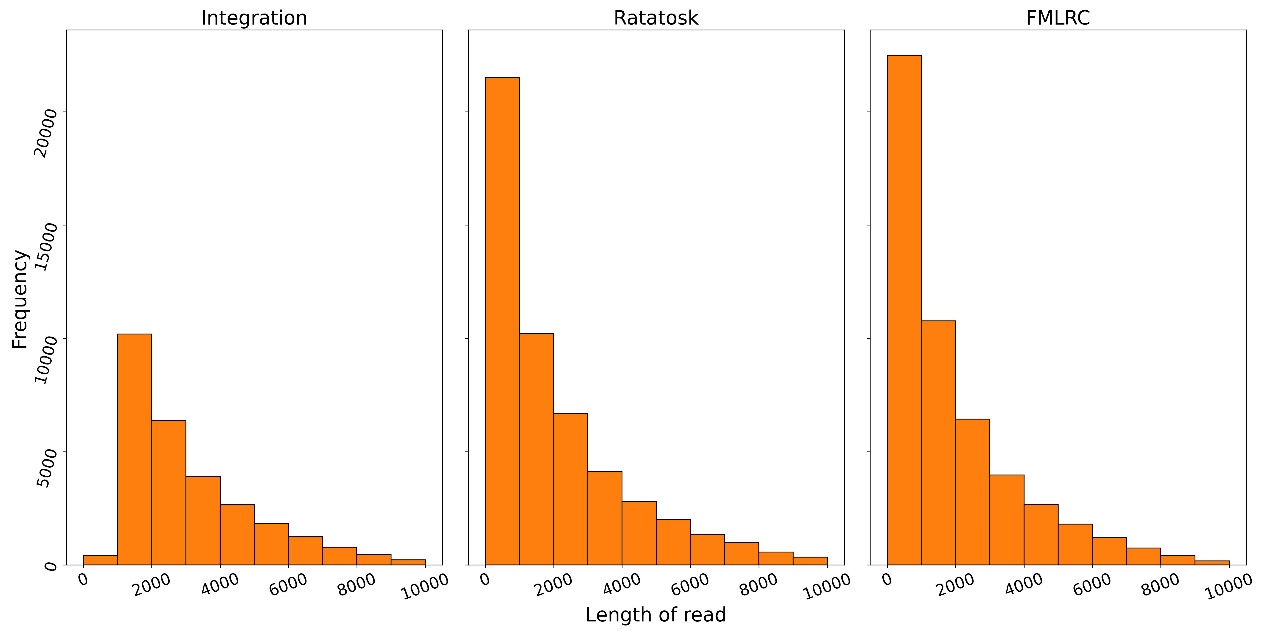
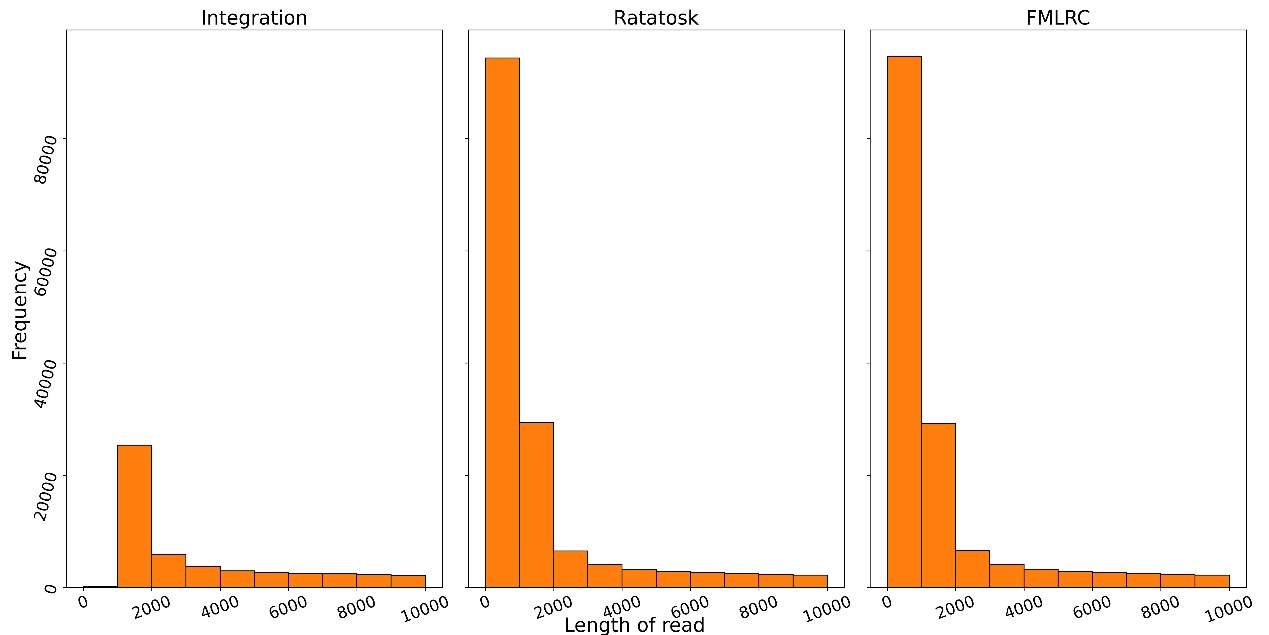
**Integration of hybrid and self-correction method improves the quality of long-read sequencing data**

Tao Tang, Yiping Liu, Binshuang Zheng, Rong Li, Xiaocai Zhang, Yuansheng Liu

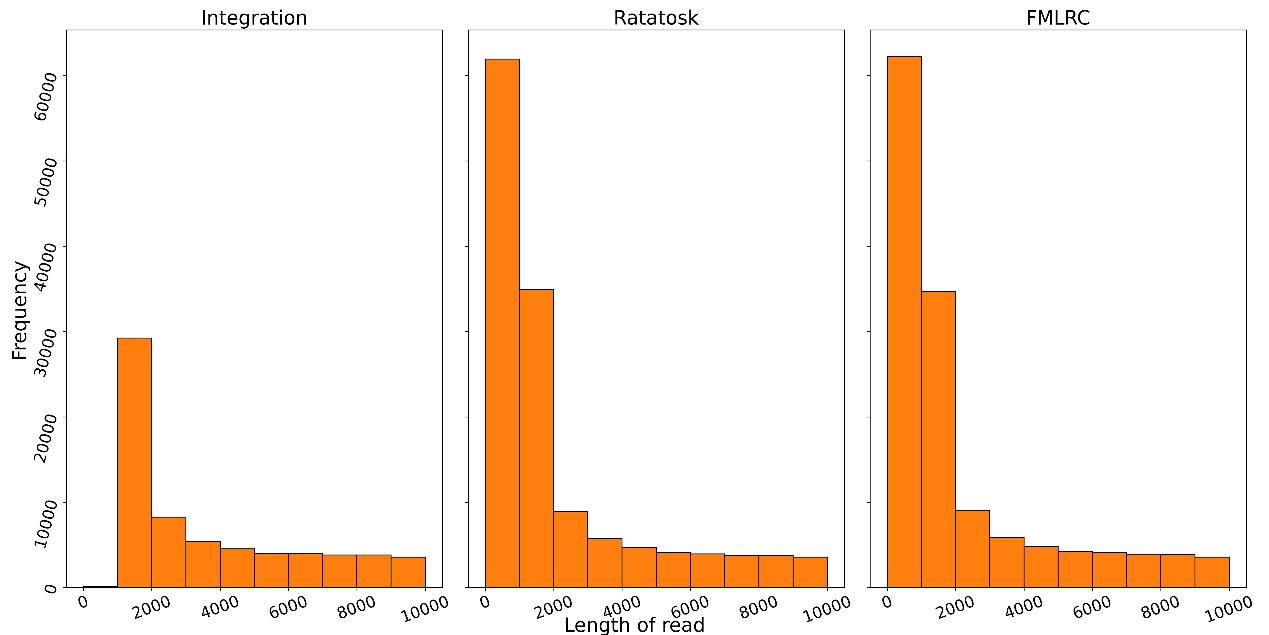
**Section 1: Distribution of Length of corrected dataset**

****

Frequency of read length in D1 after correction with different tools.

****

Frequency of read length in D2 after correction with different tools.

****

Frequency of read length in D3 after correction with different tools.

**Section 2: The intermediate result of integration approach**

Table S1: The result after the self-correction phase of integration approach

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Dataset | Relative Size | Error rate | Matched base | Matched Rate | Aligned rate | Average length | Read number | Base number |
| D1 | 92.00% | 6.63% | 75125546 | 83.14% | 96.13% | 3204 | 28204 | 90360847 |
| D2 | 81.83% | 6.43% | 237845804 | 56.50% | 59.25% | 6337 | 66423 | 422932756 |
| D3 | 90.44% | 6.22% | 429604573 | 51.41% | 64.83% | 8323 | 100398 | 835567079 |
| D4 | 99.93% | 4.00% | 84118783 | 72.69% | 84.20% | 7680 | 15067 | 115715963 |
| D5 | 97.06% | 1.06% | 83463479 | 92.63% | 96.96% | 9210 | 9783 | 90102346 |

**Section 3: The assembly result of error correction tools**

After correction, the error rate of D1,D4,D5 still differs a lot from D4 and D5. In order to examine the impact of difference in error rate for the further downstream analysis. We utilized Canu to assemble corrected result and quast (<https://github.com/ablab/quast>). The following metrics are used to measure the quality of correction:

•N50: the sequence length of the shortest contig at 50% of the total assembly length.

•NG50: the sequence length of the shortest contig at 50% of the known genome size that must be of the NG50 length or longer.

•NA50: the sequence length of the shortest aligned blocks at 50% of the total assembly length.

•N90: the sequence length of the shortest contig at900% of the total assembly length.

•Misassembled Length: total number of bases in misassembled contigs.

•Aligned Length: The total length of unaligned regions in the assembly.

As demonstrated in Table S2, the NG50 of corrected data of D2 and D3 can not be computed. In terms of the other metric, it is consistently observed that the datasets with lower error rates yield superior results compared to those with higher error rates.

Canu failed to assembly the corrected result of D1 by Lordec and D4 by Lordec+Canu, we will investigate the underlying cause of this failure..

Table S2: The comparison of assembly result between error correction tools.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Dataset | Method | Error rate | N50 | NG50 | NA50 | N90 | Misassembled Length | Aligned Length |
| D1 | Canu | 7.44% | 15671 | 13064 | 15671 | 5867 | 12512 | 3758496 |
| Lordec\* | 2.88% | -- | -- | -- | -- | -- | -- |
| Lordec+Canu | 1.96% | 250536 | 250536 | 204333 | 104446 | 681141 | 4652681 |
| Integration | 0.70% | 229829 | 229829 | 229829 | 107989 | 355893 | 4655644 |
| D2 | Canu | 11.31% | 18310 | -- | 9292 | 6089 | 450621 | 1004407 |
| Lordec | 12.52% | 14240 | -- | 8558 | 2586 | 85157 | 206668 |
| Lordec+Canu | 10.73% | 18772 | -- | 9550 | 6453 | 639260 | 1298735 |
| Integration | 10.45% | 18728 | -- | 9000 | 6627 | 645542 | 1206530 |
| D3 | Canu | 11.02% | 13991 | -- | 10160 | 6888 | 1274407 | 3589863 |
| Lordec | 13.13% | 14596 | -- | 5040 | 1708 | 27519 | 188933 |
| Lordec+Canu | 9.82% | 15257 |  | 11407 | 6982 | 1739351 | 5730772 |
| Integration | 10.39% | 14459 |  | 11213 | 7272 | 1502019 | 4678474 |
| D4 | Canu | 4.00% | 1497714 | 2472171 | 33903 | 679132 | 5416858 | 4114759 |
| Lordec | 3.80% | 630708 | 647319 | 35308 | 100395 | 5061082 | 4213279 |
| Lordec+Canu\* | 3.77% | -- | -- | -- | -- | -- | -- |
| Integration | 3.76% | 4640854 | 4640854 | 34806 | 662303 | 5406596 | 4217127 |
| D5 | Canu | 1.06% | 3691823 | 3691823 | 3691823 | 937869 | 0 | 4629692 |
| Lordec | 0.23% | 4636560 | 4636560 | 4636560 | 4636560 | 0 | 4636560 |
| Lordec+Canu | 0.12% | 3763136 | 3763136 | 3762976 | 872834 | 0 | 4635080 |
| Integration | 0.10% | 4636600 | 4636600 | 4636600 | 4636600 | 0 | 4636600 |

\*The assembly tool encountered errors while processing the corresponding data.