

Abstract

Although AI-driven basecalling has propelled genome sequencing forward, the reliance on imperfect training data often introduces noise and systematic biases. This compromises data fidelity, and prevents the integration of genome sequencing into clinical applications, such as personalized medicine. In contrast, our research champions a paradigm shift toward single-molecule sequencing (SMS) accuracy, grounded in precisely engineered plasmid-based DNA standards and a robust computational framework. This strategy lays the groundwork for standardizing single-molecule sequencing error assessment, with the ultimate goal of realizing clinically robust genome analyses. By integrating precise plasmid standards and detailed error profiling, we move closer to achieving the single molecule sequencing accuracy required for widespread adoption of genome sequencing in healthcare.

Introduction

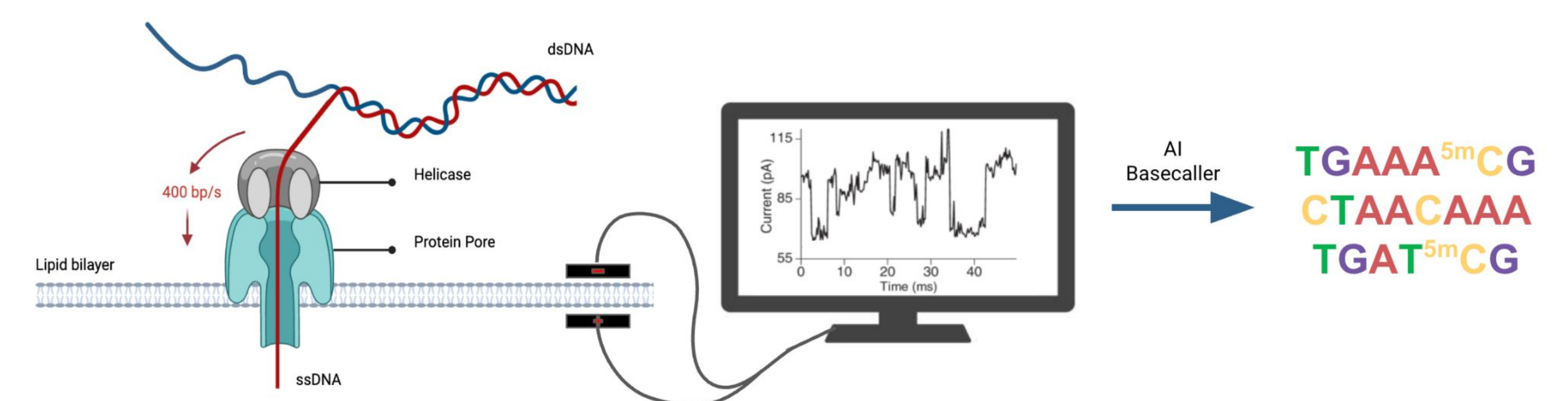


Figure 1. Diagram of Nanopore Mechanism¹

Background

Long-read sequencing enables the analysis of extended DNA fragments without fragmentation, making it essential for accurately sequencing complex genomic regions. Currently the accuracy of basecalling long sequences is measured by how well the sequences align to a portion of the reference. However, when sequencing an entire DNA molecule, the aim should be single molecule sequencing accuracy, how well the basecalled sequence matches the entire actual sequence. In this study, we employed Oxford Nanopore MinION, which detects nucleotide sequences by measuring electrical current as DNA strands pass through a protein pore. These signals are recorded in POD5 files and translated back into sequences using AI

Objective: Our study aims to quantify the empirical single molecule sequencing error rate of Nanopore basecallers using highly pure molecular standards with empirically determined sequence and fragment size. We hypothesize that our metric will provide a more valid and accurate assessment of Nanopore's basecallers.

Methodology

Creating the Dataset

We utilized Nanopore POD5 data collected from standardized CYP2D6 plasmids from a purified E.Coli cell line. The plasmids were sequence verified using Sanger sequencing to generate the known reference sequence. These plasmids were cut with Eco53KI to produce blunt ends, which were then dA tailed using Taq polymerase. This substrate was ligated to the ligation adapters from the ONT Ligation Sequencing Kit (SQK-LSK114). These were then sequenced using the Nanopore MinION Flow cell (FLO-MIN114).

References:

[1] Wang, Y., Zhao, Y., Bollas, A. *et al* Nanopore sequencing technology, bioinformatics and applications. *Nat Biotechnol* 39, 1348–1365 (2021). <https://doi.org/10.1038/s41587-021-01108-x>

Bioinformatic Workflow

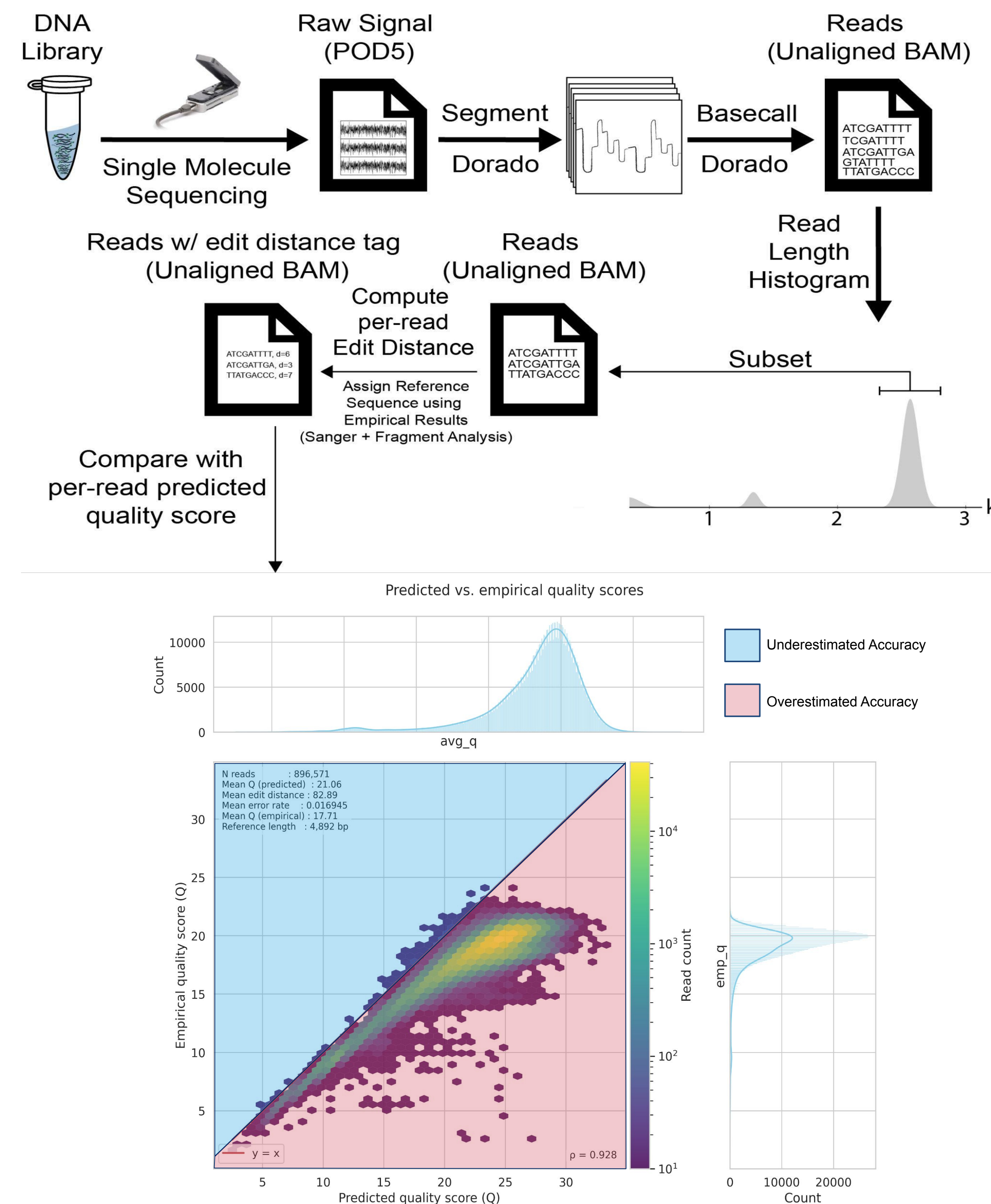
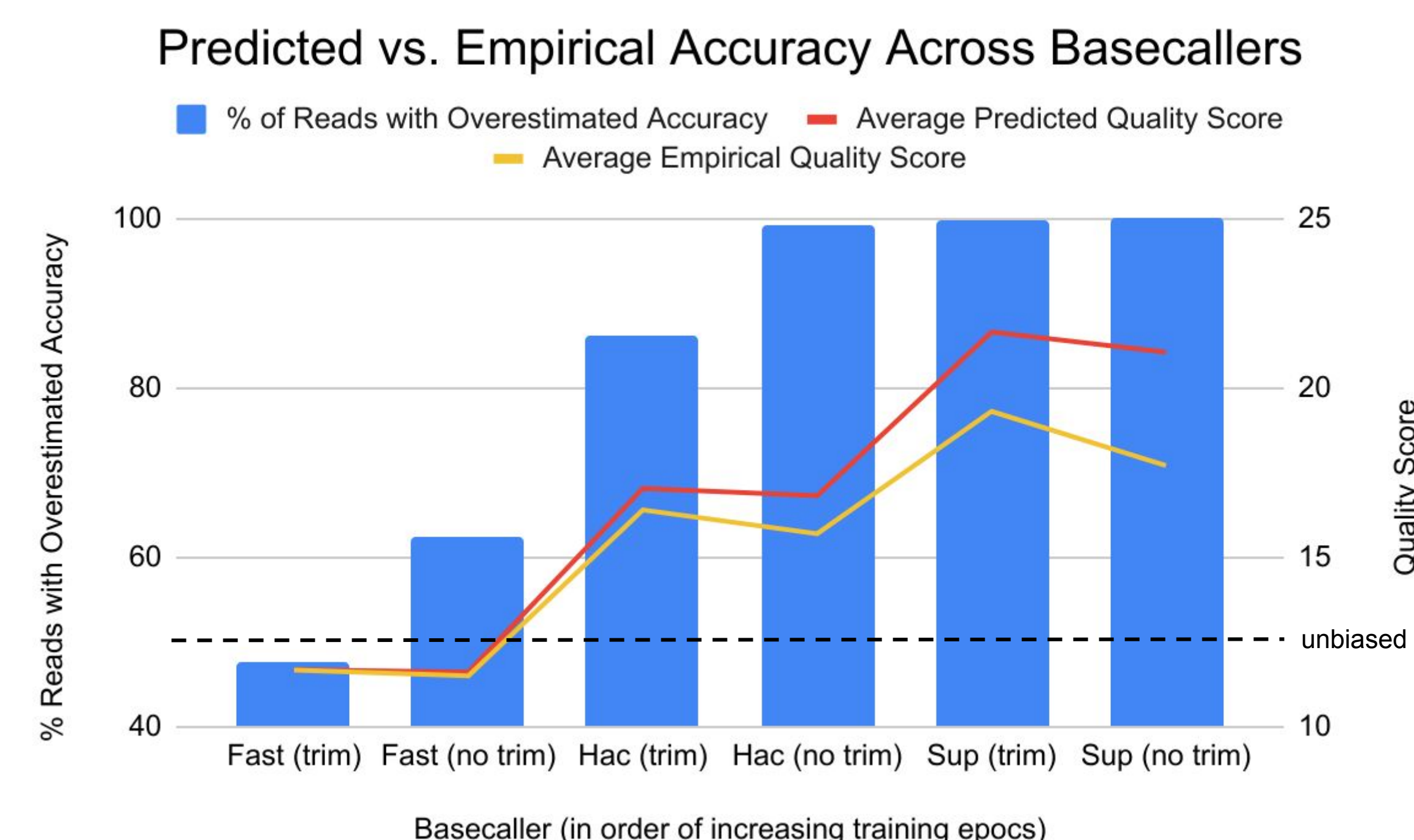


Figure 2. Workflow for Deriving Empirical Error Rates from Nanopore Sequencing Data and Generating 2D Kernel Plots. After initial basecalling, the reads were subsetting by length to ensure one molecular species. This subsetting BAM file was compared against the reference sequence assigned through empirical results to generate the edit distances. Finally, these edit distances were translated to a standard Phred Quality Score and compared against Nanopore's predicted quality scores. Red shading indicates reads where predicted quality exceeds empirical (overestimated accuracy); blue indicates the opposite (underestimated accuracy).



Results

Significant Results

- Majority of basecallers produced more reads with overestimated quality scores (red) than underestimated quality scores (blue)
- The sup model had the highest proportion of overestimated quality scores and fast model had the lowest with it reads appearing to be approximately evenly distributed between over and underestimated quality scores

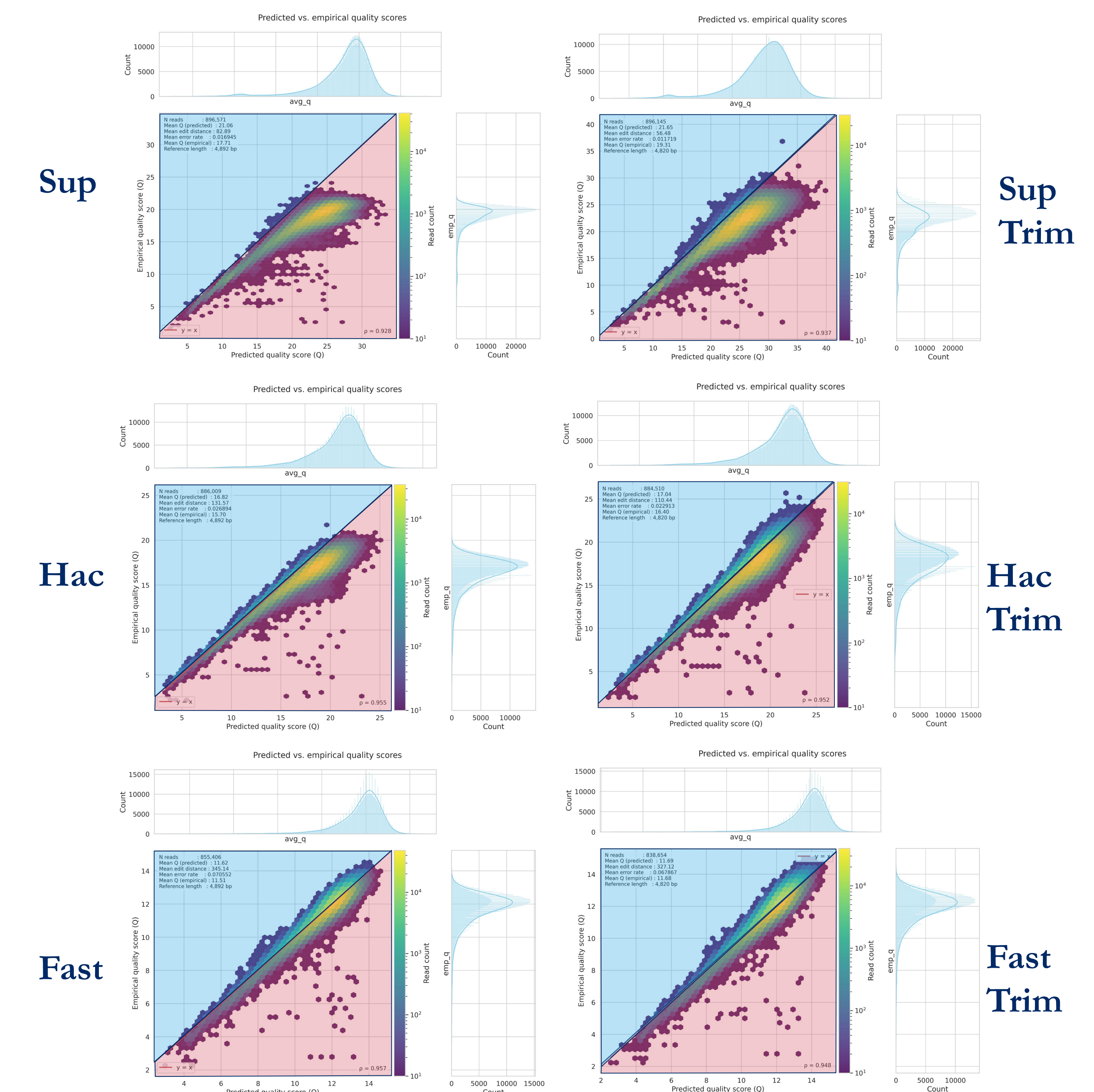


Figure 4. Comparison of Empirical vs. Predicted Quality Scores Across Multiple Basecallers Super Accurate (Sup), High Accuracy (Hac), and Fast are the three primary basecallers from Nanopore. Fast has the least parameters and is typically used in live basecalling. Sup has the most parameters and is commonly used in detailed analysis. Each has an option to trim in which the basecaller will remove the signal corresponding to the ligation adapter sequence to improve the overall quality score

Discussion and Conclusions

- None of the basecallers were able to call a perfect sequence, even sup (the model with the most parameters and most extensive training)
- Our analysis on this plasmid suggests that the more trained the models were the more inflated the estimated accuracy (quality score) was
- By reducing these graphs to a single metric of over or underestimation when retraining the basecallers, we can quantitatively assess our retraining program's effectiveness while minimizing the risk of overfitting.

Acknowledgements: Supported by Phenomics Health Inc.