DNAModAnnot: a R toolbox for DNA modification filtering and annotation

Alexis Hardy, Mélody Matelot, Amandine Touzeau, Christophe C. Klopp, Céline Lopez-Roques, Sandra Duharcourt, Matthieu Defrance

Bioinformatics, 37(17), 2021, 2738–2740
doi: 10.1093/bioinformatics/btab032
Advance Access Publication Date: 20 January 2021
Applications Note

Genome analysis

DNAModAnnot: a R toolbox for DNA modification filtering and annotation

Alexis Hardy[1], Mélody Matelot[1], Amandine Touzeau[1], Christophe Klopp[2], Céline Lopez-Roques [3], Sandra Duharcourt [1,*] and Matthieu Defrance [4,*]

[1]Université de Paris, CNRS, Institut Jacques Monod, F-75006 Paris, France, [2]Plateforme bioinformatique Genotoul, UR875 Mathématique et Informatique Appliquée de Toulouse, INRA, 31326 Castanet-Tolosan, France, [3]INRAe, US 1426, GeT-PlaGe, Genotoul, 31326 Castanet-Tolosan, France and [4]Université Libre de Bruxelles, Interuniversity Institute of Bioinformatics in Brussels (IB2), Brussels 1050, Belgium

*To whom correspondence should be addressed.
Associate Editor: Peter Robinson

Received on September 24, 2020; revised on December 17, 2020; editorial decision on January 10, 2021; accepted on January 13, 2021

Abstract

Motivation: Long-read sequencing technologies can be employed to detect and map DNA modifications at the nucleotide resolution on a genome-wide scale. However, published software packages neglect the integration of genomic annotation and comprehensive filtering when analyzing patterns of modified bases detected using Pacific Biosciences (PacBio) or Oxford Nanopore Technologies (ONT) data. Here, we present DNA Modification Annotation (DNAModAnnot), a R package designed for the global analysis of DNA modification patterns using adapted filtering and visualization tools.

Results: We tested our package using PacBio sequencing data to analyze patterns of the 6-methyladenine (6mA) in the ciliate *Paramecium tetraurelia*, in which high 6mA amounts were previously reported. We found *P. tetraurelia* 6mA genome-wide distribution to be similar to other ciliates. We also performed 5-methylcytosine (5mC) analysis in human lymphoblastoid cells using ONT data and confirmed previously known patterns of 5mC. DNAModAnnot provides a toolbox for the genome-wide analysis of different DNA modifications using PacBio and ONT long-read sequencing data.

Availability and implementation: DNAModAnnot is distributed as a R package available via GitHub (https://github.com/AlexisHardy/DNAModAnnot).

Contact: sandra.duharcourt@ijm.fr or matthieu.dc.defrance@ulb.ac.be

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Recent progress in sequencing methods enables genome-wide detection and localization of DNA modifications at single-base resolution. Pacific Biosciences (PacBio) and Oxford Nanopore Technologies (ONT), two long-read sequencing technologies, can now directly detect modified bases (Gouil and Keniry, 2019). Modification detection is based either on raw electric signals alterations (ONT) or on DNA polymerase slowing-down events (PacBio). After sequencing and alignment, modified bases positions can be retrieved using dedicated detection pipelines (Amarasinghe *et al.*, 2020).

However, pipelines developed for DNA modification patterns analysis using PacBio (e.g. SMRTPortal, SMRTER) (Amarasinghe *et al.*, 2020) or ONT data [e.g. MethplotLib, (De Coster *et al.*, 2020), pycoMeth (https://a-slide.github.io/pycoMeth/)] do not integrate annotation or thorough filtering steps (Supplementary Table S1). Furthermore, recent studies highlighted that the unproper filtering of PacBio data could cause high amounts of artifacts (O'Brown *et al.*, 2019).

Here, we present the DNA Modification Annotation (DNAModAnnot) package, a modular collection of R tools allowing comprehensive filtering of PacBio and ONT data and, modified base pattern analysis at the genome level. DNAModAnnot provides a valuable addition to existing pipelines for the analysis of DNA modification patterns (see Supplementary Table S1 for a detailed comparison). It includes customized data visualization functions. Thanks to its flexibility, it can also be easily extended with existing analysis functions provided by external Bioconductor packages [https://rdrr.io/bioc/hiAnnotator, (Peters *et al.*, 2015), https://rdrr.io/bioc/methyAnalysis/].

2 Package description

The DNAModAnnot package is designed for the filtering and global analysis of DNA modifications detected from PacBio or ONT data. To increase the detection specificity of PacBio data, the package

provides a set of filters based on False Discovery Rate (FDR). The package also provides computational and visualization tools to compare modified bases distribution and genomic annotations.

The modularity of this toolbox allows the user to design its own flow of analysis using a combination of the following modules (Fig. 1A and Supplementary Fig. S1; see Supplementary Description):

'Data loading' module—Preprocessed PacBio or ONT data, using dedicated modification detection software, are converted to standard Bioconductor formats to facilitate the interoperability with other tools.

'Sequencing quality' module—Sequencing data are compared to the provided reference genome to check for minimal coverage.

'Global DNA Mod analysis' module—Parameters describing the genome-wide distribution of DNA modifications, such as the Modification Ratio or modification-associated motifs percentages, are computed using the provided reference genome (Supplementary Table S2).

1. 'FDR estimation' module—Based on Zhu *et al.* (2018), two methods can be used to estimate the FDR, and choose the appropriate filter(s), using either (i) a modification-depleted negative control DNA sample or (ii) DNA modification enriched motifs. We provide an empirical test in the Supplementary Analyses to validate the use of our FDR estimation algorithm (Supplementary Fig. S2).
2. 'Filter' module—It can be used to remove specific contigs or bases identified using modules 2 to 4 as shown in Supplementary Fig. S1.

'DNA Mod annotation' module—Modified bases distribution is analyzed using the provided genomic annotations. For example, modified base proportions can be computed for any user-defined genomic features (Fig. 1B and C). Additional files, such as MNase-seq data, can be imported and compared with the annotation and the modification distribution. For local visualization of DNA modification patterns, DNAModAnnot can export compliant files to the Gviz package (Hahne and Ivanek, 2016) or other genomic browser software.

3 Application examples

To illustrate the utility and features of DNAModAnnot, we used PacBio sequencing data generated to detect 6-methyladenine (6mA) modifications in the ciliate *Paramecium tetraurelia* (see Supplementary Methods). In this organism, high amounts of 6mA were reported using thin-layer chromatography (Cummings *et al.*, 1974).

PacBio data were imported along with the *Paramecium* reference genome and insufficiently covered contigs were removed (Supplementary Fig. S1B, see Supplementary Methods). A Mod ratio of ~1.6% was estimated (~0.8% if corrected with the mean fraction) (Supplementary Table S2), which is lower than the ~2.5% 6mA percentage reported by Cummings *et al.* (1974). Approximately 81.5% of 6mA were found in AT motifs (Supplementary Table S2 and Supplementary Fig. S3A): the same motif was observed in two evolutionary distantly related ciliate species, *Tetrahymena thermophila* (Wang *et al.*, 2017) and *Oxytricha trifallax* (Beh *et al.*, 2019).

High-confidence 6mAT sites were selected for the following analyses (Supplementary Fig. S1B; see Supplementary Methods): we used FDR estimations to remove potential false positives resulting from a low methylation level or low coverage. 197,154 6mAT sites were removed using 5% FDR-associated filters on ipdRatio and score, resulting in 490,222 high-confidence 6mAT sites.

By comparing 6mAT distribution with genomic annotations, we observed a relative 6mAT enrichment in genes (data not shown). To understand where 6mAT sites were localized inside the genes,

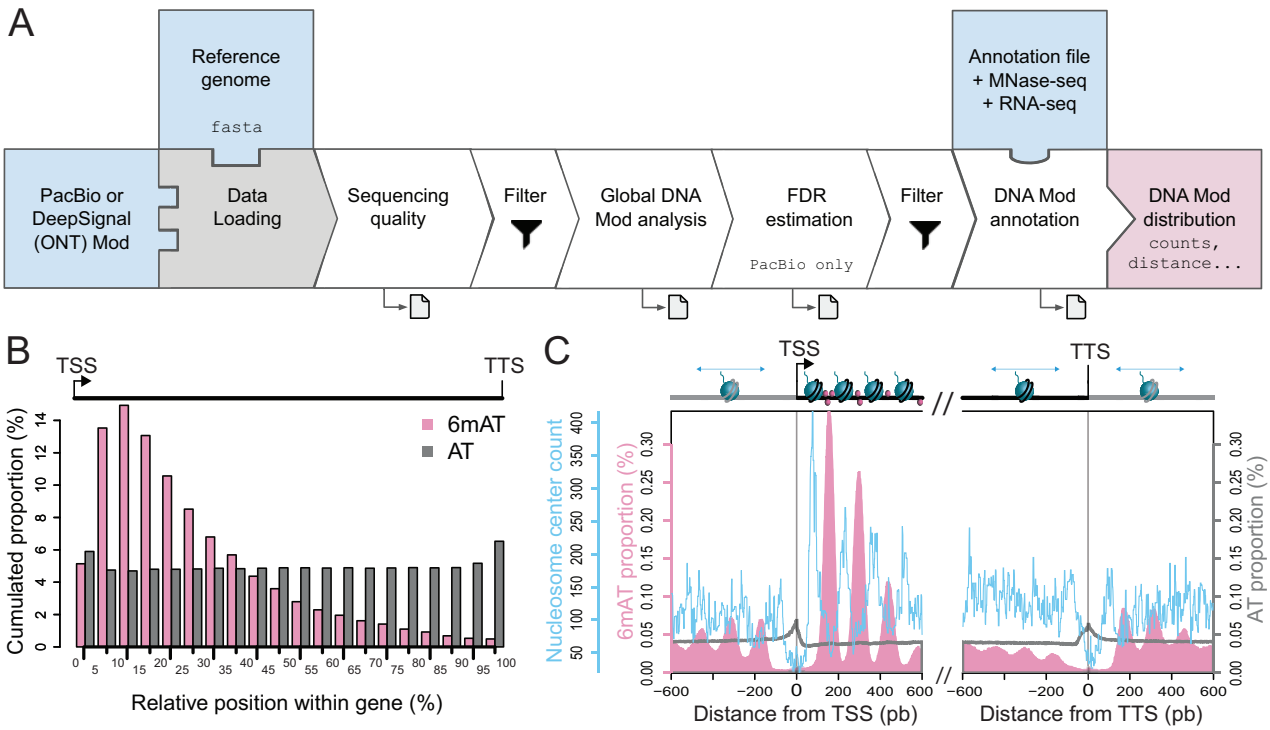

Fig. 1. DNAModAnnot toolbox. (**A**) Pipeline processing example using PacBio or ONT data for DNA Modification (Mod) analysis in a provided genome. Input modules are displayed in blue, output modules in red. Required modules are in in gray and optional in white (the full module description is provided in the Supplementary Description). (**B** and **C**) Application example: 6-methyladenine (6mA) DNA Modification (Mod) analysis in the *P. tetraurelia* somatic genome. (B) Cumulated 6mAT (pink bars) and AT (gray bars) proportions along averaged genes from TSSs to TTSs. 6mAT signal is enriched downstream TSS (Transcription Start Site) and progressively decreases until TTS (Transcription Termination Site). (C) Distribution of 6mAT sites (pink), AT sites (gray) and nucleosome centers (blue) in 1.2 kb windows centered on TSSs (vertical gray line, left panel) and TTSs (vertical grey line, right panel). Top: 6mAT sites are detected between well-positioned nucleosomes (blue circles without arrows) downstream TSSs

6mAT and AT proportions were computed by chunks from Transcription Start Site (TSS) to Transcription Termination Site (TTS) (Fig. 1B). A global enrichment of 6mAT signal was found downstream TSS and progressively decreasing until TTS.

To look closely at the distribution of 6mAT sites near TSSs, 6mAT and AT proportions were computed at each base position around TSSs and TTSs (Fig. 1C; see Supplementary Methods). 6mAT enrichment can be observed downstream TSS into peaks, indicating that 6mAT sites are distributed into periodic clusters downstream TSSs. In *T.thermophila* (Wang *et al.*, 2017) and *O.trifallax* (Beh *et al.*, 2019), the same 6mA pattern downstream TSSs was observed and was anticorrelated with nucleosome positioning.

To assess for a potential link between nucleosome positioning and 6mAT sites, MNase-seq data from *P. tetraurelia* was used and central positions of nucleosomes (nucleosome centers) were counted at each base position around TSSs and TTSs (Fig. 1C). 6mAT peaks can be observed between peaks of nucleosome signal, suggesting that 6mAT sites are enriched between well-positioned nucleosomes downstream TSSs: this 6mAT-nucleosome pattern can also be easily observed on a local example (Supplementary Fig. S3C). Additional analysis also revealed a positive correlation between 6mA enrichment and gene expression (Supplementary Fig. S3B; see Supplementary Analyses). Similar 6mA patterns were observed in *T.thermophila* (Wang *et al.*, 2017) and *O.trifallax* (Beh *et al.*, 2019), thus demonstrating DNAModAnnot's capacity for efficient genome-wide analysis.

To expand the use of our package, we tested another DNA modification and another long-read sequencing technology. We used human ONT data (Jain *et al.*, 2018) preprocessed with the DeepSignal software to detect 5-methylcytosine in CpGs (5mCpG) and analyze their distribution in the genome (see Supplementary Analyses). Analysis of 5mCpG patterns allowed us to highlight the high proportion of CpG islands at TSSs that are unmethylated (Deaton and Bird, 2011) (Supplementary Fig. S4). This result thus shows that DNAModAnnot is also adapted for ONT data and 5mC DNA modification.

4 Conclusion

DNAModAnnot is a modular package capable of analyzing patterns of different DNA modifications using different long-read sequencing technologies. We demonstrated that our package can be used to study 6-methyladenine or 5-methylcytosine using PacBio or ONT input data.

Acknowledgements

We thank Olivier Arnaiz for the preprocessing of the MNase-seq data and Hélène Pérée for the preliminary analysis of the PacBio data. The MNase sequencing benefited from the facilities and expertise of the high throughput sequencing core facility of I2BC (Centre de Recherche de Gif—http://www.i2bc.paris-saclay.fr/).

Funding

This work has been supported by intramural funding from the CNRS, the Fondation de la Recherche Médicale (Equipe FRM DEQ20160334868), the Agence Nationale de la Recherche (ANR-18-CE12-0005-03; ANR-19-CE12-0015-01) to S.D. This work was performed in collaboration with the GeT core facility, Toulouse, France (http://get.genotoul.fr), and was supported by France Génomique National infrastructure, funded as part of 'Investissement d'avenir' program managed by Agence Nationale pour la Recherche (contract ANR-10-INBS-09) and by the GET-PACBIO program ('Programme operationnel FEDER-FSE MIDI-PYRENEES ET GARONNE 2014-2020'). A.T. was recipient of PhD fellowships from 'Ministère de l'Enseignement Supérieur et de la Recherche' and 'Fondation ARC'.

Conflict of Interest: none declared.

References

Amarasinghe,S.L. *et al.* (2020) Opportunities and challenges in long-read sequencing data analysis. *Genome Biol.*, **21**, 30.

Beh,L.Y. *et al.* (2019) Identification of a DNA N6-adenine methyltransferase complex and its impact on chromatin organization. *Cell*, **177**, 1781–1796.e25.

Cummings,D.J. *et al.* (1974) Methylated bases in DNA from *Paramecium aurelia*. *Biochim. Biophys. Acta*, **374**, 1–11.

Deaton,A.M. and Bird,A. (2011) CpG islands and the regulation of transcription. *Genes Dev.*, **25**, 1010–1022.

De Coster,W. *et al.* (2020) Methplotlib: analysis of modified nucleotides from nanopore sequencing. *Bioinformatics*, **36**, 3236–3238.

Gouil,Q. and Keniry,A. (2019) Latest techniques to study DNA methylation. *Essays Biochem.*, **63**, 639–648.

Hahne,F, and Ivanek,R. (2016) Visualizing genomic data using Gviz and bioconductor. In: Mathé,E. and Davis,S. (eds.), *Statistical Genomics: Methods and Protocols, Methods in Molecular Biology*. Springer, New York, NY, pp. 335–351.

Jain,M. *et al.* (2018) Nanopore sequencing and assembly of a human genome with ultra-long reads. *Nat. Biotechnol.*, **36**, 338–345.

O'Brown,Z.K. *et al.* (2019) Sources of artifact in measurements of 6mA and 4mC abundance in eukaryotic genomic DNA. *BMC Genomics*, **20**, 445.

Peters,T.J. *et al.* (2015) De novo identification of differentially methylated regions in the human genome. *Epigenet. Chromatin*, **8**, 6.

Wang,Y. *et al.* (2017) N6-adenine DNA methylation is associated with the linker DNA of H2A.Z-containing well-positioned nucleosomes in Pol II-transcribed genes in Tetrahymena. *Nucleic Acids Res.*, **45**, 11594–11606.

Zhu,S. *et al.* (2018) Mapping and characterizing N6-methyladenine in eukaryotic genomes using single-molecule real-time sequencing. *Genome Res.*, **28**, 1067–1078.