Quantitative Detection of DNA Modifications from Nanopore Sequencing Data without Raw Signals

Zhixing Feng

Department of Clinical Genetics, Xinhua Hospital affiliated to Shanghai Jiao Tong University School of Medicine, Shanghai, China.

Email: fengzhixing@shsmu.edu.cn. Twitter: @zxfeng_thu. WeChat: zhixingfeng.

Motivation

The state of the art: All the existing methods require ONT raw-signal FAST5 files to detect DNA modifications.

The bottleneck: Difficult to repurpose public ONT data to detect DNA modifications because over 85% of public ONT data do NOT include FAST5 files or BAM files with modifications (**Fig 1**).

The solution: Developing a novel method to detect DNA modifications using only FASTQ files.

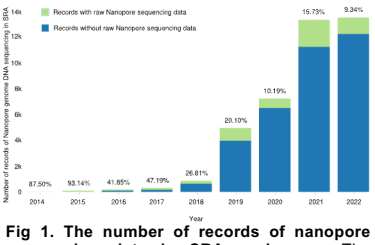

Fig 1. The number of records of nanopore sequencing data in SRA each year. The percentage on the top of each bar is the ratio of records with raw signals.

Error pattern and base QV are effective features to predict DNA modifications

As shown in **Fig 2**, We found that hypermethylated the error pattern and base QV (Quality Value) are significantly different between hypermethylated CpG and hypomethylated CpG in HG01109 of Human Pangenome Project (R9.4.1 + Guppy 4.2.2).

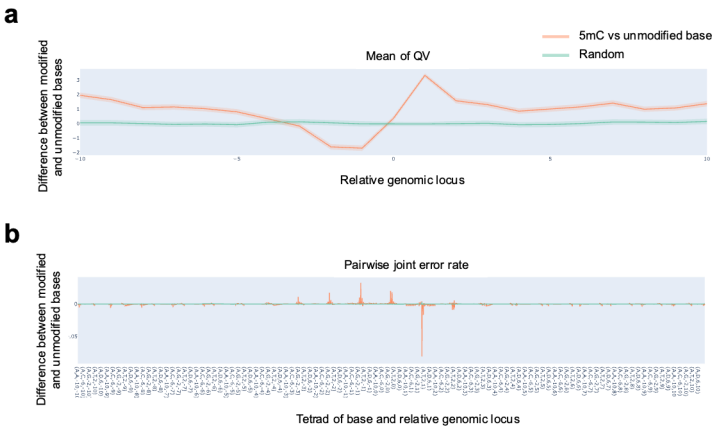

Fig 2. The impact of 5mC on different features. a, Mean of QV. b, Pairwise joint error rates.

NanoFreeLunch: a novel framework to detect DNA modifications quantitatively from nanopore sequencing data without raw signal

We developed a novel framework termed NanoFreeLunch, which extracts features from the sequencing errors and base QVs around the genomic locus of interest and use the known DNA modification level as response to train a Boosting model to predict DNA modifications quantitatively (**Fig 3**).

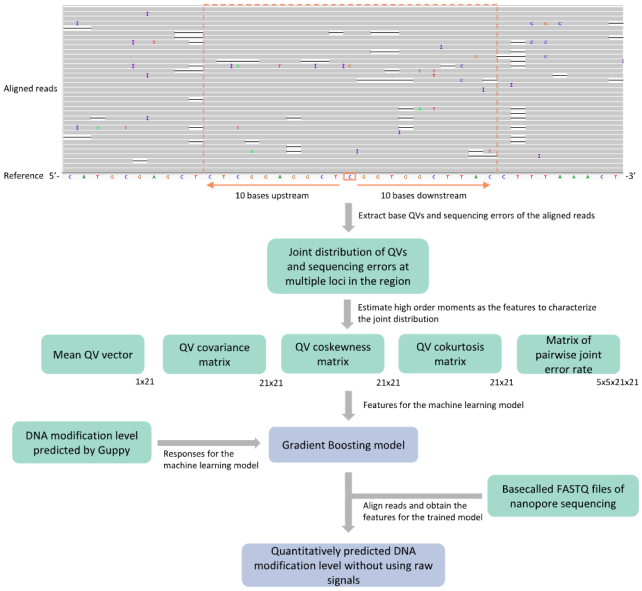

Fig 3. The workflow of NanoFreeLunch. The aligned reads are shown in an IGV snapshot. IGV = Integrative Genomics Viewer.

The results of NanoFreeLunch are consistent with BS-seq and Guppy

We evaluated NanoFreeLunch on the GM24385 data from ONT Open Data and Human Pangenome data (R9.4.1 + Guppy 6.3.8). On the GM24385 data, the PCC of locus-level methylation level between NanoFreeLunch and BS-seq (Bisulfite Sequencing) is 0.67 and the PCC is 0.70 between NanoFreeLunch and Guppy 6.3.8. The PCC of average methylation level of CpG islands is 0.95 for both of Guppy and BS-seq. On the Human Pangenome data, the PCCs of locus-level range from 0.67 to 0.73 if basecalling is performed by Guppy 5.0.15 or Guppy 6.3.8 and range from 0.75 to 0.81 if basecalling is performed by Guppy 4.2.2. The PCCs of average methylation level of CpG islands range from 0.95 to 0.99.

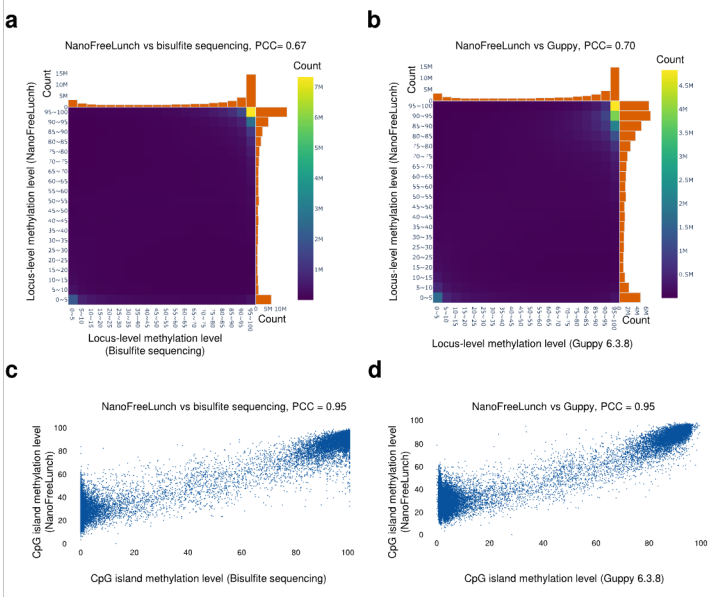

Fig 4. Evaluating the performance of NanoFreeLunch. a-b, Comparing the methylation level of each CpG locus estimated by NanoFreeLunch with bisulfite sequencing and Guppy 6.3.8 on the GM24385 data. PCC = Pearson Correlation Coefficient. c-d, Comparing the average methylation level of CpG islands estimated by NanoFreeLunch with bisulfite sequencing and Guppy 6.3.8 on GM24385 data.

The DNA modifications restored from FASTQ files are accuracy enough to provide meaningful biological insights

We evaluated NanoFreeLunch on imprinting control regions (ICRs), regions with histone modification H3K9me3, and the overlaps between H3K4me3 regions and DNase hypersensitive regions. The results demonstrate that NanoFreeLunch recaptures the established epigenomic knowledge: ICRs are partially methylated (about 50%); H3K9me3 regions are associated with cytosine hypermethylation and repressed transcription; The overlaps between H3K4me3 regions and DNase hypersensitive regions are associated with hypomethylation and active transcription.

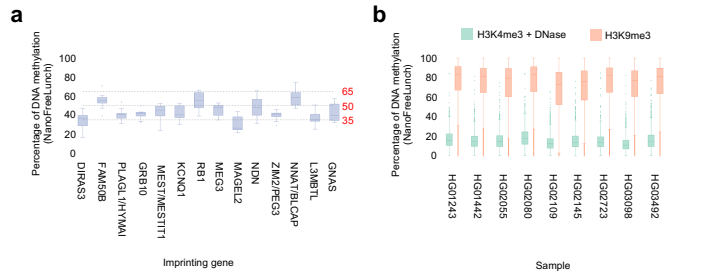

Fig 5. Evaluating the performance of NanoFreeLunch. a, The average methylation level estimated by NanoFreeLunch for each imprinting control region (ICR). b, The average methylation level estimated by NanoFreeLunch in H3K9me3 regions and the overlaps between H3K4me3 regions and DNase hypersensitive regions reported in GM12878 from ENCODE.

Conclusions

- NanoFreeLunch can detect DNA modifications quantitatively from nanopore sequencing data without raw signals.
- NanoFreeLunch enables the construction of large-scale databases of DNA modifications by repurposing nanopore genome sequencing data.

Acknowledgements

This work is supported by the National Key R&D Program of China (No. 2022YFC2703400).