Bioinformatics, 2023, **39(10)**, btad631
https://doi.org/10.1093/bioinformatics/btad631
Advance Access Publication Date: 18 October 2023
Applications Note

Sequence analysis

Evaluation of haplotype-aware long-read error correction with hifieval

Yujie Guo [1,2], Xiaowen Feng [1], Heng Li[1,2,*]

[1]Department of Data Science, Dana-Farber Cancer Institute, Boston, MA 02215, United States
[2]Department of Biomedical Informatics, Harvard Medical School, Boston, MA 02215, United States

*Corresponding author. Department of Data Science, Dana-Farber Cancer Institute, 3 Blackfan St, Boston, MA 02115, USA. E-mail: hli@ds.dfci.harvard.edu

Associate Editor: Inanc Birol

Abstract

Summary: The PacBio High-Fidelity (HiFi) sequencing technology produces long reads of >99% in accuracy. It has enabled the development of a new generation of *de novo* sequence assemblers, which all have sequencing error correction (EC) as the first step. As HiFi is a new data type, this critical step has not been evaluated before. Here, we introduced hifieval, a new command-line tool for measuring over- and under-corrections produced by EC algorithms. We assessed the accuracy of the EC components of existing HiFi assemblers on the CHM13 and the HG002 datasets and further investigated the performance of EC methods in challenging regions such as homopolymer regions, centromeric regions, and segmental duplications. Hifieval will help HiFi assemblers to improve EC and assembly quality in the long run.

Availability and implementation: The source code is available at https://github.com/magspho/hifieval.

1 Introduction

The PacBio High-Fidelity (HiFi) sequencing technology produces long reads of ∼15 kb in length and >99% in accuracy (Wenger *et al.* 2019). This new data type leads to the development of several recent sequence assemblers, including HiCanu (Nurk *et al.* 2020), hifiasm (Cheng *et al.* 2021, 2022), mdBG (Ekim *et al.* 2021), LJA (Bankevich *et al.* 2022), and Verkko (Rautiainen and Marschall 2021, Rautiainen *et al.* 2023), which leverage the high accuracy of HiFi reads and could generate phased assemblies of higher quality than previous methods (Jarvis *et al.* 2022). HiFi has also played a central role in the complete assembly of the first human genome (Nurk *et al.* 2022, Rhie *et al.* 2023).

The power of HiFi read assembly mainly comes from resolving similar repeats or haplotypes by requiring exact sequence matches between reads (Nurk *et al.* 2020, Cheng *et al.* 2021). Although HiFi reads are accurate, there are still a small number of sequencing errors per read. These errors would result in inexact matches and fail the assembly algorithms. To obtain near error-free reads, all the HiFi assemblers mentioned above start with error correction (EC), a critical step to correct away most sequencing errors in reads. Errors at the EC step may be propagated to the assembly step and lead to broken contigs or misassemblies downstream.

Zhang *et al.* (2020) evaluated many EC tools developed for correcting noisy long reads. However, most of these tools disregard phasing and would collapse reads originated from different repeat copies or from different parental haplotypes in a diploid sample. They are not used by modern assemblers and are largely obsolete.

To better understand the accuracy of modern EC tools, we conducted a new benchmark using real human data. Different from Zhang *et al.* (2020), who did evaluation on small homozygous model organisms, we took high-coverage HiFi reads both from a homozygous cell line CHM13 and from a diploid individual HG002. More importantly, we developed a user-facing tool, hifieval, to measure and record under-corrected and over-corrected bases in reads. This allows other assembler developers to associate EC errors with different genomic features and to study the behavior of EC tools in detail.

2 Methods

Suppose we have the perfect assembly of a sample and a set of reads sequenced from the same sample. After EC, we expect corrected reads to be exactly aligned to the assembly. Any substitutions and gaps in the alignment would be EC errors. Hifieval makes use of this observation to evaluate correction errors.

In practice, the assembly may have errors and the sample may contain somatic mutations. We would overestimate errors due to imperfect assembly. Nonetheless, the assembly tends to be more accurate than EC as the assembly is built from multiple sequencing technologies and involves other downstream algorithms such as graph cleaning and consensus which can prune uncorrected errors. We can capture the majority of EC errors even if the assembly has minor errors.

Received: 5 June 2023; **Revised:** 18 September 2023; **Editorial Decision:** 12 October 2023**Accepted:** 17 October 2023

2.1 Evaluating tools without homopolymer compression

We align the raw reads and the corrected reads to the ground truth assembly using minimap2 v2.24+ (Li 2018). For each read, hifieval compares the alignment of the raw read and the alignment of the corrected read and calculates three metrics: correct corrections (CC), errors that are in raw reads but not in corrected reads; under-corrections (UC), errors present in both raw and corrected reads; and over-corrections (OC), new errors found in corrected reads but not in raw reads. These are analogous to true positive (TP), false negative (FN), and false positive (FP), respectively (Fig. 1). Therefore, we have the following definitions on False Discovery Rate (FDR) and False Negative Rate (FNR):

$$FDR = \frac{FP}{TP + FP} = \frac{OC}{CC + OC}$$

$$FNR = \frac{FN}{TP + FN} = \frac{UC}{CC + UC}.$$

Occasionally, raw reads and corrected reads may be mapped to different positions. Hifieval computes CC, OC, and UC anyway but we did not count them in our evaluation.

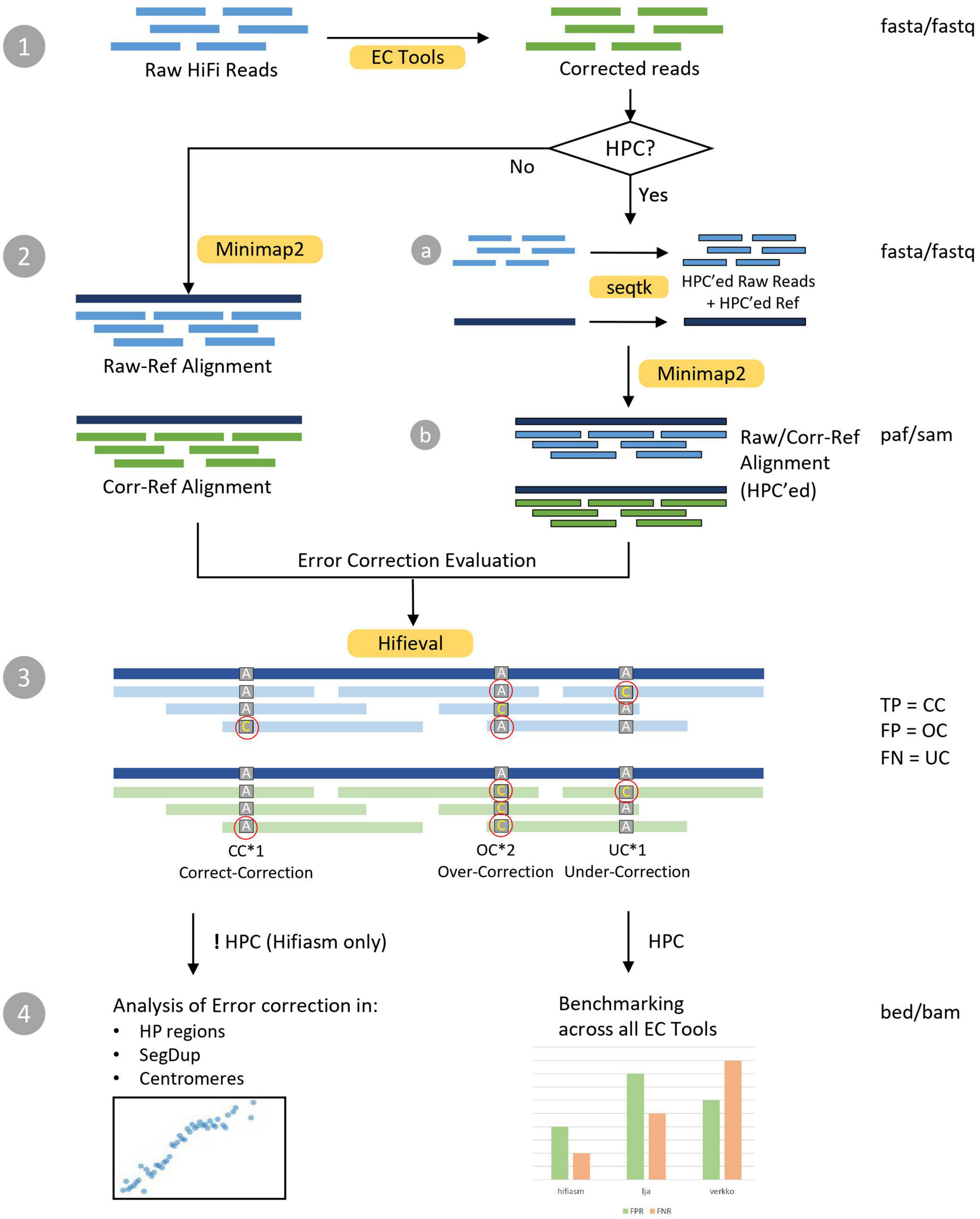

Figure 1. Hifieval workflow. Version 0.4.0.

2.2 Evaluating tools with homopolymer compression

Homopolymer compression (HPC) converts a run of identical bases, called a homopolymer, to a single base. It is a simple and widely used method for reducing the negative impacts of homopolymer expansion/contraction error from long-read sequencing technologies. Some EC tools such as Verkko and LJA only report HPC reads. To make the alignment consistent, we also apply HPC to both the truth assembly and the raw reads and perform minimap2 alignment all in the HPC encoding. CC, UC, and OC can be measured in a similar manner to the previous section.

3 Results

3.1 Datasets

3.1.1 Simulated reads

We used *Escherichia coli* str. K-12 substr. MG1655 genome as our reference to simulate the reads (AC: NC_000913.3). We simulated HiFi reads at 30-fold coverage using PBSIM2 with a uniform error rate of 0.2% across the reference (Ono *et al.* 2021). This resulted in 8970 reads from one reference chromosome.

3.1.2 Real reads

We included two real human datasets: T2T-CHM13 and HG002. For CHM13, We took the published assembly (Nurk *et al.* 2022) without the Y chromosome as ground truth and obtained reads from SRR11292120–SRR11292123 at 30-fold coverage. For HG002, we took the Verkko assembly (Rautiainen *et al.* 2023) as the ground truth and downloaded reads from SRR10382244, SRR10382245, SRR10382248, and SRR10382249. These four runs gave 34-fold coverage in total.

3.2 Evaluated EC tools

We evaluated three EC tools: hifiasm v0.19 (Cheng *et al.* 2021), LJA v0.2 (Bankevich *et al.* 2022), and Verkko v1.3 (Rautiainen *et al.* 2023). While LJA and Verkko only output corrected reads in the HPC encoding, hifiasm preserved homopolymer runs. For a consistent comparison between the three tools, we emulated hifiasm HPC correction, called hifiasmhpc, by applying hifiasm to homopolymer-compressed raw reads. For tools that correct reads sequenced from large human diploid genome, hifieval is able to evaluate the performance within an hour after the input alignment files are generated.

HiCanu (Nurk *et al.* 2020) and mdBG (Ekim *et al.* 2021) are also optimized for HiFi assembly and both have an EC step. We did not evaluate HiCanu because it used the same EC module as Verkko. mdBG corrected reads in the minimizer space. It did not output in the base space. In theory, EC tools developed for noisy long reads (Zhang *et al.* 2020) could work for accurate PacBio HiFi reads. Nonetheless, most of these older EC tools were developed for small genomes and did not scale well to human data. More importantly, they disregarded phasing and would not be useful for high-quality phased assembly.

3.3 Evaluating simulated *E. coli* reads

As a sanity check, we started with the simulated *E. coli* dataset. Both hifiasm and LJA performed well. Verkko reported many over-corrections with 25% of its corrections being wrong (Supplementary Fig. S1). We looked at the results in Integrative Genomics Viewer (IGV) but could not figure out why it made these many errors (Supplementary Fig. S2).

3.4 Evaluating on the CHM13 dataset

For each EC tool, only ∼0.2% of corrected reads were mapped to positions different from raw reads. These reads contributed to ∼0.5% of total number of corrections and thus would have little effect on the total number of OC, UC, and CC. We excluded these reads when calculating FNR and FDR.

On this dataset, hifiasm appeared to have lower FNR and FDR than hifiasmhpc (Fig. 2A). This was because hifiasm made nine times as many CC as hifiasmhpc and thus had a much larger denominator when we calculated FNR and FDR. EC tools applying HPC are not directly comparable to EC tools without HPC. Between the three EC tools applying HPC, hifiasmhpc and Verkko had similar FDR at $<0.4\%$. LJA had a much higher FDR. It also has more sparse miscorrections (Supplementary Fig. S3). Verkko on the other hand had a much higher FNR (Fig. 2A). As we looked at the raw summary statistics of Verkko performance on EC outputted by hifieval, 77% of Verkko UCs came from with reads with 10 or more UCs. For these reads, Verkko often corrected part of a read perfectly without touching the errors on the rest of the read. By looking at its log file, we found out Verkko effectively threw away 18% of data though the errors it corrected were mostly right.

For CHM13, we also stratified hifiasm FNR and FDR by annotation (Nurk *et al.* 2022). Centromeric satellites, segmental duplication, and homopolymer regions are some of the most challenging regions to assemble across a genome, yet, they might contain significant genetic information that is biologically meaningful. It is important for EC tools to develop strategies to approach the complex characteristics of such regions. Most EC steps are carried out in HPC encoding to simplify the problem and make the whole assembly pipeline more efficient, but it would sacrifice the improved accuracy in difficult regions. As expected, hifiasm had higher error rate in complex regions such as centromeric satellites and segmental duplications. FNR and FDR were only slightly elevated in homopolymer runs (Fig. 2B and Supplementary Fig. S4).

3.5 Evaluating on the HG002 dataset

All three EC tools had higher error rates on the HG002 dataset than on the CHM13 dataset (Fig. 2A) probably because the two similar parental haplotypes confused the tools. Note that the HG002 assembly was not as polished as the CHM13 assembly. Nonetheless, the contig base accuracy was about two orders of magnitude higher than HiFi base accuracy. Therefore, the great majority read-to-contig differences we saw were sequencing errors. Our evaluation method was still applicable.

For this diploid sample, we counted reads that were mapped to different positions before and after EC and classified the mapping differences into three categories: haplotype difference, chromosome difference, and position difference. The correction of a read would lead to a haplotype difference if the raw read and the corrected read were mapped to opposite parental haplotypes, and the alignment of the raw or the alignment of the corrected read had a mapping quality 2 or higher. Similarly, we counted a chromosome difference if the raw read and the corrected read were mapped to different

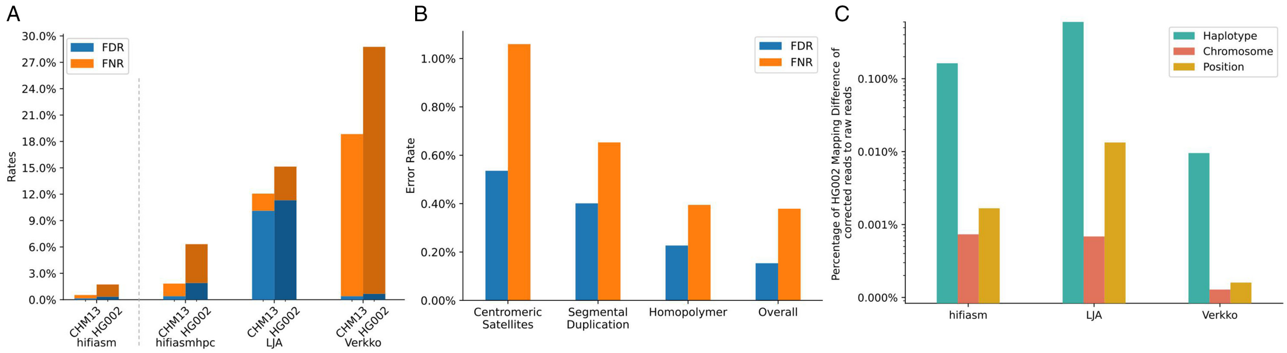

Figure 2. (A) The FNR and FDR of reads EC of each tool on CHM13 (haploid) and HG002 (diploid), respectively. Note that hifiasm corrects reads without HPC, while LJA and Verkko do. Hifiasm can perform EC using HPC reads to make the results more comparable. (B) Hifiasm EC performance in CHM13 challenging regions. (C) Percentage of HG002 mapping difference of corrected reads to raw reads against Verkko assembled reference. Reads are filtered such that at least either one of raw or corrected read has mapping quality ≥2.

chromosomes on the same haplotype and counted a position difference if the raw read and the corrected read were mapped to different positions on the same chromosome.

Figure 2C showed that Verkko was least likely to correct reads onto the opposite haplotype, at an error rate an order of magnitude lower than hifiasm and LJA (Supplementary Fig. S5). We checked the haplotype errors made by hifiasm and found that most of the errors occurred in long regions where there were a few heterozygous insertions and deletions (INDELs) but no heterozygous SNPs. These haplotype differences were likely to cause missing INDELs. As we increased the mapping quality threshold to 10, the large haplotype difference between hifiasm and Verkko was much reduced (Supplementary Fig. S6), suggesting hifiasm and Verkko would rarely correct a read to a very different haplotype. We also performed similar analysis on CHM13 dataset (Supplementary Figs S7–S9).

In Fig. 2A and C, we used the Verkko assembly as the ground truth. This could potentially be biased toward Verkko. To evaluate this possibility, we also did the same set of analyses but taking a hifiasm assembly as the ground truth. The results remained similar, and changing assembly has minor effect (Supplementary Fig. S10). This has shown the quality of the assembly between hifiasm and Verkko would not change the overall result that Verkko has fewer correction errors and haplotype mapping differences in HG002. In fact, consensus accuracy with a 50–60 Phred score is generally much higher than read accuracy, whose Phred score is around 23 (Wenger *et al.* 2019, Rautiainen *et al.* 2023). This means that a typical assembly is a thousand folds more accurate than the reads. Although in some cases, large-scale mis-assembly may indeed affect the results of hifieval, we don't need perfect assembly in order to sufficiently evaluate the read EC performance.

4 Conclusion

Hifieval is a fast tool that performs systematic benchmark of haplotype-aware long read EC tools. Unlike earlier studies, hifieval evaluates phased assemblies and can distinguish under-corrections and over-corrections. It is perhaps the first user-facing EC evaluation tool that can be easily deployed to users' own datasets. In our analyses, it is apparent that current EC tools all have their weakness in different aspects. We hope hifieval can help the development of more accurate EC tools which would be essential to high-quality assembly.

Funding

This work is supported by US National Institute of Health grant R01HG010040 and U01HG010961 to H.L.

Author contributions

H.L. conceived the project. Y.G. and X.F conducted the experiments. Y.G. analyzed the results. Y.G. and H.L. wrote the manuscript.

Supplementary data

Supplementary data are available at *Bioinformatics* online.

Conflict of interest

None declared.

Data availability

The CHM13 reference genome and HiFi reads are available at https://github.com/marbl/CHM13. HG002 reference genome and HiFi reads are available at https://github.com/marbl/HG002. Hifieval test data on *E. coli* is available at https://zenodo.org/record/7799845, and raw output data of hifieval on CHM13 data described in the results section is available at https://zenodo.org/record/8316047.

References

Bankevich A, Bzikadze AV, Kolmogorov M *et al.* Multiplex de Bruijn graphs enable genome assembly from long, high-fidelity reads. *Nat Biotechnol* 2022;**40**:1075–81.

Cheng H, Jarvis ED, Fedrigo O *et al.* Haplotype-resolved assembly of diploid genomes without parental data. *Nat Biotechnol* 2022;**40**:1332–5.

Cheng H, Concepcion GT, Feng X *et al.* Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nat Methods* 2021;**18**:170–5.

Ekim B, Berger B, Chikhi R. Minimizer-space de Bruijn graphs: whole-genome assembly of long reads in minutes on a personal computer. *Cell Syst* 2021;**12**:958–68.

Jarvis ED, Formenti G, Rhie A *et al.* Semi-automated assembly of high-quality diploid human reference genomes. *Nature* 2022;**611**:519–31.

Li H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* 2018;**34**:3094–100.

Nurk S, Walenz BP, Rhie A *et al.* Hicanu: accurate assembly of segmental duplications, satellites, and allelic variants from high-fidelity long reads. *Genome Res* 2020;**30**:1291–305.

Nurk S, Koren S, Rhie A *et al.* The complete sequence of a human genome. *Science* 2022;**376**:44–53.

Ono Y, Asai K, Hamada M. PBSIM2: a simulator for long-read sequencers with a novel generative model of quality scores. *Bioinformatics* 2021;**37**:589–95.

Rautiainen M, Nurk S, Walenz BP *et al.* Telomere-to-telomere assembly of diploid chromosomes with verkko. *Nat Biotechnol* 2023;**41**:1474–82.

Rautiainen M, Marschall T. Mbg: minimizer-based sparse de bruijn graph construction. *Bioinformatics* 2021;**37**:2476–8.

Rhie A, Nurk S, Cechova M *et al.* The complete sequence of a human Y chromosome. *Nature* 2023;**621**:344–54.

Wenger AM, Peluso P, Rowell WJ *et al.* Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. *Nat Biotechnol* 2019;**37**:1155–62.

Zhang H, Jain C, Aluru S. A comprehensive evaluation of long read error correction methods. *BMC Genomics* 2020;**21**:889.