Hi-C Library QC Report

Input Data

File Type	File Name
BAM File	P_citronellum.asm.hic.hap1.p_ctg_vs_HiC.sorted.bam
Assembly File	P_citronellum_OmniC.hic.bp.hap1.p_ctg.fa
Forward Hi-C Reads	**forward Hi-C reads not found**
Reverse Hi-C Reads	**reverse Hi-C reads not found**

Genome Scaffolding Sufficiency

Label	Library statistics	Expected values
Subjective Hi-C library judgment	**SUFFICIENT**	See Judgment
Same strand high-quality* (HQ) read pairs (RPs)	**14.48%**	> 1.5%
Informative RPs**	**13.40%**	> 5.0%

*High-quality (HQ) read pairs have minimum mapping quality >= 20, maximum edit distance <= 5, and are not duplicates.
**Informative read pairs are read pairs which have MAPQ > 0, are not PCR duplicates, and map to different contigs or >10kb apart.

Metrics Demonstrating Strong Proximity Signal

Label	Library statistics	Expected values
Fraction of HQ RPs >10kb apart (CTGs >10kb)*	**13.59%**	> 3.0%
Fraction of HQ RPs Intercontig (CTGs >10kb)**	**28.42%**	> 2.5%
Informative HQ RPs per-contig per-1M-RPs (CTGs >5kb)***	**40.25**	> 100.0

*The proportion of *read pairs that span at least 10kb*, out of all read pairs that map (a) with high-quality, (b) to the same contig, (c) where that contig is at least 10kb long.
**The proportion of *read pairs mapping to two different contigs each greater than 10kb*, out of all read pairs that map with high-quality.
***The average number of HQ reads, per contig *among contigs at least 5kb in length*, per million read pairs, that are informative for identifying contigs which belong on the same chromosome.

Noninformative Read Pair Breakdown

Label	Library statistics	Expected values
Noninformative RPs*	**67.80%**	<= 50.0%
Duplicate reads	**0.00%**	< 20.0%
Zero map distance read pairs	**0.34%**	<= 20.0%
Zero MAPQ reads	**60.78%**	<= 20.0%
Unmapped reads	**0.00%**	<= 10.0%

*Note that the sum of informative and noninformative read pairs is not 100% because read pairs with mapping distance between 1b and 10kb are not classified as either informative or noninformative.
Because noninformative reads can belong to more than one category, these numbers may sum to a value larger than the overall noninformative read pair amount at the top of the report.

Assembly Statistics

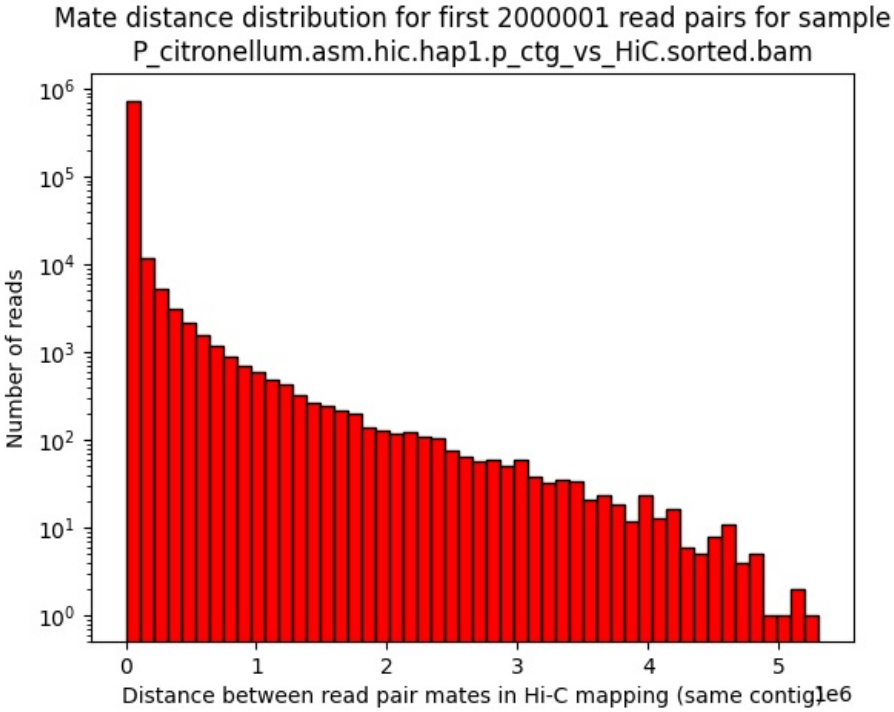
Label	Assembly statistics
BAM file	P_citronellum.asm.hic.hap1.p_ctg_vs_HiC.sorted.bam
Forward Hi-C Reads	**forward Hi-C reads not found**
Reverse Hi-C Reads	**reverse Hi-C reads not found**
Assembly file	P_citronellum_OmniC.hic.bp.hap1.p_ctg.fa
Assembly size	514,186,233
Contig (CTG) N50	760,978
CTGs	1,566
CTGs > 10kb	1,566
CTGs > 5kb	1,566

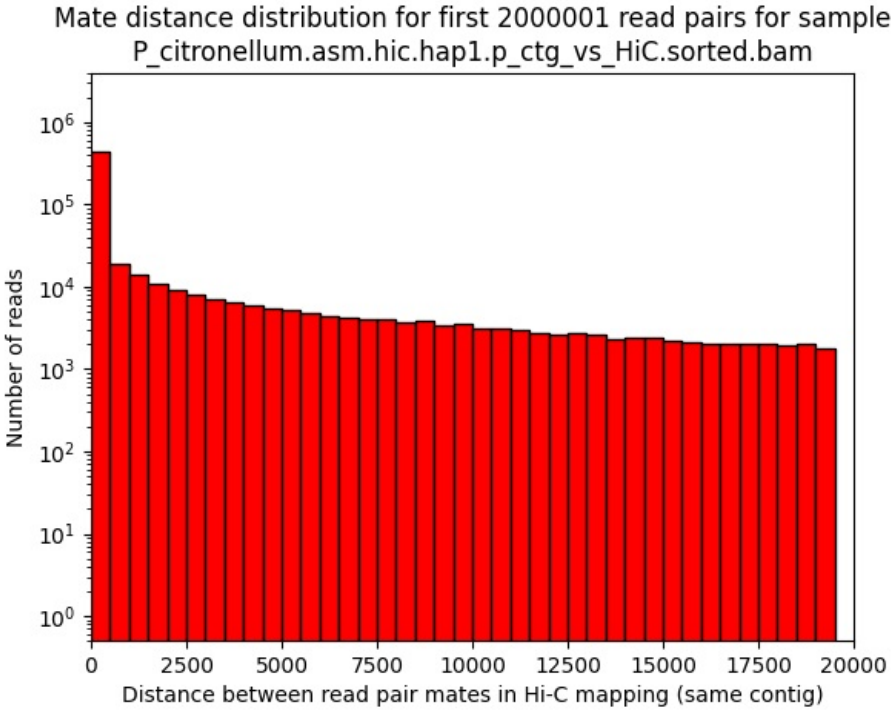
Extended Library Statistics

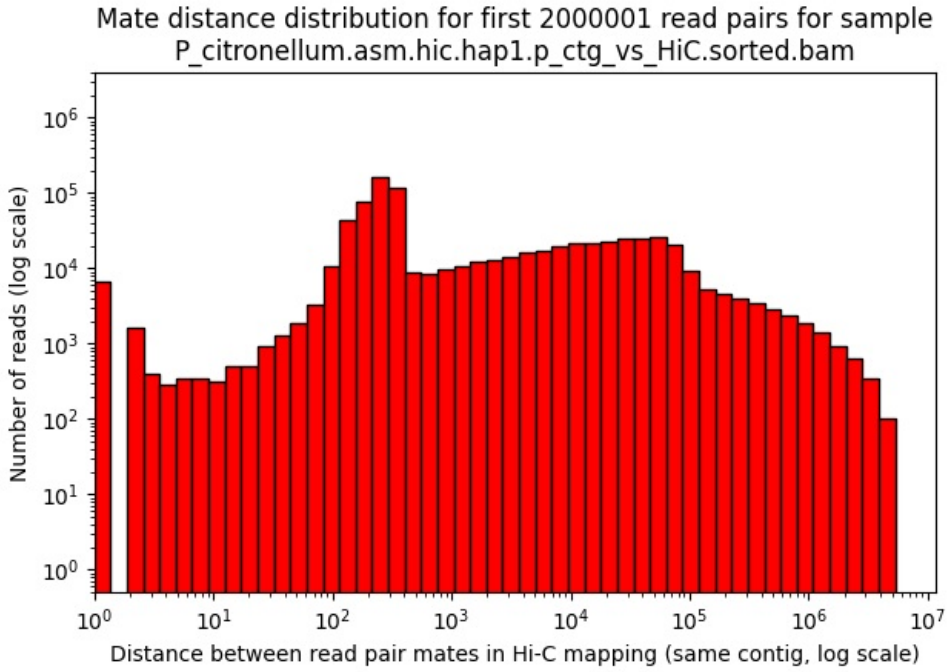
Label	Library statistics	Expected values
Total read pairs (RPs) analyzed	2,000,001	N/A
High-quality (HQ) RPs	22.18%	N/A
Clustering usable HQ reads per contig (CTGs >5kb)*	80.50	> 600.0
RPs >10kb apart	9.79%	1-15%
RPs >10kb apart (CTGs >10kb)	25.77%	1-15%
Intercontig RPs	62.03%	10-60% (contigs) 1-20% (chromosomes)
Intercontig HQ RPs	28.42%	10-60% (contigs) 1-20% (chromosomes)
Same strand RPs	22.92%	2-50%
Split reads	19.56%	1-10% (PG libraries) 30%+ (other libraries)
Alignment Parameters	bwa mem -t 24 -5SP	N/A
Samblaster Parameters	samblaster command not found	N/A
Restriction Enzyme(s)	unspecified	N/A

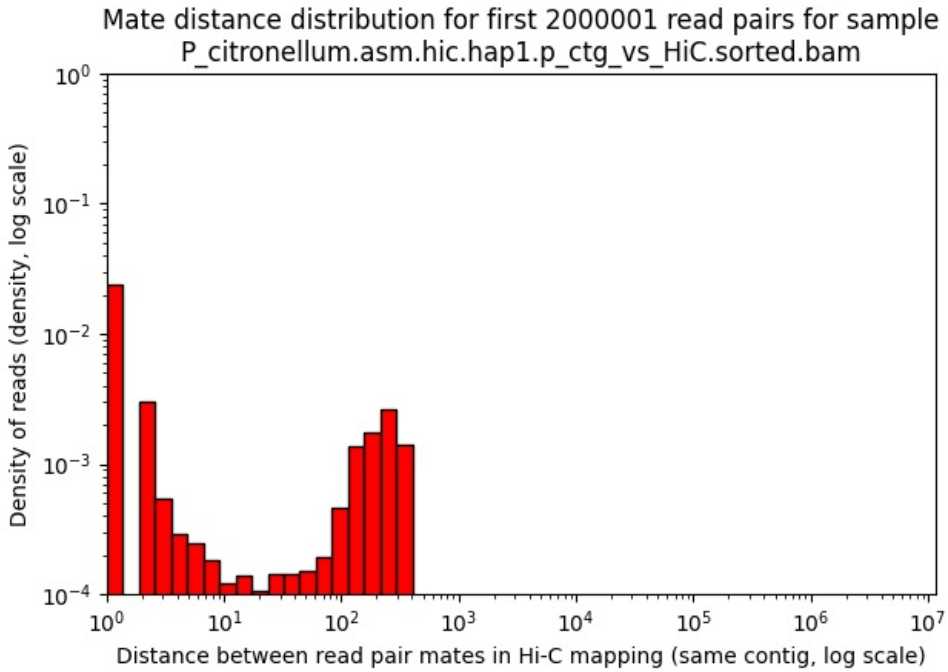
*The average number of usable high-quality read pairs per contig, for contigs greater than 5kb. Read pairs are "usable" if they map (a) with high-quality, (b) to different contigs, (c) where each of those contigs are greater than 5kb and (d) both mappings are high-quality.

Aligned mate distance histograms






Mate distance distribution for first 2000001 read pairs for sample
P_citronellum.asm.hic.hap1.p_ctg_vs_HiC.sorted.bam
Number of reads (log scale)
Distance between read pair mates in Hi-C mapping (same contig, log scale)


Mate distance distribution for first 2000001 read pairs for sample
P_citronellum.asm.hic.hap1.p_ctg_vs_HiC.sorted.bam
Density of reads (density, log scale)
Distance between read pair mates in Hi-C mapping (same contig, log scale)

Alignment distance statistics and plots

We briefly describe some of the statistics we compute below to aid interpretation of this report.

Subjective Hi-C Library Judgment

While Hi-C data is nuanced and some analyses are more sensitive to data quality than others, a basic quality assessment can usually be made by examining the mapping characteristics of the Hi-C library. Based on our experience working with Hi-C data, we classify libraries into one of four QC categories:
- **Sufficient** means that from everything we can tell, the library looks to be in great shape. Proceed to full sequencing or analysis with confidence.
- **Mixed Results** means that the library is good in some ways, but not in others. Perhaps it has a good amount of long-range data, but there are also an elevated number of read pairs with MAPQ 0. Usually, data generated from Mixed Results libraries works out just fine (a high MAPQ 0 number can be due to repetitiveness in the assembly or unpurged haplotigs, for example), but it is good to know there may have been a few hiccups in the library prep in case troubleshooting is needed down the line.
- **Low Signal** means that the library contains good Hi-C signal, but it's in lower percentage than usual. These libraries are generally good for generating useful Hi-C data, but you may need to sequence a little deeper than normal to get enough of it. You might consider size selecting the library to discard reads outside the 300-700bp range, as these are unlikely to be good Hi-C junctions. Alternatively, you might just want to prep a new library.
- **Insufficient** means that the library, or perhaps the library in combination with a low-contiguity or error-prone assembly, does not look useful. Sometimes size selection can rescue such libraries, but sometimes a new prep is the only way forward.

Contact us if your library is not sufficient and we will help you out.

Read pairs on same strand

This is the percentage of reads mapping to the same contig in the same orientation (FF or RR). For shotgun libraries, this should be ~1%, but for a pure Hi-C library, it could be as high as 50%. This is a primary metric of Hi-C library quality because it is minimally affected by assembly contiguity.

Read pairs > 10kb apart

This is the percentage of read pairs which map to the same contig, with at least 10kb separating them. More is always better, but because this number is affected by assembly contiguity, there is not a specific target threshold. Note that for some analyses, such as scaffolding or metagenomic deconvolution, read pairs that map to the same contig are not useful because they do not provide information that the assembly doesn't already contain. This statistic is more useful for these projects because it correlates with library prep success. These reads are useful for analyses like structural variant analysis or assembly misjoin detection, because they provide detailed structural information about existing assembled sequences.

Read pairs > 10kb apart mapping to contigs >10kb

This is the percentage of read pairs which map to the same contig, with at least 10kb separating them, but only considering read pairs mapping to contigs that are at least 10kb long. This attempts to corrects for assembly contiguity differences. More is always better, but typically at least 5% is desired. Note that for some analyses, such as scaffolding or metagenomic deconvolution, read pairs that map to the same contig are not useful because they do not provide information that the assembly doesn't already contain. This statistic is more useful for these projects because it correlates with library prep success. These reads are useful for analyses like structural variant analysis or assembly misjoin detection, because they provide detailed structural information about existing assembled sequences.

Read pairs mapping to different contigs or chromosomes

This is the percentage of read pairs which map to different contigs, which is particularly important. More is always better, but because this number is affected by assembly quality, there is not a specific target threshold, although at least 20% on *de novo* assembly projects is helpful. These reads are the primary source of information for Hi-C scaffolding or metagenomic deconvolution analyses. This statistic useful on most *de novo* projects because it also correlates with library prep success. These reads may be useful for analyses like structural variant analysis or assembly misjoin detection if those are performed on lower contiguity assemblies, because they provide detailed structural information about sequences which were not assembled together into contigs.

Split reads

Traditionally, split reads have been a favored measure of Hi-C library quality because they directly exhibit Hi-C junctions. Most traditional Hi-C library preparations produce many reads that sequence through junctions because their Hi-C junctions tend to occur randomly on the proximity ligated chimeric molecules. However, Phase Genomics libraries, whether produced in our laboratory or by means of our Plant, Animal, Human, or Microbe Hi-C kits, will have a generally lower percentage of split reads. This is because we have optimized our Hi-C protocol to enrich for slightly longer fragments around Hi-C junctions, such that each read is less likely to read through a junction even when a junction is present.

This innovation improves mappability and increases the amount of useful data and reduces the utility of split read measurements to assess library quality. We therefore rely more heavily on metrics that directly relate to the usefulness of Hi-C reads for proximity analysis, such as the percentage of read pairs that map to the same strand or to different contigs.

Duplicate reads

IMPORTANT NOTE: THE DUPLICATE FLAG IS NOT SET BY DEFAULT IN A BAM FILE. YOU NEED TO EXPLICITLY SET IT BY E.G. RUNNING SAMBLASTER OR PICARD MARKDUPLICATES ON YOUR BAM FILE. IF THE PERCENT OF DUPLICATES IS EXACTLY ZERO, IT PROBABLY MEANS THAT THE FLAG HAS NOT BEEN SET.

Sequencing libraries frequently contain duplicate reads due to PCR or optical issues. These are generally considered to be non-informative because they are chemical artifacts rather than biological signal and are thus typically excluded from further analysis. Higher percentages of duplicate reads are also correlated with low library complexity and poor library performance, making the percentage of duplicate reads a useful quality control measure.

Unmapped reads

A high percent of unmapped reads may indicate sequence is missing from the reference, the reads are mapped to the wrong reference, or the sample is contaminated.

REPORT VERSION: 0+untagged.261.g6881c33