Post-Sequencing Quality Control Process Of Dovetail Proximity Ligation Libraries

Table of Contents

Introduction

A key component of working with any NGS-based assay is the processing and quality control of the data that come off the sequencer. For proximity ligation libraries, the main goal is to classify and assess the distance information captured by ligation events from high-quality read pairs. To make the data processing and QC of proximity-ligation libraries easier, Dovetail Genomics has designed a workflow that incorporates 4D Nucleome best practices to help you accurately assess the quality of libraries generated with Dovetail™ Micro-C and Omni-C™ kits. A detailed breakdown of the tools can be found here:

- Micro-C: https://micro-c.readthedocs.io/en/latest/index.html
- Omni-C: https://omni-c.readthedocs.io/en/latest/index.html

We recommend shallow sequencing of your library to 1-2 million read pairs to get an initial assessment of library quality. This document walks you through the consecutive post-sequencing QC process while clarifying what the different QC metrics indicate.

How Is A Valid Proximity-Ligation Read Pair Defined?

Before we can discuss the QC process, we must first define a valid read pair as not all read pairs produced in a proximity ligation library are of equal interest. Read pairs result from one of three ligation events:

1. Self-ligation
2. Re-ligation

} Invalid read pair

3. Valid ligation — Valid read pair

The first two ligation events are of low interest while the third ligation event - the desired class - yields a valid read pair. Figure 1 provides a detailed schematic defining each class and how it is generated for both restriction enzyme (RE) and DNase/MNase-based approaches. The percentage of read pairs that fall into the valid ligation class is, therefore, an important QC metric. Depending on the chromatin fragmentation approach, these classes may require different data filtering strategies. It should be noted that self-ligation products are not a concern when working with Dovetail™ Micro-C and Omni-C™ Kits as the workflow does not require sonication, and thus, these products cannot physically be converted into sequenceable molecules.

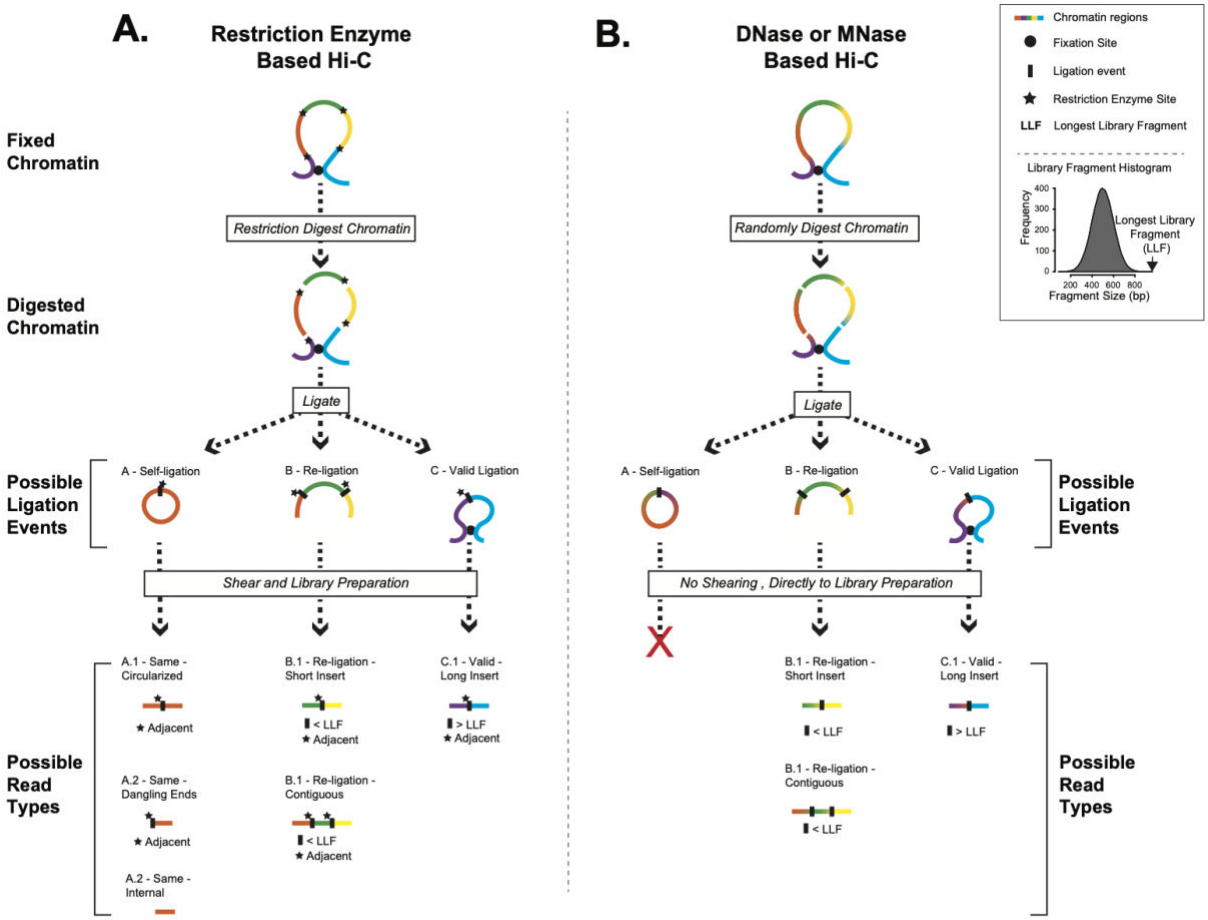

Figure 1. Classification schematic of reads generated from restriction enzyme (RE)-based and RE-free proximity ligation assays. Possible ligation events and resulting read types are depicted. The RE-based (A) and DNase/MNase (B) proximity ligation workflows are shown in parallel for direct comparison. 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 depicted in the inset. *Trans* read pairs, where each read from a pair maps to two different chromosomes, are considered valid read pairs but not pictured.

Post-Sequencing QC Analysis Workflow Overview

After sequencing the library to 1-2 M read pairs, the QC analysis workflow is completed in three serial steps (Figure 2):

I. Align raw reads and filter out unmapped, low mapping quality and PCR duplicate read pairs.

II. Classify remaining read pairs as *cis* or *trans* and characterize insert distance to identify valid read pairs.

III Estimate complexity (this step is not applicable to deeply sequenced data sets).

The above workflow is outlined in detail in the readthedocs pages for each product.

- Micro-C: https://micro-c.readthedocs.io/en/latest/index.html
- Omni-C: https://omni-c.readthedocs.io/en/latest/index.html

The workflow follows these steps which are outlined in Figure 2. We will discuss the metrics according to the step in which they are computed. To clarify each group of metrics, graphical representations of the data are included, however, these graphs are not part of the QC analysis output file.

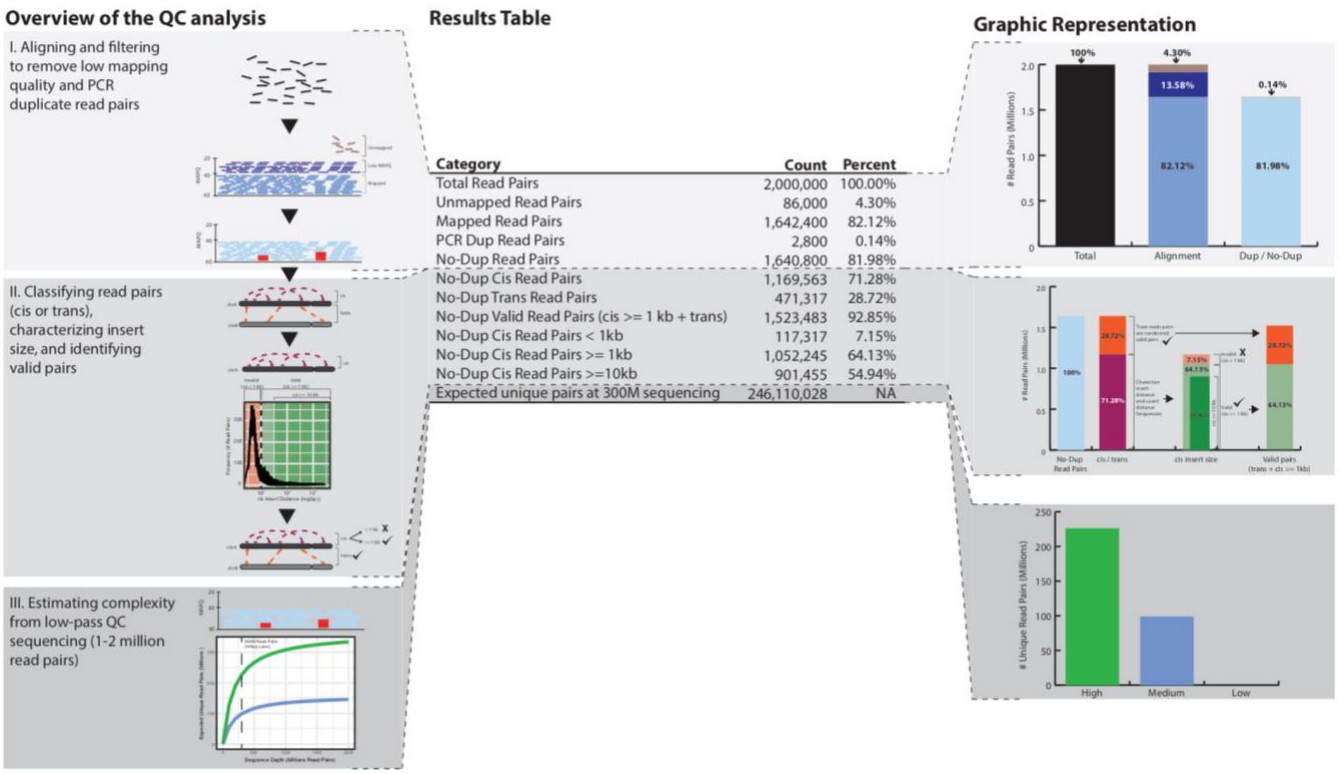

Figure 2. Overview of the post-sequencing QC analysis workflow and outputs.

QC Analysis Step-By-Step

I: Aligning raw reads and filtering for unmapped, low mapping quality and PCR duplicate read pairs.

After sequencing, the read pairs are aligned using **BWA MEM** to the appropriate reference genome. The alignment step results in:

a) Unmapped read pairs
b) Mapped read pairs with a mapping quality (MAPQ) value < 40
c) Mapped read pairs with a mapping quality (MAPQ) value ≥ 40

Unmapped and low MAPQ read pairs are removed from the subsequent steps. (Note: Low MAPQ read pairs are not reported in the QC table output by the script.)

Mapped read pairs with MAPQ ≥ 40 are processed by **pairtools** to flag and remove PCR duplicates. Only non-duplicate mapped read pairs with MAPQ ≥ 40 (referred to as No-Dup Read Pairs) progress into step 2 of the QC analysis.

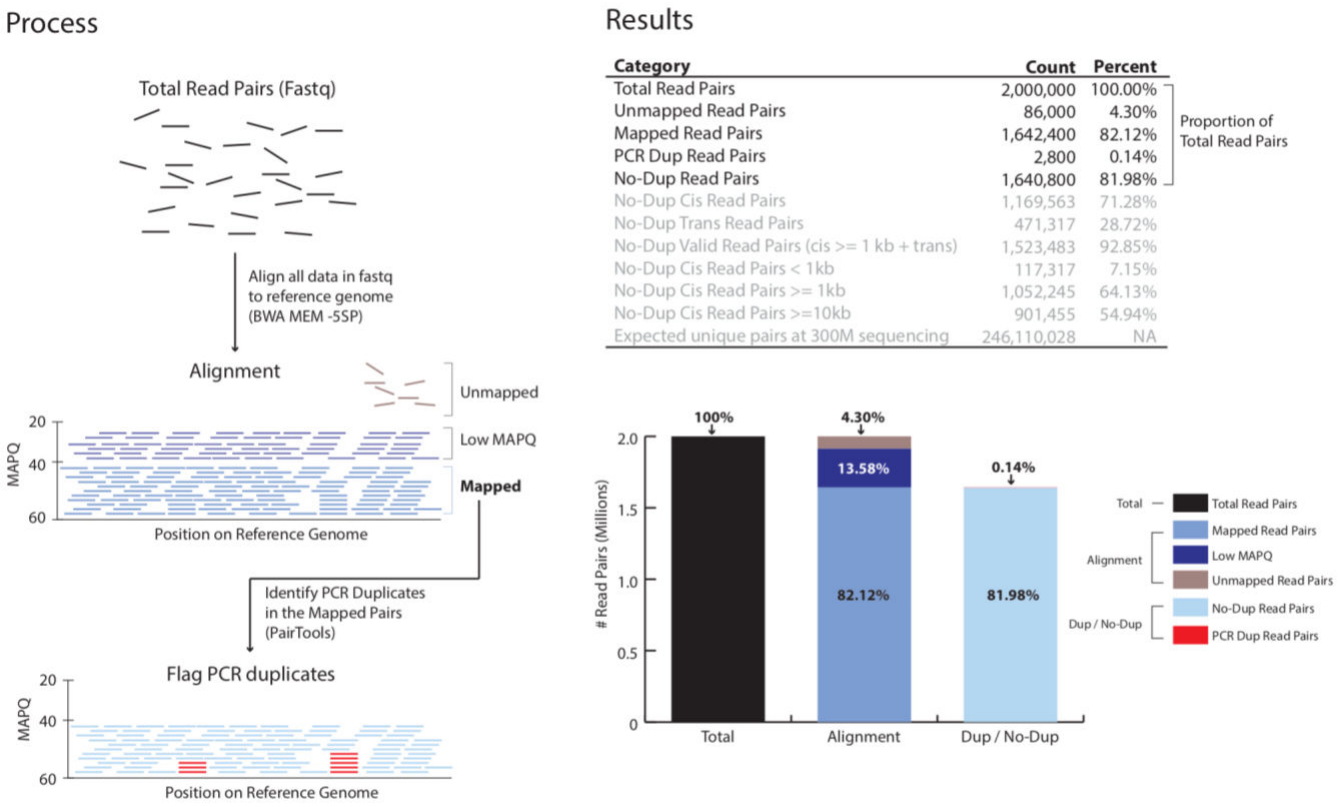

Figure 3. Step 1: process, results, and graphical representation.
Process – Total reads are aligned to a reference genome. The reads are then characterized as unmapped, low MAPQ (< 40), or mapped read pairs (> 40). PCR duplicates are then flagged and filtered from the mapped read pairs using **pairtools**.
Results – The results of this step are captured in the first 5 rows of the QC table.
Graphical Representation – The three bars represent each step in the alignment and filtering process with number of read pairs on the y-axis. Total Read Pairs represents the denominator used to calculate percentages.

II: Classifying filtered read pairs as *cis* or *trans* and characterizing insert distance to identify valid read pairs.

The non-duplicate mapped read pairs with MAPQ ≥ 40 (No-Dup Read Pairs) from step 1 are categorized by **pairtools** as valid if they meet one of the following criteria:

a) the pair maps to different chromosomes (*trans*).
b) the pair maps to the same chromosome (*cis*) and the distance between the interacting points is > 1 kbp.

In addition to looking at the percentage of valid read pairs as a QC metric, another consideration is how these valid read pairs are partitioned across the two valid categories of *trans* and *cis* > 1 kbp.

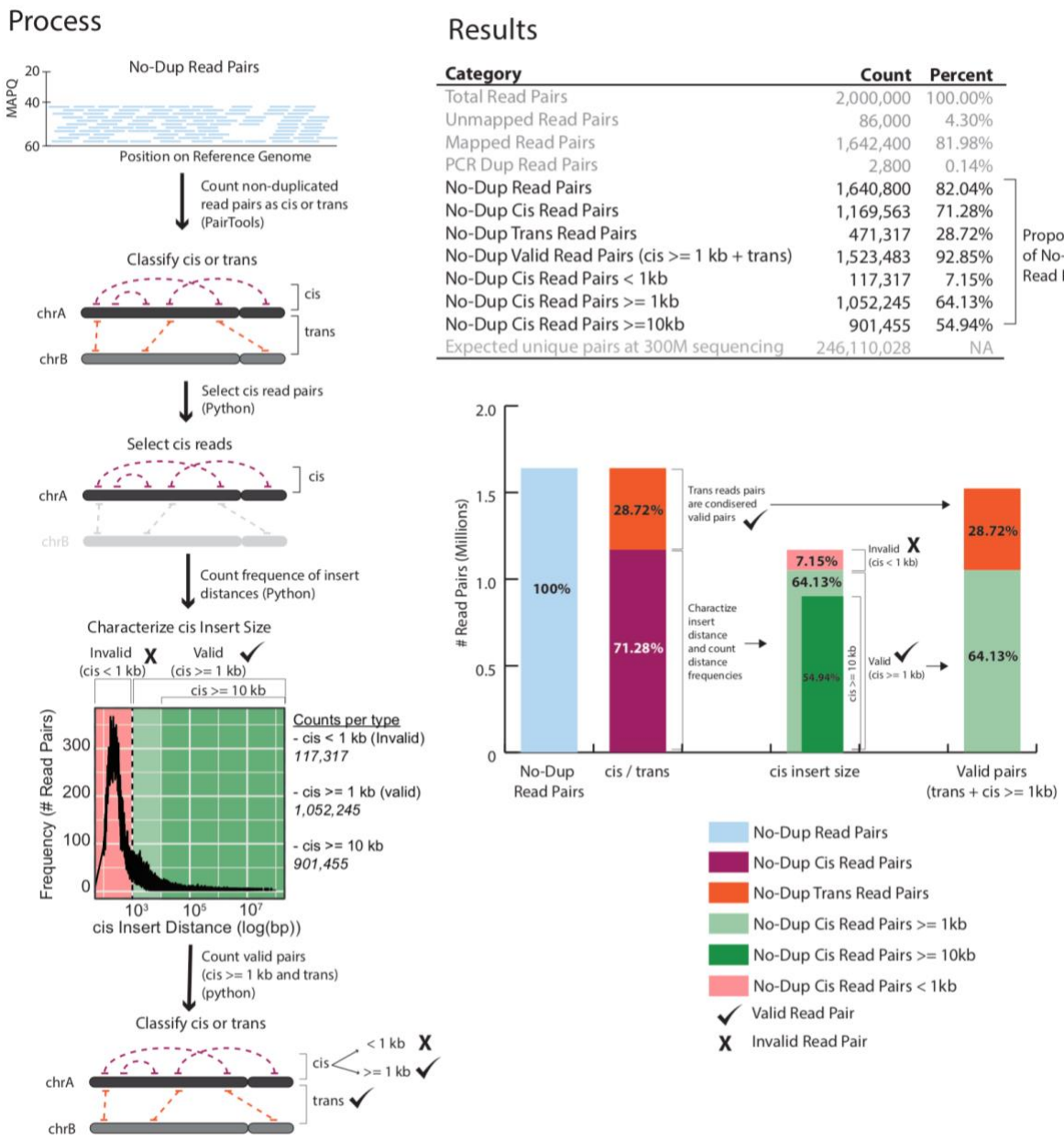

Figure 4. Step 2: process, results, and graphical representation.
Process – No-Dup read pairs are classified as *cis* or *trans* using **pairtools**. All *trans* read pairs are considered valid. By contrast, valid *cis* reads must have an insert size greater than the Longest Library Fragment (LLF). 1 kbp is used as the LLF cut-off for both Omni-C and Micro-C libraries. Since the libraries are size selected, physical insert sizes range from 350 bp to 1 kbp. Therefore, mapped insert size < 1 kbp represent re-ligation events (and are invalid).
Results – The results of this step are captured in rows 6 – 11 of the QC table.

Graphical Representation – On the left is a plot of *cis* read pair insert size (frequency); color changes mark the 1 kb, 10 kb and >10kb insert size bins. The bar chart on the right plots the reads classified in the QC table. The No-Dup read pairs count is the denominator for the percentages calculated.

III: Estimating Library Complexity (this step is not applicable to deeply sequenced data sets).

When sequencing PCR-amplified DNA, some amount of duplication will occur. Since most proximity ligation experiments require deep sequencing in excess of 200 M read pairs, we are interested in ensuring sufficient diversity exists in the library despite our shallow QC dataset. Therefore, the last step of the QC analysis estimates library complexity at 300 M reads from the 1-2 M read pair input. Based on the expected Micro-C and Omni-C library complexity, we would not recommend sequencing the library beyond a maximum of 300 M read pairs. Estimates of complexity are only applicable for low pass sequencing. For an already deeply sequenced library, the complexity is reflected in the percentage of No-Dup Read Pairs. We recommend shallow sequencing libraries to 1-2 M read pairs. Using less than 1 M read pairs can lead to incorrect estimates.

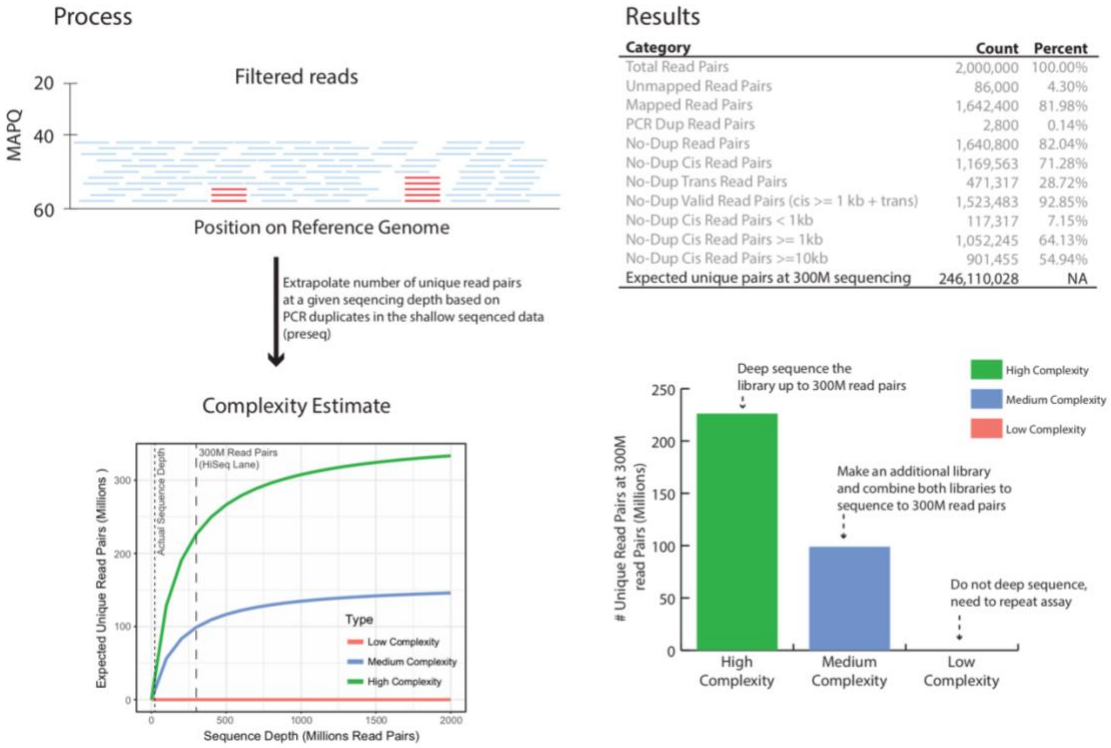

Figure 5. Step 3: process, results, and graphical representation.
Process – **Preseq** is used to estimate the complexity of the library. **Preseq** uses the number of mapping reads and the PCR duplicate rate to extrapolate how many unique reads would be generated if the library was to be sequenced to a given depth.

Results - The number of unique reads expected at 300 M read pairs (or the equivalent of 30X coverage for a human genome or 1 Illumina HiSeq lane) is reported in the last row of the QC table.
Graphical Representation – The assessment of unique molecules will inform whether 1) the library is of sufficient complexity for deep sequencing; 2) additional libraries should be generated and several medium complexity libraries combined for deep sequencing (additional libraries can be prepared from any remaining proximity ligated DNA recovered at the end of Stage 3), or 3) if the library should not be deep sequenced and the assay should be repeated.

Summary

The post-sequencing QC process is a three-step process designed to filter out low quality and unmapped read pairs, eliminate PCR duplicates, identify valid read pairs, and estimate library complexity. It follows best practices in processing and QC of proximity ligation libraries. Get started with the QC analysis of your proximity ligation library here:

- Micro-C https://micro-c.readthedocs.io/en/latest/index.html
- Omni-C https://omni-c.readthedocs.io/en/latest/index.html