DNA methylation calling tools for Oxford Nanopore sequencing: a survey and human epigenome-wide evaluation

Yang Liu[1,#], Wojciech Rosikiewicz[1,#,†], Ziwei Pan[1,2,#], Nathaniel Jillette[1], Aziz Taghbalout[1], Jonathan Foox[3,4], Christopher Mason[3,4,5,6], Martin Carroll[7], Albert Cheng[1,2,], Sheng Li[1,2,8,9,*]

1. The Jackson Laboratory for Genomic Medicine, Farmington, CT, USA.
2. Department of Genetics and Genome Sciences, UConn Health Center, Farmington, CT, USA.
3. Department of Physiology and Biophysics, Weill Cornell Medicine, New York, New York, USA.
4. The HRH Prince Alwaleed Bin Talal Bin Abdulaziz Alsaud Institute for Computational Biomedicine, Weill Cornell Medicine, New York, New York, USA.
5. The Feil Family Brain and Mind Research Institute, New York, New York, USA.
6. The WorldQuant Initiative for Quantitative Prediction, Weill Cornell Medicine, New York, NY, USA.
7. Department of Medicine, University of Pennsylvania, Philadelphia, PA, USA.
8. The Jackson Laboratory Cancer Center, Bar Harbor, ME, USA.
9. Department of Computer Science and Engineering, University of Connecticut, Storrs, CT, USA.

[#]These authors contribute equally.

[†]Present address: Center for Applied Bioinformatics, St. Jude Children's Research Hospital, Memphis, TN, USA.

*Correspondence should be addressed to S.L. (sheng.li@jax.org).

Abstract

Background: Nanopore long-read sequencing technology greatly expands the capacity of long-range single-molecule DNA-modification detection. A growing number of analytical tools have been actively developed to detect DNA methylation from Nanopore sequencing reads. Here, we examine the performance of different methylation calling tools to provide a systematic evaluation to guide practitioners for human epigenome-wide research.

Results: We compare five analytic frameworks for detecting DNA modification from Nanopore long-read sequencing data. We evaluate the association between genomic context, CpG methylation-detection accuracy, CpG sites coverage, and running time using Nanopore sequencing data from natural human DNA. Furthermore, we provide an online DNA methylation database (https://nanome.jax.org) with which to display genomic regions that exhibit differences in DNA-modification detection power among different methylation calling algorithms for nanopore sequencing data.

Conclusions: Our study is the first benchmark of computational methods for mammalian whole genome DNA-modification detection in Nanopore sequencing. We provide a broad foundation for cross-platform standardization, and an evaluation of analytical tools designed for genome-scale modified-base detection using Nanopore sequencing.

Keywords: DNA methylation, base modification, long-read sequencing, Nanopore sequencing, methylation calling

Background

DNA methylation, the process by which methyl groups are added to DNA molecules, is a fundamental epigenetic modification process in gene transcription regulation [1]. Several DNA modifications, such as N6-methyladenine (6mA), N4-methylcytosine (4mC), and 5-methylcytosine (5mC) and its oxidative derivatives, are diversely distributed in genomes and play important roles in genomic imprinting, chromatin structure modulation, transposon inactivation, stem cell pluripotency and differentiation, inflammation, and transcription repression regulation [2-4]. DNA methylation measurement has traditionally depended on the combination of bisulfite conversion (which can damage DNA) and next-generation sequencing, which only detects short-range methylation pattern [5].

Recently, third-generation sequencing technologies, including single molecule real-time (SMRT) sequencing by Pacific Biosciences (PacBio), and Nanopore sequencing by Oxford Nanopore Technologies (ONT), have overcome the length limitation to achieve ultra-long read, single-base detection at a genome-wide level [6, 7]. SMRT sequencing can detect 5mC based on polymerase kinetics at 250x coverage [8]. This is due to the subtle impact of 5mC on polymerase kinetics [8]. Thus, the high coverage requirement and direct single-molecule 5mC detection by SMRT is still challenging [9]. Single molecule real-time bisulfite sequencing allows to sequence up to ~2kb amplicons but it relies on bisulfite conversion [10].

Nanopore sequencing, instead of using a sequencing-by-synthesis method to detect signal for the amplified DNA fragment population, is able to directly detect DNA or RNA translocation through a voltage-biased Nanopore sensor, enabling rapid long-read sequencing and single-base and single-molecule sensitivity [11]. Several different versions of Nanopore chemistry have been developed by ONT to improve the accuracy of single-cell molecular identification (**Figure 1A**). The initial pore version of flow cells, termed R6/R7, was replaced by R9 pore series. R9 pore series were derived from the bacterial amyloid secretion pore gene Curlin sigma S-dependent growth (CsgG) to yield a modal (i.e., most commonly observed) accuracy of up to 95% at the single-molecule level [12, 13]. Q scores, also known as Phred quality scores, are logarithmically linked to the error probability (P) of each called base: $Q = -10 \times log_{10}(p)$. Q scores measure the accuracy of nucleobase identification in DNA sequencing. Higher Q values correspond to lower error probability and higher quality [18,19]. For example, Q30 indicates that the chance that a specific base is called incorrectly is 1 in 1000. R9-series pores, R9.4 and its slightly updated, broadly used version R9.4.1, are the most favored version and can achieve the best consensus accuracy at 99.99% (Q45) [14, 15]. Recently, ONT released Nanopore R10 with a predicted model accuracy of 94% [16, 17], and introduced the newest version R10.3 with of 99.995% single molecule consensus accuracy, which has a longer barrel and a dual-reader head inside the pore [15, 18]. The current study is conducted on R9.4 series version.

Nanopore sequencing techniques enables DNA modification detection due to the difference in the electric current intensity produced from a nanopore read, termed “squiggles”. Specifically, the ionic-current resulting from the passage of modified bases through the pores differs from the current produced by the passage of unmodified bases

[19, 20]. The difference can be determined after nanopore read base calling and alignment by: (1) statistical tests comparing to an in silico reference or a non-modified control sample [21, 22]; (2) pre-trained supervised learning models such as neural network [23-25] and Hidden Markov Model (HMM) [9, 26]. However, DNA-methylation detection using Nanopore data presents a methodological challenge, i.e., accurate detection of DNA modifications in CpG sites (CpGs) termed non-singletons. A 10-base-pair (bp) region that contains only one CpG site is defined as a singleton, while a 10-bp region that contains more than one CpG site are called non-singletons [9]. The primary difficulty is the capacity to detect modifications in different CpGs that are in close proximity to one another on a DNA fragment, as it is assumed that all CpGs within a 10-bp region share the same methylation status. Several methylation calling tools have been developed to handle singletons to improve DNA-methylation detection accuracy (**Table 1**), but DNA-methylation detection power for non-singletons containing both methylated and unmethylated states remains difficult [9, 27]. Also, DNA methylation level is not linearly distributed across the genome and is dependent on genomic context [28-30]. Therefore, the accuracy of methylation callers likely differ among various types of genomic regions within which CpGs are located. However, there is no published guideline and systematic comparison of current DNA methylation calling tools for Nanopore sequencing using human natural DNA [31], especially at whole epigenome scale [32, 33].

Here, we present the first benchmark of computational methods for detecting of DNA 5-methylcytosine (5mC) from human Nanopore sequencing data at whole genome scale. We assess the impact of CpG locations on detection accuracy using human whole genome nanopore sequencing data, with a focus on the impact of genomic context and singletons vs. non-singletons. It has been reported that even homogeneous cell populations can exhibit cell-to-cell variations in epigenetic pattern (epiallele) such as gain or loss of cytosine methylation at specific loci [34]. Such epigenetic heterogeneity is increasingly recognized as a contributor to biological variability in tumors and worse clinical outcomes in malignancies [5]. Thus, to enable assessment of this critical epigenetic heterogeneity, we have evaluated the DNA methylation accuracy at single-molecule and single-base resolution, which is critical for epigenetic heterogeneity assessment [35-38]. This comprehensive survey and systematic comparison offer user-specific, best-practice recommendations to maximize accurate detection of 5mC using current methylation calling tools and provide guidance for next generation calling tools. We also generated and made available a R Shiny database to distribute the modification-detection power associated with different genomic regions using different tools for development of future algorithms and analytic tools development.

Results

Benchmarking dataset

We used four datasets for benchmarking: Nanopore sequencing of the human B-lymphocyte cell line NA19240 (hereafter referred to as NA19240 in the following text) [39], human leukemia cell lines K562 and HL-60 (referred to as K562 and HL-60), and a human primary acute promyelocytic leukemia clinical specimen (referred to as APL).

NA19240 was sequenced at ~32x coverage by the 1000 Genomes Project [39] as a high-coverage dataset. We take the union of sites from two reduced representation bisulfite sequencing (RRBS) replicates for NA19240 as corresponding DNA methylation ground truth. We generated nanopore sequencing data for K562, HL-60, and APL with ~1-3x coverage and whole genome bisulfite sequencing (WGBS) for APL. We used the published WGBS for K562 and RRBS for HL-60 as ground truth.

Overall strategy to compare DNA methylation calling tools

Several methylation calling tools have been developed to detect DNA methylation using Nanopore direct DNA sequencing data (**Table 1**). Among the nine tools, seven tools are compatible with R9.4 flow cells and six of these tools can predict 5-methylcytosine (5mC). To compare the performance of these state-of-the-art methylation calling tools, we developed a three-step standardized workflow to compare five methylation calling tools targeting 5mC in CpG context compatible with the most favored Nanopore flow cell version (R9.4.1 pores): Nanopolish [9], Megalodon [24], DeepSignal [27], Tombo/Nanoraw (referred to as Tombo) [21], and DeepMod [25] (**Figure 1B, Figure S1**). Nanopolish, Megalodon, DeepSignal and DeepMod, is model-based while Tombo is statistics-based. We excluded methBERT [40], as its repository is still under active development.

Step 1. Base-calling and quality control. To translate raw signal data into nucleotide sequences, we conducted the base calling step for Nanopore reads with Guppy (v4.2.2). Then we used NanoPack [41] for data visualization and processing, in order to assess the read length and quality of the base-called and to demultiplex sequencing data for downstream analysis. The APL, K562, and HL-60 ONT datasets exhibited comparable read length and base quality compared to the published NA19240 ONT dataset [39] (**Figure 2A-B**). Distribution of CpG sites distribution based on singletons/non-singletons is shown in (**Table S1, Figure S2**), while the number of CpG sites in various genomic contexts distribution is shown in **Figure 2C and 2D**.

Step 2. Genome assembly and polishing. We aligned the base-called reads to the human genome assembly GRCh38/hg38 using minimap2 [42] for all five tools. The electric current signal level data of a nanopore read produced by an ONT sequencer is called a squiggle. Base calling a squiggle, i.e., translating the current signal into a DNA sequence, typically contains some errors when comparing to a reference sequence [43]. The Tombo re-squiggle algorithm refines the assignment from a squiggle to a reference sequence after base-calling and alignment. The re-squiggle algorithm is required by Tombo and DeepSignal for DNA methylation calling.

Step 3. Methylation calling. We detected 5mCs in CpG context with five methylation calling tools: Nanopolish [9], Megalodon [24], DeepSignal [27], Tombo [21], and DeepMod [25]. We then designed three performance evaluation criteria (**Figure 1B and S1**) to compare the performances of each methylation calling tools. First, we evaluated the predictions of 5mCs at single-molecule, single-base resolution based on per-read prediction accuracy of fully methylated or fully unmethylated CpG sites determined by bisulfite sequencing data. We examined various biologically relevant genomic regions and singletons vs. non-singletons. Singletons are CpG sites with only one CpG within the 10-bp region, while non-singletons contain more than one CpG with the 10-bp region [32]. We further divided non-singletons into two sub-categories: (1) concordant non-singletons: all CpGs within the region share the same absolute methylation state (i.e., all fully methylated or all fully unmethylated), (2) discordant non-singletons: the methylation states of CpGs appearing in a close neighborhood (10bp) were mixed with both fully methylated and fully unmethylated sites present. Second, we measured the 5mC methylation correlation coefficient between ONT output and bisulfite sequencing data across all CpG sites at genome level. Third, we assessed the running speed and per-read resource usage evaluation. Further details on performance criteria used in evaluation are shown in **Methods**.

Predictions of 5mC at single-molecule, single-base resolution

To understand the impact of various DNA methylation callers on 5mC prediction at single-molecule, single-base resolution, from different genomic contexts, we assessed the per-read accuracy in singletons and non-singletons. We compared methylation-calling performances on fully methylated/unmethylated CpGs in bisulfite sequencing (BS-seq) (coverage>=5) at the singleton and non-singleton levels across four datasets (**Table S2**). For NA19240, there are 30,377 singleton CpGs and 224,645 non-singleton CpGs in BS-seq (coverage>=5) that overlap with the ONT data. The comparison performance metrics include accuracy, F1 score, receiver operating characteristic curves (ROC curves) and area under the ROC curve (AUC) (**Figure 3, Figure S3, and Table S3**). DeepMod performance is much lower than other four tools when applied to all four human ONT datasets (**Table S3**). While DeepMod robustness is comparable to other tools when using 5mC positive control dataset from *E. coli* [33] (**Table S4**). Thus, for clarity, we only keep display the other four tools in **Figure 3-5**. Specifically, Nanopolish, Megalodon, and DeepSignal outperformed the other two tools on all datasets (**Figure 3A** and **Table S3**). While Nanopolish, Megalodon, DeepSignal, and Tombo exhibit lower accuracy (less than 0.90) at discordant non-singletons, consistent across four datasets (**Figure 3A, Figure**

S3A). Next, we assessed the performance using the ROC curves and AUC in singletons and non-singletons (**Figure 3B**). Again, Nanopolish, Megalodon, and DeepSignal achieved the highest AUC values (singletons AUC: 0.92 - 0.93, non-singletons AUC: 0.96 – 0.98, concordant non-singletons AUC: 0.96 – 0.98, discordant non-singletons AUC: 0.81 – 0.82). We further confirmed the performance assessment using the F1 score (**Figure S3B-S3C**), which is the harmonic mean of precision and recall, and addresses any imbalanced classes. Overall, Nanopolish, Megalodon, and DeepSignal are consistently the top three performers at singleton and non-singleton 5mC prediction at single-read, single-base resolution.

Predictions of 5mC at single-molecule, single-base resolution across various biologically relevant genomic regions

Since different genomic contexts display various CpG density and DNA methylation levels [44], we overlapped CpG islands, promoters, exons, introns, and intergenic regions (referred as intergenic) with BS-seq (coverage>=5) to evaluate the impact of biologically relevant genomic contexts on 5mC predictions (**Figure 4A-B, Figure S4, Table S3**). Specifically, we define the region 2000 bp around transcription start site (TSS) as the promoter. Nanopolish, Megalodon, and DeepSignal exhibit higher overall accuracy on genome wide CpG sites across all datasets, and overall, intergenic regions display the lowest accuracy (**Figure 4A**). Whereas the overall F1 scores for all tools are not as accurate at CpG island and promoters. Megalodon and DeepSignal are less accurate on CpG islands and promoters (F1 score <0.88) than other regions **(Figure 4B**). The decreased F1 at these two regions may be caused by highly imbalanced distribution of

5mC and 5C on: CpG islands (9,054:166,126) and promoters (11,495:164,989) regions for NA19240 (**Table S3**). In comparison, DeepMod exhibits lower accuracy and F1 score across all genomic regions (**Table S3**). In summary, we concluded that Nanopolish, Megalodon, and DeepSignal achieved better methylation calling performance across genomic contexts.

Methylation calls by Nanopolish, Megalodon, and DeepSignal show high concordance with ground truth BS-seq.

To assess the performance of 5mC prediction of these tools for CpG sites with full range of methylation levels, we evaluated the Pearson's correlation coefficient between methylation patterns of the predicted DNA methylation percentage (read coverage>=3) and the corresponding BS-seq data (coverage>=5) at single-base resolution. We found that the methylation levels for all CpG sites predicted by Nanopolish, Megalodon and DeepSignal showed highest correlation (**Figure 5A**) with NA19240 reduced representation bisulfite sequencing (RRBS) data. We also observed that the results of Nanopolish, Megalodon, and DeepSignal are highly correlated (R >= 0.94) for NA19240. Similar correlation coefficients of these three tools can be found for APL, K562, and HL-60 (**Figure S5**) datasets. Ideally, as ground truth BS-seq suggested, a bimodal distribution of DNA methylation is expected (0 for unmethylated, 1 for methylated). The histogram of the DNA methylation output of Nanopolish, Megalodon, and DeepSignal displayed bimodal distribution as the BS-seq data. In contrast, Tombo and DeepMod exhibit different data distributions. The DNA methylation level histogram of Tombo output had multiple peaks between 0% and 100% methylation levels, while DeepMod did not display the peak around 100% methylation level. The Pearson's correlation between BS-seq and DeepMod with (R = -0.07) for NA19240 data indicates that DeepMod cannot effectively predict methylation distribution at whole-genome level for human cells. We further evaluated the Pearson's correlation coefficient of methylation percentage achieved by methylation-calling tools with BS-seq across different genomic contexts (**Table S5**). Nanopolish, Megalodon and DeepSignal consistently produce the highest correlation coefficients at all genomic regions for NA19240 data (**Figure S6**).

To assess the biological context of the methylation calls, we explored the relationship between CpG methylation percentage and distance to annotated TSS (**Figure 5B-C and S7A-B**). As expected, CpG sites near TSS tend to be unmethylated. Methylation level gets higher as the distance from the TSS increased. DNA methylation patterns from Nanopore sequencing closely resemble the pattern for the WGBS data (**Figure 5C and S7A**). Nanopolish displayed the lowest DNA methylation levels at TSS.

Transcriptional factors CCCTC-binding factor (CTCF) binding sites are featured with low DNA methylation [45]. CTCF plays a critical role in long-range chromatin interactions, the formation and maintenance of the topologically associated domains, and transcription. Thus, we further assessed the relationship between CpG methylation percentage and distance to the center of the CTCF binding peaks from the ChIP-seq data of the matching cell lines (NA19240, K562, and HL-60). Indeed, DNA methylation is the lowest at the center of the CTCF binding peaks (**Figure 5D** and **S7C-D**) and the ONT 5mC calls by

Nanopolish, Megalodon, and DeepSignal closely track the pattern of WGBS data (**Figure S7C**).

Overall, Nanopolish, Megalodon, and DeepSignal had high correlations with the background truth BS-seq, and they closely tracked the methylation pattern for the background truth BS-seq at whole genome level. The correlation coefficient of DNA methylation across CpG sites between the five tools and BS-seq is consistent with the read-level accuracy (**Figure 3-4**).

Megalodon and DeepSignal covered more CpG sites than Nanopolish.

Lastly, we evaluated the capacity of the five tools to make 5mC prediction for CpG sites by evaluating the number of CpG sites (read coverage>=3) covered by each tool. Megalodon and DeepSignal covered more CpG sites than other tools on four datasets (**Figure 6 and S8, Table S6**). The UpSet diagram shows the number of overlapped sites by the five tools (**Figure 6 and S8**). 52% of the predicted CpG sites in NA19240 were predicted by all five tools (**Table S6**). Furthermore, for all the CpG sites predicted by any of the three top performers (i.e., Nanopolish, Megalodon, and DeepSignal), 92% CpGs were predicted by all three tools, shown by proportional Venn diagram (**Figure 6**). Megalodon and DeepSignal covered more CpG sites that were not covered by the other three tools (Megalodon and DeepSignal predicted 99% of the union of CpG sites using NA19240). Nanopolish covered 93% of the union CpG sites due to the more stringent criteria of log-likelihood ratio used to predict 5mC for non-singletons [9]. Megalodon and DeepSignal covered 6% more CpG sites than Nanopolish and the differences increases greatly for lower sequencing-depth ONT datasets (**Figure 6 and S8, Table S6**). Therefore, Megalodon and DeepSignal predicted the most CpG sites, while Tombo and DeepMod predicted the least CpG sites.

Running time and memory usage on benchmarking datasets.

To evaluate the running time and peak memory of each methylation-calling tool, we ran five pipelines starting from the initial stage of taking input of raw fast5 files to the final output of the read level and genome level prediction results using the same High-Performance Computing (HPC) platform and environment (See **Methods**). In order to parallelize methylation calling, we split the raw reads of the benchmarking dataset, and start 50 running jobs on each part of reads for each methylation-calling tool. A GPU and eight processors of hardware resources were allocated to each job running GPU accelerated computing supported tools (Guppy, Megalodon, DeepSignal and DeepMod) to minimize run time. The SLURM resource and job management system effectively monitor the usage of computing resources on HPC clusters [46]. Therefore, for each tested dataset we ran all jobs managed by SLURM and calculated the sums of run time totals (hours) and the peak memory usage (GB) based on reported logs of SLURM jobs for each pipeline (**Figure 7, Table S7**). Nanopolish and Megalodon had the shortest run times (703 and 704 hours) to process the fast5 raw signal file for NA19240 (32x coverage). While Tombo, DeepSignal, and DeepMod were much longer (9, 32, and 40x longer, respectively) for the same file. Furthermore, Nanopolish required the lowest peak memory usage (~21 GB) while Megalodon required the highest peak memory usage (21 times). The same analysis of run time and memory usage for other benchmarking datasets also confirmed the ranking for these tools (**Table S7**). In conclusion, Nanopolish requires the least CPU time and the lowest peak memory usage resource. For other tools, there is a trade-off between prediction performance and running resources. Thus, Nanopolish is more appealing for high-coverage mammalian ONT dataset for 5mC prediction.

Discussion

Enhanced detection of DNA methylation in the human genome is critical to improve our understanding of the functional impacts of epigenetic modifications. Recently, ONT nanopore-based sequencers have made possible direct DNA sequencing to generate long single-molecule reads at base resolution. ONT long-read contributes to the phasing of base modifications with genetic variants, along individual nucleic acids. Therefore, it allows exploration of epigenetic heterogeneity at single-molecule resolution and can improve our ability to detect DNA modification in long range.

ONT has released multiple commercialized platforms and pore-chemistry versions (see timeline in **Figure 1A**). In 2015, ONT released its first commercialized platform, MinION™ [47, 48], a portable device enabling simultaneous sequencing using up to 512 pores, with the capacity to generate up to 30 GB of DNA data [49]. In 2017, ONT introduced a scaled-up platform, GridION™, allowing analysis of up to five MinION flow cells and generation of up to 100GB of data per run [50]. In 2018, ONT introduced the ultra-high-throughput platform PromethION™ with up to 48 flow cells [51], and later offered PromethION24/48 for much larger-scale sequencing [52]. Nanopore sequencing is considered a paradigm among recent sequencing approaches, because of its unique design enabling significant portability and relatively low cost [11, 53].

In the past, the advantages of long-reads and real-time sequencing have made Nanopore sequencing an effective tool to detect genomic and genetics aberrations such as DNA structural variants and RNA alternative splicing events [54]. Nanopore sequencing have demonstrated its powerful capability of detecting structural variation in lung cancer [55, 56], leukemia [57], and neuron disorder [58-60], and it has been applied to clinical samples for molecular etiology or diagnosis of genomic variants relevant disease [58-63]. Meanwhile, Nanopore sequencing of splicing changes has been utilized in cancer research such as breast cancer [64], leukemia [65, 66], and brain tumor [67]. Such research with Nanopore sequencing has improved our understanding of evolutionary process in human diseases.

Nanopore sequencing also provides new opportunity for epigenetic research. For example, Miga et al provides telomere-to-telomere assembly and DNA methylation maps of human X-chromosome [68] using Nanopore sequencing and Ewing et al. developed a new computational tools and long-read nanopore sequencing for transposable element epigenomic profiling [69]. Recently, some efforts have been made to combine Nanopore sequencing and other methods to epigenomics profiling and chromosome structures exploration. For example, Wongsurawat et al utilized Nanopore Cas9-targeted sequencing to simultaneously assess *IDH* mutation status and *MGMT* methylation level in both cell lines and fresh biopsies of diffuse glioma [70]. And, Lee et al developed a new method based on Nanopore sequencing to evaluate CpG methylation and chromatin accessibility simultaneously [71]. Also, several preprint papers utilized Nanopore sequencing to enhance the understanding of epigenetic heterogeneity and mechanism [72-74].

In this study, we benchmarked state-of-art methylation calling tools for Nanopore sequencing data. Based on our systematic comparison analysis, we revealed four key observations. First, the choice of methylation calling critically affects the level of accuracy, F1 score and AUC score on different Nanopore data sets and at different genomic regions. Second, both the HMM model-based Nanopolish and deep learning-based tools Megalodon and DeepSignal, are comparable in terms of overall accuracy, F1 score and AUC values, at single-base and single-read resolution. Notably, Nanopolish has the lowest memory usage, and both Nanopolish and Megalodon are faster than DeepSignal. Third, the methylation detections in discordant regions with mixed DNA methylation and intergenic regions exhibit lower accuracy and F1 score across all five tools. Nanopolish is fast and accurate, at the same time, it outputs the methylation levels of 6% fewer CpG sites than DeepSignal and Megalodon, due to the more stringent log-likelihood ratio cutoff for predicting non-singleton CpG sites. Nanopolish can be used for quick prediction, and future algorithm development can focus on increasing the accuracy and higher CpG coverage, which leads to higher overall performance. When high-performance clustering or cloud computing is available, Nanopolish, Megalodon, and DeepSignal can each produce high-quality methylation predictions on the largest number of CpG sites. In the absence of an HPC or cloud it is feasible to run Nanopolish on a laptop for DNA methylation calling due to its short run-time and low memory for in-field analysis that also makes it compatible with ONT MinION's portability.

We believe that our benchmarking of methylation calling tools will guide researchers and practitioners to make conscious and effective choices when designing the analytic plan for epigenomic profiling using ONT sequencing, including Nanopore Cas9-targeted sequencing data analysis. The bottlenecks revealed by our analysis can help developers to improve ONT sequencing data methylation-calling training data generation and tool design. We note that one recent preprint [33] proposed a consensus random forest model to improve accuracy by combining read level methylation predictions of some tools (i.e., DeepSignal and Megalodon). Our analysis demonstrates that a training dataset covering discordant non-singletons and intergenic regions would improve the overall robustness of DNA methylation prediction at single-molecule, single-base resolution for human epigenome-wide study.

Conclusion

Oxford Nanopore long-read sequencing technology poses a challenge for accurate methylation predictions. The past few years have witnessed rapid development of both the sequencing technology and analytical tools. For DNA methylation analysis, many algorithms are emerging for ONT sequencing data. We comprehensively surveyed current publicly available computational tools for direct ONT DNA sequencing data methylation detections. We systematically evaluated the advantages, disadvantages, and identified performance bottlenecks that affect the robustness of DNA methylation detection at single-molecule and single base resolution. Using a standardized workflow we assessed the performance of five DNA methylation calling tools and found that methylation callers vary in their accuracy in diverse genomic contexts, epigenome coverage, peak memory usage, and run time, for both single-read and single-base resolution. For initial DNA methylation analysis, we recommend Nanopolish given its short run-time, low memory requirement and overall high performance in calling DNA methylation in whole genome level, singleton, non-singleton, promoter, CpG islands, exonic, and intronic regions. For systematic analysis, we recommend integrating the preliminary results with Megalodon, or DeepSignal output for optimal performance, e.g., more comprehensive epigenome coverage. Comprehensive and balanced training datasets that cover various genomic contexts is desirable for more robust prediction of DNA methylation in discordant and intergenic regions and will help improve our understanding of epigenetic mechanisms underlying many different biological processes, such as aging and cancer development.

Methods

Sample collection and processing

In the study we provided four independent human datasets - one normal B-Lymphocyte cell line (NA19240) [39], one primary acute promyelocytic leukemia clinical specimen (APL), two cancer cell lines (K562, HL-60).

For APL, sample was obtained from the Stem Cell and Xenograft Core of the University of Pennsylvania. The Core maintains a tissue bank of cells from patients with Hematologic Malignancies. This is Institutional Review Board (IRB) approved research (IRB protocol #703185). The patient sample was collected at the time of clinical presentation and prior to therapy. The sample was collected as leukopheresis and viably frozen using standard techniques. The de-identified specimen was then provided to the Jackson Laboratory for Genomic Medicine (JAX-GM). Diagnosis (Dx) of acute promyelocytic leukaemia (APML) was confirmed by Fluorescence in situ hybridization (FISH) analysis for t(15;17). K562 and HL-60 were cultivated in Roswell Park Memorial Institute (RPMI) 1640 Medium (Gibco, A10491-01) with 10% fetal bovine serum (FBS) (Gibco, 26140079). K562 medium was additionally supplemented with 1% Antibiotic-Antimycotic (Gibco, 15240062). HL-60 medium was additionally supplemented with 1.2% of penicillin-streptomycin (Gibco, 15140-163), GlutaMAX (Gibco, 35050-061), Sodium Pyruvate (Gibco, 11360-070), MEM Nonessential Amino Acids (Corning, MT25025CI) and MEM Vitamin Solution (Corning, MT25020CI). Incubator conditions were 37°C and 5% CO_2.

Bisulfite sequencing (BS-seq) dataset and analysis

We generated whole genome bisulfite sequencing (WGBS) for APL. DNA was extracted using AllPrep DNA/ RNA kit (Qiagen) following manufacturer's recommendation. Two 500ng DNA were sheared to 500bp using a LE220 focused-ultrasonicator (Covaris) and purified using 0.9X SPRI beads (Beckman Coulter). The libraries were prepared using the KAPA Hyper Prep Kit for Illumina (Roche) and bisulfite conversion was performed using the TrueMethyl Seq Kit (CEGX). Briefly, the fragmented DNA was first spiked in with CEGX sequencing controls, followed by end-repair and A-tailing, and then ligated with SeqCap indexed adaptor (Roche). Sample destined for 5hmC library was first subjected to oxidation whereas samples destined for 5mC library was treated as mock. This is then followed by a bisulfite conversion. The treated DNA were cleaned up and amplified with 15 cycles of PCR and purified. The final library was quantified by real time qPCR for an accurate concentration. Libraries were sequenced paired end 2x150bp on the Illumina HiSeq 2500 instrument.

We utilized the published whole genome bisulfite sequencing (WGBS) for K562 (ENCODE accession number: ENCFF721JMB, ENCFF867JRG), and reduced representation bisulfite sequencing (RRBS) for HL-60 (ENCODE accession number: ENCFF000MDA, ENCFF000MDF) and NA19240 (ENCODE accession number: ENCFF000LZS, ENCFF000LZT).

All BS-seq data were analyzed with Bismark [75] with the human reference genome (GRCh38/hg38) to get the cytosine methylation frequency at each CpG site. Region-specific analysis and local smoothing for samples was performed using the BS-seq package (https://github.com/TheJacksonLaboratory/BS-seq-pipleine). Then, we only select high-confidence CpG sites with coverage >=5, where a CpG is considered as fully methylated when its methylation frequency is 100% and considered as unmethylated when its methylation frequency is zero. For each dataset, we take the union of high-confidence sites from all tools and BS-seq as our final high-confidence set and the selected high-confidence sites (**Table S2**). In total, 30,377 singleton CpGs (10,815 fully methylated and 19,562 unmethylated) and 224,645 non-singleton CpGs (29,432 fully-methylated and 195,213 unmethylated) were selected from NA19240 RRBS, and 42,137 singleton CpGs (21,738 fully-methylated and 20,399 unmethylated) and 276,411 non-singleton CpGs (56,444 fully-methylated and 219,967 unmethylated) were selected from HL-60 RRBS. For K562 and APL, the total selected high-confidence CpG sites are 25,382,453 and 8,707,630 respectively from WGBS.

Nanopore sequencing dataset and analysis

We generated Nanopore sequencing dataset for APL, K562, and HL-60 at JAX-GM.

For APL, genomic libraries were prepared using the Rapid Sequencing Kit (SQK-RAD004, ONT) according to manufacturer's recommendation. Briefly 1200ng DNA was incubated with 2.5ul of FRA at 30°C for 1 min and 80°C 1 min. This is followed by an addition of 3ul of adaptor (RAP) to the reaction mix and incubated at 5 min at room temperature. The libraries were sequenced on the flowcell R9.4.1 (FLO-MIN106, ONT) on GridION (ONT) using the MinKNOW software for 48hr.

For K562 and HL-60, HMW genomic DNA were extracted from 5m cells using phenol chloroform approach (PMID30933081). Libraries were prepared using the Rapid Sequencing Kit (SQK-RAD004, ONT) according to manufacturer's recommendation. Briefly 1200ng DNA was incubated with 2.5ul of FRA at 30°C for 1 min and 80°C 1 min. This is followed by an addition of 3ul of adaptor (RAP) to the reaction mix and incubated at 5 min at room temperature. The libraries were sequenced on the flowcell R9.4.1 (FLO-MIN106, ONT) on a GridION (ONT) using the MinKNOW software for 48hr.

For NA19240, we request Nanopore raw data from previously published research [39].

For *E.coli*, we utilized an example dataset on Github (https://github.com/comprna/METEORE/tree/master/data/example), which contains 50 single-read fast5 files from the positive control dataset for *E.coli* generated by Simpson et al [9].

All Nanopore reads (.FAST5 files) were base-called by Guppy (v4.2.2) with default high-accuracy model (dna_r9.4.1_450bps_hac.cfg). The base-called reads were then aligned to human reference genome (GRCh38/hg38) for human dataset (NA19240, APL, K562, HL-60) or aligned to the *E.coli* K12 MG1655 genome for *E.coli* dataset using minimap2 [42]. Specially, R9.4-series pore is the current broadly used Nanopore flow cell and there is a slight difference between R9.4 and R9.4.1 flow cells and most computational model can work for both [76].

Additionally, for cell line authentication of K562 and HL-60, we aligned the base-called reads to the human reference genome (GRCh37/hg37) with the help of Minimap2 [42] and Samtools [77] and compared target regions in aligned reads with reported insertions/deletions (indels) derived from the Cancer Cell Line Encyclopedia (CCLE) project [78] in genome browser IGV to identify the cancer cell line information.

Experimental settings and running configurations for Nanopore sequencing analysis

We ran five tools on the benchmarking datasets. We supply FAST5 files generated by Nanopore sequencers with raw signals and base calls as input for methylation detection analysis. To compare speed performance, all tools were carried out on the same computer clusters: 32 cores, 2.6GHz HP Proliant SL Series CPU, 300 GB RAM, NVIDIA Tesla P100 Data Center and 1 TB Data Direct Networks Gridscalar GS7k GPFS storage appliance. The HPC platform software and hardware specifications are: slurm manager version: 19.05.5, CPU: Intel(R) Xeon(R) Gold 6136 CPU @ 3.00GHz, GPU: Tesla V100-SXM2-32GB.

Base calling, the process of translating raw electrical signal of the sequencer into nucleotide sequence, is the initial step of Nanopore data analysis. Both ONT and independent researchers are actively developing different tools for base calling step. Specifically, ONT provides base-calling programs including official ONT community-only software (Albacore and Guppy) and open-source software (Flappie, Scrappie, Taiyaki, Runnie, and Bonito), the latter of which are under development with new algorithms for base calling [31, 79, 80]. Only very recently has it been possible to base-call DNA modifications directly from the raw signal without genomic anchoring, which can be accomplished via specific base callers such as Scrappie [81]. Among the base calling programs, Albacore and Guppy are compatible with Nanopore R9.4 reads and offer the most stable performance [43]. Albacore [82, 83] is a general-purpose base caller that runs on CPUs. Guppy [84] is a neural network based basecaller with several bioinformatic post-processing features. Guppy supports both CPUs and GPUs for improved base-calling run time, and it is available on the ONT community site (https://community.nanoporetech.com) for internal use. Because the state-of-art basecaller Guppy using the default model showed excellent performance among ONT basecalling tools [43], we utilized Guppy (v4.2.2, with all 32 CPU threads) for base calling for all datasets and all DNA methylation calling tools.

RNN and HMM are computationally intensive algorithms. In HMM- based Nanopolish tool, the Viterbi algorithm is used for methylation prediction. The Viterbi algorithm is a sequential technique, and its computation cannot currently be parallelized with multithreading. However, in RNN-based DeepSignal and DeepMod, multiple threads can work on different sections of the neural network and thus RNN computation can be parallelized with multithreading. We choose this system for evaluation since it has a larger memory capacity than desktop systems and, with the help of a large number of cores, the tasks can be easily parallelized to accelerate data output for state-of-the-art tools.

Methylation calling Nanopore sequencing at read level and site level

We evaluated the performance of Nanopolish (v0.13.2), Megalodon (v2.2.9), DeepSignal (v0.1.8), Tombo (v1.5.1), and DeepMod (v0.1.3) to detect 5mC at CpG dinucleotides.

These five tools differ in the underlying algorithms and the modifications they are trained to detect.

Nanopolish [9] calls 5mC in a CpG context using a HMM to assign a log-likelihood ratio (LLR) for each CpG site, where a positive log-likelihood ratio (LLR) indicates support for methylation. Nanopolish groups nearby CpG sites together and calls the cluster jointly to assign the same methylation status to each site in the group. For example, on a motif such as CGCGT, Nanopolish reports a LLR for the whole group, rather than a separate LLR for the individual cytosine. We use the 2.0 as the LLR threshold for methylation calling as the Nanopolish authors suggests that the initial 2.5 shown in the paper is overly conservative [85]. To be more specific, we first called methylation at the read level: we removed ambiguous reads when the absolute value of their LLR was less than 2.0, and then called CpG sites as methylated when the LLR > 2.0 and called CpG sites as unmethylated when the LLR < -2.0. Then we calculated methylation frequency at the site level by converting the LLR to a binary call (methylated/unmethylated) for each read and calculating the fraction of reads classified as methylated.

Megalodon [24] is a new ONT-developed a research command line tool and can identify modified base and sequence variant calls from raw nanopore reads. For modified base calls, Megalodon utilizes Guppy (v>=4.0) on the backend and pre-trained models for basecalling. It anchors the intermediate basecalling neural network output to a reference genome. Megalodon performs the methylation calling at either the per-read or per-site level (aggregate per-read results) based on the log probability that the base is modified or canonical. Guppy (v>=4.0) backend and pre-trained models is recommended for base calling, so we fed Megalodon with Guppy v4.2.2 with the latest 5mC in an all context model (res_dna_r941_min_modbases_5mC_v001.cfg) from Rerio [86] as the basecalling model, and chose the default 0.8 threshold as the probability cutoff to count the called base (modified or canonical) with probability >0.8 toward the final aggregated output at per-site level.

DeepSignal [27] proposed a deep recurrent neural network with Bidirectional Long short-term memory (BiLSTM)+Inception structure to detect the methylation state of target cytosine in CpG context. DeepSignal required an extra the re-squiggle module of Tombo before methylation calling. The methylation calling output of DeepSignal is a tab-delimited text file (tsv) at read level including two probability values for each base, one for methylated (prob_1) and one for unmethylated (prob_0), as well as a binary call (unmethylated/methylated) for each base. The CpG sites is called as methylated when prob_1 > prob_0 and is called as unmethylated when prob_1 <= prob_0. We performed per-read methylation calling with the CpG model trained using HX1 R9.4 1D reads (model.CpG.R9.4_1D.human_hx1.bn17.sn360.v0.1.7+.tar.gz) provided with the latest version of DeepSignal, and calculated the fraction of reads classified as methylated at site level with their official methylation frequency script.

ONT-developed Tombo [21] performed a statistical test to identify modified nucleotides with its alternative model without the need for prior training data. Tombo computed per-read, per-genome location test statistics by comparing the signal intensity difference between modified bases and canonical bases. We chose to use the recommended CpG motif specific model with the default threshold of (-1.5, 2.5) for DNA where scores below -1.5 were considered as methylated and above 2.5 unmethylated, and scores between these thresholds did not contribute to the per-site methylation. After that, we calculated methylation percentage at each genomic position.

DeepMod [25] designed a bidirectional recurrent neural network (RNN) with an LSTM unit for genome-scale detection of DNA modifications. The input is a reference genome and FAST5 files with raw signals and base calls, and the output is a BED file with coverage, number of methylated reads, and methylation percentage information for genomic positions of interest. Since 5mC in CpG motifs has a cluster effect in the human genome [25], DeepMod provides a cluster model to generate a final output for site level predicted methylation probability in human genome. We performed DeepMod for methylation calling with the RNN model (rnn_conmodC_P100wd21_f7ne1u0_4) and cluster model (na12878_cluster_train_mod-keep_prob0.7-nb25-chr1) [87]. Also, since DeepMod aggregated methylation callings results into a per-site output BED file, we counted the number of methylated callings and unmethylated callings from BED outputs to evaluate its read level performance.

The performances of these methods that use prior knowledge about the expected deviations in signal depend notably on the training data used, which is typically composed of a fully unmodified and a fully modified sample. Motifs that are not represented in the training set or that contain mixtures of modified and unmodified bases lead to suboptimal performance.

Methylation calling performance evaluation at read level

We designed the performance evaluation process for 5-methylcytosine status prediction among five methylation calling tools.

First, we evaluated performance for the five tools on four real world Nanopore Sequencing datasets at singleton and non-singleton site levels and biologically relevant genomic context level at read level. To be more specific, we only considered CpG sites covered by >= 5 reads in BS-seq and CpGs sites covered by >= 3 reads by methylation calling tools, and joined the common sites identified by the five tools with background truth BS-seq. For the CpG sites that showed 0% or 100% methylation level, we evaluated the performance of these tools as a per-read classification model. Then we joined each tool's prediction results to a common CpG set and measured accuracy on basis of singleton and non-singleton sites, or biologically relevant genomic context. We compared the percentage of methylation calculated by the five Nanopore-based methods to that derived by BS-seq at annotated locations. On each location basis, we calculated the F1 score, accuracy, precision, recall, and assessed the tradeoff between true-positive and false-positive rates of 5mC prediction by calculating receiver operating characteristic (ROC) curve by varying the threshold for methylation calling and reported the area under the ROC curve (AUC) values. Metrics of performance are calculated as following using BS-seq as ground truth:

TP: true positive

TN: true negative

FP: false positive

FN: false negative

Precision: TP/(TP+FP)

Recall: TP/(TP+FN)

Accuracy: (TP + TN) / (TP + TN + FP + FN)

F1 Score: 2*(Recall * Precision) / (Recall + Precision). We calculated F1 score for both 5mC and 5C and used macro_F1(average F1_5mC and F1_5C) for final F1 score.

ROC AUC: the area under receiver operating characteristic curve, usually ranging from 0.5 to 1.0. It is a performance metric used to evaluate how a classifier performs on both methylated and unmethylated class predictions.

Methylation calling performance evaluation at site level

We calculated the Pearson's coefficient between predicted methylation status and BS-seq status and checked the methylation distribution structure for each tool at a genome level. Again, we first kept CpGs with >= 5 reads in BS-seq and CpGs with >= 3 reads by methylation calling tools, joined the CpG sites as overlapped sets, and calculated methylation frequencies for all DNA CpGs at a genome level for each tool from read level. For correlation analysis, we treat each pair of tools as a regression model to calculate Pearson's correlation coefficients at a genome level. We compute the relationship between CpG methylation percentage with distance to annotated transcription start site (TSS) and transcriptional factors CCCTC-binding factor (CTCF) binding sites using deepTools [88].

Memory usage and running time for methylation tools

We compared the capability of the five methylation calling tools for memory usage and running time on single-read fast5 file in each dataset. All tools have support for multi-processors, so we generated simulated data sets to compare the scalability of these tools on the same system configurations: We split each dataset into 50 batches to ensure the same scale of input, chose default tool configurations to run the program. Each job was allocated eight processors (200GB per processor) and one GPU hardware resource (GPU is allocated for running Guppy, Megalodon, DeepSignal and DeepMod). We extract running time (field name: CPU Utilized) and peak memory utilization (field name: Memory Utilized) from the SLURM job log data. These results were used as the measurement of running time and memory usage for hardware performance comparison and evaluation.

CpG sites comparison by different methylation calling tools in each dataset.

We compared the number of CpG sites covered by different methylation calling tools. For each dataset, we kept CpGs with >= 3 reads by methylation calling tools, then joined the CpG sites with BS-seq to check the overlapping sites that were also detected in BS-seq. We also checked the CpG sites detected by Nanopolish, Megalodon and DeepSignal since these three performed best among the methylation calling tools.

Availability of Data and Materials.

All source codes are publicly available at GitHub https://github.com/liuyangzzu/nanome and https://zenodo.org/record/4730517 with the DOI: 10.5281/zenodo.4730517.

APL WGBS and ONT datasets, HL-60 and K562 ONT datasets were deposited under GEO accession GSE173675, GSE173676, GSE173687, and GSE173688.

Acknowledgements

S.L. is supported by a startup fund from The Jackson Laboratory for Genomic Medicine, Leukemia Research Foundation New Investigator Grant, The Jackson Laboratory

Director's Innovation fund 19000-17-31 and 19000-20-05, The Jackson Laboratory Cancer Center New Investigator Award, The Jackson Laboratory Cancer Center Fast Forward Award, and the National Institute of General Medical Sciences of the National Institutes of Health under Award Number R35GM133562. Research reported in this publication was partially supported by the National Cancer Institute of the National Institutes of Health under Award Number P30CA034196. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. The authors thank Dr. Charles Lee for scientific discussion and providing the access to the long-read sequencing data. The authors thank Dr. Derya Unutmaz for helpful discussion. The authors thank the members of Li Lab for discussions and thank Drs. Stephen Sampson and Kevin Seburn from The Jackson Laboratory Research Program Development for editing this paper. The authors thank The Jackson Laboratory Computational Sciences and Research IT team for technical support, e.g., Shiny app deployment helped by Sandeep Namburi.

Declarations

None of the authors have any competing interests.

TABLE

Table 1. Current DNA methylation-calling tools for Nanopore sequencing

FIGURE LEGENDS

Figure 1.

(**A**) Timeline of publication and technological developments of Oxford Nanopore Technologies (ONT) methylation calling tools to detect DNA cytosine modifications. Methylation calling tools are listed acc in order by publication date instead of bioRxiv online submission date (BioRxiv date for methBERT and Github release time for Megalodon since they lack an available official publication). Chemical pore versions of Nanopore flow cell are represented as colored bars. Chemical pore version of Nanopore flow cell compatibility for each methylation calling tools is shown in corresponding colors. Relevant publication time are from multiple source [9, 15, 16, 18, 21, 40, 50-52, 79, 89-91]. (**B**) Performance evaluation on 5mC/5C prediction of methylation calling tools with Nanopore sequencing. We generated four datasets for nanopore sequencing, applied five methylations calling tools separately to detect methylation status and compare the 5mC/5C classification with the background truth BS-seq. To compare the performance of different tools, we compared methylation calling results in singletons/non-singletons regions or biological relevant genomic regions, correlated all CpGs sites with methylation distribution in BS-seq, and evaluated the running speed and computing memory usage.

Figure 2.

(**A-B**) Quality control summary for four datasets. (**A**) Violin plot of log-transformed read length (**B**) Violin plot of base call quality. Data shown are colored by dataset and plotted by Plots NanoPack [41]. **(C)** Scheme for biologically relevant annotations of genomic context and singleton/non-singleton classification. We consider biologically relevant genomic context including promoter, exon, intron, intergenic regions and CpG island. Singletons are CpG sites that contain only one CG within the 10 base pairs, including absolute state (CpG is 0% or 100% methylated) or mixed state (0% < CpG methylation frequency < 100%). Non-singletons are those CpGs that contain more than one CpG in 10bp. **(D)** CpG sites distribution based on singletons, non-singletons and biologically relevant genomic context in Nanopore sequencing of the NA19240 dataset. Only regions with coverage >= 1 were considered.

Figure 3. Comparison of Nanopore methylation calling tools for the detection of CpG methylation on four real world data sets in singletons and non-singletons.

(**A**) Prediction accuracy across four datasets based on singleton and non-singleton classification. Singletons are CpG sites that contain only one CG within the 10 base pairs, non-singletons are those CpGs that contain more than one CG. Non-singletons can be divided into two groups: i) all means fully mixed, for which methylation of all CpGs in non-singleton is > 0% and < 100% and, ii) all CpGs are in the absolute states (100% or 0% methylated). Non-singletons in absolute states include Concordant non-singletons: all CpGs inside have the same absolute state (i.e., all 100% or all 0% methylated); Discordant non-singletons: at least one CpG is fully methylated and at least one other CpG is fully unmethylated. (**B**) ROC curves on NA19240 dataset on singleton, non-singleton, concordant, discordant coordinates.

Figure 4. Comparison of Nanopore methylation calling tools for the detection of CpG methylation on four real world data sets in biologically relevant genomic contexts.

(**A**) Prediction accuracy across four datasets based on biologically relevant genomic contexts. The biologically relevant genomic contexts include Genome-wide, CpG Islands, promoters, exons, intergenic regions(intergenic) and introns. Promoter is 2000 bp upstream of the start. (**B**) F1 score across four datasets based on biologically relevant genomic context.

Figure 5. Comparison of Pearson correlation of methylation call tools across all CpG sites.

(**A**) Correlation plot showing Pearson correlation of each methylation calling tool with BS-Seq on NA19240. The upper left triangle denotes Pearson's correlation coefficients; the diagonals are distributions of 5mC percentage of BS-seq data and 5mC percentage predicted by each tool using ONT data. 2D kernel density plots are shown at the lower left triangle for each pair of comparison. (**B-C**) Relationship between CpG methylation percentage and distance to annotated TSS in (**B**) NA19240 and (**C**) APL. (**D**) Relationship between CpG methylation percentage and distance to annotated CTCF binding peaks in NA19240. Distances are binned into (**B, C**) 50-bp, and (**D**) 100-bp windows. Negative distances are upstream and positive distances are downstream of the (**B-C**) TSS and CTCF binding peaks (**D**).

Figure 6. CpG sites detected by methylation calling tools using NA19240

UpSet diagram shown at the lower left is for CpG sites detected by all methylation calling tools. Venn diagram shown at the upper right is for CpG sites detected by Top3 performance methylation calling tools (Nanopolish, Megalodon and DeepSignal). For each methylation calling tool, only CpG sites covered >= 3 reads are considered.

Figure 7. CPU utilized time and memory usage for each methylation calling tool on each dataset.

All tools were compared on the same computer clusters: 32 cores, 2.6GHz HP Proliant SL Series CPU, 300 GB RAM, NVIDIA Tesla P100 Data Center and 1 TB Data Direct Networks Gridscalar GS7k GPFS storage appliance. The HPC platform software and hardware specifications are: slurm manager version: 19.05.5, CPU: Intel(R) Xeon(R) Gold 6136 CPU @ 3.00GHz, GPU: Tesla V100-SXM2-32GB.

SUPPLEMENTARY INFORMATION

Additional file 1: Supplementary figures. This file contains supplementary figures S1-S8.

Additional file 2: Supplementary tables. This file contains supplementary tables S1 – S7.

REFERENCE

1. Chen K, Zhao BS, He C: **Nucleic Acid Modifications in Regulation of Gene Expression.** *Cell Chemical Biology* 2016, **23:**74-85.
2. Smith ZD, Meissner A: **DNA methylation: roles in mammalian development.** *Nature Reviews Genetics* 2013, **14:**204-220.
3. Barros-Silva D, Marques CJ, Henrique R, Jerónimo C: **Profiling DNA Methylation Based on Next-Generation Sequencing Approaches: New Insights and Clinical Applications.** *Genes* 2018, **9:**429.
4. Luo G-Z, Blanco MA, Greer EL, He C, Shi Y: **DNA N6-methyladenine: a new epigenetic mark in eukaryotes?** *Nature Reviews Molecular Cell Biology* 2015, **16:**705-710.
5. Li S, Chen X, Wang J, Meydan C, Glass JL, Shih AH, Delwel R, Levine RL, Mason CE, Melnick AM: **Somatic Mutations Drive Specific, but Reversible, Epigenetic Heterogeneity States in AML.** *Cancer Discovery* 2020, **10:**1934-1949.
6. Goodwin S, McPherson JD, McCombie WR: **Coming of age: ten years of next-generation sequencing technologies.** *Nature Reviews Genetics* 2016, **17:**333-351.
7. Levy SE, Myers RM: **Advancements in Next-Generation Sequencing.** *Annual Review of Genomics and Human Genetics* 2016, **17:**95-115.
8. Biosciences P: **Detecting DNA Base Modifications Using Single Molecule, Real-Time Sequencing.** 2015.
9. Simpson JT, Workman RE, Zuzarte PC, David M, Dursi LJ, Timp W: **Detecting DNA cytosine methylation using nanopore sequencing.** *Nature Methods* 2017, **14:**407-410.
10. Yang Y, Scott SA: **DNA Methylation Profiling Using Long-Read Single Molecule Real-Time Bisulfite Sequencing (SMRT-BS).** *Methods Mol Biol* 2017, **1654:**125-134.
11. Rang FJ, Kloosterman WP, de Ridder J: **From squiggle to basepair: computational approaches for improving nanopore sequencing read accuracy.** *Genome Biology* 2018, **19:**1-11.
12. Goyal P, Krasteva PV, Van Gerven N, Gubellini F, Van den Broeck I, Troupiotis-Tsaïlaki A, Jonckheere W, Péhau-Arnaudet G, Pinkner JS, Chapman MR, et al: **Structural and mechanistic insights into the bacterial amyloid secretion channel CsgG.** *Nature* 2014, **516:**250-253.
13. Oxford Nanopore Technologies: **Update: New 'R9' nanopore for faster, more accurate sequencing, and new ten minute preparation kit.** 2020.
14. Carter J-M, Hussain S: **Robust long-read native DNA sequencing using the ONT CsgG Nanopore system.** *Wellcome Open Research* 2018, **2:**23.
15. Oxford Nanopore Technologies: **Product comparison.** 2020.
16. Oxford Nanopore Technologies: **New 'R10' nanopore released into early access.** 2020.
17. Oxford Nanopore Technologies: **R10 Evaluation by GrandOmics The Road to High Accuracy of Single Nucleotide.** 2020.
18. Oxford Nanopore Technologies: **R10.3: the newest nanopore for high accuracy nanopore sequencing - now available in store.** 2020.
19. Laszlo AH, Derrington IM, Brinkerhoff H, Langford KW, Nova IC, Samson JM, Bartlett JJ, Pavlenok M, Gundlach JH: **Detection and mapping of 5-methylcytosine and 5-hydroxymethylcytosine with nanopore MspA.** *Proceedings of the National Academy of Sciences* 2013, **110:**18904-18909.

20. Schreiber J, Wescoe ZL, Abu-Shumays R, Vivian JT, Baatar B, Karplus K, Akeson M: **Error rates for nanopore discrimination among cytosine, methylcytosine, and hydroxymethylcytosine along individual DNA strands.** *Proceedings of the National Academy of Sciences* 2013, **110:**18910-18915.
21. Stoiber M, Quick J, Egan R, Eun Lee J, Celniker S, Neely RK, Loman N, Pennacchio LA, Brown J: **De novo Identification of DNA Modifications Enabled by Genome-Guided Nanopore Signal Processing.** *bioRxiv* 2017**:**094672.
22. Liu Q, Georgieva DC, Egli D, Wang K: **NanoMod: a computational tool to detect DNA modifications using Nanopore long-read sequencing data.** *BMC Genomics* 2019, **20:**31-42.
23. McIntyre ABR, Alexander N, Grigorev K, Bezdan D, Sichtig H, Chiu CY, Mason CE: **Single-molecule sequencing detection of N6-methyladenine in microbial reference materials.** *Nature Communications* 2019, **10:**1-11.
24. Oxford Nanopore Technologies: **Megalodon.** 2020.
25. Liu Q, Fang L, Yu G, Wang D, Xiao C-L, Wang K: **Detection of DNA base modifications by deep recurrent neural network on Oxford Nanopore sequencing data.** *Nature Communications* 2019, **10:**1-11.
26. Rand AC, Jain M, Eizenga JM, Musselman-Brown A, Olsen HE, Akeson M, Paten B: **Mapping DNA methylation with high-throughput nanopore sequencing.** *Nature Methods* 2017, **14:**411-413.
27. Ni P, Huang N, others: **DeepSignal: detecting DNA methylation state from Nanopore sequencing reads using deep-learning.** *Bioinformatics* 2019, **35:**4586-4595.
28. Li E, Zhang Y: **DNA Methylation in Mammals.** *Cold Spring Harbor Perspectives in Biology* 2014, **6:**a019133.
29. Almouzni G, Cedar H: **Maintenance of Epigenetic Information.** *Cold Spring Harbor Perspectives in Biology* 2016, **8:**a019372.
30. Greenberg MVC, Bourc'his D: **The diverse roles of DNA methylation in mammalian development and disease.** *Nature Reviews Molecular Cell Biology* 2019, **20:**590-607.
31. Amarasinghe SL, Su S, Dong X, Zappia L, Ritchie ME, Gouil Q: **Opportunities and challenges in long-read sequencing data analysis.** *Genome Biology* 2020, **21:**1-16.
32. Akbari V, Garant J-M, O' N, Kieran, Pandoh P, Moore R, Marra MA, Hirst M, Jones SJM: **Megabase-scale methylation phasing using nanopore long reads and NanoMethPhase.** *Genome Biology* 2021, **22:**1-21.
33. Yuen ZW-S, Srivastava A, Daniel R, McNevin D, Jack C, Eyras E: **Systematic benchmarking of tools for CpG methylation detection from Nanopore sequencing.** *bioRxiv* 2021**:**2020.2010.2014.340315.
34. Li S, Garrett-Bakelman FE, Chung SS, Sanders MA, Hricik T, Rapaport F, Patel J, Dillon R, Vijay P, Brown AL, et al: **Distinct evolution and dynamics of epigenetic and genetic heterogeneity in acute myeloid leukemia.** *Nature Medicine* 2016, **22:**792-799.
35. Chen X, Ashoor H, Musich R, Wang J, Zhang M, Zhang C, Lu M, Li S: **epihet for intra-tumoral epigenetic heterogeneity analysis and visualization.** *Sci Rep* 2021, **11:**376.
36. Li S, Chen X, Wang J, Meydan C, Glass JL, Shih AH, Delwel R, Levine RL, Mason CE, Melnick AM: **Somatic Mutations Drive Specific, but Reversible, Epigenetic Heterogeneity States in AML.** *Cancer Discov* 2020, **10:**1934-1949.

37. Li S, Garrett-Bakelman FE, Chung SS, Sanders MA, Hricik T, Rapaport F, Patel J, Dillon R, Vijay P, Brown AL, et al: **Distinct evolution and dynamics of epigenetic and genetic heterogeneity in acute myeloid leukemia.** *Nat Med* 2016, **22:**792-799.
38. Li S, Garrett-Bakelman F, Perl AE, Luger SM, Zhang C, To BL, Lewis ID, Brown AL, D'Andrea RJ, Ross ME, et al: **Dynamic evolution of clonal epialleles revealed by methclone.** *Genome Biol* 2014, **15:**472.
39. Chaisson MJP, Sanders AD, Zhao X, Malhotra A, Porubsky D, Rausch T, Gardner EJ, Rodriguez OL, Guo L, Collins RL, et al: **Multi-platform discovery of haplotype-resolved structural variation in human genomes.** *Nature Communications* 2019, **10**.
40. Zhang Y-z, Hatakeyama S, Yamaguchi K, Furukawa Y, Miyano S, Yamaguchi R, Imoto S: **On the application of BERT models for nanopore methylation detection.** *bioRxiv* 2021**:**2021.2002.2008.430070.
41. De Coster W, D'Hert S, Schultz DT, Cruts M, Van Broeckhoven C: **NanoPack: visualizing and processing long-read sequencing data.** *Bioinformatics (Oxford, England)* 2018, **34:**2666-2669.
42. Li H: **Minimap2: pairwise alignment for nucleotide sequences.** *Bioinformatics (Oxford, England)* 2018, **34:**3094-3100.
43. Wick RR, Judd LM, Holt KE: **Performance of neural network basecalling tools for Oxford Nanopore sequencing.** *Genome Biology* 2019, **20:**1-10.
44. Akalin A, Kormaksson M, Li S, Garrett-Bakelman FE, Figueroa ME, Melnick A, Mason CE: **methylKit: a comprehensive R package for the analysis of genome-wide DNA methylation profiles.** *Genome Biology* 2012, **13:**1-9.
45. Ong C-T, Corces VG: **CTCF: an architectural protein bridging genome topology and function.** *Nature Reviews Genetics* 2014, **15:**234-246.
46. Yoo AB, Jette MA, Grondona M: **SLURM: Simple Linux Utility for Resource Management.** In *{Job Scheduling Strategies for Parallel Processing}.* Berlin, Germany: Springer; 2003: 44-60
47. Deamer D, Akeson M, Branton D: **Three decades of nanopore sequencing.** *Nature Biotechnology* 2016, **34:**518-524.
48. Jain M, Olsen HE, Paten B, Akeson M: **The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community.** *Genome Biology* 2016, **17:**1-11.
49. Oxford Nanopore Technologies: **How it works.** 2020.
50. Oxford Nanopore Technologies: **Continuous development and improvement.** 2020.
51. Oxford Nanopore Technologies: **PromethION.** 2020.
52. Oxford Nanopore Technologies: **PromethION 24 and PromethION 48 now available.** 2019.
53. Leggett RM, Clark MD: **A world of opportunities with nanopore sequencing.** *PeerJ Preprints* 2017.
54. Yang M, Thompson M: *Detection Methods in Precision Medicine (ISSN).* Royal Society of Chemistry; 2020.
55. Sakamoto Y, Xu L, Seki M, Yokoyama TT, Kasahara M, Kashima Y, Ohashi A, Shimada Y, Motoi N, Tsuchihara K, et al: **Long-read sequencing for non-small-cell lung cancer genomes.** *Genome Research* 2020.
56. Suzuki A, Suzuki M, Mizushima-Sugano J, Frith MC, Maka\l o, Wojciech, Kohno T, Sugano S, Tsuchihara K, Suzuki Y: **Sequencing and phasing cancer mutations in lung cancers using a long-read portable sequencer.** *DNA Research* 2017, **24:**585-596.

57. Valle-Inclan JE, Stangl C, de Jong AC, van Dessel LF, van Roosmalen MJ, Helmijr JCA, Renkens I, de Blank S, de Witte CJ, Martens JWM, et al: **Rapid identification of genomic structural variations with nanopore sequencing enables blood-based cancer monitoring.** *medRxiv* 2019**:**19011932.

58. Ishiura H, Doi K, Mitsui J, Yoshimura J, Matsukawa MK, Fujiyama A, Toyoshima Y, Kakita A, Takahashi H, Suzuki Y, et al: **Expansions of intronic TTTCA and TTTTA repeats in benign adult familial myoclonic epilepsy.** *Nature Genetics* 2018, **50:**581-590.

59. Sone J, Mitsuhashi S, Fujita A, Mizuguchi T, Hamanaka K, Mori K, Koike H, Hashiguchi A, Takashima H, Sugiyama H, et al: **Long-read sequencing identifies GGC repeat expansions in NOTCH2NLC associated with neuronal intranuclear inclusion disease.** *Nature Genetics* 2019, **51:**1215-1221.

60. Zeng S, Zhang M-y, Wang X-j, Hu Z-m, Li J-c, Li N, Wang J-l, Liang F, Yang Q, Liu Q, et al: **Long-read sequencing identified intronic repeat expansions in SAMD12 from Chinese pedigrees affected with familial cortical myoclonic tremor with epilepsy.** *Journal of Medical Genetics* 2019, **56:**265-270.

61. Bowden R, Davies RW, Heger A, Pagnamenta AT, de Cesare M, Oikkonen LE, Parkes D, Freeman C, Dhalla F, Patel SY, et al: **Sequencing of human genomes with nanopore technology.** *Nature Communications* 2019, **10:**1-9.

62. Euskirchen P, Bielle F, Labreche K, Kloosterman WP, Rosenberg S, Daniau M, Schmitt C, Masliah-Planchon J, Bourdeaut F, Dehais C, et al: **Same-day genomic and epigenomic diagnosis of brain tumors using real-time nanopore sequencing.** *Acta Neuropathologica* 2017, **134:**691-703.

63. Lee J, Shim H-r, Lee J-Y, Kim Y, Lee J-Y, Jung M-H, Choi W-Y, Hwang J-H, Kim LK, Kim Y-J: **Transcriptome profiling of Korean colon cancer by cDNA PCR Nanopore sequencing.** 2020.

64. de Jong LC, Cree S, Lattimore V, Wiggins GAR, Spurdle AB, Miller A, Kennedy MA, Walker LC: **Nanopore sequencing of full-length BRCA1 mRNA transcripts reveals co-occurrence of known exon skipping events.** *Breast Cancer Research* 2017, **19:**1-9.

65. Minervini CF, Cumbo C, Orsini P, Anelli L, Zagaria A, Impera L, Coccaro N, Brunetti C, Minervini A, Casieri P, et al: **Mutational analysis in BCR-ABL1 positive leukemia by deep sequencing based on nanopore MinION technology.** *Experimental and Molecular Pathology* 2017, **103:**33-37.

66. Tang AD, Soulette CM, van Baren MJ, Hart K, Hrabeta-Robinson E, Wu CJ, Brooks AN: **Full-length transcript characterization of SF3B1 mutation in chronic lymphocytic leukemia reveals downregulation of retained introns.** *Nature Communications* 2020, **11:**1-12.

67. Clark MB, Wrzesinski T, Garcia AB, Hall NAL, Kleinman JE, Hyde T, Weinberger DR, Harrison PJ, Haerty W, Tunbridge EM: **Long-read sequencing reveals the complex splicing profile of the psychiatric risk gene CACNA1C in human brain.** *Molecular Psychiatry* 2020, **25:**37-47.

68. Miga KH, Koren S, Rhie A, Vollger MR, Gershman A, Bzikadze A, Brooks S, Howe E, Porubsky D, Logsdon GA, et al: **Telomere-to-telomere assembly of a complete human X chromosome.** *Nature* 2020, **585:**79-84.

69. Ewing AD, Smits N, Sanchez-Luque FJ, Faivre J, Brennan PM, Richardson SR, Cheetham SW, Faulkner GJ: **Nanopore Sequencing Enables Comprehensive Transposable Element Epigenomic Profiling.** *Molecular Cell* 2020, **80:**915-928.e915.

70. Wongsurawat T, Jenjaroenpun P, De Loose A, Alkam D, Ussery DW, Nookaew I, Leung Y-K, Ho S-M, Day JD, Rodriguez A: **A novel Cas9-targeted long-read assay for simultaneous detection of IDH1/2 mutations and clinically relevant MGMT methylation in fresh biopsies of diffuse glioma.** *Acta Neuropathologica Communications* 2020, **8:**1-13.
71. Lee I, Razaghi R, Gilpatrick T, Molnar M, Gershman A, Sadowski N, Sedlazeck FJ, Hansen KD, Simpson JT, Timp W: **Simultaneous profiling of chromatin accessibility and methylation on human cell lines with nanopore sequencing.** *Nature Methods* 2020, **17:**1191-1199.
72. Goldsmith C, Cohen D, Dubois Aid, \else\"e,\fi,lle, Martinez M-G, Petitjean K, Corlu A, Testoni B, Hernandez-Vargas H, Chemin I: **Epigenetic heterogeneity after de novo assembly of native full-length Hepatitis B Virus genomes.** *bioRxiv* 2020**:**2020.2005.2029.122259.
73. Wei Y, Iyer SV, Costa ASH, Yang Z, Kramer M, Adelman ER, Klingbeil O, Demerdash OE, Polyanskaya S, Chang K, et al: **In vivo genetic screen identifies a SLC5A3-dependent myo-inositol auxotrophy in acute myeloid leukemia.** *bioRxiv* 2020**:**2020.2012.2022.424018.
74. Yang Z, Wei Y, Wu XS, Iyer SV, Jung M, Adelman ER, Klingbeil O, Kramer M, Demerdash OE, Chang K, et al: **Transcriptional silencing of ALDH2 in acute myeloid leukemia confers a dependency on Fanconi anemia proteins.** *bioRxiv* 2020**:**2020.2010.2023.352070.
75. Krueger F, Andrews SR: **Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications.** *Bioinformatics* 2011, **27:**1571.
76. Oxford Nanopore Technologies: **Nanopore sequencing 101 Q&A.** 2020.
77. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, Subgroup GPDP: **The Sequence Alignment/Map format and SAMtools.** *Bioinformatics* 2009, **25:**2078.
78. Barretina J, Caponigro G, Stransky N, Venkatesan K, Margolin AA, Kim S, Wilson CJ, Leh\ifmmode\acutea ea, \fi,r, Joseph, Kryukov GV, Sonkin D, et al: **The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity.** *Nature* 2012, **483:**603-607.
79. Cali DS, Kim JS, Ghose S, Alkan C, Mutlu O: **Nanopore sequencing technology and tools for genome assembly: computational analysis of the current state, bottlenecks and future directions.** *Briefings in Bioinformatics* 2019, **20:**1542-1559.
80. Oxford Nanopore Technologies: **Oxford Nanopore Technologies Github repository.** 2021.
81. Oxford Nanopore Technologies: **scrappie.** 2021.
82. Course dNNT: **Basecalling with Albacore.** In *deNBI Nanopore Training Course stable documentation*; 2019.
83. Oxford Nanopore Technologies: **New basecaller now performs ‘raw basecalling’, for improved sequencing accuracy.** 2018.
84. Oxford Nanopore Technologies: **Nanopore sequencing data analysis.** 2020.
85. Simpson Lab: **nanopolish-v0.12.0.** 2020.
86. Oxford Nanopore Technologies: **rerio.** 2021.
87. Wang Genomics Lab: **DeepMod model:rnn_conmodC_P100wd21_f7ne1u0_4.** 2021.

88. Ramírez F, Ryan DP, Grüning B, Bhardwaj V, Kilpert F, Richter AS, Heyne S, Dündar F, Manke T: **deepTools2: a next generation web server for deep-sequencing data analysis.** *Nucleic Acids Research* 2016, **44:**W160-W165.

89. Quick J, Quinlan AR, Loman NJ: **A reference bacterial genome dataset generated on the MinION™ portable single-molecule nanopore sequencer.** *GigaScience* 2014, **3**.

90. Jain M, Fiddes IT, Miga KH, Olsen HE, Paten B, Akeson M: **Improved data analysis for the MinION nanopore sequencer.** *Nature Methods* 2015, **12:**351-356.

91. Oxford Nanopore Technologies: **Company history.** 2020.

Table 1. Current DNA methylation calling tools for Nanopore sequencing

Tools	DNA Modification					Input required	Support multi-thread FAST5 file?	Flow cells	Model Trained On	Algorithm	Reported Performance
	4mC	5mC	5hmC	6mC	6mA						
Nanopolish[9]		✓				Base-called FAST5	✓	R7.3, R9, R9.4 series	*E. coli*	Hidden Markov model (HMM)	Accuracy = 0.94 (5mC, *Homo sapiens*)
Tombo/ Nanoraw[21]	✓	✓			✓	Raw FAST5		R9.4 series, R9.5	no model	Mann-Whitney and Fisher's exact test	Accuracy = 0.839, AUC = 0.78
NanoMod[22]		✓				Base-called FAST5, requires control sequence		R7.3	no model	Kolmogorov-Smirnov test	Precision = 0.9
DeepMod[25]		✓			✓	FAST5 with raw signals and base calls		R9.4 series	*E. coli*	Bidirectional recurrent neural network (RNN) with long short-term memory (LSTM)	Precision = 0.99, AUC > 0.97
SignalAlign[26]		✓	✓		✓	Base-called FAST5		R7.3	Synthetic nucleotides	Hidden Markov model with a hierarchical Dirichlet process (HMM-HDP)	Accuracy = 0.76 (for 5hmC, 5mC)
									E. coli		Accuracy = 0.96 (for 5mC), Precision = 0.92
mCaller[23]					✓	Base-called FAST5		R9.4 series	*E. coli*	Neural network	Accuracy = 0.954, AUC = 0.99
DeepSignal[27]		✓			✓	FAST5 processed by Tombo re-squiggle module		R9.4 series	*E. coli*	Bidirectional RNN with LSTM+Inception structure	Accuracy = 0.92(5mC, *Homo sapiens*), 0.90(m6A), Precision = 0.97
Megalodon[24]		✓			✓	Raw FAST5[a]	✓	R9.4 series	*Homo sapiens* and *E. coli*[b]	Recurrent neural network[c]	N/A[d]
methBERT[40]		✓			✓	Raw FAST5[a]		R9.4 series	*Homo sapiens* and *E. coli*	Bidirectional encoder representations from transformers (BERT)	Precision = 0.9147(5mC, *Homo sapiens*)[e]

a. Megalodon must obtain the intermediate output from the basecall neural network, and Guppy is the recommended backend to obtain this output from from FAST5.
b.The model is trained in biological contexts only on *Homo sapiens* and *E. coli*. Users have to specify the model from the modified base models included in basecaller Guppy or research models in ONT Rerio repository.
c. Megalodon's functionality centers on the anchoring of the high-information neural network basecalling output to a reference sequence.
d. The performance for Megalodon is not available since it is still actively developed, no available published paper yet.
e. Only 5mC precision on *Homo sapiens* at genomic level is listed here, more performance parameter(AUC, Recall) of 5mC at genomic level and read-level, and 5mC/6mA performance on *E.coli* are available in the original paper.

Figure1

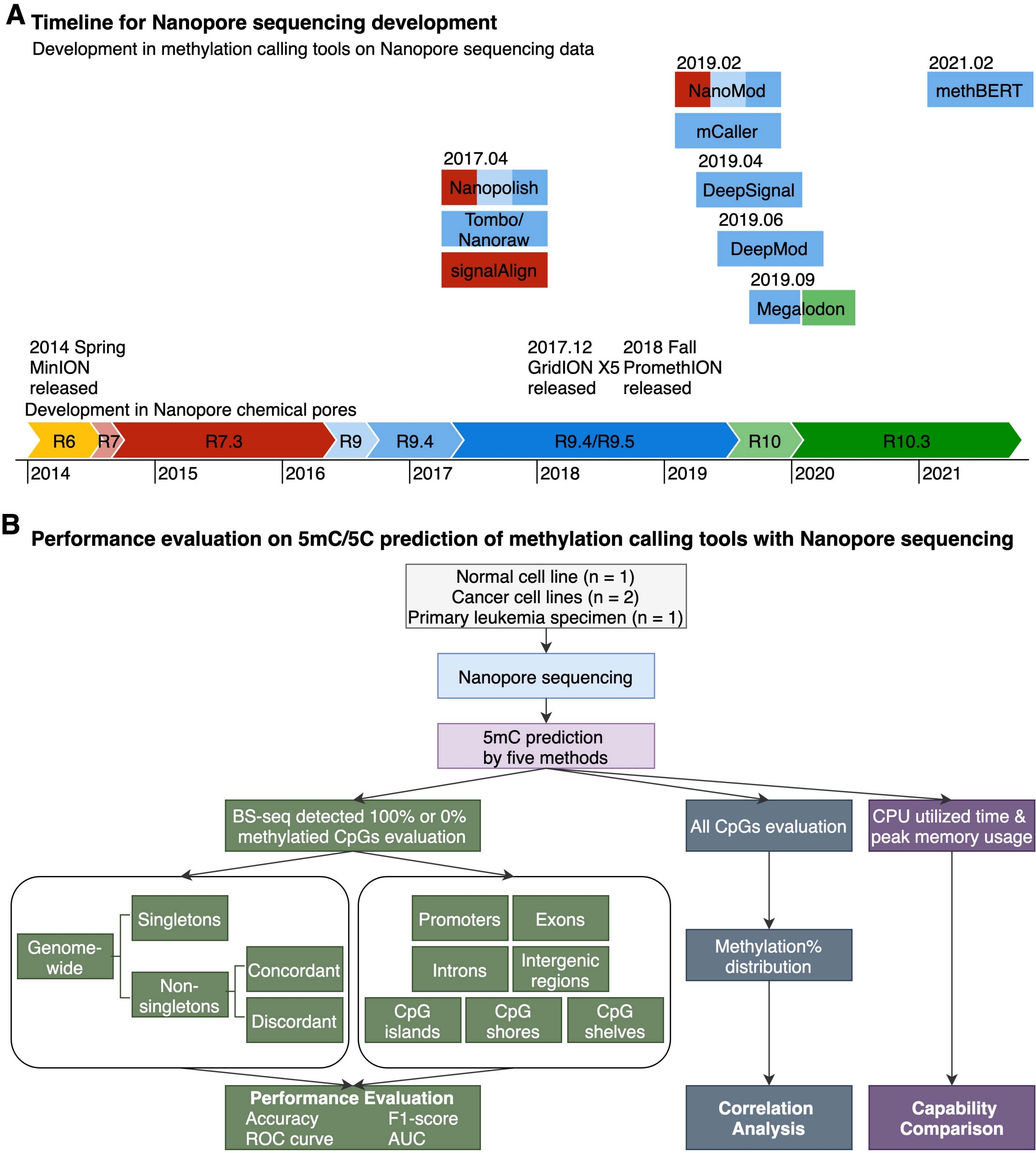

A
Timeline for Nanopore sequencing development
Development in methylation calling tools on Nanopore sequencing data
2017.04
Nanopolish
Tombo/
Nanoraw
signalAlign
2019.02
NanoMod
mCaller
2019.04
DeepSignal
2019.06
DeepMod
2019.09
Megalodon
2021.02
methBERT
2014 Spring
MinION
released
2017.12
GridION X5
released
2018 Fall
PromethION
released
Development in Nanopore chemical pores
R6
R7
R7.3
R9
R9.4
R9.4/R9.5
R10
R10.3
2014
2015
2016
2017
2018
2019
2020
2021
B
Performance evaluation on 5mC/5C prediction of methylation calling tools with Nanopore sequencing
Normal cell line (n = 1)
Cancer cell lines (n = 2)
Primary leukemia specimen (n = 1)
Nanopore sequencing
5mC prediction
by five methods
BS-seq detected 100% or 0%
methylatied CpGs evaluation
All CpGs evaluation
CPU utilized time &
peak memory usage
Genome-
wide
Singletons
Non-
singletons
Concordant
Discordant
Promoters
Exons
Introns
Intergenic
regions
CpG
islands
CpG
shores
CpG
shelves
Methylation%
distribution
Performance Evaluation
Accuracy
F1-score
ROC curve
AUC
Correlation
Analysis
Capability
Comparison

Figure2

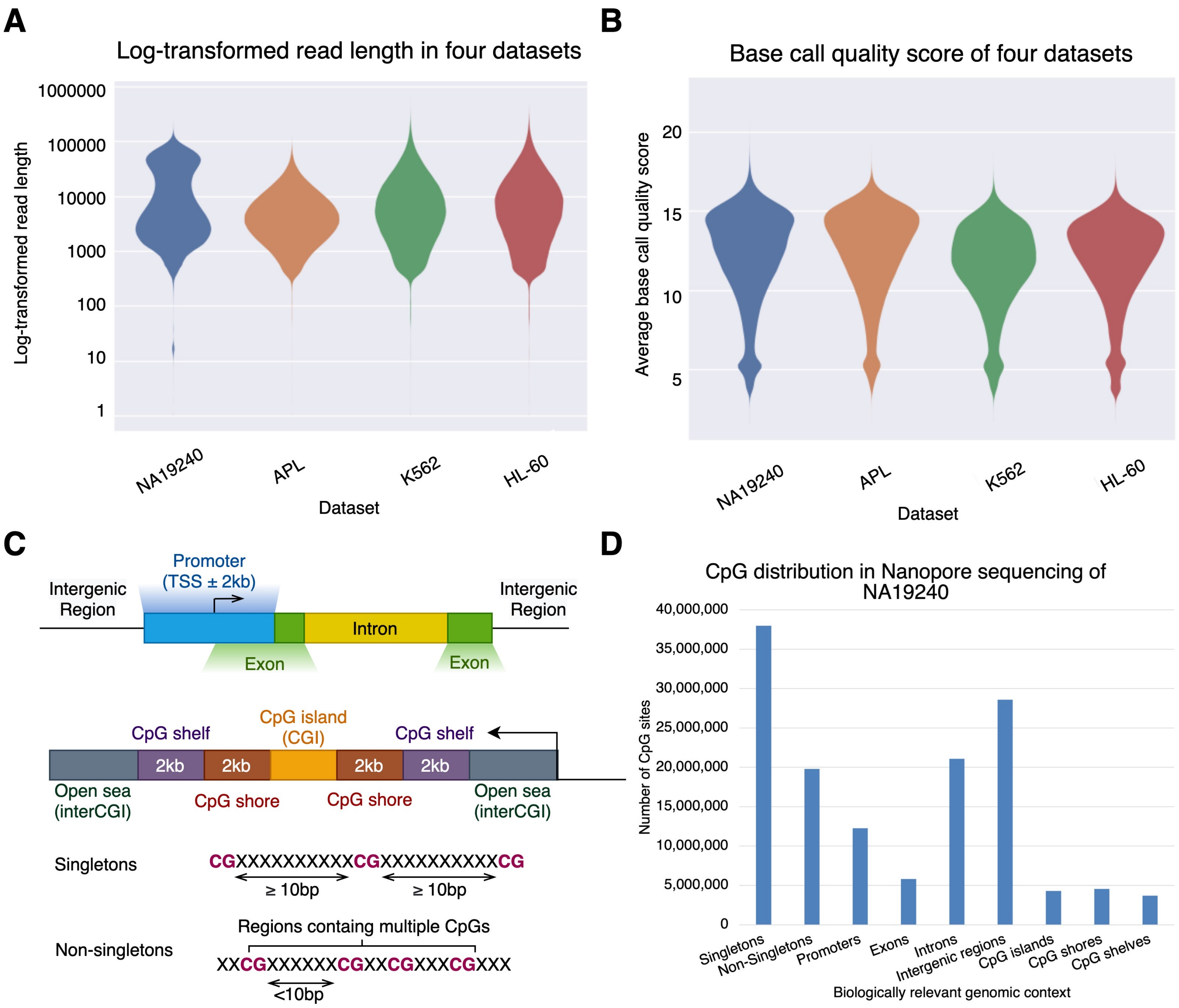

A
Log-transformed read length in four datasets
Log-transformed read length
1000000
100000
10000
1000
100
10
1
NA19240
APL
K562
HL-60
Dataset
B
Base call quality score of four datasets
Average base call quality score
20
15
10
5
NA19240
APL
K562
HL-60
Dataset
C
Promoter
(TSS ± 2kb)
Intergenic
Region
Intergenic
Region
Intron
Exon
Exon
CpG shelf
CpG island
(CGI)
CpG shelf
2kb
2kb
2kb
2kb
Open sea
(interCGI)
CpG shore
CpG shore
Open sea
(interCGI)
Singletons
CGXXXXXXXXXXCGXXXXXXXXXXCG
≥ 10bp
≥ 10bp
Non-singletons
Regions containg multiple CpGs
XXCGXXXXXXCGXXCGXXXCGXXX
<10bp
D
CpG distribution in Nanopore sequencing of NA19240
Number of CpG sites
40,000,000
35,000,000
30,000,000
25,000,000
20,000,000
15,000,000
10,000,000
5,000,000
0
Singletons
Non-Singletons
Promoters
Exons
Introns
Intergenic regions
CpG islands
CpG shores
CpG shelves
Biologically relevant genomic context

Figure3

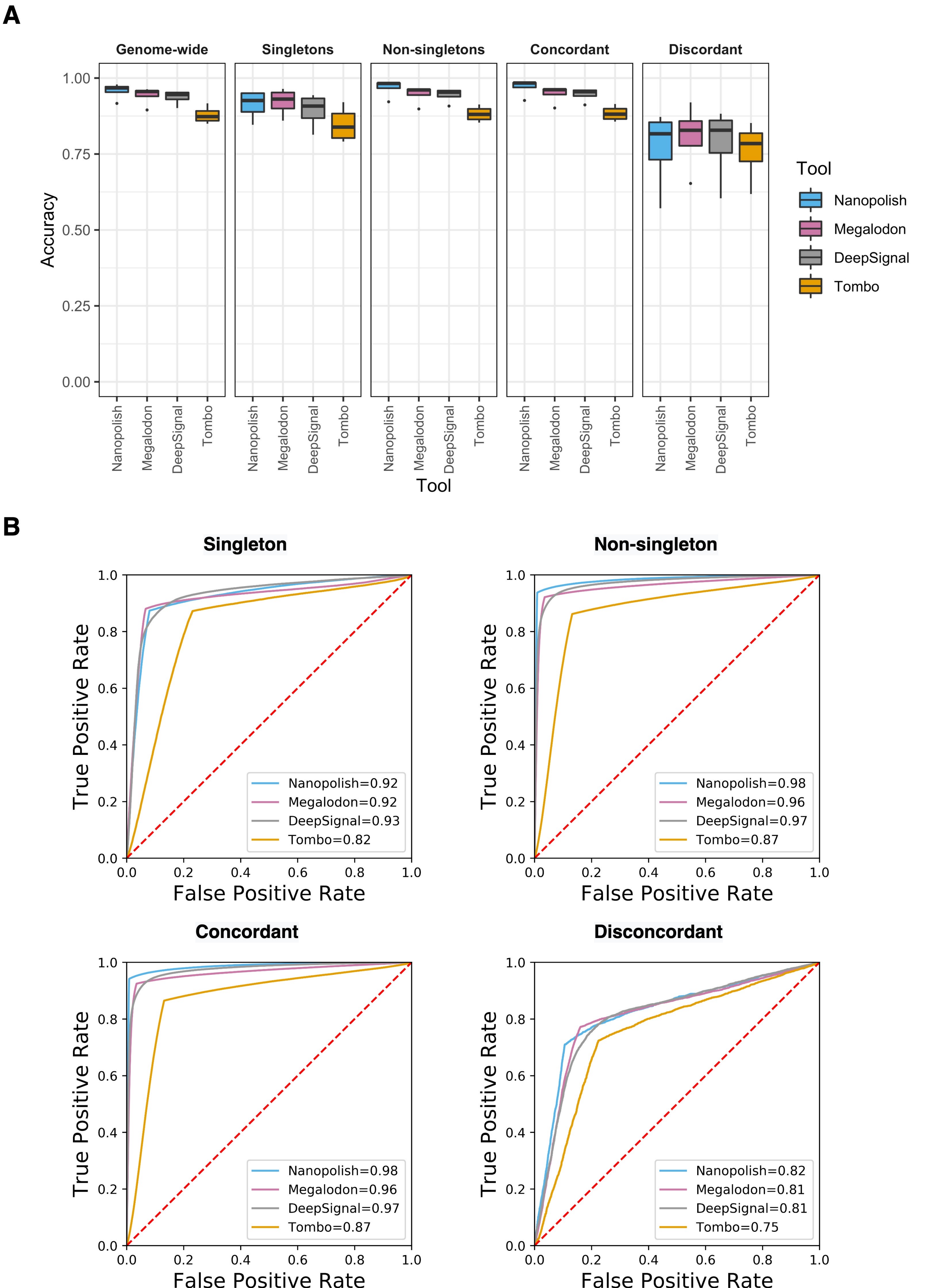

A
Genome-wide
Singletons
Non-singletons
Concordant
Discordant
Accuracy
1.00
0.75
0.50
0.25
0.00
Nanopolish
Megalodon
DeepSignal
Tombo
Tool
Tool
Nanopolish
Megalodon
DeepSignal
Tombo
B
Singleton
Non-singleton
Concordant
Disconcordant
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
Nanopolish=0.92
Megalodon=0.92
DeepSignal=0.93
Tombo=0.82
Nanopolish=0.98
Megalodon=0.96
DeepSignal=0.97
Tombo=0.87
Nanopolish=0.98
Megalodon=0.96
DeepSignal=0.97
Tombo=0.87
Nanopolish=0.82
Megalodon=0.81
DeepSignal=0.81
Tombo=0.75

Figure4

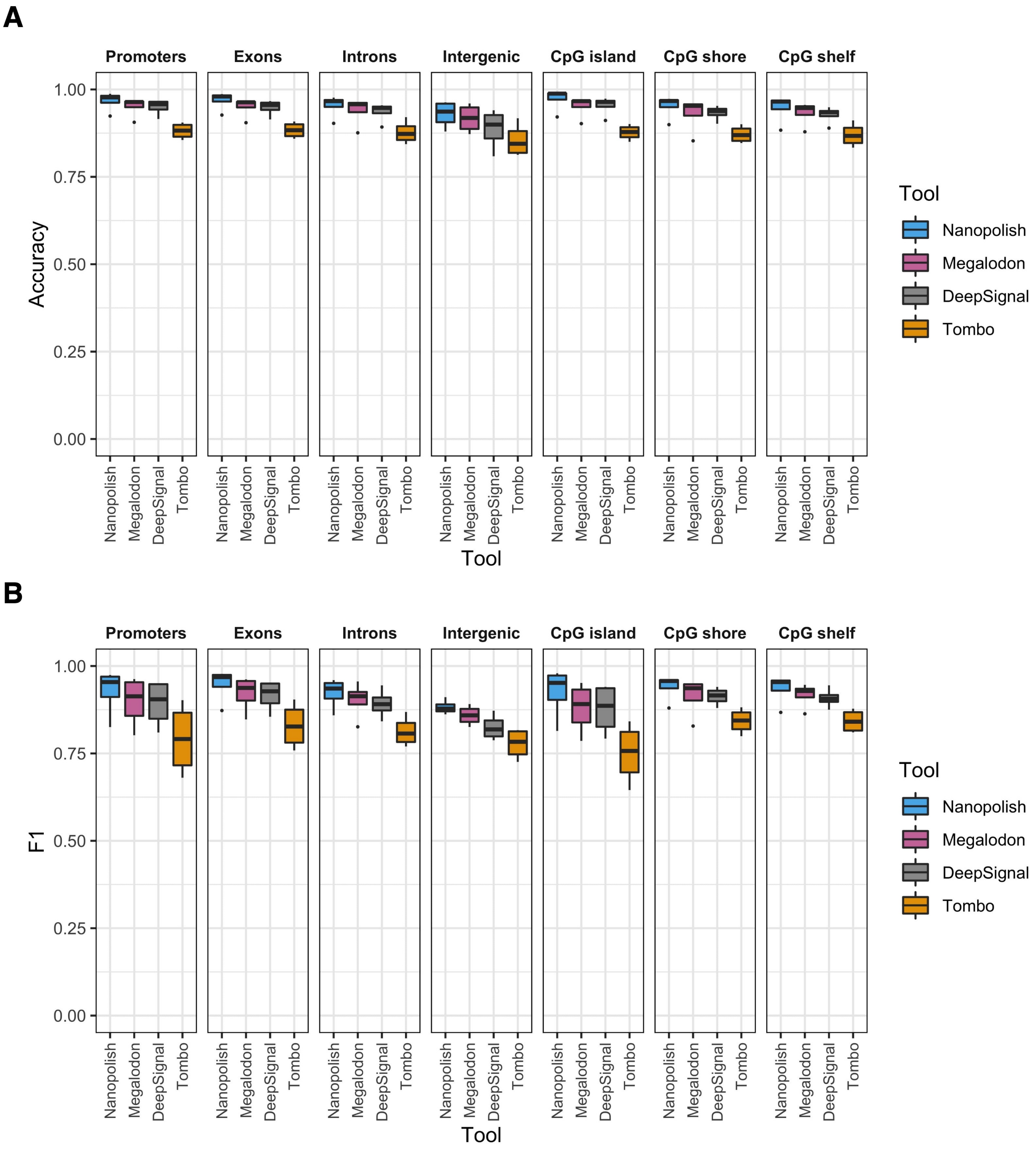

A
Promoters
Exons
Introns
Intergenic
CpG island
CpG shore
CpG shelf
Accuracy
1.00
0.75
0.50
0.25
0.00
Nanopolish
Megalodon
DeepSignal
Tombo
Tool
Tool
Nanopolish
Megalodon
DeepSignal
Tombo
B
Promoters
Exons
Introns
Intergenic
CpG island
CpG shore
CpG shelf
F1
1.00
0.75
0.50
0.25
0.00
Nanopolish
Megalodon
DeepSignal
Tombo
Tool
Tool
Nanopolish
Megalodon
DeepSignal
Tombo

Figure5

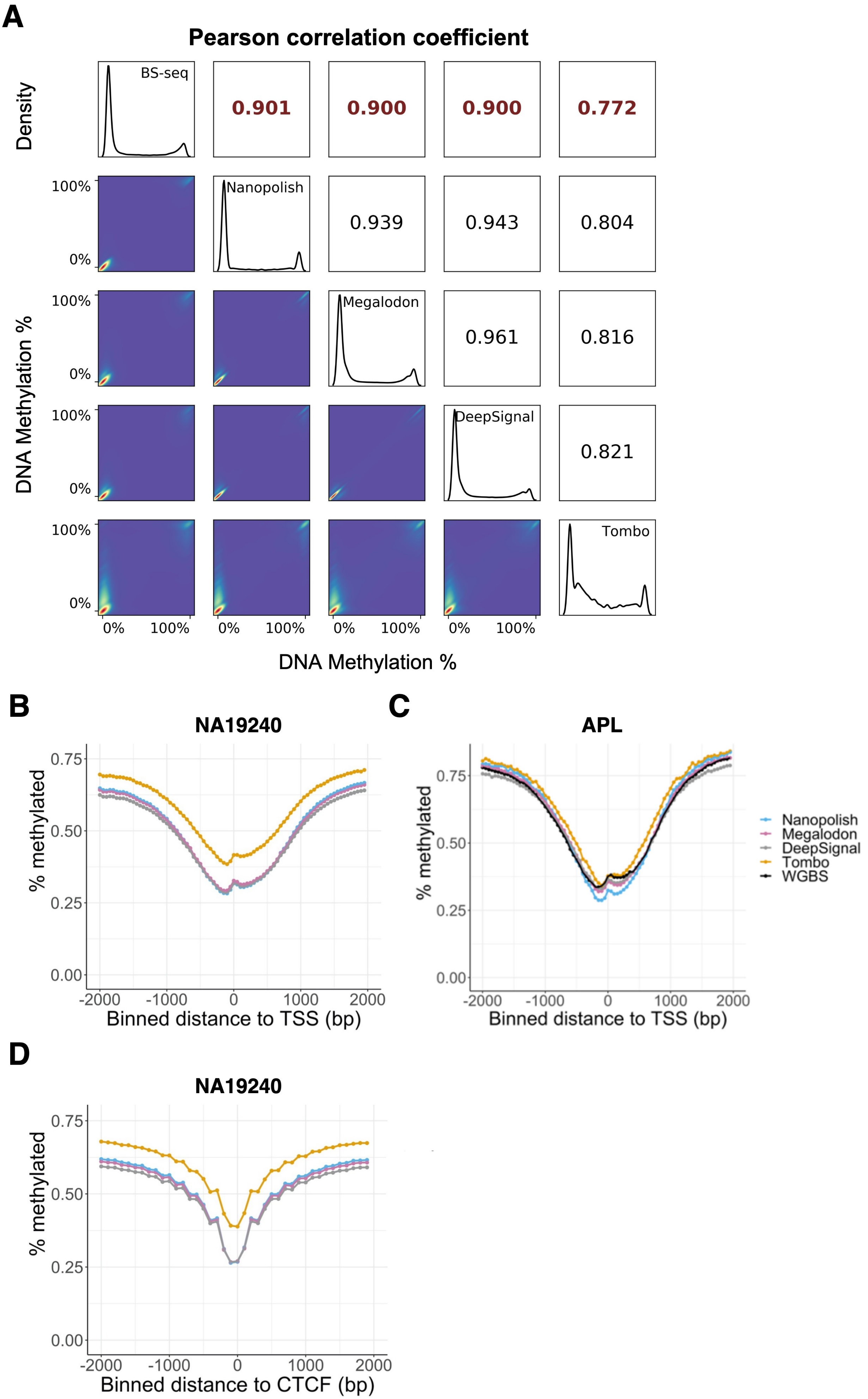

A
Pearson correlation coefficient
Density
BS-seq
0.901
0.900
0.900
0.772
Nanopolish
0.939
0.943
0.804
Megalodon
0.961
0.816
DeepSignal
0.821
Tombo
DNA Methylation %
100%
0%
0%
100%
DNA Methylation %
B
NA19240
% methylated
0.75
0.50
0.25
0.00
-2000
-1000
0
1000
2000
Binned distance to TSS (bp)
C
APL
% methylated
0.75
0.50
0.25
0.00
-2000
-1000
0
1000
2000
Binned distance to TSS (bp)
Nanopolish
Megalodon
DeepSignal
Tombo
WGBS
D
NA19240
% methylated
0.75
0.50
0.25
0.00
-2000
-1000
0
1000
2000
Binned distance to CTCF (bp)

Figure6

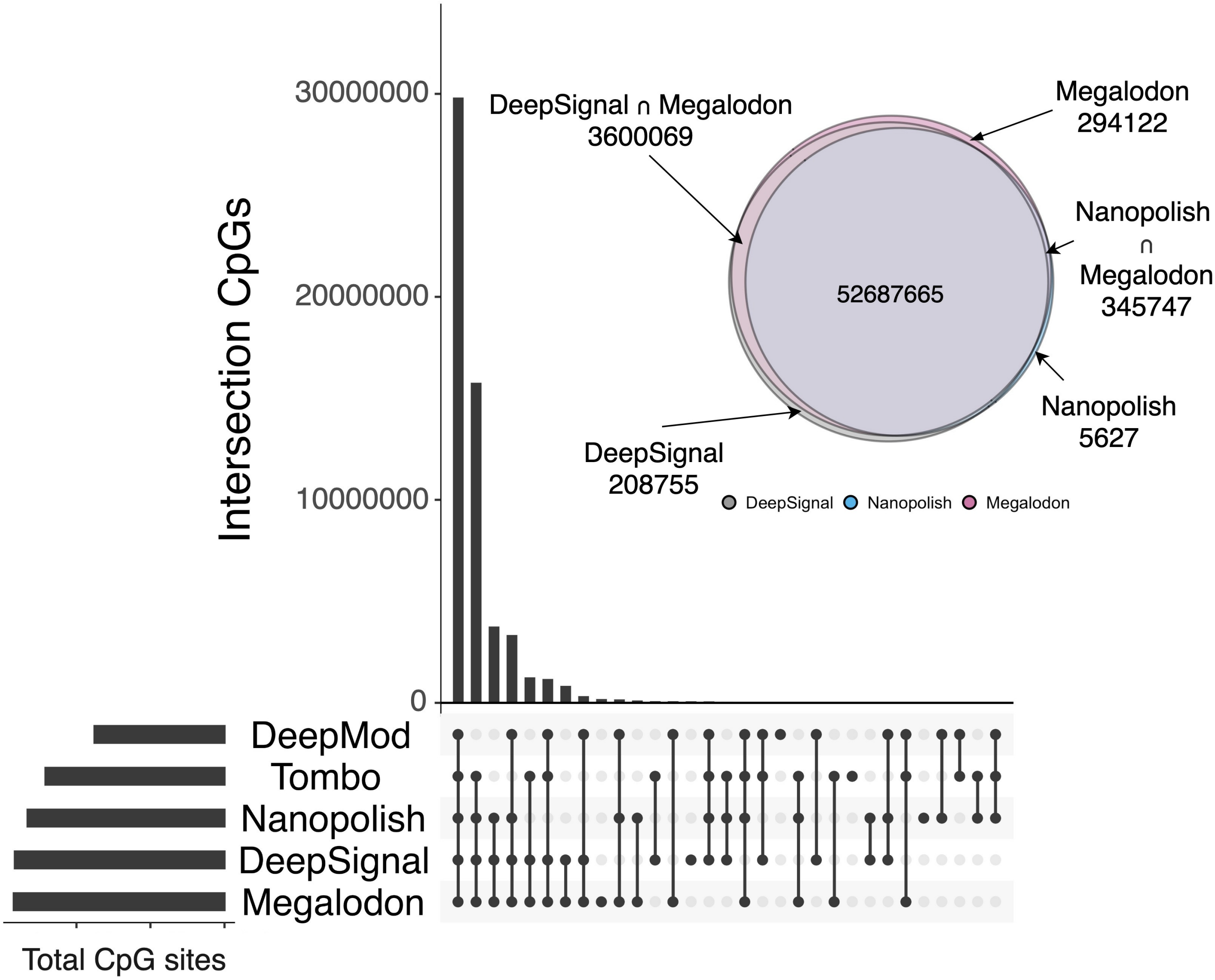
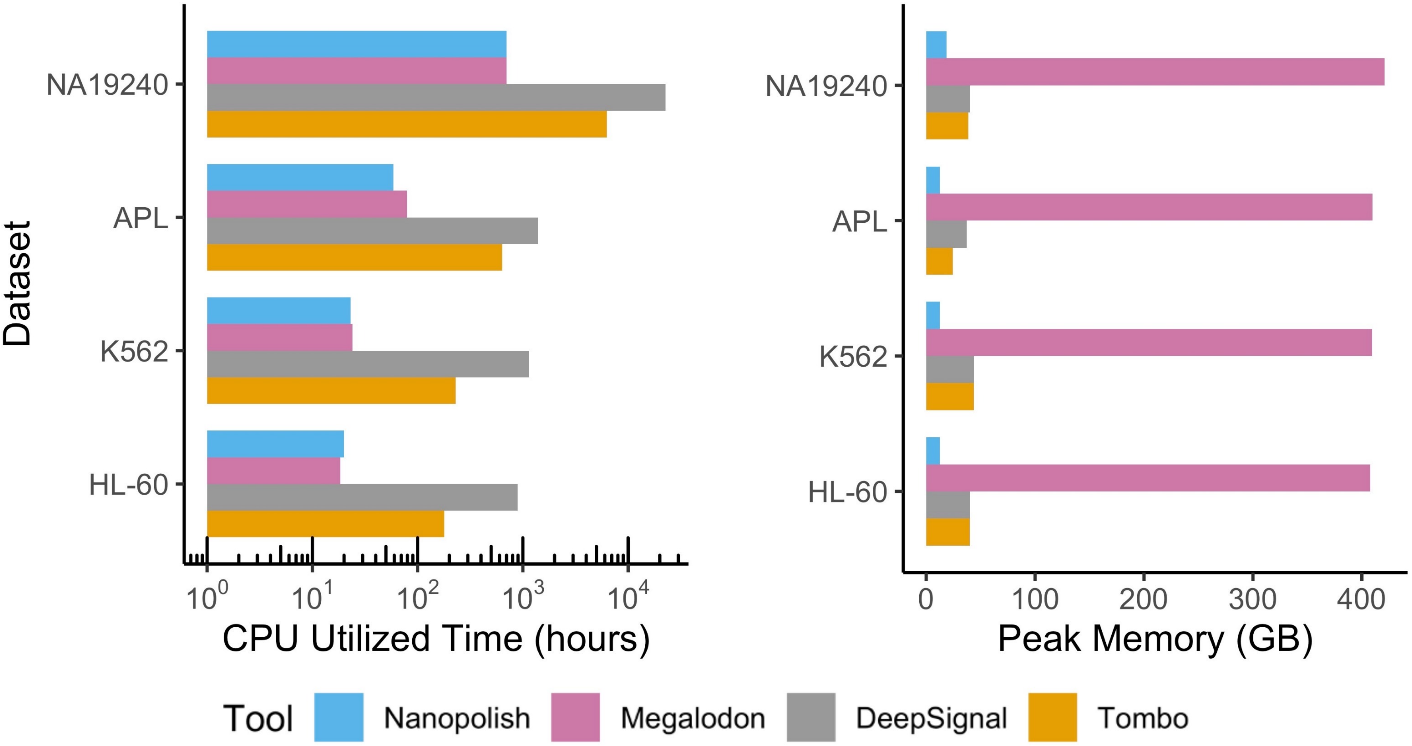

Intersection CpGs
30000000
20000000
10000000
0
DeepSignal ∩ Megalodon
3600069
Megalodon
294122
Nanopolish
∩
Megalodon
345747
52687665
Nanopolish
5627
DeepSignal
208755
DeepSignal
Nanopolish
Megalodon
DeepMod
Tombo
Nanopolish
DeepSignal
Megalodon
Total CpG sites

Figure7


Dataset
NA19240
APL
K562
HL-60
10^0
10^1
10^2
10^3
10^4
CPU Utilized Time (hours)
NA19240
APL
K562
HL-60
0
100
200
300
400
Peak Memory (GB)
Tool
Nanopolish
Megalodon
DeepSignal
Tombo