ACCESS MICROBIOLOGY

RESEARCH ARTICLE
Pai *et al.*, *Access Microbiology*
DOI 10.1099/acmi.0.000363

MICROBIOLOGY SOCIETY
OPEN ACCESS

Analysis of the long-read sequencing data using computational tools confirms the presence of 5-methylcytosine in the *Saccharomyces cerevisiae* genome

Shruta Sandesh Pai[1], Saumya Ranjan[1], Aimee Rachel Mathew[1], Roy Anindya[1,*] and Gargi Meur[2,*]

Abstract

Modification of DNA bases plays important roles in the epigenetic regulation of eukaryotic gene expression. Among the different types of DNA methylation, 5-methylcytosine (5mC) is common in higher eukaryotes. Although bisulfite sequencing is the established detection method for this modification, newer methods, such as Oxford nanopore sequencing, have been developed as quick and reliable alternatives. An earlier study using sensitive liquid chromatography tandem mass spectrometry (LC-MS/MS) indicated the presence of 5mC at very low concentration in *Saccharomyces cerevisiae*. More recently, a comprehensive study of the yeast genome found 40 5mC sites using the computational tool Nanopolish on nanopore sequencing output raw data. In the present study, we are trying to validate the prediction of the 5mC modifications in yeast with Nanopolish and two other nanopore software tools, Tombo and DeepSignal. Using publicly available genome sequencing data, we compared the open-access computational tools, including Tombo, Nanopolish and DeepSignal, for predicting 5mC. Our results suggest that these tools are indeed capable of predicting DNA 5mC modifications at a specific location from Oxford nanopore sequencing data. We also predicted that 5mC present in the *S. cerevisiae* genome might be located predominantly at the *RDN* locus of chromosome 12.

DATA SUMMARY

Analysis of nanopore sequencing data reveals the presence of 5-methylcytosine in the *Saccharomyces cerevisiae* genome.

INTRODUCTION

Epigenetic DNA methylation, 5-methylcytosine (5mC), is the most ubiquitous DNA methylation present in humans [1]. 5mC usually occurs in the context of the CpG dinucleotide clustered into regions, known as CpG islands. DNA CpG methylation is an important epigenetic modification involved in the regulation of transcription, development and genome stability. Loss of CpG modification is linked to cancer and inherited disorders such as Albright hereditary osteodystrophy (AHO), Beckwith–Wiedemann, Prader–Willi and Angelman syndromes, and pseudohypoparathyroidism (PhP) [2]. Due to the important implications of 5mC in epigenetics and disease, detection of 5mC is crucial. Among the numerous methods developed for the detection of DNA methylation, bisulfite sequencing is mostly used, as this method can yield single-base resolution of 5mC [3]. To achieve faster and cheaper sequencing, long-read PCR-free direct sequencing technologies are being developed to detect epigenetic 5mC [4], and recently Oxford Nanopore Technologies' (ONT's) nanopore sequencing and Pacific Biosciences' BacBio sequencing have been used for genome-wide analysis of methylation [5]. The ONT nanopore sequencers quantify the fluctuations of current when single-stranded nucleic acids pass through the nanopores and distinguishes the canonical bases from the 5mC from the distinct patterns of variation of current [6]. The current oscillation pattern is analysed using a basecaller algorithm that can detect the four canonical bases from the raw nanopore signals [7]. However, to detect the non-canonical bases such as 5mC, trained computational tools are employed, such as Nanopolish [8], DeepSignal [9], Tombo [10] and DeepMod [11].

Received 15 March 2022; Accepted 27 April 2022
Author affiliations: [1]Department of Biotechnology, Indian Institute of Technology Hyderabad, Kandi, Sangareddy-502284, India; [2]National Institute of Nutrition, Hyderabad-500007, India.
***Correspondence:** Roy Anindya, anindya@bt.iith.ac.in; Gargi Meur, gargimeur@gmail.com
Keywords: DNA methylation; long-read sequencing; third-generation sequencing; Nanopore sequencing.
Abbreviations: 5mC, 5-methylcytosine; ONT, Oxford Nanopore Technology.
Two supplementary figures and one supplementary table are available with the online version of this article.
000363

Although ubiquitous in most mammals, 5mC is considered to be absent in many lower eukaryotes, including the budding yeast *Saccharomyces cerevisiae*. However, using a sensitive method based on gas chromatography/mass spectrometry (GC/MS), it was recently shown to be present at very low concentration in several yeast species, including *S. cerevisiae* [12]. Whole-genome sequencing of *S. cerevisiae* with PacBio sequencing and ONT nanopore sequencing revealed that the yeast genome has ~40 5mCs, detected through the computational tool Nanopolish [13]. However, no other computational tools for nanopore sequencing data have been evaluated for the prediction of the 5mC in the *S. cerevisiae* genome. In the present study, we compared computational tools that are available in the public domain, namely, Tombo, mCaller, Nanopolish and DeepSignal, for their efficiency in predicting 5mC modifications from the deposited ONT nanopore sequencing data.

METHODS

Sequence data

We used ONT reads of yeast MCM869 available on the Sequence Read Archive (SRA) database (run ID ERR2804505) that is part of BioProject PRJEB28657. We downloaded the raw fast5 files using SRA-toolkit and they are referred to as dataset 1 in further experiments. For dataset validation, we used the raw fast5 sequence data of yeast made available on the DeepSignal GitHub repository by the authors of DeepSignal. These are referred to as dataset 2 in further experiments.

Tombo

Tombo is a Mann–Whitney U statistical comparison-based software tool, developed by ONT for the identification of non-canonical bases from nanopore sequencing data using Tombo command line interface [10]. Tombo works on the principle of resquiggling raw nanopore reads, where the squiggled raw nanopore signal is aligned to the reference genome. The latest available version of Tombo (1.5.1) was installed on Linux OS via bioconda environment (on Python 3.6 support). The resquiggle algorithm takes an input read file in fast5 format containing raw nanopore signal and associated base calls along with a reference genome. Resquiggling was the first step carried out by Tombo and the sequence to signal assignment was saved back into the read .fast5 file format. This step created a hidden file alongside the fast5s directory containing the essential genomic location for each read. The next step was to detect modifications using the Mann–Whitney U statistical test and this generated a binary statistics file in hdf5 file format that contained statistics associated with each genomic base producing a valid result. The following steps in the Tombo downstream pipeline made use of the above generated binary stats file as an input. Estimates of the fraction of modified readings can be inaccurate in low-coverage areas. In order to decrease the estimated fraction of modified reads at low-coverage sites, the coverage dampened fraction option is available. The 'tombo text_output browser_files' command generated output text files for both plus and minus strands in wiggle file format for the dampened fraction and bedGraph file format for the coverage, which can be used for visualization in the Integrated Genome Viewer (IGV) and Circos plot, and for the calculation of the ratio of methylated motifs to unmethylated motifs. Finally, the most significant modified base positions were obtained from the raw signal in the form of a plot by running the command 'tombo plot most_significant'.

Nanopolish

Nanopolish is based on the hidden Markov model (HMM), which computes the probability of observing a modified base (5mC) based on the differences in the event distributions of methylated and unmethylated DNA [8]. The latest available version of Nanopolish (0.9.0) was installed using 'apt install' along with the other prerequisites, namely samtools and minimap2. This program algorithm takes an input read file in fast5 file format containing raw nanopore signal and associated base calls and a reference genome. We first ran the 'extract' command to extract basecalled information from fast5 files into an output .fastq file. This file was further indexed using the 'index' command. With the help of minimap2 and samtools, the fastq file was aligned to a reference fasta file for the genome to produce a bam file that was further indexed. Next, methylations were predicted with the bam file as input using the 'call-methylation' command. Output was obtained as tsv file containing log_likelyhood_methylated, log_likelyhood_unmethylated and log_likelyhood_ratio for a particular position in every read. Lastly, a methylation summary file was obtained using the Python script available to give methylation frequency and coverage per position. Using this output, bedGraph files for log_likelyhood_ratio, methylation frequency and coverage were created, which were further analysed with IGV and Circos plot and used for the calculation of ratio of methylated motifs to unmethylated motifs.

DeepSignal

DeepSignal is a neural network-based tool that employs two modules to construct features from raw electrical signals of ONT reads [9]. This is performed by using the convolutional neural network (CNN) to construct features directly from raw electrical signals followed by the bidirectional recurrent neural network (BRNN) to construct features from sequences of signal information, which are then fed together into a fully connected neural network to predict the 5mC methylation states. The DeepSignal predictions were made using the Google Colaboratory platform by installing DeepSignal (0.1.8) with pip and other prerequisites, namely Python (3.6.0) with Conda, Ont-Tombo with Conda and TensorFlow (1.12) with pip. The trained model for 5mC in the context of CpG motif was downloaded from Google drive link available on the GitHub repository. We first ran the Ont-Tombo

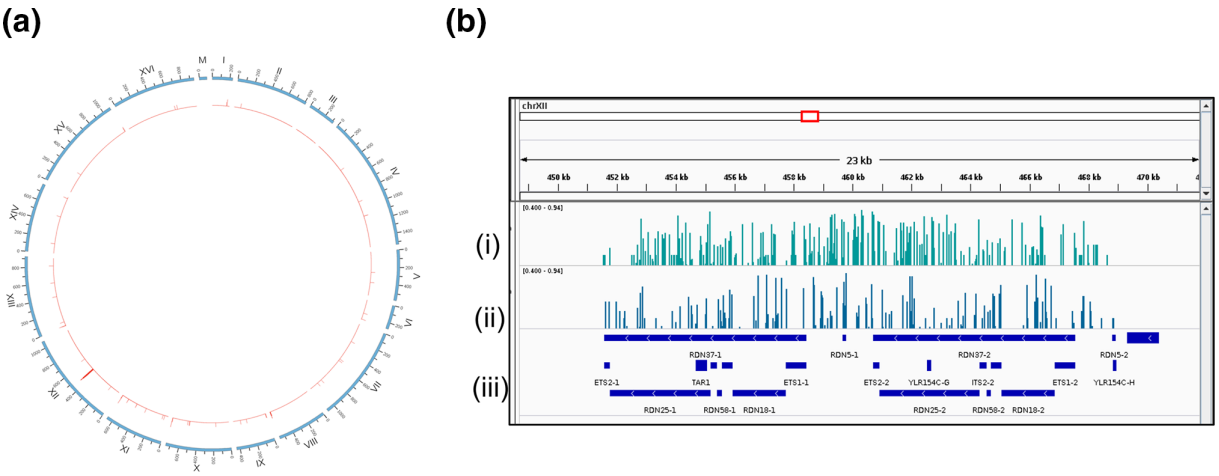

Fig. 1. Prediction of 5mC methylation by Tombo. (a) Circos plot for Nanopolish 5mC prediction in the *S. cerevisiae* genome showing all chromosomes in a circular map with ticks every 100 kbp and histogram tracks for each output. (i) Dampened fraction of 5mC on the positive strand and negative strand (threshold >0.5). (b) 5mC methylation at the *RDN* locus on *S. cerevisiae* chromosome XII; IGV analysis of Tombo output for starting with (i) dampened fraction of 5mC methylated residues on the positive strand (0.4 to 0.88), (ii) dampened fraction of 5mC methylated residues on the negative strand (0.4 to 0.92) and (iii) genes.

'resquiggle' command on the fast5 data to resquiggle the input based on a reference genome file followed by the DeepSignal 'extract' command to extract the signal features of motifs of defined length (17 mer) and sequence (CG). These motif features were saved as a tsv file, which was further used to call 5mC modifications using the 'call_mods' command. This command produced a tsv file as output that included probability_0 (unmethylated), probability_1 (methylated) and called_label, which was either 0 or 1 based on the probabilities for each position in every read. Lastly, a methylation summary file was obtained by running the Python script available to give methylation frequency and coverage per position. Using these outputs, bedGraph files for probability_0, probability_1, methylation frequency and coverage were created, which were further analysed with IGV and Circos plot and used for the calculation of the ratio of methylated motifs to unmethylated motifs. The details of all commands used to instal and run the computational tools are provided in the Supplementary Information (available in the online version of this article).

RESULTS AND DISCUSSION

Prediction of 5mC modification by Tombo, Nanopolish and DeepSignal

We first evaluated the Tombo tool for 5mC prediction in the context of CpG motifs. Tombo tool-generated outputs were completed successfully in the form of three file formats, namely, wiggle, bedGraph and PDF. The output in the wig text files, for both the forward and reverse strands, gave information regarding the base position and its dampened fraction, i.e. the estimated fraction of modified bases in the context of the GC motif at that position. The wig files were loaded as tracks into IGV, with SacCer3 being used as the genome for comparison to visualize the methylation patterns, while certain parameters, such as track height, track colour and windowing function, were adjusted for better visualization purposes. Another parameter, data range, was set arbitrarily to a minimum of 0.5 to eliminate redundant methylation values. After adjusting all the required parameters, we observed few low peaks in some of the telomeric regions of yeast chromosomes and a high peak in the middle of chromosome XII, as shown in Fig. 1(a) (i, ii) for the forward and reverse strands, respectively. The bedGraph files with coverage showed high values corresponding to the peak observed for dampened fraction, as shown in Fig. 1(a) (iii, iv) for the forward and reverse strands, respectively.

On further analysis, it was observed that the peak represented the *RDN* locus of the genome as shown in Fig. 1(b) (i, ii). Notably, the high coverage at the locus could well be due to the repetitive nature of the locus. It was also observed that methylations were present on both the forward and reverse strands of the *RDN* locus and in coding as well as non-coding regions with no positional bias to regulatory elements.

The Nanopolish computational tool gave output in terms of the log likelihood ratio of 5mC methylation in the context of CpG motifs and was used for predicting DNA methylation [13]. Positive values of the ratio indicated positions that are likely to be methylated, with negative values indicating otherwise. The summary Python script to calculate methylation frequency considered all the reads with a positive ratio value to be equally likely. As a result, we observe high peaks for methylation frequency throughout the genome, even in areas with a lower positive likelihood ratio value, as shown in Fig. 2(a) (i), owing to lower coverage in multiple

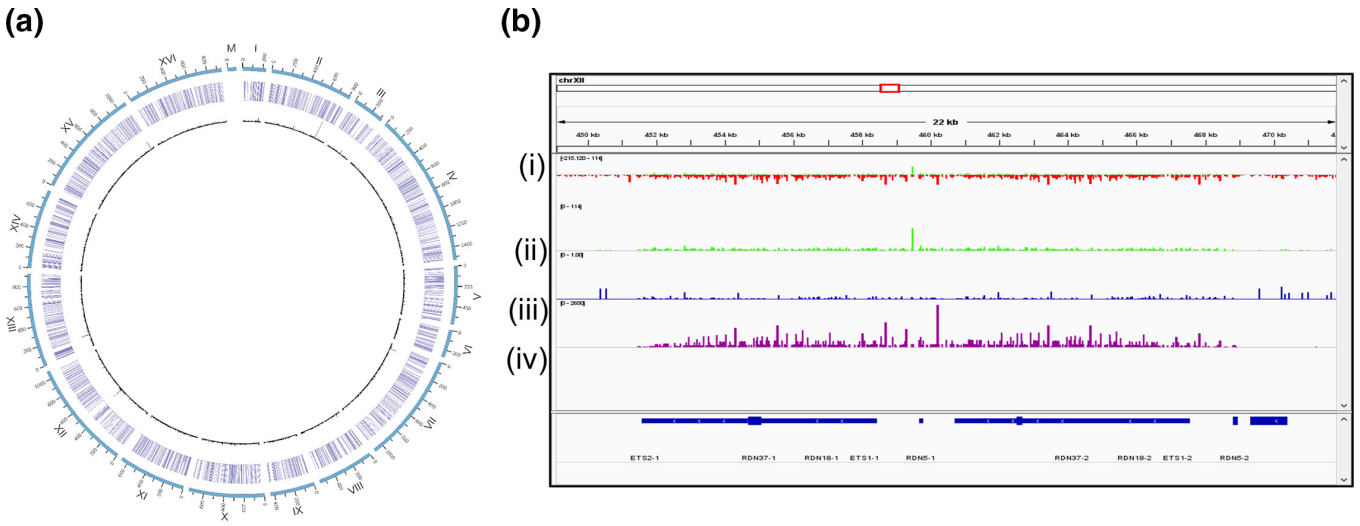

Fig. 2. Prediction of 5mC methylation by Nanopolish. (a) Circos plot for Nanopolish 5mC prediction in the *S. cerevisiae* genome showing all chromosomes in a circular map with ticks every 100 kbp and histogram tracks for each output. (i) Methylation frequency at each base (threshold >0.5). (ii) Positive values of the log likelihood ratio of the base being methylated in a particular read (0 to 11). (b) 5mC methylation at the *RDN* locus on *S. cerevisiae* chromosome XII; IGV analysis of Nanopolish output. (i) Log likelihood ratio of 5mC methylation with positive values in green and negative values in red (−215 to 11). (ii) Log likelihood ratio of 5mC methylation with positive values only (0 to 11). (iii) Methylation frequency of 5mC (blue) (0 to 1). (iv) Read coverage.

areas. We observed that very few regions are predicted to have high positive values of the likelihood ratio for 5mC methylation, with the *RDN* locus being one such region, as shown in Fig. 2(a) (ii). Within the *RDN* locus, we observed few reads showing a high positive value for the ratio, but the rest of the reads indicated a negative ratio value, as shown in Fig. 2(b) (i, ii). The summary script considered these multiple values as reads and not repeats and thus gave an average value for the locus showing significantly low methylation frequency, as shown in Fig. 2(b) (iii). Another computational tool, DeepSignal, yielded output in terms of the probability of being methylated and the probability of being unmethylated at a position in a particular read. If the value of the probability of being methylated was arbitrarily fixed >0.5, the position in the read was given a label of methylated and the summary frequency calculation script considered all the values with the methylated label to be equally probable. On applying a stringent threshold for probability (>0.9), we only observed 5mC predictions in a few regions spread throughout the genome, as shown in Fig. 3(a); interestingly, the *RDN* locus was one of these regions. Within the *RDN* locus, we observed few reads showing high 5mC

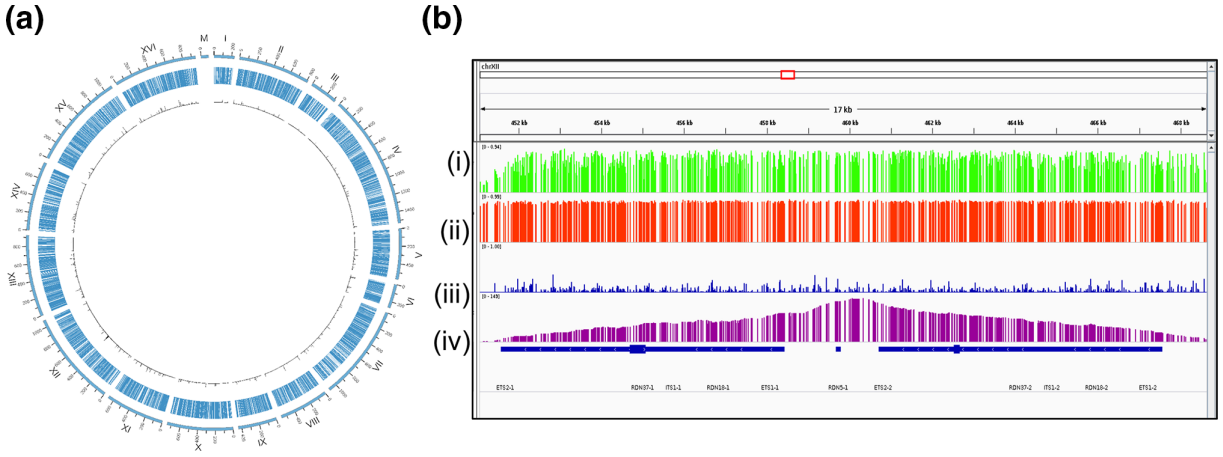

Fig. 3. Prediction of 5mC methylation by DeepSignal. (a) Circos plot for DeepSignal 5mC prediction in the *S. cerevisiae* genome showing all chromosomes in a circular map with ticks every 100 kbp and histogram tracks for each output. (i) Methylation frequency at the base (threshold >0.5) and (ii) methylation frequency at the base (threshold >0.9). (b) 5mC methylation at the *RDN* locus on the *S. cerevisiae* chromosome XII; IGV analysis of DeepSignal output. (i) Probability of being 5mC methylated (0 to 0.94). (ii) Probability of being unmethylated C (0 to 0.99). (iii) Methylation frequency of 5mC (0 to 1). (iv) Coverage.

methylation probability, but the rest of the reads indicated a high probability of unmethylated cytosine, as shown in Fig. 3(b) (i, ii). The summary script considered these multiple values as reads and not repeats and thus gave an average value for the locus showing significantly low methylation frequency, as shown in Fig. 3(b) (iii).

For the validation of the dataset, we used the raw fast5 data from nanopore sequencing of yeast available on the DeepSignal GitHub repository containing ~4000 R9.4 1D reads basecalled by Albacore (dataset 2), as mentioned previously. We observed that the variation with respect to 5mC peak for dampened fraction (Tombo), methylation likelihood (Nanopolish) and probabililty (DeepSignal) was consistent with dataset 1, which is depicted in Figs 1–3. The results for validation experiments are available in Figs S1 and S2.

The accuracy of 5mC methylation prediction

Whole-genome sequencing of *S. cerevisiae* with PacBio sequencing revealed that the yeast genome has ~40 5mCs [13]. Similarly, analysis of the yeast genome using highly sensitive mass spectrometry revealed that 5mC is present at a low level in *S. cerevisiae* [12]. Clearly, our analysis revealed that 5mC predictions made with dampened fraction (Tombo), methylation likelihood (Nanopolish) and methylation probability (DeepSignal) were overestimated and thus require stringent thresholds.

To obtain a realistic prediction of the methylation of the *S. cerevisiae* genome, the quantitative analysis was performed again by calculating a ratio that gave the percentage of the motifs predicted to be methylated out of all the motifs present in the sequence according to each computational tool. First, the methylation prediction for the whole genome was carried out at a different dampened fraction for Tombo, log likelihood ratio for Nanopolish and probability threshold for DeepSignal.

Tombo was used at a different dampened fraction to calculate the total CpG motifs that are methylated in the genome sequence. For the dampened fraction, the 'tombo text output browser files' command generated output text files in wiggle file format for both plus and minus strands. The wiggle file was converted into the tsv format, which has a column containing dampened fractions. The value of dampened fractions ranged from 0 to 0.93. The total CpG motifs were equal to 216 724 (Table S1). Total methylated CpG motifs were calculated by adding both plus and minus strands. At dampened fractions 0.7, there were 42 methylated CpG motifs, which is approximately equal to the ~40 5mC methylated sites predicted previously by PacBio sequencing. The ratio of methylated motifs to unmethylated motifs was found to be 0.019 at dampened fractions set to 0.7. A spike plot of Tombo output illustrating the variation of total methylated sites at different dampened fractions is shown in Fig. 4(a).

The Nanopolish tool yields the predictions by considering groups of motifs. As a result, the total number of lines did not correlate to the total number of CpG motifs present. Thus, the groups were split into individual CpG motifs and the methylation prediction for the whole group was considered to be true for all the CpG motifs in that group. The values for the total number of motif groups, the number of motif groups predicted to be methylated and all the values in the initial predictions are provided in the Table S1. The 'call methylation command' was used to predict the methylations with the bam file as an input file. The tsv file produced has a log likelihood ratio column whose values range from −215.12 to 113.55. A log likelihood ratio >1 is considered to be methylated. A total of 683541 sites are analysed in all reads. The different log likelihood ratios were calculated by setting different threshold and total readings that were methylated in all reads. The total methylated motif group, the total CpG motifs present in the genome sequence and the total methylated CpG motifs were calculated by using a methylation summary file containing methylation frequency and coverage per position. The total number of reads present in the methylated summary file gives the total analysed motif group, i.e. 188 037. Next, different log likelihood ratios were set and the total methylated motif group was calculated. The methylation summary file has a column named 'num_cpgs_in group', whose sum gives the total CpG motifs present in the genome, i.e. 256 946 (Table S1). The total methylated CpG motifs were calculated by setting different log likelihood thresholds and then obtaining the sum of 'num_cpgs_in group', whose values vary depending upon the log threshold log likelihood ratio. From Table S1, we can see that at the loglikelihood ratio 14 the number of CpG motifs that are methylated is 39, which is very close to the ~40 5meC sites predicted by PacBio sequencing. The ratio of methylated motifs to unmethylated motifs was 0.015. A spike plot of the Nanopolish output illustrating the variation of total methylated sites at different loglikelihood ratios is shown in Fig. 4(b).

The threshold values and corresponding 5mC methylation prediction sites were next calculated for DeepSignal. In DeepSignal, call mods command produced a tsv file that includes a probability_1(methylated) column, whose value ranged from 0.1 to 0.94. The total number of sites analysed in all reads is equal to the 683 541 obtained from the tsv file. The total number of methylated readings in all reads was calculated by setting different probability thresholds from 0.1 to 0.94. A methylation summary file containing the methylation frequency obtained using the tsv file as the input file has a total of 387 912 reads. The methylation frequency varied from 0 to 1. We were interested to determine the threshold at which the number of methylated sites was close to the number of 5meC calculated by PacBio sequencing. As shown in Table S1, DeepSignal predicted 51 5meC sites with the probability threshold set to 0.92, which is close to the ~40 5mC methylation sites determined by PacBio. The ratio of methylated motifs to unmethylated motifs at a probability threshold of 0.92 is equal to 0.013. A spike plot of the DeepSignal output illustrating the variation of total methylated sites at different probability thresholds is shown in Fig. 4(c).

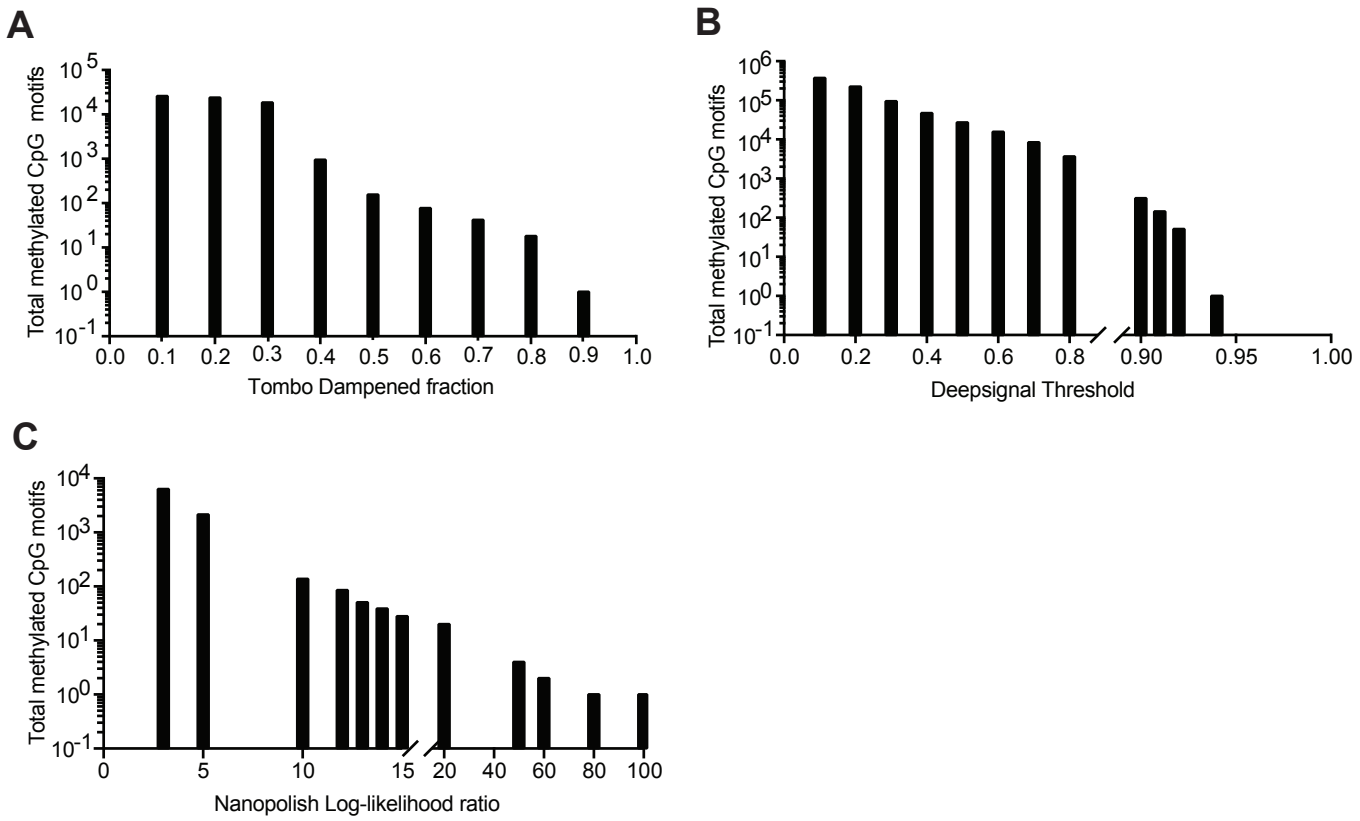

Fig. 4. Spike plot showing total methylated CpG motifs plotted against respective thresholds. (a) Tombo output illustrating the variation of total 5mC sites at different dampened fractions. (b) Variation of total methylated sites at different probability threshold, as determined by DeepSignal. (c) Nanopolish result depicting the variation of total methylated sites at different log likelihood ratios.

Localization of 5mC methylation in the *S. cerevisiae* genome

We have also predicted the chromosomes involved in 5mC methylation after setting thresholds for different tools that give values close to the ~40 5mC sites predicted by PacBio sequencing [13]. The chromosome number was studied with the dampened fraction set to 0.7 for Tombo, the log likelihood ratio set to 14 for Nanopolish and the probability threshold set to 0.92 for DeepSignal. The 5mC sites predicted by the three tools are present in different chromosomes. The values for all the 5mC methylation sites present in different chromosome are given in Table 1. The majority of methylation sites are found on chromosome 12, as shown in the Circos plot (Fig. 5a). DeepSignal predicts 14 5meC sites, Nanopolish predicts 10 sites and Tombo predicts 40 sites on chromosome 12. The CpG methylation in chromosome 12 is found to be localized at the *RDN* locus, as shown in the IGV analysis (Fig. 5b); Tombo predicts a high methylation fraction, whereas Nanopolish and DeepSignal predict high 5mC methylation probability. Nanopolish and DeepSignal provide a clearer picture for the repetitive locus by providing predictions of cytosine being methylated as well as unmethylated. At the repetitive *RDN* locus, Tombo shows a high methylation fraction, whereas Nanopolish and DeepSignal find high probability but low frequency. The *RDN* locus not only harbours the repeated units encoding ribosomal RNAs (rRNA), but is also the site of DNA replication, transcription and recombination [14]. Here, approximately 150 repeated units are arranged in tandem and each single unit of contains two genes, the RNA pol I-transcribed 35S rRNA gene and the RNA pol III-transcribed 5S rRNA gene. The 35S rRNA is further processed to generate the 18S, 5.8S and 25S rRNAs. Therefore, the prediction of 5mC in the *RDN* locus is intriguing. Whether 5mC modification regulates transcription in the RDN locus remains to be explored.

The *S. cerevisiae RDN* locus encoding ribosomal RNA is present as approximately 150 repeats and has even been reported to have 200 repeats in some strains. Remarkably, RNA polymerase II transcription is silenced within the rDNA locus, even though this region of the yeast genome is very actively transcribed by Pol I and III. Earlier studies indicated that Set1-mediated H3 Lys4 methylation is present at the rDNA locus and crucial for repression of RNA polymerase II transcription within rDNA [15]. It was also reported that Sir2 is present in the RDN locus contributing to RNA polymerase II transcriptional repression [16]. Interestingly, a recent study demonstrated that actively transcribed rRNA genes of the *RDN* locus are largely devoid of histone proteins, but instead associate with the high-mobility group protein Hmo1 [17]. To the best of our knowledge, there are no reports on the impact of endogenous DNA methylation on chromatin modification of yeast. However, our results indicate that methylation sites

Table 1. Number of 5mC found using Tombo, Nanopolish and DeepSignal in different *S. cerevisiae* chromosomes

Chromosome	No. of 5mC		
	Tombo	Nanopolish	DeepSignal
Chromosome 1	0	1	0
Chromosome 2	0	5	2
Chromosome 3	0	1	0
Chromosome 4	0	3	1
Chromosome 5	0	1	2
Chromosome 6	0	0	3
Chromosome 7	0	3	4
Chromosome 8	0	3	3
Chromosome 9	0	0	1
Chromosome 10	1	1	1
Chromosome 11	1	5	4
Chromosome 12	40	10	14
Chromosome 13	0	2	5
Chromosome 14	0	0	0
Chromosome 15	0	3	4
Chromosome 16	0	1	7
Total 5meC sites	**42**	**39**	**51**

are present on the forward and reverse strands of the RDN locus and in coding as well as non-coding regions, with no positional bias to regulatory elements. Considering the present preliminary findings, it is worth exploring further whether the observed pattern of methylation in this region is linked with any transcription regulation or chromatin organization.

Taken together, we demonstrated that the Tombo, Nanopolish and DeepSignal software tools can indeed be used to predict DNA 5mC modifications. Our study shows that Tombo predicts high methylation fractions, whereas Nanopolish and DeepSignal predict high 5mC methylation probability. Since Nanopolish and DeepSignal give a measure of both frequency and probability, either of the two tools would be suitable for analysing methylation, but DeepSignal could provide a specific advantage with a lower read coverage requirement. If prior information regarding the total number of methylation sites is available, then accurate localization of the methylation sites could be achieved by the nanopore sequence analysis. Recent advances in highly sensitive mass spectrometry methods for detecting epigenetic DNA modification might be helpful for this.

GitHub links

(1) https://nanoporetech.github.io/tombo/ (for Tombo).
(2) https://github.com/al-mcintyre/mCaller (for mCaller).
(3) https://github.com/jts/nanopolish (for Nanopolish).
(4) https://github.com/bioinfomaticsCSU/deepsignal (for DeepSignal).

Funding information
The work was funded by the Science and Engineering Research Board (SERB), Government of India, (extra-mural research grant EMR/2016/005135). The fellowship support of A.R.M. and S.S.P. from the Ministry of Education, Government of India, is acknowledged.

Author contributions
R.A., conceived the study; S.S.P., S.R. and A.R.M., performed the experiments and analysis; R.A., S.R., S.S.P. and A.R.M., wrote the manuscript.

Conflicts of interest
The authors declare that there are no conflicts of interest.

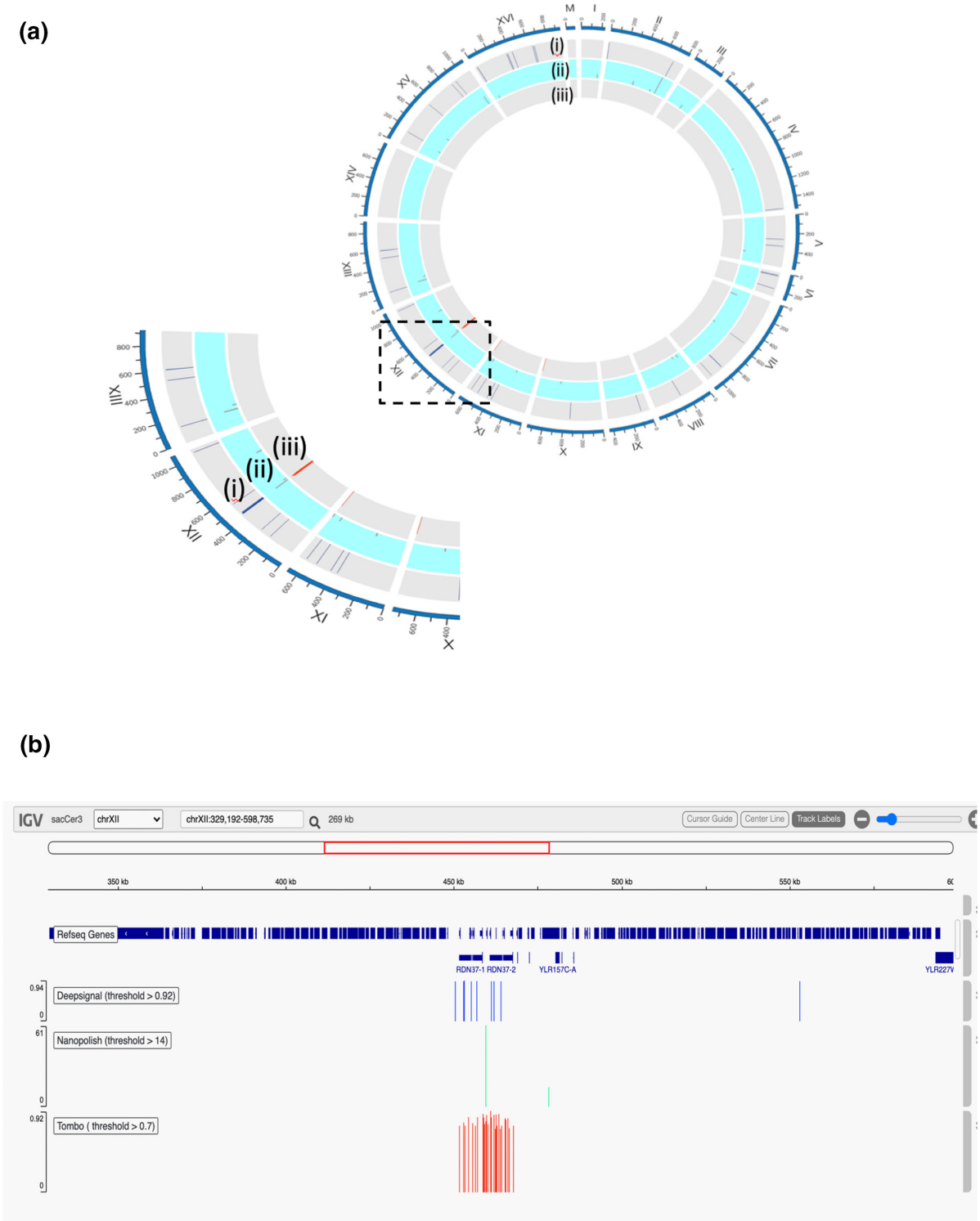

Fig. 5. Comparison of 5mC methylation predictions in the *S. cerevisiae* genome. (a) Circos plot for comparison of 5mC predictions in the *S. cerevisiae* genome showing all chromosomes in a circular map with ticks every 100 kbp and histogram tracks for output data from each computational tool. (i) Probability of methylation predicted by DeepSignal (threshold >0.92 ; scale, 0–1). (ii) Log likelihood ratio predicted by Nanopolish (threshold >14 ; scale, 0–14). (iii) Dampened fraction predicted by Tombo (threshold as >0.7; scale, 0–1). (b) IGV analysis of 5mC methylation at the *RDN* locus on *S. cerevisiae* chromosome XII. (i) DeepSignal output showing the probability of being 5mC methylated at threshold >0.94. (ii) Nnaopolish output showing the log likelihood ratio at threshold >14. (iii) Tombo output showing the dampened fraction at threshold >0.7.

AQ1 References

1. Smith ZD, Meissner A. DNA methylation: roles in mammalian development. *Nat Rev Genet* 2013;14:204–220.
2. Robertson KD. DNA methylation and human disease. *Nat Rev Genet* 2005;6:597–610.
3. Li Y, Tollefsbol TO. DNA methylation detection: bisulfite genomic sequencing analysis. *Methods Mol Biol* 2011;791:11–21.
4. Schreiber J, Wescoe ZL, Abu-Shumays R, Vivian JT, Baatar B, *et al*. Error rates for nanopore discrimination among cytosine, methylcytosine, and hydroxymethylcytosine along individual DNA strands. *Proc Natl Acad Sci U S A* 2013;110:18910–18915.
5. Schatz MC. Nanopore sequencing meets epigenetics. *Nat Methods* 2017;14:347–348.
6. Gouil Q, Keniry A. Latest techniques to study DNA methylation. *Essays Biochem* 2019;63:639–648.
7. Xu L, Seki M. Recent advances in the detection of base modifications using the Nanopore sequencer. *J Hum Genet* 2020;65:25–33.
8. Simpson JT, Workman RE, Zuzarte PC, David M, Dursi LJ, *et al*. Detecting DNA cytosine methylation using nanopore sequencing. *Nat Methods* 2017;14:407–410.
9. Ni P, Huang N, Zhang Z, Wang D-P, Liang F, *et al*. DeepSignal: detecting DNA methylation state from Nanopore sequencing reads using deep-learning. *Bioinformatics* 2019;35:4586–4595.
10. Stoiber M, Quick J, Egan R, Eun Lee J, Celniker S, *et al*. *De novo* identification of DNA modifications enabled by genome-guided nanopore signal processing. *Bioinformatics* 2016;094672.
11. Liu Q, Fang L, Yu G, Wang D, Xiao CL, *et al*. Detection of DNA base modifications by deep recurrent neural network on Oxford Nanopore sequencing data. *Nat Commun* 2019;10:2449.
12. Tang Y, Gao X-D, Wang Y, Yuan B-F, Feng Y-Q. Widespread existence of cytosine methylation in yeast DNA measured by gas chromatography/mass spectrometry. *Anal Chem* 2012;84:7249–7255.
13. Jenjaroenpun P, Wongsurawat T, Pereira R, Patumcharoenpol P, Ussery DW, *et al*. Complete genomic and transcriptional landscape analysis using third-generation sequencing: a case study of *Saccharomyces cerevisiae* CEN.PK113-7D. *Nucleic Acids Res* 2018;46:e38.
14. Egidi A, Di Felice F, Camilloni G. *Saccharomyces cerevisiae* rDNA as super-hub: the region where replication, transcription and recombination meet. *Cell Mol Life Sci* 2020;77:4787–4798.
15. Briggs SD, Bryk M, Strahl BD, Cheung WL, Davie JK, *et al*. Histone H3 lysine 4 methylation is mediated by Set1 and required for cell growth and rDNA silencing in *Saccharomyces cerevisiae*. *Genes Dev* 2001;15:3286–3295.
16. Smith JS, Caputo E, Boeke JD. A genetic screen for ribosomal DNA silencing defects identifies multiple DNA replication and chromatin-modulating factors. *Mol Cell Biol* 1999;19:3184–3197.
17. Merz K, Hondele M, Goetze H, Gmelch K, Stoeckl U, *et al*. Actively transcribed rRNA genes in S. cerevisiae are organized in a specialized chromatin associated with the high-mobility group protein Hmo1 and are largely devoid of histone molecules. *Genes Dev* 2008;22:1190–1204.