Data quality report generated by PPL

This is a plain statistics report about **Pore-C library quality** based on the results from **PPL data processing pipeline.**

# 1. Basic statistical information

|  |  |
| --- | --- |
| Metrics | Values |
| Reads count | 800000 |
| Average length of reads | 4735 |
| N50 length of reads | 5439 |
| Mapped Reads count | 793143 |
| Unmapped Reads count | 6857 |
| Fragments count | 3617606 |
| Average fragments count by read | 4.561101844181945 |
| Average length of fragments | 960 |
| N50 length of fragments | 1392 |

# 2. Restriction Digestion quality

## 2.1 Length distribution comparision between real fragment and virtual fragments

Please note: The more consistent the length real fragment is with the virtual, the more sufficient restriction enzyme digestion is represented.

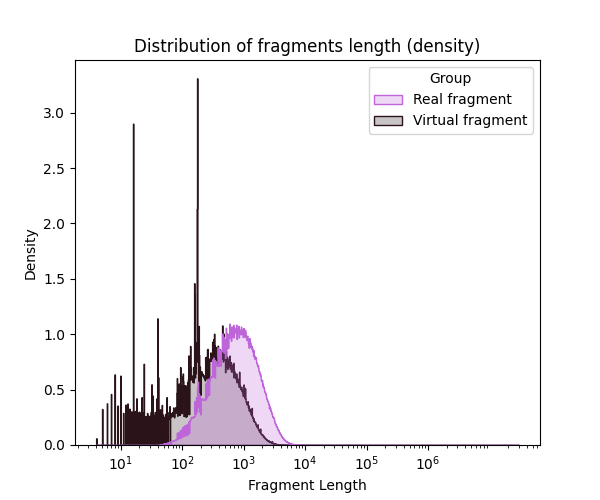


Figure 1. Length distribution of real and virtual fragments.

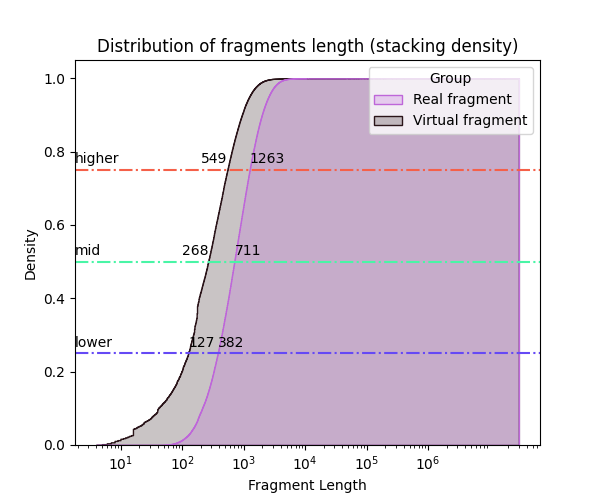


Figure 2. Cumulative distribution of real and virtual fragments length. The 1/4, 1/2, and 3/4 quantile lines are represented by blue, green, and red colors, respectively with length number.

# 3. Reads length and distribution of contact fragments

## 3.1 Distribution comparision between reads length and virtual fragments length

Please note: The longer reads are compared to fragments, the more fragments are contained in one read. For human sample, N50 from 3000 to 6000 is a proper metric for Pore-C experiment

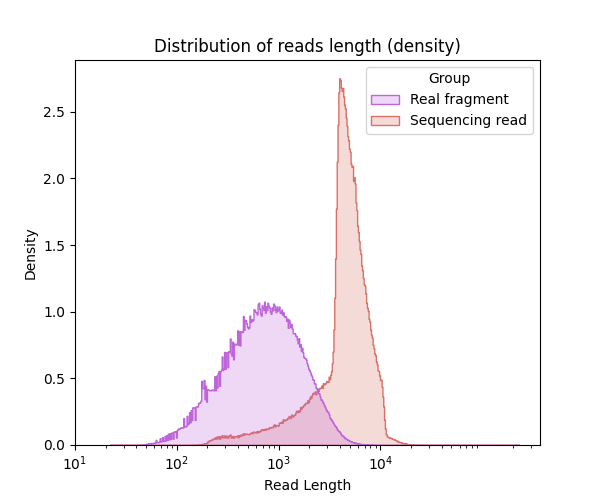


Figure 3. Length distribution of reads and fragments.

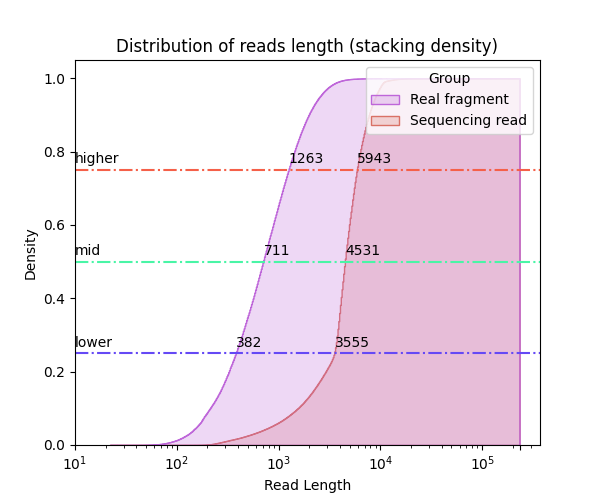


Figure 4. Cumulative distribution of fragments and reads length. The 1/4, 1/2, and 3/4 quantile lines are represented by blue, green, and red colors, respectively with length number.

## 3.2 Distribution of fragment number of multi-way contacts

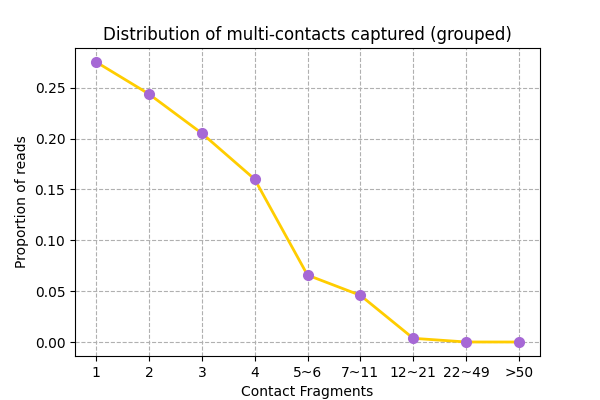


Figure 5. Distribution of fragment number of multi-way contacts. All contacts are grouped by fragment number.

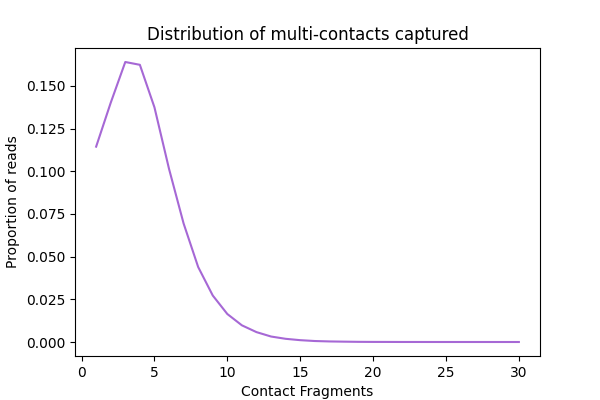


Figure 6. Distribution of fragment number of multi-way contacts.

# 4. Proximity Ligation quality

Please note: The quality of proximity ligation is evaluated by the proportion of three types of VPCs (virtual pairwise contacts) by genomic distance. In Hi-C process pipeline, one metric of high-quality datasets is the proportion inter-chromosomes contacts less than 60% and the proportion of the long-range intra-chromosomes contacts more than 40%. The inter VPCs usually be generated from very high order contacts, which probably be created by noisies, such as random ligation issues. You can use FilterHyper fuction to filter these contacts in PPL.jar

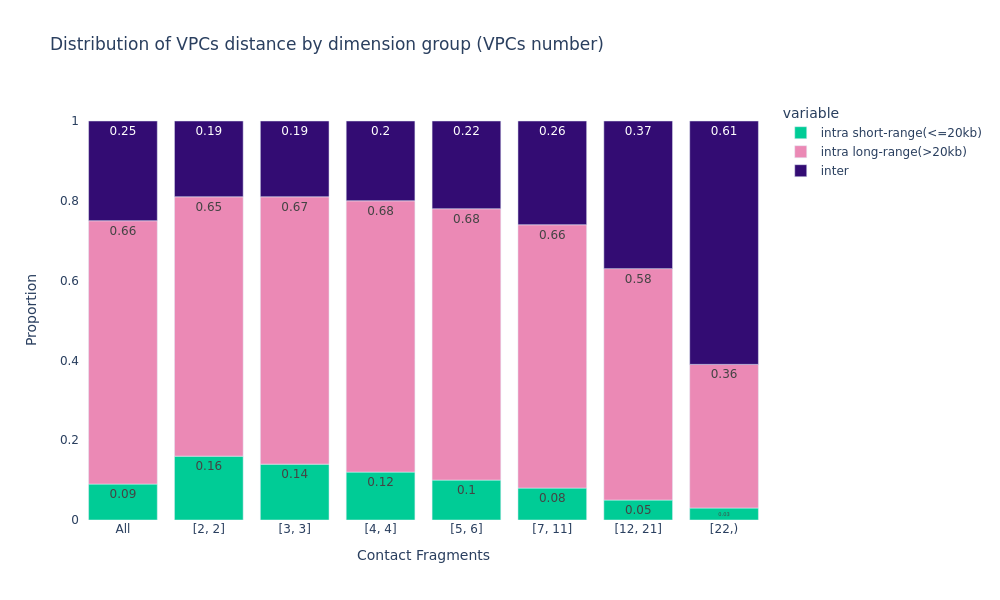


Figure 7. Distribution of VPCs by genomics distance.

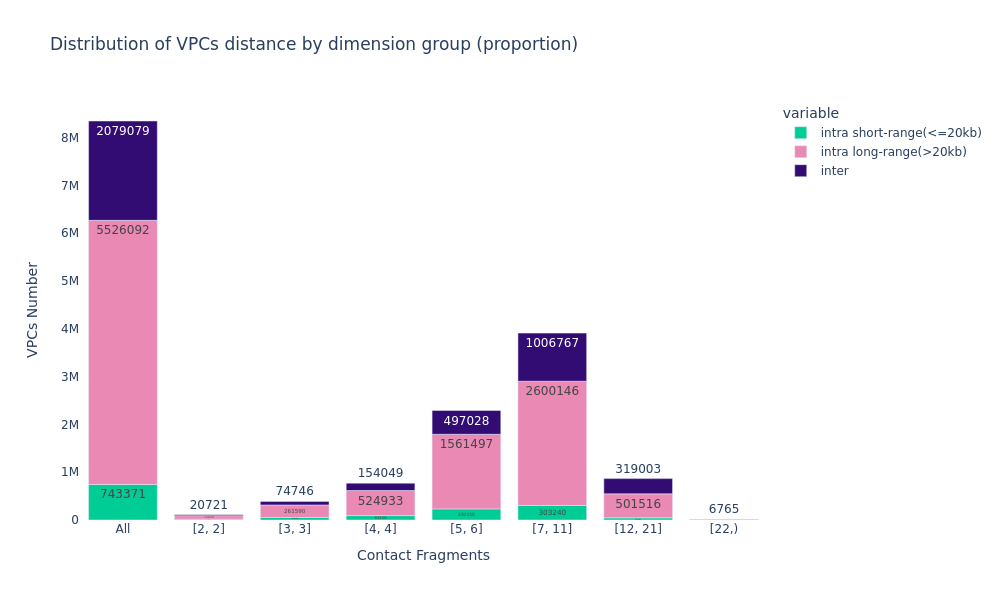


Figure 8. Distribution of VPCs by genomics distance. Real counts are shown.

# 5. Mapping quality

Please note: The mapping quality by three strategies: mapping quality score, boundary check, normalized penalty.

## 5.1 Mapping quality score check.

Please note: This score was generated by aligner. minimap2 is recommended.

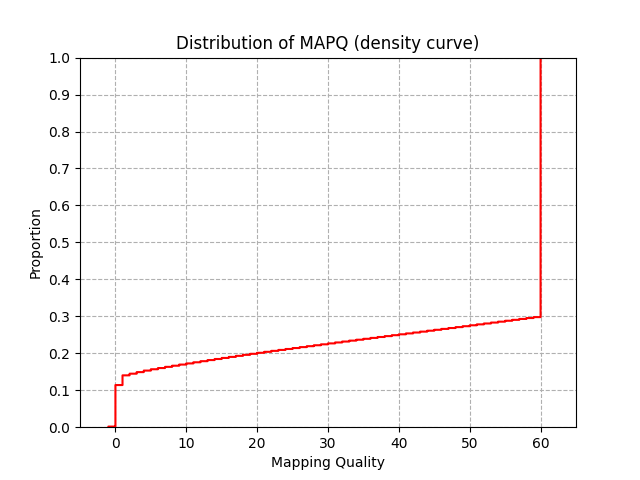


Figure 9. Distribution of mapping quality score.

## 5.2 Mapping boundaries check.

Please note: BoundaryCheck is a PPL function to evaluate the rightness of mapping by to check the mapping position and virtual restriction sites.

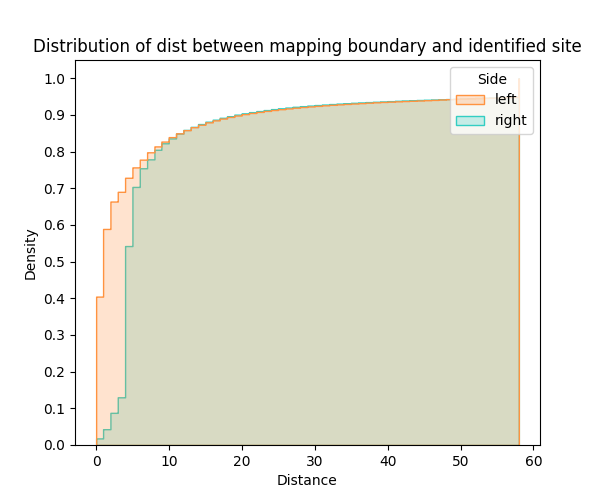


Figure 10. Distance between mapping boundaries and identified sites. Right side usually have extra several-bases distance because of the strategy of virtual digestion.

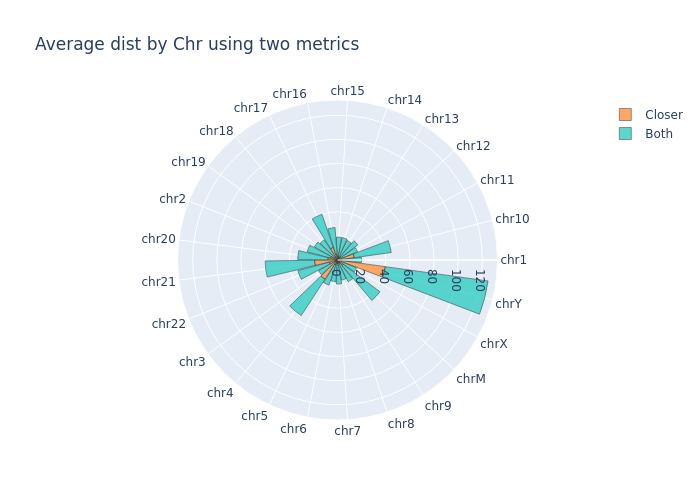


Figure 11. Distance between mapping boundaries and identified sites by chromoses. Two average distances were computed on all chromosomes using two strategies. In strategy 1 (both), we compute the average value of both sides of one read. In strategy 2 (closer), we only take the sides closer to identified site compared to another.

## 5.3 Mapping set penalty check.

Please note: In PPL mapping step, three penalties were calculated to get all right mapping records. Here, they were normalized by being divided by reads length. Overaly penalty is the symbol of the overlap region between mappings. Gap penalty is ratio sum of unmapped regions length to the sum length of all reads. The indel/mismatch is edit count, which means the similarity between reads and references.

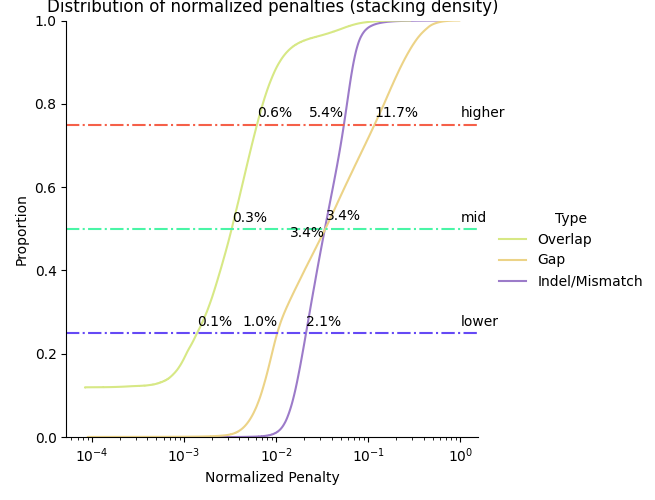


Figure 12. Normalised penalty score.

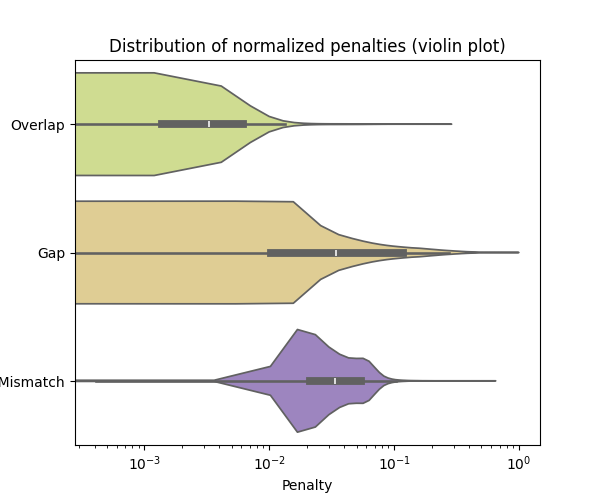


Figure 13. Normalised penalty score. Same to Fig 12, just presented by violin plot.

Code Accessibility: https://github.com/versarchey/PPL

Powered by python-docx