CALLING VALID READS IN RESTRICTION ENZYME FREE HI-C

Introduction

Chromosome conformation capture (3C), and its many derivatives (Hi-C, DNase Hi-C, Micro-C XL, Capture-C, HiChIP, etc.) are used to interrogate hierarchical three-dimensional chromatin structure on a genome scale. Integral to the technique, the following features are shared among these different methods (**Figure 1**):

- Fixation of chromatin structure to create a three-dimensional scaffold
- Chromatin fragmentation
- Ligation of free ends

Due to their relative simplicity, the use of restriction enzymes (RE) for the chromatin fragmentation step has been broadly adopted. During ligation, free ends, held in close proximity within the chromatin scaffold, are free to ligate – any ligation events yielding new chimeric sequences reflect 3-D chromatin architectural features.

Initially, processing the high-throughput sequence data from these assays was a heavily involved process that included mapping, characterizing, and filtering ligation events and insert sizes. The development of many open-source pipelines has aided in the wide-spread adoption of 3C methods by reducing the computational expertise required to use these data types.

A key step in processing 3C reads is filtering for valid ligation events (valid read-pairs). Two parameters define traditional valid read pairs [1,2]:

1. **RE site proximity**: Ligation events should only occur adjacent to the RE sites defined by the RE used for chromatin fragmentation
2. **Library insert size**: Ligation events in inserts larger than the longest library fragment (LLF) are only of interest, generally ~1 kbp

Such filtering minimizes the inclusion of non-Hi-C mediated ligation events in downstream analyses. Such non-valid events typically result from:

- Fragments created during shearing and library prep[3]
- RE cut ends that have re-ligated

These metrics are defined by the assay chemistry and may not always be relevant. In particular, the requirement that a valid ligation event occur proximal to a RE site is driven by the chromatin digestion method with REs (**Table 1, Figure 1A**).

Advances to 3C chemistry include the use of sequence-independent methods to digest chromatin, including DNase[4] or MNase[5] digestion. By not using a RE to fragment the chromatin, the definition of valid reads must change as requiring ligation events to be adjacent to a RE site is no longer applicable. Here we demonstrate that filtering valid reads

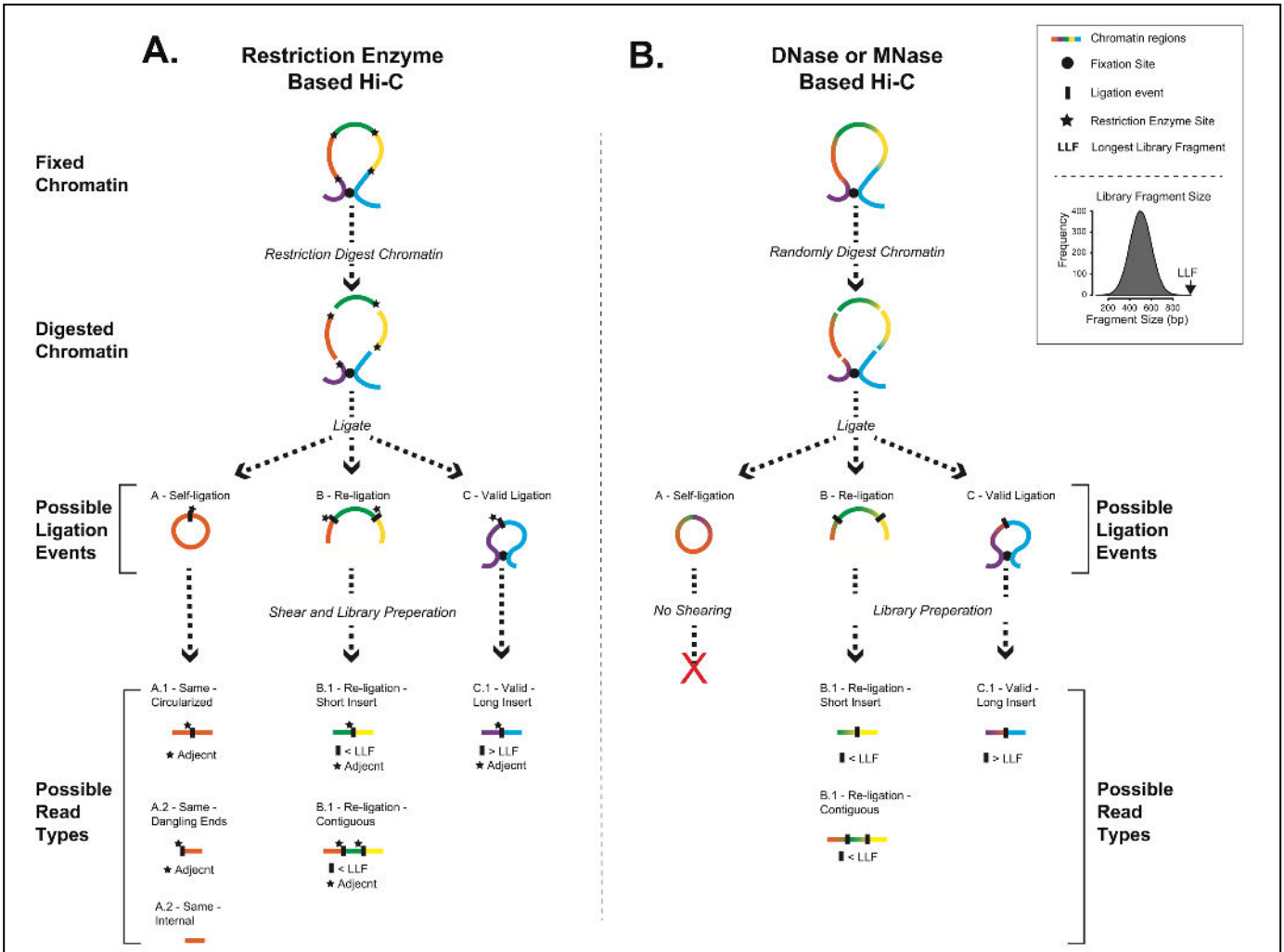


Figure 1 – **Schematic of read classifications for different chromatin digestion profiles**. The workflow is orientated vertically, starting with the fixed chromatin. Following digestion, possible ligation events and resulting read types are depicted. The RE-based (A) and DNase/MNase Hi-C (B) workflows are shown in parallel for direct comparison. Chromatin regions are denoted by different colors, the change in color either abrupt at a RE site (star) or blended to represent a sequence independent view of chromatin digestion. Ligation events are shown as a vertical black bar. The longest ligation fragment (LLF) is defined as the upper limit of the library size distribution as shown in the library fragment size histogram.

solely on insert size larger than the LLF[2] (**Table 1, Figure 1B**) is sufficient for processing Omni-C™ data, a DNase-based Hi-C assay. We describe the implementation of the open-source 3C data processing pipeline, Hi-C Pro, in a configuration compatible with RE free Hi-C approaches[2] using the DNase based library generated by the Dovetail Omni-C™ Kit.

Methods

To test the Hi-C Pro pipeline on RE-free Hi-C data, a DNase-based Hi-C library was generated

Table 1 – **Read type classification applicable to Hi-C chemistries**. Ligation type and resulting read type are listed (also illustrated in **Figure 1**). Each read type is denoted as applicable (checkmark) or not (dash) for different approaches to chromatin digestion. Definitions for each read type[1,2] are described for RE-based Hi-C, and a description of how this differs for DNase or MNase Hi-C is delineated[2].

Ligation Type	Figure ID	Read Type	RE	DNase or MNase	RE-based Hi-C Description	Non RE-based Hi-C Difference
A. Self-Ligation	A.1	Circularized	✓	-	DNA fragment cut with RE circularizes, ligating to itself, and is then linearized by sonication during library preparation	Not shearing the DNA for library preparation prevents inclusion of self-ligated read types in the library
	A.2	Dangling Ends	✓	-	Read-pairs map to the same restriction fragment and at least one end overlaps the RE site	
	A.3	Internal	✓	-	Read-pairs map to a single restriction fragment but neither end of the di-tag overlaps the RE cut site	
B. Re-ligation	B.1	Short Ligation Event	✓	✓	Read-pairs map to adjacent restriction fragments which have re-ligated in the same orientation as found in the genome, where the ligation event is smaller than the LLF	Does not require mapping near RE site, where ligation event is smaller than the LLF
	B.2	Contiguous	✓	✓	Calculated read-pair length is not within the library fragments length distribution, where the ligation event is smaller than the LLF	Does not require mapping near RE site, where ligation event is smaller than the LLF
C. Valid Ligation	C.1	Long Ligation Event	✓	✓	Read-pairs map to adjacent restriction fragments which have ligated to a different region of the genome, where the ligation event is larger than the LLF	Does not require mapping near RE site, where the ligation event is larger than the LLF

Table 2 – **Description of the two-step workflow used for DNase Hi-C (Omni-C data) processing through Hi-C Pro**. For each step, the name of the configuration file used (hyperlinked to download the .txt file) and a description of the key alterations are noted.

Step	Config File Used	Description
1. Mapping	DNase_config_motif_step1.txt	Under the 'digestion Hi-C' section the variable 'LIGATION_SITE' is populated with the Omni-C bridge sequence
2. Filtering	DNase_config_no_motif_step2.txt	The 'digestion Hi-C' section is left blank because RE's were not used to digest the chromatin

using the Omni-C Kit following the vendor provided protocol and recommendations on the human cell line GM12878. To execute the HiC-Pro pipeline for Omni-C data, the pipeline needs to be run in two steps, with each step using a separate configuration file (**Table 2**). A two-step approach is required since Omni-C™ proximity ligation events are captured using a custom oligonucleotide bridge. Since Hi-C Pro uses bowtie2[6] to map reads, the presence of the bridge sequence in the read pair results in read pairs being flagged as an unaligned read. A modified Hi-C Pro workflow compatible with the Omni-C™ library conducts the processing in two steps:

1. Initial global mapping followed by trimming and re-mapping of unaligned reads with a

ligation event defined for trimming, the resulting alignments are merged into a single bam file

2. Filtering the merged bam with no 'digestion Hi-C' variables populated

Each step defined above uses its own configuration file. The first step requires a ligation motif (Omni-C™ ligation bridge sequence) to be described in the configuration file while the second step requires the 'digestion Hi-C' variable to be disabled. If any variable of the 'digestion Hi-C' is populated in the configuration, downstream read characterization and filtering fails because the pipeline looks for a RE digested reference genome which is not used for RE-free Hi-C. The commands used to execute the two-step Hi-C Pro workflow are captured in **Box 1**.

Results and Discussion

Hi-C Pro was able to process the Omni-C library using the two-step approach. The total mapping rate is ~82% (**Figure 2A**), which is consistent with other mappers such as BWA using the -5SP (flag recommended for mapping Hi-C data). Of the aligned sequences, 94.04% are classified as valid reads (*cis* > LLF + *trans* reads), with the remaining 5.96% flagged as

Box 1: Hi-C Pro: Two Step command line

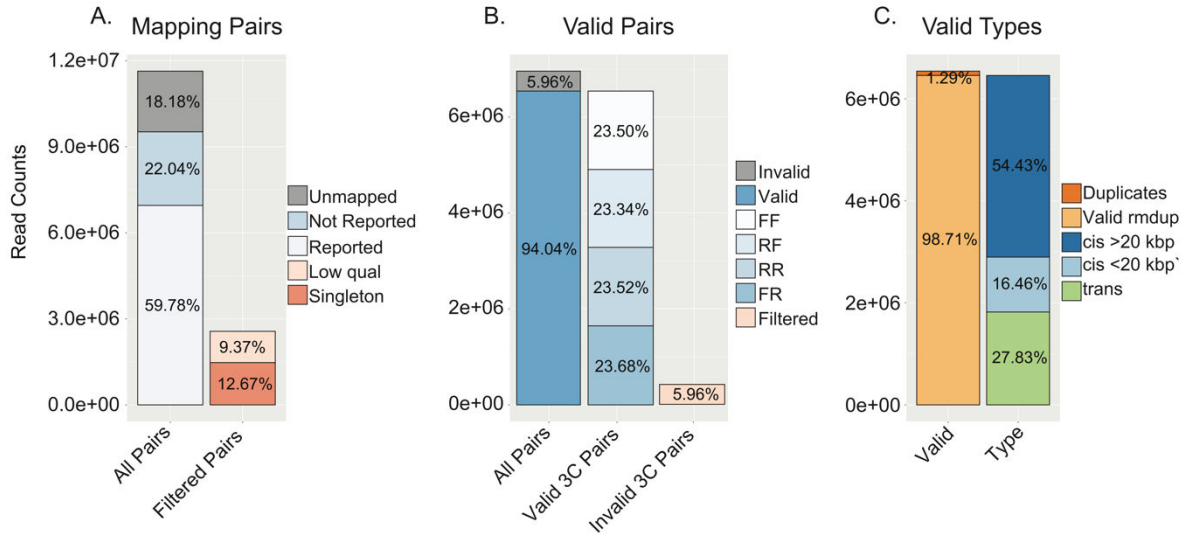
Step 1 – Mapping

HiC-Pro -I HiC_data/ -o pro_out/motif_trimming –c pro_out /DNase_config_motif_step1.txt -s mapping

Step 2 – Characterizing and Filtering

HiC-Pro -i pro_out/motif_trimming/bowtie_results/bwt2 -o /pro_test/motif_trimming/ -c DNase_config_no_motif_step2.txt -s proc_hic -s quality_checks

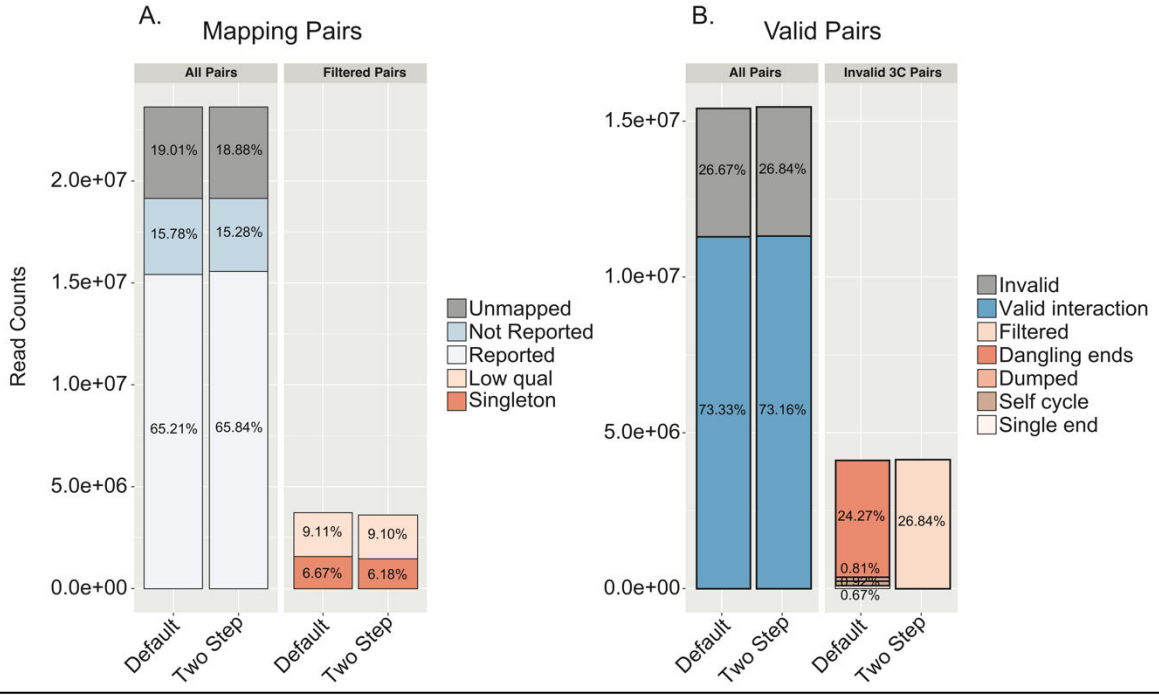
Figure 2 – Results from Hi-C Pro on the Omni-C library using the two-step approach. Hi-C Pro results for mapping pairs (A), valid pair filtering (B), and valid read type classification (C) are shown as bar graphs. The x-axes for all plots are the classification of read type reported by Hi-C pro, with the y-axes indicating the number of read counts. The percentage of each read type (# read pairs / # total number included in that category) is reported on the corresponding bar section.



invalid. The invalid reads are equal to the frequency of *cis* reads < LLF (1 kbp). The valid reads consist of the equal proportion directionality expected of random ligation assays. The library contains a low duplication rate, indicative of high complexity (**Figure 2B**). Finally, the library is enriched (54.43%) in long *cis* reads >20 kbp and contains ~16% of *cis* reads between 1 kbp - 20 kbp, with *trans* reads making up the remains fraction of valid reads at ~28% (**Figure 2C**).

To demonstrate that this two-step method does not alter the results of the Hi-C Pro pipeline, we performed the same workflow on a *DpnII* digested Hi-C library and compared the results to the default Hi-C Pro pipeline. Both the default and two-step approach produced nearly identical mapping statistics (**Figure 3A**) and the fraction of valid/invalid reads. The main difference is how the invalid reads are classified; the two-step approach classifies all the invalid reads as "filtered", meaning the *cis* ligation event < LLF of 1 kbp (**Figure 3B**). This is consistent with the fact that all possible invalid read types of RE-based Hi-C are essentially read-pairs with an insert distance less than the LLF **(Figure 1, Table 1**). This is reflected in the filtered fraction on the two-step approach to be equal to the sum of all the invalid types flagged in the default run (**Figure 3B**).

Figure 3 – ***DpnII* Hi-C comparison between the default Hi-C Pro pipeline and the two-step approach**. Hi-C Pro results for mapping pairs (A), valid pair filtering (B), and valid read type classification (C) are shown as bar graphs. The x-axes categories are default and two-step approach. The y-axes depicts the number of read counts. The histogram plots are grouped by read type classification reported by Hi-C pro. The percentage of each read type (# read pairs / # total number included in that category) is reported on the corresponding bar section.



Summary

Due to the alternate approach to chromatin digestion, the definition of a valid read changes slightly but does not impact the proportion of read pairs classified as valid or invalid. This is highlighted in the results comparing the default and two-step mapping approaches on a single *DpnII* Hi-C dataset demonstrating that this does not effectively change the Hi-C Pro results. This indicates that the Hi-C Pro results on the Omni-C, a DNase-based Hi-C, library are an accurate representation of the valid/invalid read types. In conclusion, Hi-C Pro is an open-source tool that provides an appropriate means to process RE-free Hi-C data.

References

1. Wiggett et al., 2015. HiCUP: pipeline for mapping and processing Hi-C data.
2. Servant et al., 2015. HiC-Pro: An optimized and flexible pipeline for Hi-C processing.
3. HOMER Hi-C documentation.
4. Ramani et al., 2016. Mapping 3D genome architecture through in situ DNase Hi-C.
5. Hsieh et al., 2016. Micro-C XL: assaying chromosome conformation from the nucleosome to the entire genome.
6. Langmead et al., 2012. Fast gapped-read alignment with Bowtie 2.