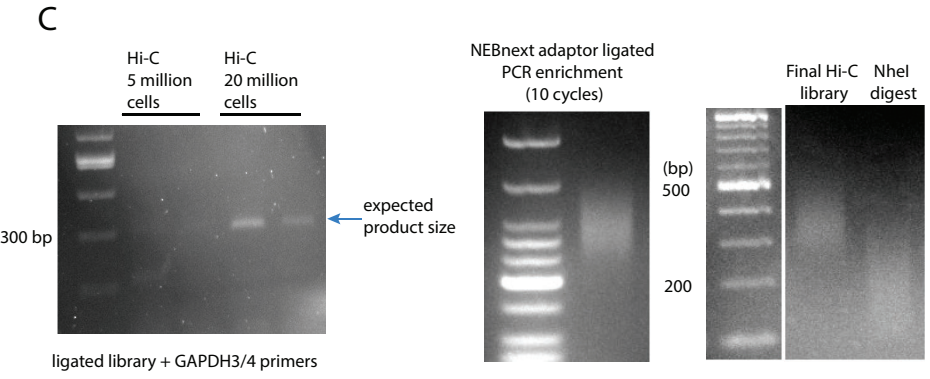

C
Hi-C
5 million
cells
Hi-C
20 million
cells
300 bp
expected
product size
ligated library + GAPDH3/4 primers
NEBnext adaptor ligated
PCR enrichment
(10 cycles)
(bp)
500
200
Final Hi-C
library
NheI
digest

Figure 5.

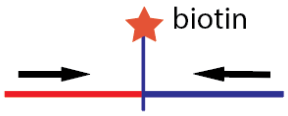
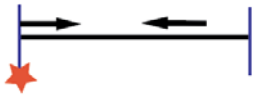
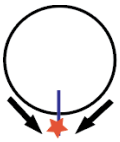
	Hi-C Term	Depiction	Definition
Interaction types for all reads	Valid Pair		Reads from different restriction fragments and facing toward restriction site
	Single Side Reads		Only one side of the interacting pair maps to the genome
	Unmapped Reads		Neither side of the interacting pair maps to the genome
	Dangling End		Reads facing inward in the same fragment
	Self Circle		Reads facing outward in the same fragment
Classification of Valid Pairs	Bottom Strand		Reads both on the (-) strand from non-adjacent fragments
	Top Strand		Reads both on the (+) strand from non-adjacent fragments
	Inward		Reads on opposite strands from non-adjacent fragments facing inward
	Outward		Reads on opposite strands from non-adjacent fragments facing outward
	Cis		Interactions within the same chromosome
	Trans		Interactions between different chromosomes

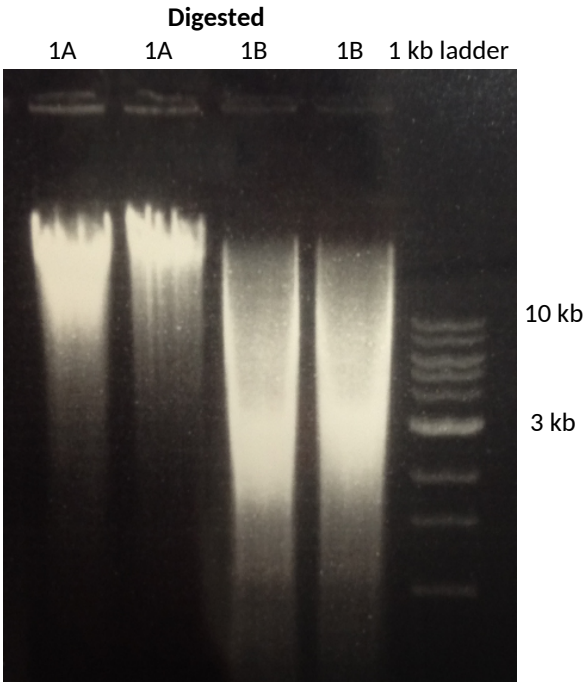
Table 1. Definitions of Hi-C molecule types

Appendix A. Digestion Efficiency and Resulting Hi-C Library Statistics

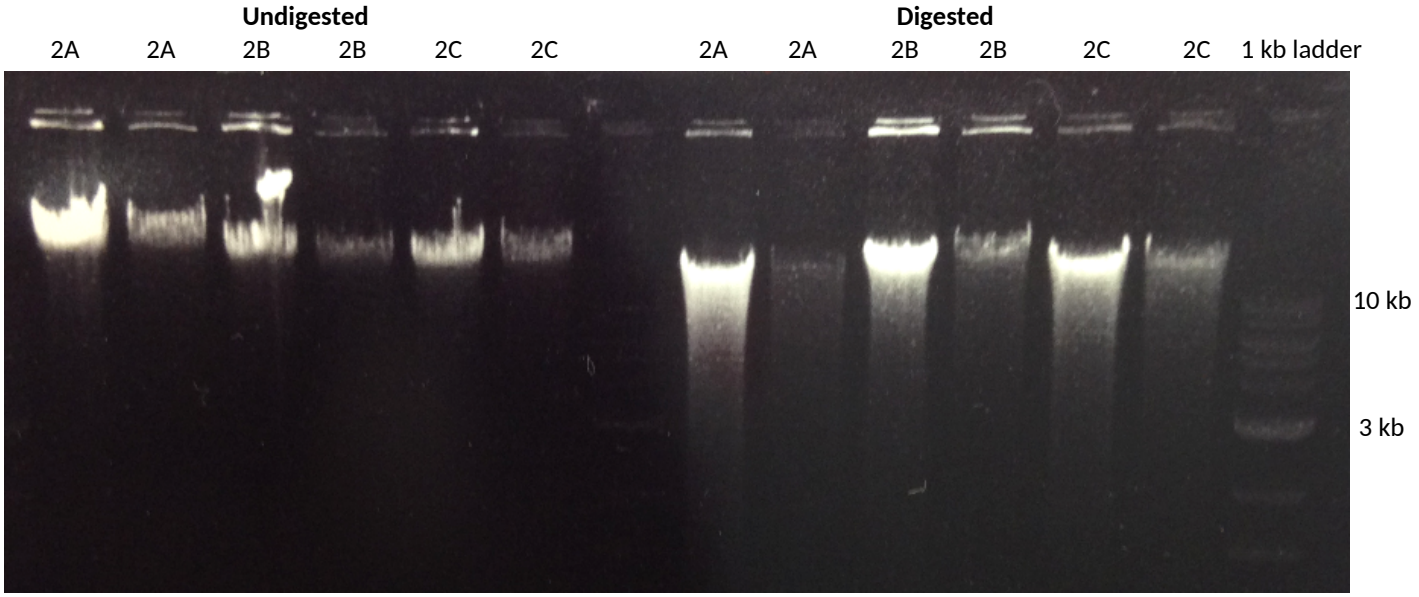
The table below summarizes results from three different problematic Hi-C experiments in which inter-sample variation existed in the extent of HindIII digestion visible on a 0.8% agarose gel (actual gel images shown below the table) immediately after the digestion step of the Hi-C protocol. In each case, the samples within each experiment that had less digestion / more intact DNA resulted in higher percentages of uninformative dangling end molecules in the final Hi-C library. In most cases, more evident digestion was related to a much higher fraction of the molecules ligating to others in cis rather than in trans. The high proportions of trans interactions relate to higher levels of overall background noise. Note that Experiment 3A (starred) corresponds to the low complexity replicate pictured in Figure 4A.

	Digestion gel appearance	Dangling end %	% Cis
Expt 1A	Still quite intact, high MW	69	18
Expt 1B	clearly digested	45	48
Expt 2A	More clearly digested	33	60
Expt 2B	less digested	43	33
Expt 2C	More clearly digested	22	61
Expt 3A*	Still quite intact, high MW	56	64
Expt 3B	More clearly digested	29	64

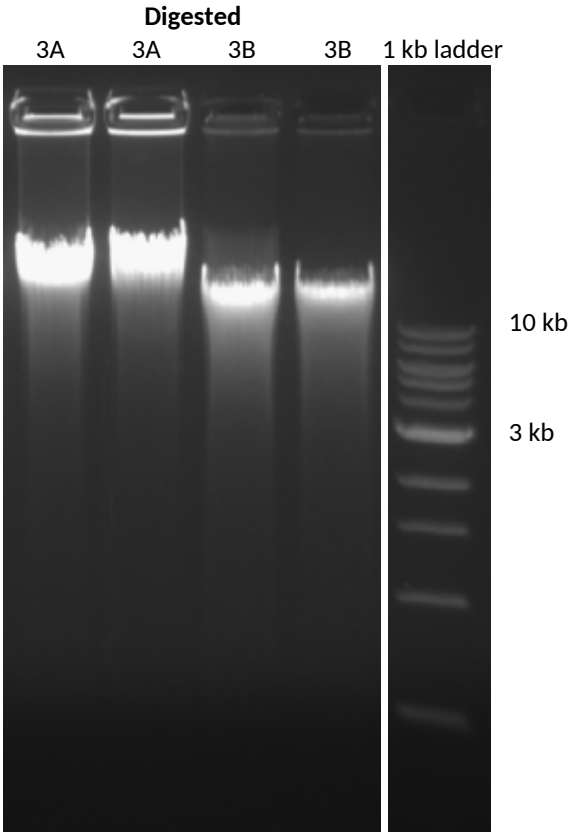
Experiment 1



Experiment 2



Experiment 3



Appendix B.

Detailed protocols for Hi-C 1.0 and Hi-C 2.0 variations tested with HindIII digestion. NEBNext library preparation adaptations detailed (page 10) as well.

This protocol is modeled on the Hi-C 2.0 protocol published by Belaghzal and Gibcus, 2017. The complete Hi-C protocol we used is included, with variations we tested for Hi-C 1.0 and Hi-C 2.0 noted. The 5 million cell aliquot used for the DpnII Hi-C 2.0 was processed according to the protocol below through step 2.3 and then exactly as described in Belaghzal and Gibcus, 2017.

Day 1. CROSSLINKING OF CELLS, CELL LYSIS, AND CHROMATIN DIGESTION

Grow the cells in appropriate culture medium. 5 million cells per Hi-C library are used. The degree of cell confluence, morphology, etc. should be documented carefully before each Hi-C experiment for future reference.

1. CROSSLINKING ADHERENT CELLS (~90 min):

Volumes below are for a single T-75 Plate
Prepare 10 mL 1x HBSS + protease inhibitor cocktail per plate, place on ice
1.1) Aspirate the medium, wash with 10 ml 1x HBSS (no serum) per plate.
1.2) Immediately before adding to the plate, mix 275 µl of 37% formaldehyde with 10 mL 1x HBSS to obtain 1% final formaldehyde concentration. Aspirate 1x HBSS wash from cell plate and add this pre-mixed 1% formaldehyde solution.
1.3) Incubate at room temperature (RT) for exactly 10 min on a gently shaking platform.
1.4) To quench the crosslinking reaction, add 554 µL of 2.5 M glycine, mix well
1.5) Incubate for 5 min at RT (rocking platform) and then incubate on ice for at least 15 min to stop crosslinking completely.
1.6) Aspirate liquid off of plates (discard as hazardous waste) and add 10 mL ice cold 1x HBSS + protease inhibitors
1.7) Scrape the cells from the plates with a cell scraper and transfer to a 50 ml tube.
1.8) Centrifuge aliquots of 5 million cells each at 800x*g* for 10 min, 4°C. Check to make sure all cells are pelleted. If there are visible floating cells, spin again.
1.9) Discard the supernatant by aspiration.
1.10) Cells can be snap-frozen in liquid nitrogen and stored at -80°C for at least 1.5 years or one can continue with cell lysis.

2. CELL LYSIS AND CHROMATIN DIGESTION (WITH HINDIII) (~75 min):

2.1) Resuspend one crosslinked cell aliquot (~$5x10^6$cells) in 1 ml of ice-cold lysis buffer (see Belaghzal et al., 2017 or Belton et al., 2012 for buffer formulations) containing 10 µl protease inhibitor cocktail (100x, Thermo)
    2.1.1) Incubate on ice for 15 min.
2.2) Dounce homogenize the cells on ice with pestle A.
    2.2.1) Slowly move pestle up and down 30 times
    2.2.2) Incubate on ice 1 min to let the cells cool down
    2.2.3) Then do 30 more strokes.
    2.2.4) Image lysed cells under the microscope to check for intact nuclei and document cell appearance
2.3) Transfer the lysate to a 1.7ml tube
2.4) Centrifuge for 5 minutes at (Hi-C 1.0: 2,000xg /Hi-C 2.0: 2,500x*g*) at RT to pellet nuclei.
2.5) Discard the supernatant and then wash the pellet twice by resuspending it in 500 µl of ice cold 1x CutSmart Buffer and then centrifuging the sample for 5 min at (Hi-C 1.0: 2,000xg /Hi-C 2.0: 2,500x*g*).
2.6) Resuspend the pellet in 1x CutSmart Buffer, so that the total volume of the suspension is 360 uL. (Add 100 uL of buffer first, check total volume, and then add remaining amount necessary to get 360 uL total)

Save 18 µl of lysate as a chromatin integrity control:

Add 50 µl of 1x CutSmart Buffer and 10 µl of Proteinase K (10 mg/ml). Incubate for 30 minutes at 65°C. Purify DNA by single phenol-chloroform extraction without ethanol precipitation. Add 1µl of RNAseA (1 mg/ml) to the aqueous phase and incubate for 15 min at 37°C. Store at -20°C overnight and run on a 0.75% agarose gel at the same time as the digested DNA sample (see Day 2 below). The sample is good if DNA is either stuck in the well or runs as a single high molecular weight band (>23 kb)

After taking 18µl *lysate control we have 342 µl* remaining *total volume.*

2.7) Add 38 µl of 1% SDS to the Hi-C tube (*380 µl total).* Mix carefully by pipetting up and down. Avoid making bubbles (0.1% SDS final).
2.8) Incubate at 65°C for 10 minutes exactly to open chromatin
2.9) Place tubes on ice immediately after
2.10) Add 43 µl of 10% Triton X-100 to the Hi-C-tube (*423 µl total*) to quench the SDS (1% Triton final). Mix gently by pipetting up and down Avoid making bubbles.
2.11) Add 12 µl of 10x CutSmart Buffer to the Hi-C tube to compensate for added components (*435 µl total before restriction enzyme*)
2.12) Add 400U (4 µl of 100,000 units/ml) HF HindIII to the Hi-C tube (*439 µl total*) Mix gently.
2.13) Digest the chromatin overnight at 37°C on a Nutator.

Day 2. BIOTINYLATION, LIGATION, and CROSSLINK REVERSAL

DNA INTEGRITY / DIGESTION CHECK

Incubate tube at 65˚C for 20 mins in order to deactivate the endonuclease enzyme.

After this incubation put the tubes on ice until cooled to room temperature

This inactivation depends on the enzyme used. Please check enzyme datasheet

Save 10 µl of lysate as a digestion control *(429 uL left; replace with 10 uL 1x CutSmart Buffer)*

Add 50 µl of 1x CutSmart Buffer and 10 µl of Proteinase K (10 mg/ml). Incubate for 30 minutes at 65°C. Purify DNA by single phenol-chloroform extraction without ethanol precipitation. Add 1µl of RNAseA (1 mg/ml) to the aqueous phase and incubate for 15 min at 37°C.

Check undigested and digested aliquots by running on a 0.75% gel, 75-100 V for 1 h.
(while gel is running, proceed with preparing biotinylation mastermix with everything but expensive Klenow and biotin reagents. If gel results look good, immediately proceed with biotinylation)

3. BIOTIN FILL-IN

3.3) Prepare a Biotin Fill-in MasterMix as follows:

FILL-IN MIX (ADD IN THIS ORDER)	PER TUBE
milliQ water	2 uL
10x CutSmart Buffer	6 µl
10 mM dATP	1.5 µl
10 mM dGTP	1.5 µl
10 mM dTTP	1.5 µl
0.4 mM biotin-14-dCTP	37.5 µl
5U/ µl Klenow	10.0 µl
TOTAL	60 µL

3.4) Add 60 µl of the Fill-in master mix to the digested Hi-C chromatin *(should be 500 µl total; if some volume decrease occurs during digestion step above, bring total volume up to 500 uL with 1x CutSmart Buffer).*

3.5) Mix gently by pipetting up and down without producing any bubbles

3.6) Incubate the tubes at (Hi-C 1.0: 37°C for 2 hours / Hi-C 2.0: 23°C for 4 hours) in a ThermoMixer (900 RPM mixing; 10 sec every 5 min)

3.6.1) Prepare ligation mixture (below) during incubation.
3.6.2) Place on ice immediately after incubation

4. BLUNT END DNA LIGATION

4.1.1) Prepare the 665 µl ligation mix during biotinylation:

LIGATION MIX	PER TUBE
Hi-C 1.0: 10x ligation buffer [NEB] Hi-C 2.0: 5x ligation buffer [Invitrogen]	120 µl 240 µl
10% Triton X-100	120 µl
10 mg/ml BSA	12 µl
T4 DNA ligase [Invitrogen]	Hi-C 1.0: 10 ul (5 U/uL) Hi-C 2.0: 50 µl (1 U/uL)
Water	To 700 µl
TOTAL	700 µL

Note: we have tried the Hi-C 2.0 protocol with both a total of 50 U and a total of 250 U of ligase and find that additional ligase appears neither to help nor hurt the outcome.

4.2) Add 700 µl of the mix to the Hi-C tube (*1,200 µl total*).
4.3) Incubate all tubes at 16°C for 4 hours in ThermoMixer with interval shaking (900 RPM mixing, 10 sec every 30 min).

5. REVERSE CROSSLINKING

5.1) Add 25 µl of 20 mg/mL proteinase K to the ligated Hi-C tube and incubate at 65°C overnight in the ThermoMixer
5.2) Add another 25 µl of proteinase K to the Hi-C tube and incubating an additional 2 hours (*1,250 µl total Hi-C)*

Day 3. DNA PURIFICATION

6. DNA PURIFICATION

6.1) Remove tubes from 65 C and allow to cool down.
6.2) Add 2.4 ml saturated phenol:chloroform (1:1, pH=8.0) to 2 x 15 ml tubes (label Hi-C PC1 and Hi-C PC2) for Hi-C
Use fume hood, appropriate gloves, and labcoat when handling phenol:chloroform

6.3) Extract the DNA:
6.3.1) Transfer Hi-C DNA into Hi-C PC1 tube
6.3.2) Vortex for 30 seconds to obtain a homogenous, milky solution
6.3.3) Centrifuge at 1500xg for 5 minutes (room temperature)
6.3.4) Carefully transfer the aqueous phase to Hi-C PC2. Repeat vortex and spin.

6.4) Transfer the aqueous phase with Hi-C DNA (*~1.2 ml*) to a clean 15 ml tube (adequate for high-speed centrifugation at 16,000 xg)
6.4.1) Add 1x TLE (pH=8.0) up to 2 ml.
6.4.2) Add 1/10 volume of 3M sodium acetate, pH=5.2 (*200µl*) and vortex briefly.
6.4.3) Add 2.5 volumes of ice-cold 100% ethanol (*5 ml*) and mix well by inverting the tubes several times.

6.5) Incubate the tubes at -80˚C for 45 min – 1 h

6.6) Centrifuge tubes at 16,000x*g* for 30 min at 4°C.

6.7) Decant the supernatant with extra caution as the pellet detaches from the tube wall quite easily and air dry the DNA pellets very briefly (~1 min).

6.8.1) The pellet might become invisible while drying out

6.8) Dissolve Hi-C pellet in 450µl of 1x TLE (pH=8.0) and transfer to 2 mL, 30kD Amicon columns.

6.9) Wash the column at least 3 times with the following steps:

6.10.1) Centrifuge at 14,000xg in tabletop centrifuge for ~5 minutes (room temperature)

6.10.2) Remove flowthrough (to a temporary save tube). DNA stays in the volume left in the column.

6.10.3) Add 450 µl TLE

6.10) On the last spin, spin for 10 minutes instead of 5 minutes at 14,000xg

6.11) After the last wash, flip the column onto a new Amicon container tube and spin at 1,000xg for 2 min to recover the DNA from the column

6.12) Add 1 µl of RNAseA (1 mg/ml) to the sample and incubate for 30 mins at 37°C.

DNA samples can now be stored at -20°C.

7. QUALITY CONTROL OF HI-C LIBRARIES

7.1) Make a 1:10 dilution of the Hi-C library and run 2µl and 6 µl on a 0.8% agarose gel for quality control at 100V for 1 hr.

7.2) Quantify amount of DNA by densitometry. Use several dilutions of the NEB 1kb DNA ladder as a standard curve to estimate the DNA concentration more accurately.

7.3) Measure the Biotin Incorporation by a PCR digest:

7.3.1) Make 1:20 dilution of Hi-C DNA (2.5 uL Hi-C + 47.5 uL H2O)

7.3.2) Dilute primers to 10 uM in TLE

7.3.4) Perform a PCR to amplify one neighboring interaction as follows:

INGREDIENT	1 REACTION
Diluted Template DNA	20 uL
MILLIQ	33.9 uL
10X PCR buffer	7.5 µl
25mM dNTP	0.6 µl
50 mM MgSO4	6.0 uL
Primer 1 (10 µM)	3 µl
Primer 2 (10 µM)	3 µl
Taq DNA polymerase	1 µl
Total	75 uL

7.4) Combine 20 uL diluted Hi-C DNA or water (neg ctrl) with 55 uL mastermix in a PCR tube.

7.5) Run the following PCR program:

Step	Temperature	Time
1	95°C	5 minutes
2	95°C	30 seconds
3	65°C	30 seconds
4	72°C	30 seconds
5	Go to 2	34 times
6	95°C	30 seconds

7	65°C	30 seconds
8	72°C	8 minutes

7.6) Run the samples in a 2% gel at 160 V for 45 min.

7.7) Digest the HiC PCR product with HindIII and NheI as follows:

	Control (µl)	HindIII (µl)	NheI (µl)	HindIII/ NheI (µl)
PCR product	16.0	16.0	16.0	16.0
10X Cutsmart buffer	2.0	2.0	2.0	2.0
HindIII	---	1	---	1
NheI	---	---	1	1
MilliQ	2	1	1	---

7.8) Incubate samples at 37°C for at least 30 minutes:

7.8.1) Combine each with 4 uL 6x dye and load 17 uL on the top row of wells in the 2% gel. Run 1 hr at 175 V.

Successful biotin fill-in of HindIII sites (AAGCTT) followed by blunt-end ligation create sites for the NheI restriction enzyme (GCTAGC).
Successful biotin fill-in of DpnII sites (GATC) followed by blunt-end ligation create sites for ClaI restriction enzyme (ATCGAT).

8. REMOVAL OF BIOTIN FROM UN-LIGATED ENDS

8.1) Quantify DNA from Hi-C ligated library gel (step 7.2). Per 5 ug of DNA recovered, prepare reactions as follows:

INGREDIENT	Per 5 ug
Hi-C DNA sample	**5 µg**
10X NEBuffer 2.1	5 µl
1 mM dATP	1.25 µl
1 mM dGTP	1.25 µl
3,000 U/ml T4 DNA polymerase	5 µl
MILLIQ water	TO 50 µL

8.1.1) Place each 50 µl reaction in a different well of a PCR strip
8.1.2) Incubate at 20°C for 4 hours in the PCR machine.

8.2) (Hi-C 1.0: Add 2 µl of 0.5M EDTA to each tube to stop the reaction ; Hi-C 2.0: Inactivate the enzyme for 20 mins at 75°C)

8.3) Cool down/ keep at 4°C

8.4) Pool the reactions together and add water up to 500 uL. Perform an Amicon wash:

8.4.1) Add volume to Amicon column and centrifuge at 14,000xg for ~ 5 min
8.4.2) Wash twice with 400 µl of milliQ water (spinning 5 min at 14k x g each time)
8.4.3) Invert the Amicon column onto a fresh tube and centrifuge for 2 minutes at 1,000xg to collect the DNA
Note that Amicon concentration does not remove proteins (they are only inactivated by the heat or EDTA treatment above, and then also by sonication below). Concentration by volume reduction may result in DNA coming out of solution. If this occurs, sonication below will help re-solubilize DNA.

8.5) Add milliQ to a total of 132 µl. Save 2 uL for pre-sonicated gel check and transfer the rest to a Covaris microtube for sonication.

9. DNA SONICATION

9.1) Shear the DNA to a size of 200 – 400 bp using the Covaris M220 sonicator and the following parameters:

- 20% Duty Factor, 200 cycles/burst, 50 W Peak Incident Power for 110 seconds at 4 degrees C.

(note that this differs from the Hi-C 2.0 recommended settings for reasons discussed in the manuscript)

9.2) Check the results of sonication on a 2% gel. Run the gel at 160 V for 45 min.

10. SIZE FRACTIONATION USING AMPURE XP

This step can be omitted for low concentrations of DNA and sonicated DNA can be taken directly to step 11.

The liquid phase of AMpure mixture precipitates DNA onto the AMpure XP magnetic beads. The size of the DNA molecules precipitated depends on the ratio of the volume of AMpure XP liquid (PEG8000) to the volume of the sample when mixed together. Increasing the proportion of AMpure XP liquid decreases the cutoff of the size of the DNA molecules precipitated onto the beads. Here, the 0.7x Ampure step removes fragments above ~400 bp and the 1.2x Ampure captures the remaining fragments between 100-400 bp.
Note that the 0.7x and 1.2x recommendation differs from Hi-C 2.0 in order to capture the larger fragments coming out of sonication conditions above.

10.1) Bring volume of Hi-C sample up to 500ul with 1x TLE

10.2) Allow AMpure XP mixture to come to RT and vortex prior to use. Add 350 µl of AMpure XP mixture to the Hi-C tube. Label the tube as 0.7X *(the ratio of AMpure XP solution volume to Hi-C sample volume is 350/500 = 0.7).*

10.3) Vortex and spin down tubes briefly.

10.4) Incubate tubes for 10 min at RT.

10.5) Place tubes on the Magnetic Particle Separator (MPS) for 5 min at RT.

10.5.1) During above incubation, prepare a fresh tube with 500 µl AMpure XP mixture; label them as 1.2X.

10.5.2) Incubate tubes on the MPS for 5 min

10.5.3) Remove supernatant

10.5.4) Resuspend beads in 250 µl AMpure mixture

This step increases the number of beads present in a smaller volume of AMpure XP mixture. This prevents saturation of the beads with DNA, ensuring an efficient precipitation of the DNA.

10.6) Collect the 0.7X supernatants from MPS

10.6.1) Add the supernatants to the 1.2X tubes.

The ratio of AMpure XP solution volume to the original sample volume is now: (250 + 350)/500 = 1.2. Under these conditions beads bind DNA fragments >100 bp.

10.6.2) Vortex and spin down tubes briefly.

10.6.3) Incubate 1.2X tubes for 10 min at RT

10.6.4) Place 1.2X tubes on the MPS for 5 min at RT

10.7) Discard supernatant from the 1.2X tubes.

10.8) Wash the beads in 0.7X and 1.2X tubes twice with 1 ml fresh 70% ethanol, reclaiming beads against the MPS for 5 min each time.

10.9) Air dry beads on the MPS briefly (too much drying may decrease elution efficiency)

10.10) Resuspend the 0.7X and 1.2X beads in 150 µl of 1x TLE buffer in each tube to elute the DNA

10.10.1) Incubate 10 min at RT

10.10.2) Separate AMpure beads from eluate for both 0.7X and 1.2X tubes on the MPS for 5 min

10.10.3) Keep the eluate

10.10.4) For the gel, combine:

7 uL 6x dye + 10 uL 1.2x eluate + 10 uL water

2 uL 6x dye + 4.5 uL 0.7x eluate + 4.5 uL water

10.11) Desalt and concentrate the Hi-C sample on Amicon:

10.11.1) Centrifuge at 14,000xg for ~5 min.

10.11.2) Add 450 µl of 1x TLE buffer

10.11.3) Centrifuge again until at 15,000xg for about 5~10min until ~30 µl remains

10.12) Place the Amicon unit upside down onto a new tube provided with the Amicon kit. Collect Hi-C sample by spinning at 1,000xg for 2 min

10.13) Bring volume of the sample to 50 µl with milliQ water

10.14) Check the quality and estimate the quantity of DNA on a 2% agarose gel. Run one lane of the 0.7X sample and a dilution series of the 1.2X sample along with a low molecular weight DNA marker. Quantify the amount of DNA in the 1.2X sample and compare to sonication gel to evaluate degree of DNA loss due to Ampure.

NEBNext Protocol Adapted for Hi-C library Preparation on Beads

At this step, the Hi-C 2.0 protocol recommends end repair and A tailing, then bead pulldown, followed by Illumina adapter ligation. We have adapted the entire NEBNext protocol to work after bead pulldown.

11. BIOTIN PULLDOWN WITH STREPTAVIDIN COATED BEADS

11.1) Vortex the MyOne™ Streptavidin C1 beads and then transfer 20 µl of beads to a 1.7ml LoBind tube.

11.2) Wash the beads with 400 µl of Tween wash buffer (TWB) by pipetting up and down and incubating for 3 min at RT on a rocking platform.
11.3) Reclaim beads against the MPS for 1 min, discard the supernatant.

11.4) Resuspend beads in 400 µl of TWB and transfer to a new LoBind tube. Incubate 3 min again.
11.5) Reclaim beads against the MPS for 1 min, discard the supernatant.

11.6) Resuspend beads in 400 µl of 2X Binding Buffer (BB) and add the DNA (in 400 µl TLE buffer)
11.7) Incubate the sample for 15 min at RT with rotation.
11.8) Reclaim the beads against the MPS for 1 min; discard the supernatant to a "BB sup" save tube.

11.9) Resuspend the beads in 400 µl of 1X BB and transfer them to a new tube
11.10) Reclaim the beads against the MPS for 1 min, discard the supernatant.

11.11) Wash beads with 100 µl of TLE and transfer to a new tube
11.12) Reclaim the beads against the MPS for 1 minute; discard the supernatant

11.13) Finally resuspend the beads in 50 uL TLE buffer.

12. NEBNext End Prep

12.1. Add the following components to a sterile nuclease-free tube:

(green) NEBNext Ultra II End Prep Enzyme Mix	3 µl
(green) NEBNext Ultra II End Prep Reaction Buffer	7 µl
Fragmented DNA	50 µl
Total Volume	60 µl

12.2. Set a 100 µl or 200 µl pipette to 50 µl and then gently pipette the entire volume up and down at least 10 times to mix throughly. Perform a quick spin to collect all liquid from the sides of the tube.

Note: It is important to mix well. The presence of a small amount of bubbles will not interfere with performance.

12.3. Place in a thermocycler, with the heated lid set to ≥ 75°C, and run the following program:
30 minutes @ 20°C
30 minutes @ 65°C
Hold at 4°C

If necessary, samples can be stored at -20°C; however, a slight loss in yield (~20%) may be observed. We recommend continuing with adaptor ligation before stopping.

13. Adaptor Ligation

13.1. Perform a 1:15 or 1:20 dilution (or more if a first attempt gives excess adaptors) of the NEBNext Adaptor for Illumina in 10 mM Tris-HCI, pH 8.0 with 10mM NaCl.

13.2. Add the following components directly to the End Prep Reaction Mixture:

End Prep Reaction Mixture (Step 12.3)	60 µl
(red) NEBNext Ultra II Ligation Master Mix*	30 µl
(red) NEBNext Ligation Enhancer	1 µl
(red) NEBNext Adaptor for Illumina **	2.5 µl
Total volume	93.5 µl

* Mix the Ultra II Ligation Master Mix by pipetting up and down several times prior to adding to the reaction.
Note: The Ligation Master Mix and Ligation Enhancer can be mixed ahead of time and is stable for at least 8 hours @ 4°C. We do not recommend adding adaptor to a premix in the Adaptor Ligation Step.

13.3. Set a 100 µl or 200 µl pipette to 80 µl and then pipette the entire volume up and down at least 10 times to mix thoroughly. Perform a quick spin to collect all liquid from the sides of the tube.

(Caution: The NEBNext Ultra II Ligation Master Mix is very viscous. Care should be taken to ensure adequate mixing of the ligation reaction, as incomplete mixing will result in reduced ligation efficiency. The presence of a small amount of bubbles will not interfere with performance).

13.4. Incubate at 20°C for 15 minutes in a thermocycler with the heated lid off.

13.5. Add 3 µl of (red) USER™ Enzyme to the ligation mixture from Step 13.4.

Note: Steps 13.5 and 13.6 are only required for use with NEBNext Adaptors. USER enzyme can be found in the NEBNext Singleplex (NEB #E7350) or Multiplex (NEB #E7335, #E7500, #E7600, #E7710, #E7730 and #E6609) Oligos for Illumina.

13.6. Mix well and incubate at 37°C for 15 minutes with the heated lid set to ≥ 47°C.

Note: Samples can be stored overnight at -20°C.

13.7) Reclaim the beads against the MPS for 1 min. Discard the supernatant.
13.8) Wash the beads twice with 400 µl of TWB. Add buffer to the beads, mix carefully by pipetting and incubate at RT for 5 min with rotation.
13.9) Reclaim the beads against the MPS for 1 min. Discard the supernatant. (to a save tube)
13.10) Resuspend the beads in 200 µl of 1X BB and transfer to a new tube
13.11) Reclaim the beads against the MPS for 1 minute. Discard the supernatant.
13.12) Wash the beads twice with 200 µl TLE. Resuspend beads then transfer to a new tube
13.13) Reclaim the beads against the MPS for 1 min. Discard the supernatant.
13.14) After the last wash, resuspend the beads in 20 µl of TLE and transfer to a new tube.

Note: Samples can be stored at -20°C.

14. PCR Enrichment of Adaptor-ligated DNA

Note: Check and verify that the concentration of your oligos is 10 µM.

14.1. PCR TITRATION and PRODUCTION PCR

14.1.1) Use the following HiCPCR_6CYC and HiCPCR_3CYC programs for PCRs below:

HiCPCR_6CYC: (~15 min)
• Preheat lid at 100°C
— 98°C for 30 seconds
• 6 cycles of:
— 98°C or 10 seconds
— 65°C for 75 seconds
• 65°C for 5 minutes
**Hold at 4°C

HiCPCR_3CYC: (~10 min)
• Preheat lid at 100°C
— 98°C for 30 seconds
• 3 cycles of:
— 98°C or 10 seconds
— 65°C for 75 second
• 65°C for 5 minutes

14.1.2. Add the following components to a sterile strip tube:

14.1.2A. **Forward and Reverse Primer not already combined**

Adaptor Ligated DNA Fragments (Step 13.14)	3 µl
(blue) NEBNext Ultra II Q5 Master Mix	12.5 µl
(blue) Index Primer/i7 Primer*,**	2.5 µl
(blue) Universal PCR Primer/i5 Primer*, ***	2.5 µl
Add water to Total Volume of:	25 µl

* The primers are provided in NEBNext Singleplex (NEB #E7350) or Multiplex (NEB #E7335, #E7500, #E7710, #E7730, #E7600) Oligos for Illumina. For use with Dual Index Primers (NEB #E7600), look at the NEB #E7600 manual for valid barcode combinations and tips for setting up PCR reactions.
** For use with NEBNext Multiplex Oligos (NEB #E7335 or #E7500, #E7710, #E7730) use only one index primer per PCR reaction. For use with Dual Index Primers (NEB #E7600) use only one i7 primer per reaction.
*** For use with Dual Index Primers (NEB #E7600) use only one i5 Primer per reaction.

14.1.2. Set a 100 µl or 200 µl pipette to 40 µl and then pipette the entire volume up and down at least 10 times to mix thoroughly. Perform a quick spin to collect all liquid from the sides of the tube.

14.1.3) Run the following PCR programs. (Note that this must be run when you have time to take aliquots every ~10 min for 1 hour)
Set up 5 small tubes for gel loading labeled 6, 9, 12, 15, 18 and combine 2 uL 6x dye + 7 uL water in each.
After each PCR below, take out tube in the last 10 seconds of the 65 C final step, put tube directly on ice, and remove a 3 uL aliquot. Combine each aliquot with dye and water in tubes prepared above.
1) HiCPCR_6CYC (total 6 cycles)
2) HiCPCR_3CYC (total 9 cycles)
3) HiCPCR_3CYC (total 12 cycles)
4) HiCPCR_3CYC (total 15 cycles)
5) HiCPCR_3CYC (total 18 cycles)

14.1.4) Run the titration PCR on a 2% gel. Run the gel at 160 V for 45 min

14.1.5) Build a calibration curve showing the DNA quantity vs. number of cycles.
Choose an optimal number of cycles and number of PCR reactions for the final amplification of the library for deep sequencing. The cycle number should be chosen so that the PCR amplification is in the linear range and the expected size distribution is preserved (over cycling will shift the size distribution of the library towards higher molecular weight products). The number of PCR reactions should be calculated to produce the amount of DNA desired for sequencing (usually 50-100 ng). It is always better to reduce the number of PCR cycles while increasing the number of PCR reactions. The optimal amount is usually 1-2 cycles below the lowest amount visible on the gel.

Note: One can also use the KAPA Library Quantification kit that uses a qPCR system to accurately estimate the concentration of the library and number of cycles for the final amplification of the library for deep sequencing.

15. PRODUCTION PCR

15.1) Decision from titration: Perform _____ PCR reactions (maximum of 5) with _____ cycles.

15.2) Set up N PCR master mix reactions on ice as follows:

Adaptor Ligated DNA Fragments (Step 3.1.14 or 3.2.11)	3 µl
(blue) NEBNext Ultra II Q5 Master Mix	12.5 µl
(blue) Index Primer/i7 Primer*,**	2.5 µl
(blue) Universal PCR Primer/i5 Primer*, ***	2.5 µl
Add water to Total Volume of:	25 µl

15.3) Run the following PCR program with the selected number of cycles.

• Preheat lid at 100°C
— 98°C for 30 seconds
• N cycles of:
— 98°C or 10 seconds
— 65°C for 75 seconds
• 65°C for 5 minutes

15.4) Pool all PCR reactions together in a 1.7 mL Lo-Bind Tube and incubate on MPS for 2 min.

14.10.1) Remove supernatant to a new tube (this contains amplified DNA)

14.10.2) Resuspend beads in N*3 uL TLE and freeze at -20 in case of future need.

14.10.3) Take a 3 uL "pre-purified" aliquot of supernatant + 2 uL dye + 7 uL H2O for gel check.

15.6) To remove primer dimers, purify the amplified Hi-C library from the supernatant using AMpure XP beads.

15.6.1) Allow AMpure XP mixture to come to RT and vortex prior the use.

15.6.2) Add 1.5x volumes (_____ uL) of AMpure XP mixture to the supernatant.

Note: if the Bioanalyzer trace at the end of step 16 indicates high amounts of adapter-dimer or primer-dimer, re-purifying with 1.3x or 1.2x Ampure to remove those products is recommended.

15.6.3) Vortex and spin down briefly.

15.6.4) Incubate for 10 min at RT.

15.6.5) Place on the MPS for 5 min at RT. Discard supernatant to a save tube.

15.6.6) Wash the beads twice with 1 ml of freshly made 70% ethanol

15.6.7) Air-dry the beads briefly and then resuspend in 35 µl of TLE buffer

15.6.8) Incubate the beads for 10 min at RT, tapping the tube every 1-2 min.

15.6.9) Collect the beads with the MPS for 5 min

15.6.10) Transfer the supernatant, containing the final Hi-C library, to a new tube.

15.7) Check final HiC library on a 2% agarose gel at 160 V for 45 min.

16. QUALITY CONTROL OF HI-C LIBRARY BY RESTRICTION DIGEST

16.1) Digest a small aliquot of the final Hi-C library with NheI to estimate the portion of molecules with valid biotinylated junctions.

INGREDIENT	1X
10x NEB Cutsmart buffer	1 µl
milliQ water	6.5 uL
NheI-HF	1 µl
Hi-C product	1.5 uL
TOTAL	10 µL

16.2) Allow to digest at 37°C for at least 1 hour

16.3) Combine entire volume of digestion reaction + 2 uL dye and 1.5 uL uncut Hi-C DNA + 8.5 uL H2O + 2 uL 6x dye and run on a 2% agarose gel at 160 V for 45 min

16.4) Quantify approximate % of uncut DNA (unshifted smear) from the gel or a Bioanalyzer analysis. This correlates well to the fraction of non-ligation products (dangling ends) that will be obtained in the final sequencing

Run 1 uL of final Hi-C library on Agilent DNA Bioanalyzer. Dilute an aliquot of the library to 10 nM in 10 uL with Tris-HCl 10 mM, pH 8.5 + 0.1% Tween 20 for later pooling and sequencing.

Buffer formulations for Hi-C

2.5M Glycine
46.92 g Glycine in 250 mL dH_20
Autoclave

Lysis Buffer
10 mM Tris-HCl pH 8.0
10 mM NaCl
0.2% NP-40 (Igepal CA-630)
In MilliQ H_2O
Filter sterilize
Keep at 4°C

10% SDS
50 g SDS in 500 mL dH_20
Filter sterilize

10% Triton X-100
1 mL 100% TX-100 + 9 mL dH_20

10mg/mL BSA
100 mg BSA
Dissolve in 10 mL of H_2O
Aliquot and store at -20°C

Phenol Chloroform (1:1)
Add Tris buffer to Phenol, allow to stand overnight at 4 degrees
Add 100 mL Phenol to 100 mL Chloroform + some buffer from top of phenol
Allow to stand for several days at 4 degrees

1x TLE
10 mM Tris
0.1 mM EDTA
Add sterile H_20 up to 100 mL

10X PCR Buffer
600 mM Tris ph 8.9 with H_2SO_4,
Adjust to pH 8.9 with H_2SO_4 (make prior to mixing here)
180 mM $(NH_4)_2SO_4$

Tween Wash Buffer
5 mM Tris-HCl pH8.0
0.5 mM EDTA
1 M NaCl
0.05% Tween
In MilliQ H_2O
Filter Sterilize

2x Binding Buffer
10 mM Tris-HCl pH8.0
1 mM EDTA
2 M NaCl
In sterile H_2O
Filter sterilize

Appendix C. Further Data Describing and Interpreting Sequence Copying Artifact (Related to Figure 3)

The artifactual copying of a piece of the 5' end of a sequencing library molecule onto the 3' end described in Figure 3 is not only observed after Illumina sequencing, but can be detected when a molecule of a Hi-C library is cloned and sequenced. In the example below, the sequence highlighted in blue aligns to the genome as part of the 5' end fragment, but was artificially copied as a reverse complement onto the 3' end fragment. Thus, the 3' end fragment sequence does not fully align to the genome. A 4 base pair microhomology between the two ligated fragments lies just before the non-mapping copied sequence on the 3' end, supporting the idea that this copying occurs due to improper annealing and extension from single stranded microhomologies such as this.

Cloned and sanger sequenced Hi-C library molecule showing copying artifact:

5'-AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTCTACCTCTATCATTCACT
AGCTATAAA**(ATCT)**TAGGCAGATTATTTCCTCATTTTTAAAATGTAGGAAATGGATAATGTACTTAAGGTTCCCAGT
TTCATGCGAGGCACATAGTAAGCACTTAATAGAATATCATGGTTCTTTGAATCTTCAAGTTCTATCATACTGACTAA
AATACAGAGCATGCTAGCAAGCTAGCTTTAAATACAACCATGATGACAATCAAGATTGACCCTAGGACCAGCAGC
TGAAAGAAATTAGACTGAACTTTAAAGAAATCAGACCGT**(AGAT)**TTTATAGCTAGTGAATGATAGAGGTAGAGAT
CGGAAGAGCGGTTCAGCAGGAATGCCGAGACCGATCTCGTATGCCGTCTTCTGCTTG**-3'**

red = Illumina adapters
blue = sequence copied from 5' to 3' end
green = HindIII junction
parentheses = 4 bp microhomology between the two fragments

Results of UCSC BLAT with 5' end fragment and 3' end fragment to hg38:

Fragment 5' of junction

```
CTACCTCTAT CATTCACTAG CTATAAAATC TTAGGCAGAT TATTTCCTCA  50
TTTTTAAAAT GTAGGAAATG GATAATGTAC TTAAGGTTCC CAGTTTCATG  100
CGAGGCACAT AGTAAGCACT TAATAGAATA TCATGGTTCT TTGAATCTTC  150
AAGTTCTATC ATACTGACTA AAATACAGAG CATGCTAGCA AGCT
```

Genomic chr1 (reverse strand):

```
caagtgaaac tgaaccacta aaataccctc tgcccagtgt gtctcatgct  213103540
tatggtgaag tgtctatttc acgataacta aaaggacctg aattcaaata  213103490
CTACCTCTAT CATTCACTAG CTATAAAATC TTAGGCAGAT TATTTCCTCA  213103440
TTTTTAAAAT GTAGGAAATG GATAATGTAC TTAAGGTTCC CAGTTTCATG  213103390
CGAGGCACAT AGTAAGCACT TAATAGAATA TCATGGTTCT TTGAATCTTC  213103340
AAGTTCTATC ATACTGACTA AAATACAGAG CATGCTAGCA AGCTtttatt  213103290
gactgtgtga gtgaataaca aaaactttta aaagccatgt caaaataaat  213103240
aaataaatgt cttacaatga aacattatca cagataatct tttt
```

Fragment 3' of junction
(note unmapped part is reverse complement of beginning of 5' fragment)

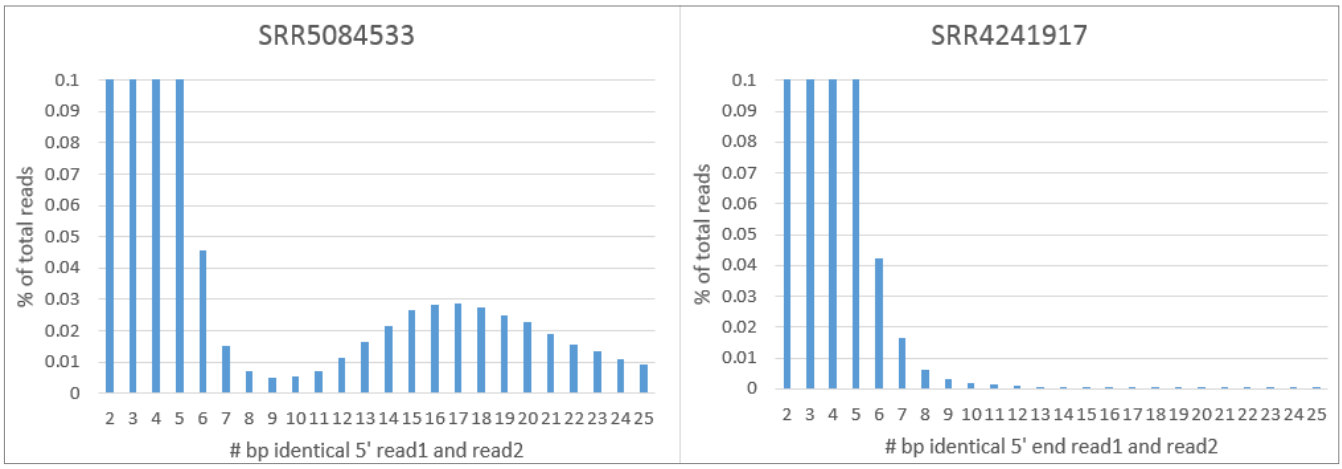
```
AGCTTTAAAT ACAACCATGA TGACAATCAA GATTGACCCT AGGACCAGCA  50
GCTGAAAGAA ATTAGACTGA ACTTTAAAGA AATCAGACCG TAGATtttat  100
agctagtgaa tgatagaggt ag
```

Genomic chr18 :

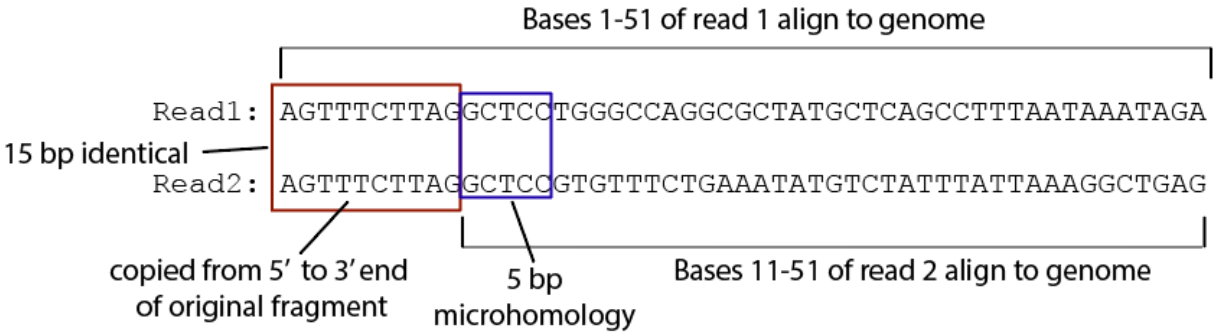
```
agaagcacct tccagaaata catatcagca aaaacaaatc acaaaggaaa  25703321
gacatcagat ttagttccga accacaccca gcacatgaat ggcacccaaa  25703371
AGCTTTAAAT ACAACCATGA TGACAATCAA GATTGACCCT AGGACCAGCA  25703421
GCTGAAAGAA ATTAGACTGA ACTTTAAAGA AATCAGACCG TAGATaaaaa  25703471
agaatttcaa taaaattatt aaacattatc caccaatact gctgtcagcc  25703521
acccatatgt agtccactga gttatctttc agttatttga tttct
```

Some ChIP-Seq Datasets Exhibit Copying Artifact

If the copying artifact indeed occurs during end repair, as proposed in Figure 3, then this artifact should be expected to occur on occasion in other types of sequencing experiments that employ this method of library preparation. Indeed, below, we show an example of a ChIP-Seq library (left; SRR5084533 from GEO Accession: GSE90992 [1]) in which a small, but detectable, proportion of reads show copying of the 5' end fragment (read1) onto the 3' end (read2) just as we observed in Hi-C.



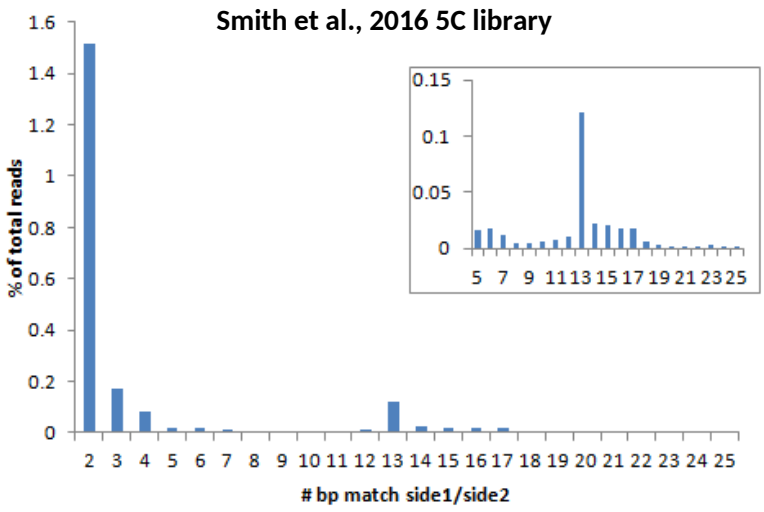
A sample set of paired reads from SRR5084533 shows that these identical bases follow the same pattern as observed in Hi-C libraries:



As with Hi-C, not all libraries prepared with End Repair show this copying artifact (for example, see SRR4241917, right; from GEO accession GSE86903 [2]). Both of these two libraries were prepared with sonication followed by end repair, A-tailing, and adapter ligation with NEBNext reagents. We hypothesize that, as with Hi-C libraries, variations in the extent of DNA single stranded overhangs and library size may influence whether this artifact is seen or not.

Lack of Artifact in Libraries that do not include End Repair and Adapter Ligation

As shown in Figure 3C, a Hi-C library prepared by tagmentation does not exhibit the copying artifact that we describe here. Further evidence that sequencing libraries prepared by methods other than end repair and adpater ligation are free from this copying artifact comes from analysis of a 5C (chromosome conformation capture carbon copy) library (from [3]). 5C adapters are added by PCR and so there is no sonication step, biotin removal step, or end-repair step. In the case shown below, the high number of 13 bp matched sequences contain the SOLiD P1 adapter sequence, demonstrating that this is an adapter-related artifact, not a copying artifact.

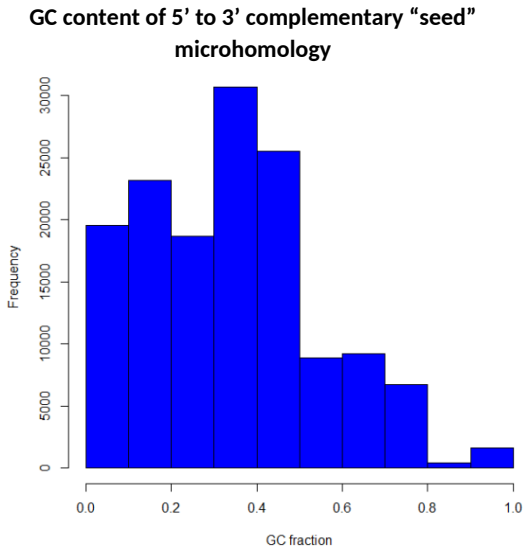


Analysis of Enriched Sequences in Overlap and Microhomologies

When raw FASTQ files are analyzed for sequences that match between read1 and read2 in the paired end set, untrimmed adapter sequence at both ends of the molecule will result in a 12 bp identity, as shown in Figure 3C and expected by the overlap in the TruSeq adapter sequences.

PE Adapter2: GATCGGAAGAGCGGTTCAGCAGGAATGCCGAG
PE Adapter1: GATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT

However, adapter sequence cannot account for the pieces of identical read1/read2 sequences, as described here, that align perfectly on read 1 but not on read2. Thus, we conclude that aberrant adaptor ligation or sequencing does not contribute to the artifact we observe in these cases.



There are no notable enriched sequence motifs among the microhomology seed sequences or copied regions, though we do notice that the short complementary sequences tend to be AT rich, as shown in the histogram at right. We hypothesize that the AT rich DNA is enriched because the A-T base pair is more fragile and likely to be damaged and result in single stranded overhangs after sonication.

Analysis Code Availability:

This copying artifact was analyzed using in house Perl scripts that processed Bowtie2 mapping Samfiles in comparison with original FASTQ files. For others wishing to assess this artifact in their own data, scripts are available upon request.

Appendix B References:

[1] V.B. Teif, J.P. Mallm, T. Sharma, D.B. Mark Welch, K. Rippe, R. Eils, J. Langowski, A.L. Olins, D.E. Olins, Nucleosome repositioning during differentiation of a human myeloid leukemia cell line, Nucleus 8(2) (2017) 188-204.
[2] F. von Meyenn, R.V. Berrens, S. Andrews, F. Santos, A.J. Collier, F. Krueger, R. Osorno, W. Dean, P.J. Rugg-Gunn, W. Reik, Comparative Principles of DNA Methylation Reprogramming during Human and Mouse In Vitro Primordial Germ Cell Specification, Dev Cell 39(1) (2016) 104-115.
[3] E.M. Smith, B.R. Lajoie, G. Jain, J. Dekker, Invariant TAD Boundaries Constrain Cell-Type-Specific Looping Interactions between Promoters and Distal Elements around the CFTR Locus, Am J Hum Genet 98(1) (2016) 185-201.