Bioinformatics Report for Nanopore Sequencing Pipeline

Chunyu Zhao

April 2, 2018

Contents

1 Assembly Stats and GAGE report

run	barcode	total	n	ave	largest	N50
cdiff_run9_20180321	barcode01	912231885	176479	5169	101202	13261
cdiff_run9_20180321	barcode02	603358129	196654	3068	93243	7173
cdiff_run9_20180321	barcode03	6.11e+08	128700	4748	95331	10818

Assembly GAGE report for cdiff_run9_20180321:barcode01

Assembly	draft2
Contigs #	4
Min contig	5006
Max contig	4261958
Not corrected N50	4261958 COUNT: 1
Genome size	4715305
Assembly size	4641733
Chaff bases	0
Missing reference bases	2082(0.04%)
Missing assembly bases	1374(0.03%)
Missing assembly contigs	0(0.00%)
Duplicated reference bases	4196
Compressed reference bases	62902
Bad trim	91
Avg idy	99.73
SNPs	3181
Indels < 5bp	8835
Indels >= 5	32
Inversions	2
Relocation	4
Translocation	0
Corrected contig #	38
Corrected assembly size	4655248
Min correct contig	620
Max correct contig	500767
Corrected N50	316223 COUNT: 6

Assembly GAGE report for cdiff_run9_20180321:barcode02

Assembly	draft2
Contigs #	1
Min contig	4184602
Max contig	4184602
Not corrected N50	4184602 COUNT: 1
Genome size	4191469
Assembly size	4184602
Chaff bases	0
Missing reference bases	1310(0.03%)
Missing assembly bases	1739(0.04%)
Missing assembly contigs	0(0.00%)
Duplicated reference bases	2003
Compressed reference bases	10815
Bad trim	227
Avg idy	99.79
SNPs	2482
Indels < 5bp	6244
Indels >= 5	28
Inversions	0
Relocation	3
Translocation	0
Corrected contig #	26
Corrected assembly size	4180619
Min correct contig	3139
Max correct contig	764784
Corrected N50	277373 COUNT: 5

Assembly GAGE report for cdiff_run9_20180321:barcode03

Assembly	draft2
Contigs #	4
Min contig	6935
Max contig	4190307
Not corrected N50	4190307 COUNT: 1
Genome size	4277632
Assembly size	4272191
Chaff bases	0
Missing reference bases	1567(0.04%)
Missing assembly bases	10787(0.25%)
Missing assembly contigs	0(0.00%)
Duplicated reference bases	340
Compressed reference bases	11563
Bad trim	14
Avg idy	99.84
SNPs	320
Indels < 5bp	6197
Indels >= 5	37
Inversions	2
Relocation	4
Translocation	0
Corrected contig #	44
Corrected assembly size	4266346
Min correct contig	575
Max correct contig	475257
Corrected N50	198090 COUNT: 7

2 Assess Reads and Assembly Accuracy

2.1 Background

In this section, we evaluate the accuracy of raw long reads, as well as the draft assembly using alignment to the reference genome **CD630**, using `Minimap2`.

Read accuracy is interesting to better understand the nanopore sequencing error, and assembly accuracy is more interesting to show whether the read errors can **average out** with high sequencing depth.

2.2 Metrics

2.2.1 Read Length Distribution

2.2.2 Read Accuracy

We only uses the **aligned** parts of the reads to calcuate the `reads identity`.

To be specific, the definition of `identity` is same with `Blast`: the number of matches in the alignment divided by alignment length (including gaps).

note: If less than 50% of a read aligned, it is assigned as **unaligned** and given an identity of 0%.

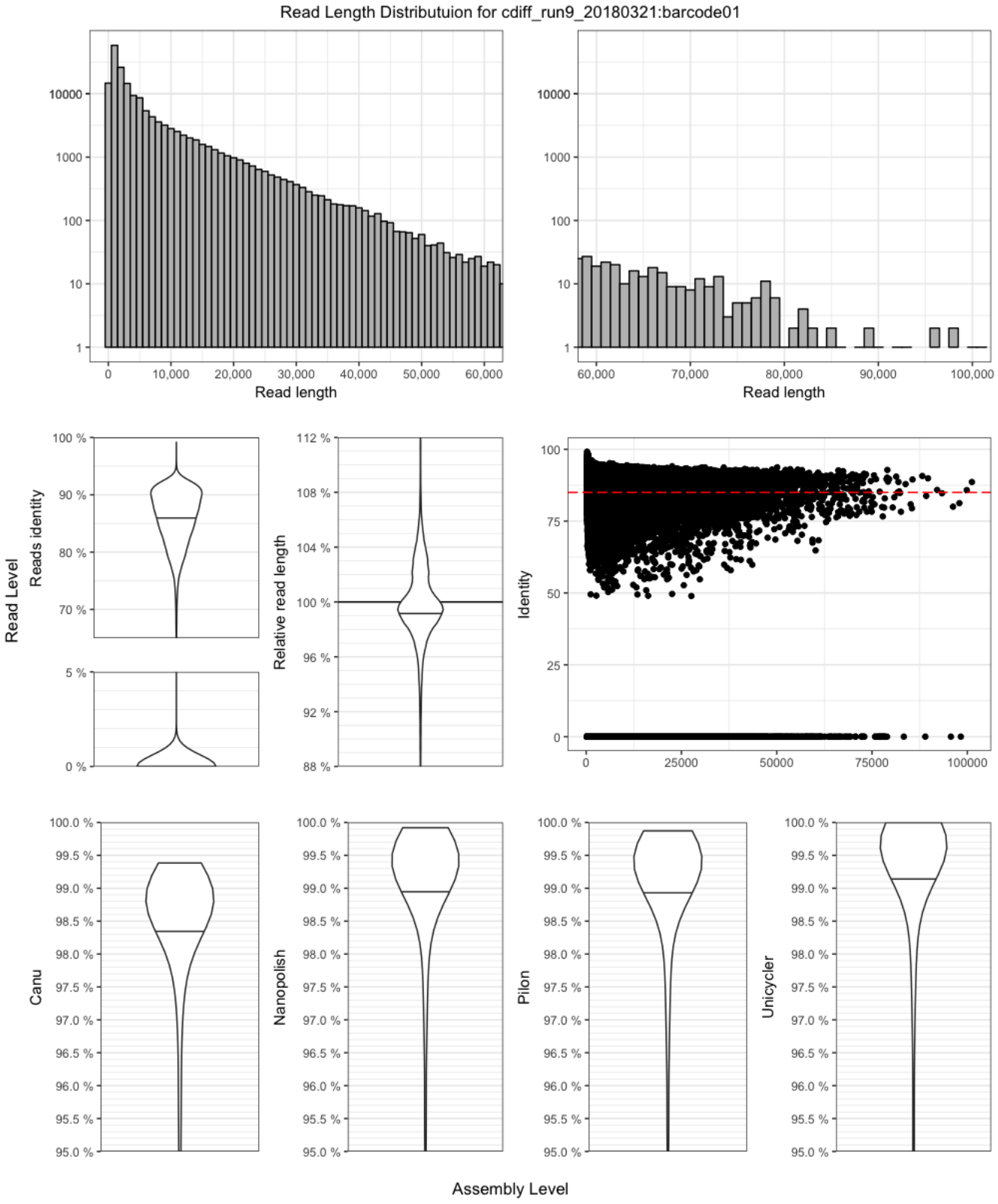
2.2.3 Relative read length

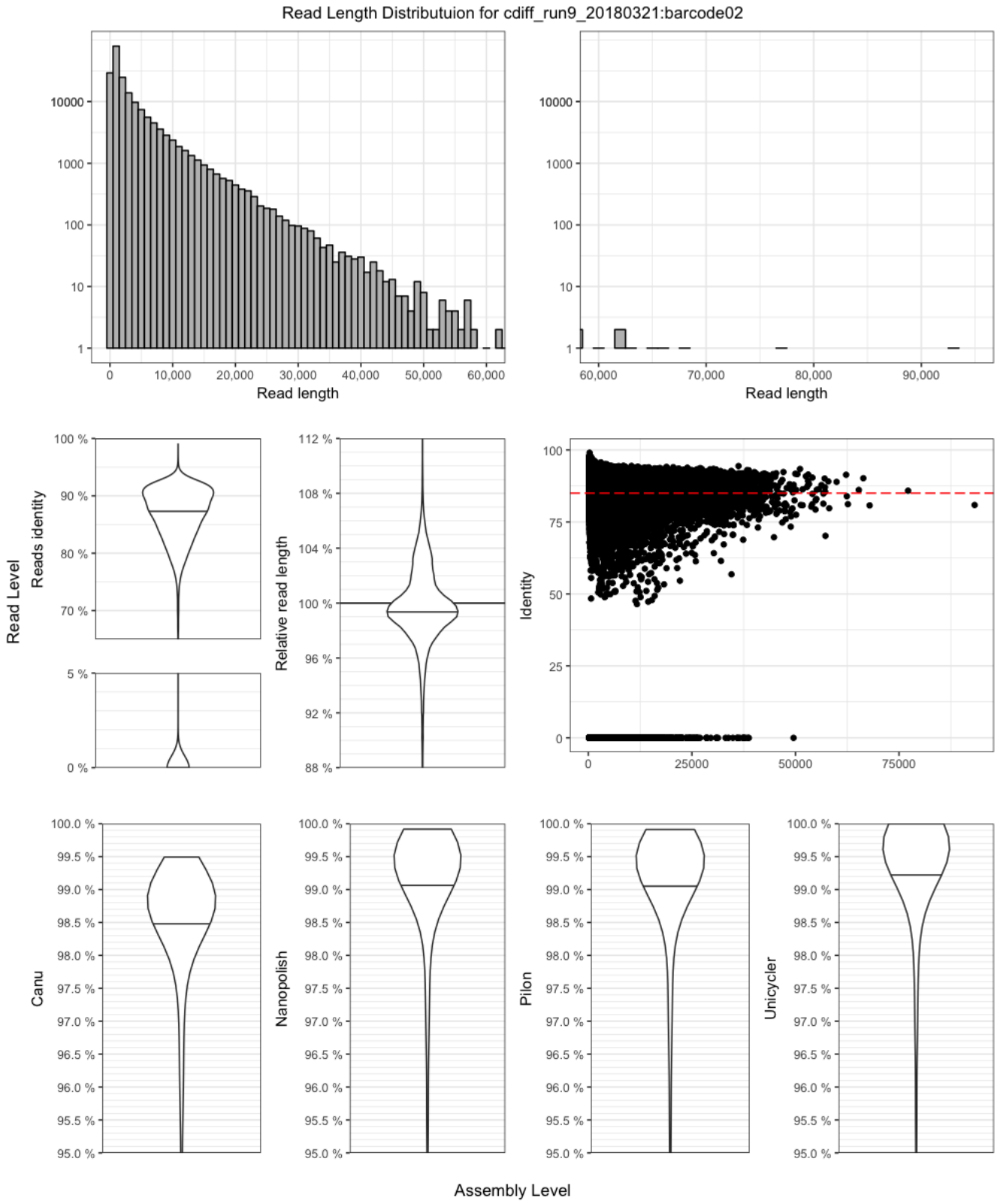
We also show the distribution of `relative read length`: read length to reference length for each alignment. This number shows whether the basecaller is more prone to interstions or deletions.

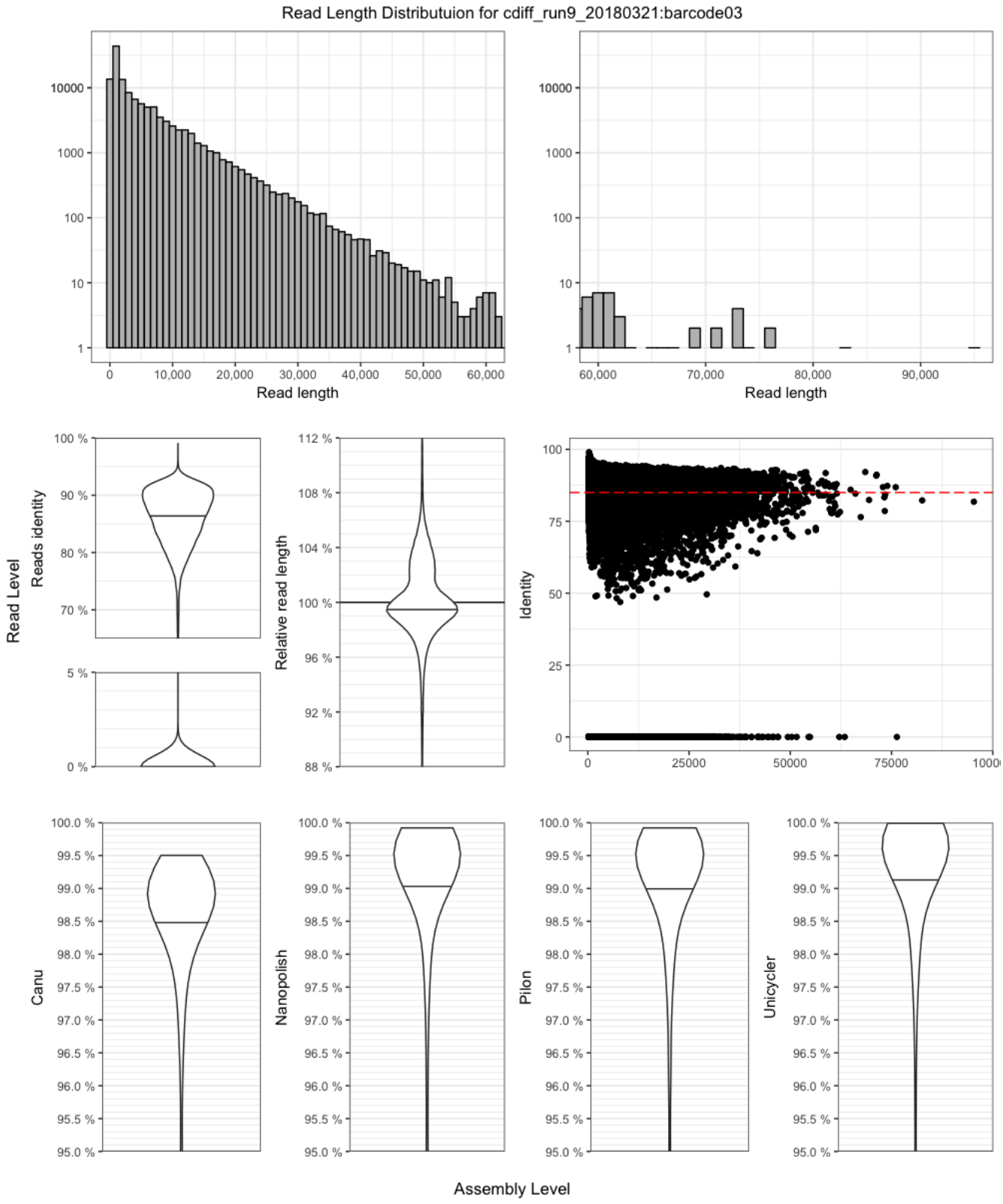
- `100%` (same length): means insertions and deletions are equally likely
- `<100%`: deletions are more common than insetions
- `>100%`: insertions are more common than deletions

2.2.4 Assembly accuracy

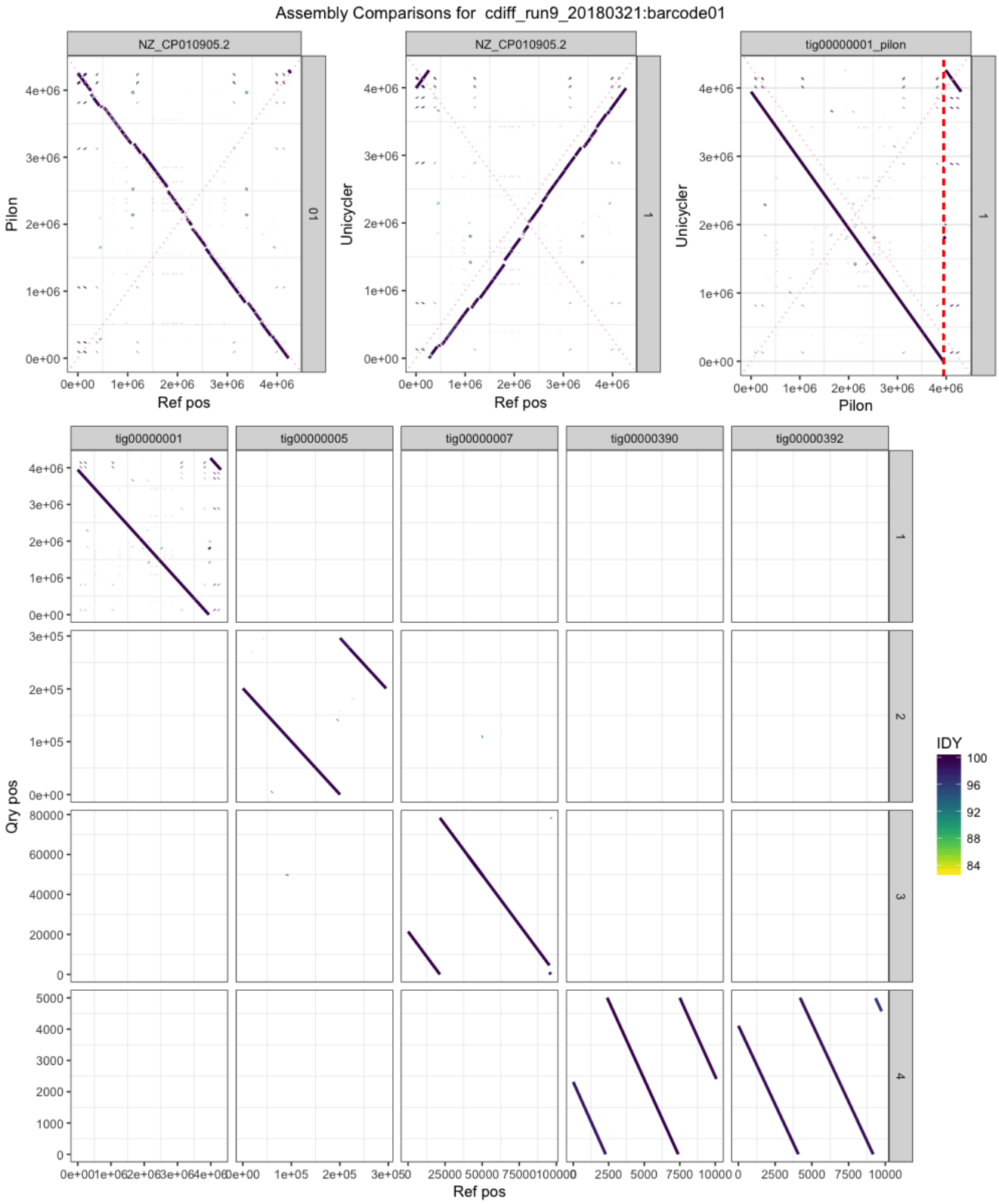
How accurate are the consensus sequences?

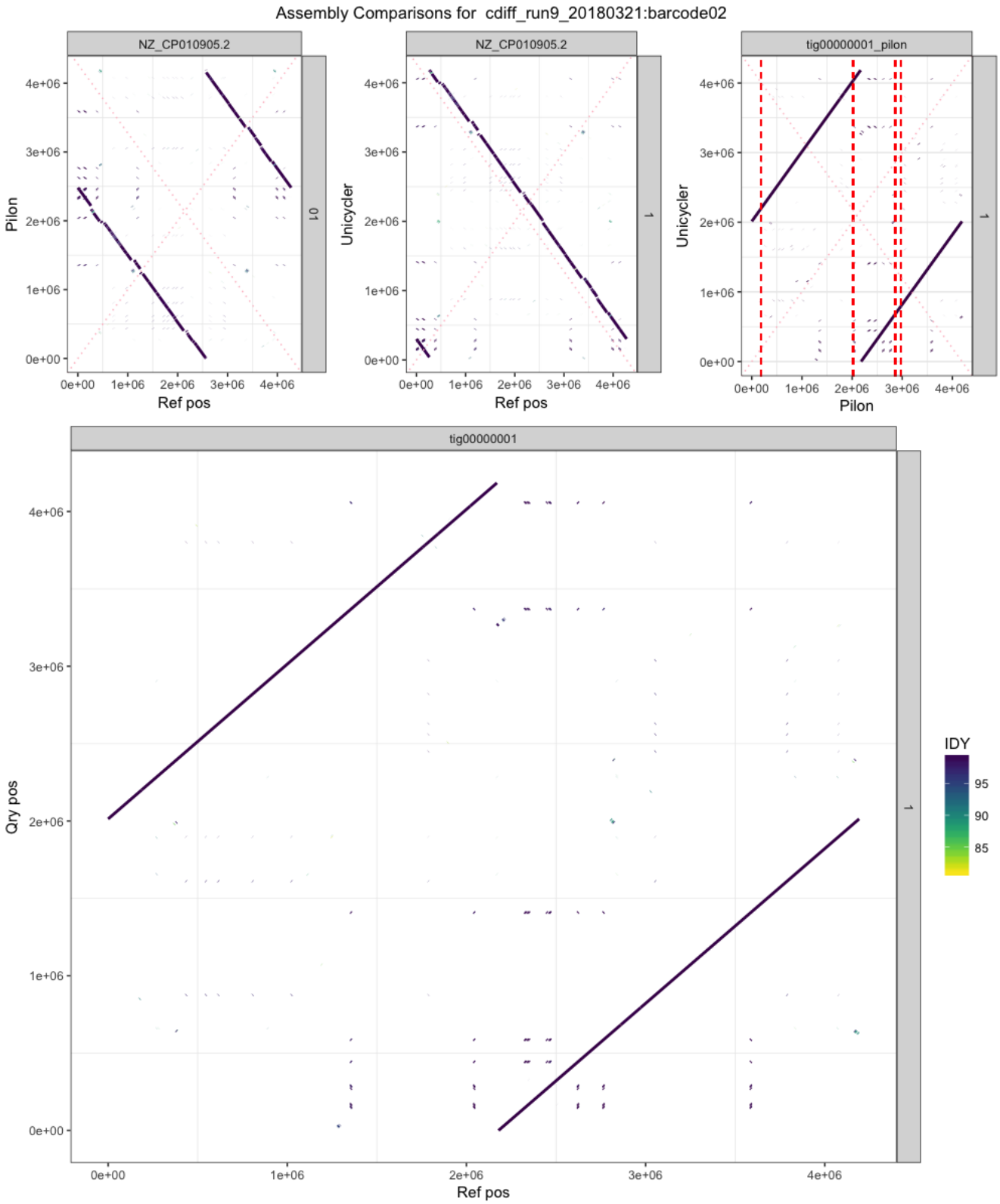

Read Length Distributuion for cdiff_run9_20180321:barcode01
Read length
Read length
Read Level
Reads identity
Relative read length
Identity
Assembly Level
Canu
Nanopolish
Pilon
Unicycler

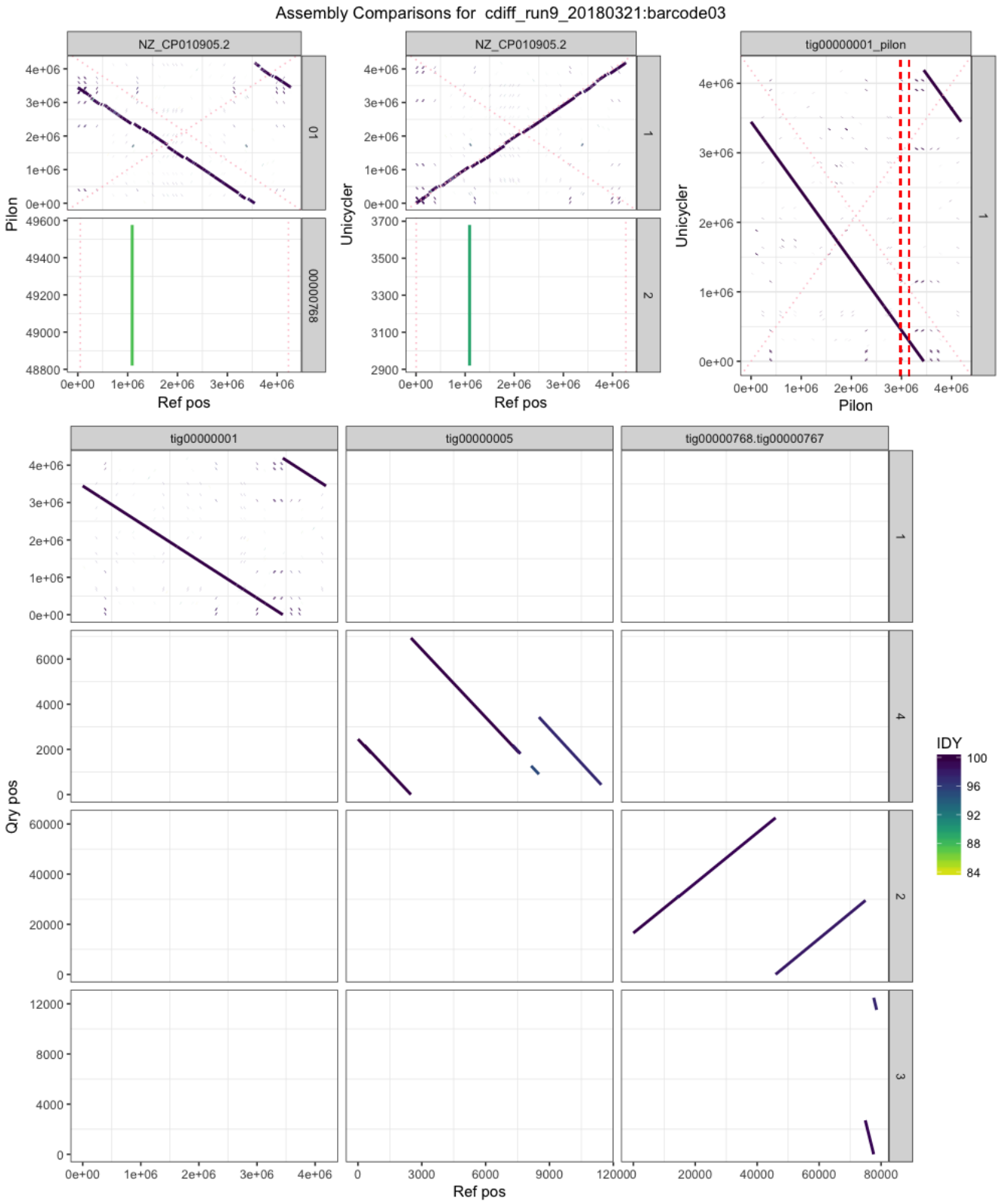

Read Length Distributuion for cdiff_run9_20180321:barcode02
Read length
Read length
Read Level
Reads identity
Relative read length
Identity
Assembly Level
Canu
Nanopolish
Pilon
Unicycler



3 Assembly Comparison by Dotplot


Assembly Comparisons for cdiff_run9_20180321:barcode01
NZ_CP010905.2
NZ_CP010905.2
tig00000001_pilon
Pilon
Unicycler
Unicycler
Ref pos
Ref pos
Pilon
01
1
1
tig00000001
tig00000005
tig00000007
tig00000390
tig00000392
Qry pos
Ref pos
IDY
100
96
92
88
84


Assembly Comparisons for cdiff_run9_20180321:barcode02
NZ_CP010905.2
NZ_CP010905.2
tig00000001_pilon
tig00000001
Pilon
Unicycler
Unicycler
Qry pos
Ref pos
Ref pos
Pilon
Ref pos
IDY
95
90
85
0e+00
1e+06
2e+06
3e+06
4e+06


Assembly Comparisons for cdiff_run9_20180321:barcode03
NZ_CP010905.2
NZ_CP010905.2
tig00000001_pilon
Pilon
Unicycler
Unicycler
Ref pos
Ref pos
Pilon
01
1
1
00000768
2
tig00000001
tig00000005
tig00000768.tig00000767
Qry pos
Ref pos
IDY
100
96
92
88
84