Polishing the Oxford Nanopore long-read assemblies of bacterial pathogens with Illumina short reads to improve genomic analyses

ABSTRACT

-**Oxford Nanopore** sequencing has been widely used to achieve complete genomes of bacterial pathogens. However, the error rates of long reads are high **(10%–15%).**

-**Polishing algorithms** using Illumina short reads to correct the errors in Oxford Nanopore long-read assemblies have been developed. (10 Strains) were selected for simulated reads, while real reads were tested on (11 Strains).

-Oxford Nanopore long reads were assembled with **Unicycler** to produce a draft assembly, followed by three rounds of polishing with Illumina short reads using two polishing tools, Pilon and NextPolish.

-One round of NextPolish polishing generated genome completeness and accuracy parameters similar to the reference genomes, whereas two or three rounds of Pilon polishing were needed, though contiguity remained unchanged after polishing.

-The polished assemblies of Escherichia coli O157:H7 , Salmonella Typhimurium, and Cronobacter sakazakii with simulated reads did not provide accurate plasmid identifications.

-One round of NextPolish polishing was needed for accurately identifying plasmids in Staphylococcus aureus and E. coli O26:H11 with real reads, whereas one and two rounds of Pilon polishing were necessary for these two strains, respectively.

-Polishing failed to provide an accurate antimicrobial resistance (AMR) genotype for S. aureus with real reads.

-One round of polishing recovered an accurate AMR genotype for Klebsiella pneumoniae with real reads.

-The reference genome and draft assembly of Citrobacter braakii with real reads differed, which carried blaCMY-83 and fosA6, respectively, while both genes were present after one round of polishing.

-Polishing did not improve the assembly of E. coli O26:H11 with real reads to achieve numbers of virulence genes similar to the reference genome.

-The draft and polished assemblies showed a **phylogenetic tree topology** comparable with the reference genomes.

-For multilocus sequence typing and pan-genome analyses, one round of NextPolish polishing was sufficient to obtain accurate results, while two or three rounds of Pilon polishing were needed.

-*Overall*, NextPolish outperformed Pilon for polishing the Oxford Nanopore long-read assemblies of bacterial pathogens, though both polishing strategies improved genomic analyses compared to the draft assemblies.

INTRODUCTION

-Advances in next-generation sequencing (**NGS**) have ushered for bacterial pathogens. The low-cost Illumina reads make them high-throughput genomics, allowing Illumina platforms to become the most dominant technology for whole-genome sequencing (**WGS**) of bacteria .The major shortcoming of Illumina reads is their length: maximally **300** bp but more commonly ≤**150** bp .

-Using Illumina short reads alone cannot resolve repeated sequences longer than the reads Illumina sequencing can span, and thus may fail to assemble genomic duplications.
While Illumina short reads are sufficient for some genomic analyses such as strain typing and outbreak tracing,
-Incomplete bacterial genomes can lead to mapping artifacts, missed gene calls, and inaccurate repeat construction .
-It may be impossible to tell from an incomplete Illumina short-read assembly whether genes of interest such as antimicrobial resistance genes (ARGs) reside on the chromosome or a plasmid,as the locations of these genes have substantial values for epidemiological implications.

-**Oxford Nanopore sequencing** can overcome these limitations by generating long reads on the principle of real-time nanopore strand sequencing, which increases read lengths **(100- to 1000)** fold and spans much longer repeat regions compared to Illumina platforms.
-The long-read lengths of Oxford Nanopore sequencing and improvements to bioinformatic tools to assemble bacterial genomes from long reads make this technology complete genomes of bacterial pathogens and the full structure of a bacterial genome can reveal the locations of all genes,but have a high error rate of

-Although there are concerns about the high error rate with Oxford Nanopore sequencing, integrated approaches that efficiently utilize both long and short reads can overcome this problem.
One solution is to polish the Oxford Nanopore long-read assembly with highly accurate Illumina short reads to improve the accuracy .This approach does not rely on specific standalone assemblers but can be accomplished using long-read assemblers to create a draft assembly, followed by a polishing step with Illumina short reads.

-So far, Pilon has been the most widely used polishing tool in the published literature beyond the initial release of the software.
Recently, Hu et al. has developed a fast and efficient polishing tool, NextPolish, that can correct sequence errors in long-read assemblies.
Vera et al. used NextPolish to error-correct the Oxford Nanopore long-read assemblies of six *Bacillus megaterium* strains with Illumina short reads. However, NextPolish has not been reported to polish the Oxford Nanopore long-read assemblies of bacterial pathogens.

-*In the current study*, we used Pilon and NextPolish to polish the Oxford Nanopore long-read assemblies of bacterial pathogens with Illumina short reads.After each round of polishing, we assessed and compared the genome completeness and accuracy to the reference genomes.
The impact of iterative polishing on improving genomic analyses of bacterial pathogens was evaluated.