Nanopore methylation detection in the non-model organism *Verticillium dahliae*

By

Doğukan Bayraktar

Supervisor: Michael F Seidl

A Master's Thesis
Bioinformatics and Biocomplexity
Utrecht University

29/11/2023

Contents

List of Figures

List of Tables

List of Supplementary Figures

1 Layman's summary

DNA is a set of instructions that tells an organism how to build and operate. For example, the leaf of a plant does not look the same as the root. As such, DNA is not interpreted in the same way by every cell of an organism. Cells can interpret DNA in different ways through DNA methylation. DNA methylation is a modification of the DNA where molecules are attached to it. These modifications do not change the information encoded in the DNA, but they control how the information is used.

To study DNA methylation, we must first be able to detect it. The standard for detection is a chemical process called bisulfite conversion. In this process, a chemical is used to physically convert parts of the DNA that are not affected by these modifications, revealing their locations. Another approach to detect DNA methylation is Nanopore sequencing. In this method, DNA is pulled through a tiny pore and every time a part of the DNA is pulled through, a signal is recorded. To understand this signal, we train machine learning models. Most researchers are interested in methylation in humans and for that purpose, models are trained to detect methylation. However, it has not been tested how well these models apply to fungi, where the modifications on the DNA look very different.

We investigated how appliable these models are for detecting methylation in fungi by analysing the *Verticillium dahliae*. This fungus is a plant-pathogen that poses a threat to over 400 different plant species. It lives in the soil and infects plants through their roots, leading to significant crop losses. We found that these models are comparable to bisulfite conversion at the global level, but not at the local level in fungi. This implies that current models are not applicable to fungi.

2 Abstract

The DNA of an organism encodes the blueprint necessary for all cellular functions. However, the expression of this code varies in each cell. This variation is, in part, attributed to DNA methylation. Which is a modification of the DNA where methyl groups are added on top of it. These modifications influence the expression and regulation of the genetic code. DNA methylation has been studied in a limited number of fungi, usually through bisulfite sequencing. Nanopore sequencing is a third-generation sequencing method that can detect methylation through the use of deep learning models. Thus far, it has not been studied how well these models apply to fungi. In this study, we investigated Nanopore methylation detection in *Verticillium dahliae* by comparing it to bisulfite sequencing. *Verticillium dahliae* is a plant-pathogen that infects over 400 plant species and causes significant crop loses. Both bisulfite and Nanopore tools are capable of detecting 5mC methylation. Analysis at the global level shows agreement between different Nanopore tools and bisulfite data, with Nanopore demonstrating higher levels of methylation over longer stretches of repeating areas in the genome. Nanopore basecalls from different tools have similar quality but were not shown to agree with each other or with bisulfite data on the exact locations of methylated cytosines. An analysis of a human Nanopore sample showed a significant improvement in local methylated cytosine overlap between Nanopore tools. Overall, our study demonstrates that various Nanopore tools and bisulfite data align on the location of methylation when observed from a global perspective, with Nanopore tools having the advantage. However, discrepancies arise among Nanopore tools themselves and when compared to bisulfite data at the level of specific cytosines. Subsequent analysis of Nanopore tools with human data suggests that Nanopore models may not be sufficiently generalized for precise methylation position detection in fungi. As such, Nanopore currently cannot be used to analyse exact methylated cytosine locations in fungi. Revealing a need for tools designed

specifically for methylation detection in fungi.

3 Introduction

3.1 DNA Methylation is important for genome regulation and functioning

The DNA of an organism encodes the blueprint necessary for all cellular functions. Although the genetic code across all cells is the same, its interpretation varies from one cell to another and from tissue to tissue. This variation in expression can be attributed, in part, to DNA methylation. At a fundamental level, DNA methylation can be considered "tags" that effectively "sit" atop specific regions in the genome. The addition or removal allows or obstructs the binding of enzymes to these regions. Therefore, cells are capable of differential expression of the genetic information encoded in this locus. By studying this mechanism, we can gain an understanding of DNA regulation and its role in various biological processes, including disease susceptibility, development, and evolution.

Research over the last decade has shown that DNA methylation does not randomly occur, but is associated with specific genomic regions such as promoters, gene bodies, transposable elements (TEs), and centromeres. Methylation on promoters and gene bodies play a role in gene regulation [Jjingo et al., 2012] [Jones and Takai, 2001]. Gene promoters are specific DNA sequences located near the beginning of a gene; they provide a binding site for RNA polymerase and other regulatory proteins, enabling the transcription of genes. Methyl groups on DNA can prevent the binding of transcription factors and RNA polymerases. The methylation of promoters is often associated with gene silencing [Jjingo et al., 2012]. The gene body refers to the coding region of a gene, from the transcription start site to the stop site. In contrast to promotor methylation, methylation on gene bodies has been observed on actively transcribed genes [Jones, 2012]. It has been suggested that gene body methylation may have an impact on splicing [Laurent et al., 2010]. Methylation on genes also plays a role in genomic imprinting, where specific parental alleles can be methylated to silence them. Contrary to promoters and gene bodies, methylation on centromeres and TEs play a role in genome integrity [Deniz et al., 2019] [Scelfo and Fachinetti, 2019]. Centromeres, are repeat rich regions important for chromosomal stability; they play roles in chromosome segregation and cell division [Moarefi and Chédin, 2011]. Methylation of centromeres has been suggested to contribute to genome stability, and controlling the binding of centromeric proteins [Moarefi and Chédin, 2011] [Scelfo and Fachinetti, 2019]. The last region, TEs, are mobile genetic elements that can duplicate and change their position throughout a genome. Methylation of TEs contributes to genome integrity by preventing them from uncontrollably duplicating throughout the genome [Deniz et al., 2019] [Jones and Takai, 2001].

DNA methylation consists of a spectrum of chemical modifications that can occur on different bases and in various contexts. DNA Modifications can be added to both cytosine and adenine molecules in the DNA, but typically occur on cytosine bases in eukaryotes. The most well-known type of modification is 5-methylcytosine (5mC) Other known types of methylation include 5-hydroxymethylcytosine (5hmC), 5-formylcytosine (5fc), 5-caboxylcytosine (5caC) and N6-methyladenine (6mA) (Fig. 1). The function of 5mC has been well researched and plays roles in processes such as genomic imprinting, and tissue-specific gene expression [Moen et al., 2014]. The roles of the other modifications, such as 6mA, 5fC and 5caC, are less clear and represent an active area of research [Xie et al., 2023] [Neri et al., 2015].

While various forms of cytosine methylation exist, a distinction can also be made within the

context of a DNA strand. These contexts are symmetric CG (also referred to as CpG), symmetric CHG, and asymmetric CHH where H stands for A, C, or T [Law and Jacobsen, 2010]. In these contexts, Symmetric and asymmetric refer to the presence or absence of a cytosine on the opposite strand at the site of this motif. Symmetric contexts are stable over cell division as a cytosine is present on the replicated strand, allowing for replication of the methylation, while in an asymmetric context there would be no cytosine present to methylate [Ming et al., 2021]. With the understanding that methylation occurs in varied genomic regions and comes in diverse forms, we might wonder if certain methylation patterns are unique to specific organisms.

3.2 Genomic methylation varies greatly between eukaryotes

In mammals, methylation predominantly occurs in inter- and intra- genic regions, while occurring less often on promoters [Suzuki and Bird, 2008] [Jin et al., 2011]. Mammals are particularly well studied in the context of DNA methylation and their genomes exhibit high levels of methylation, largely restricted to CG sites, where 60% to 80% are methylated [Hernando-Herraez et al., 2015]. These sites are not uniform across the genome and are often found on "CpG islands" which are regions with a high frequency of CG dinucleotides [Han et al., 2008]. To give an example, the genome of the house mouse (*Mus musculus*) has a methylated cytosine composition of 74.2% CG, 0.30% CHG, and 0.12% CHH [Feng et al., 2010]. Looking beyond mammals, however, the extent of genome methylation varies in the animal kingdom, as honey bees are known to only have 1% of their genome methylated [Feng et al., 2010]. In contrast to animals, plants do not only have methylation present on the gene body, but also a significant amount on transposons (Fig. 2) [Zemach and Zilberman, 2010] [Zemach et al., 2010]. The extent of genome methylation in plants varies from species to species. The model organism thale cress (*Arabidopsis thaliana*), a mustard plant, has a methylated cytosine composition of 22.3% CG, 5.92% CHG, and 1.51% CHH, whilst the black cottonwood (*Populus trichocarpa*), a broadleaf tree, has a composition of 41.9% CG, 20.9% CHG, and 3.25% CHH [Feng et al., 2010]. Taking a look at fungi, it has been observed that methylation occurs predominantly on highly repeating areas of the genome such as centromeres and on transposable elements [Nai et al., 2021]. They lack any gene body methylation, and the levels of methylation across fungi are very low and occur in all cytosine contexts [Bewick et al., 2019] [He et al., 2020]. Why do animals, plants, and fungi exhibit such different patterns of methylation, and why do methylation levels vary so much even within their kingdoms?

Methylation has been mainly studied in a few model animals and plants, and the general trends reported above have been derived from there. Importantly, however, the organisms that are studied are not a representative set. For example, many model organisms such as the larvacean (*Oikopleura dioicam*), roundworm (*Caenorhabditis elegans*), fruit fly (*Drosophila melanogaster*), and bakers yeast (*Saccharomyces cerevisiae*) lack known DNA methyltransferases (Fig. 2) [Zemach and Zilberman, 2010] [Suzuki and Bird, 2008]. Lastly, many important species have not been studied at all. For example, the fungal kingdom contains millions of species whose DNA methylation has never been studied [Wu et al., 2019]. Even though our understanding is limited by this, there must be drivers behind such variation.

3.3 DNMT1 and DNMT3 methylate cytosines in animals and plants, while DIM2 methylates cytosines in fungi

DNA methylation on cytosine residues is catalysed by DNA methyltransferases, which transfer methyl groups to bases in the DNA. DNMT1 and DNMT3 are families of eukaryotic DNA methyltransferases that existed before the divergence of plants and animals (Fig. 2) [Goll and Bestor, 2005]. The DNMT1 family are maintenance methyltransferases, that methy-

late the strand of asymmetrically methylated CG sites (hemimethylated) after DNA replication [Željko M. Svedružić, 2011]. This process is achieved in partnership with the protein UHRF1, which contains a domain that has been shown to prefer binding to hemimethylated CG sites [Bostick et al., 2007]. DNMT3 enzymes perform *de novo* methylation by methylating previously unmethylated sequences in plants and animals (Fig. 2) [Zemach and Zilberman, 2010]. In animals, the DNMT3 family methylates CG sites, while in land plants DNMT3s can methylate cytosine in all contexts [Goll and Bestor, 2005].

In contrast to animals, cytosine methylation in fungi is performed by the CMT/DIM2 family, which is present in both fungi and plants [Kouzminova, 2001] [Zemach and Zilberman, 2010]. CMT/DIM2s perform both maintenance and *de novo* methylation and are homologues to DNMT1s [Kouzminova, 2001] [Goll and Bestor, 2005]. Most knowledge about DNA methylation in fungi is derived from a single species, the *Neurospora crassa* [Honda et al., 2016] [Adhvaryu et al., 2015] [Gessaman and Selker, 2017].

DIM2 methylates by forming a complex with HP1, which is a family of proteins that are associated with chromatins [Lewis et al., 2010]. HP1 acts both in heterochromatic regions associated with centromeres and in shorter methylated regions within euchromatin [Freitag et al., 2004]. It recognizes sites marked by H3K9me3 and guides DIM2 to these sites, enabling methylation [Lewis et al., 2010]. H3K9me3 is a modification of histone proteins where three methyl groups are added to the lysine 9 residue of histone H3, these modifications are placed by DIM5 [Tamaru and Selker, 2001] [Lewis et al., 2010].

In *Cryptococcus neoformans* DNMT5 (not to be confused with DIM5) is required for the maintenance of methylation [Catania et al., 2020]. However, DNMT5 has not been shown to methylate in the *Verticillium dahliae* [Kramer et al., 2021]. There is also a protein belonging to the DNMT4 class, called RID. It plays an important role in the operation of the RIP mechanism, which is a defence mechanism against TEs [Lewis et al., 2008] [Cambareri et al., 1989]. In this process, RID induces mutations in TEs and repetitive regions, thereby preventing duplication and preserving genome integrity [Lewis et al., 2008] [Cambareri et al., 1989]. Importantly, RID has not been demonstrated to engage in DNA methylation activity [Cambareri et al., 1989] [Lewis et al., 2008]. Given the diversity of methyltransferases across eukaryotes and the desire to study them, what techniques are most effective in detecting and characterizing DNA methylation types and patterns?

3.4 Bisulfite conversion and Oxford Nanopore sequencing are modern methylation detection methods that come with downsides

To identify DNA methylation, different techniques have been developed with the capability to distinguish between the presence and absence of methyl groups attached to specific individual bases within a DNA molecule.

In the past, restriction enzymes were used for methylation detection. These restriction enzymes can recognize and cut DNA at specific sequences [Lippman et al., 2005]. By digesting DNA with these enzymes and then analysing the resulting fragments, the presence of DNA methylation at specific sites could be determined [Lippman et al., 2005]. This method is limited to known recognition sites, restriction enzyme bias, and a low resolution [Oakeley, 1999]. Another older method is methylated DNA immunoprecipitation, this method employs antibodies by raising them against 5m-cytosines which results in 5m-cytosines being pulled down. This separates the methylated regions from the unmethylated ones, allowing them to be analysed [Jacinto et al., 2008]. Similarly to restriction enzyme use, this method suffers from, low reso-

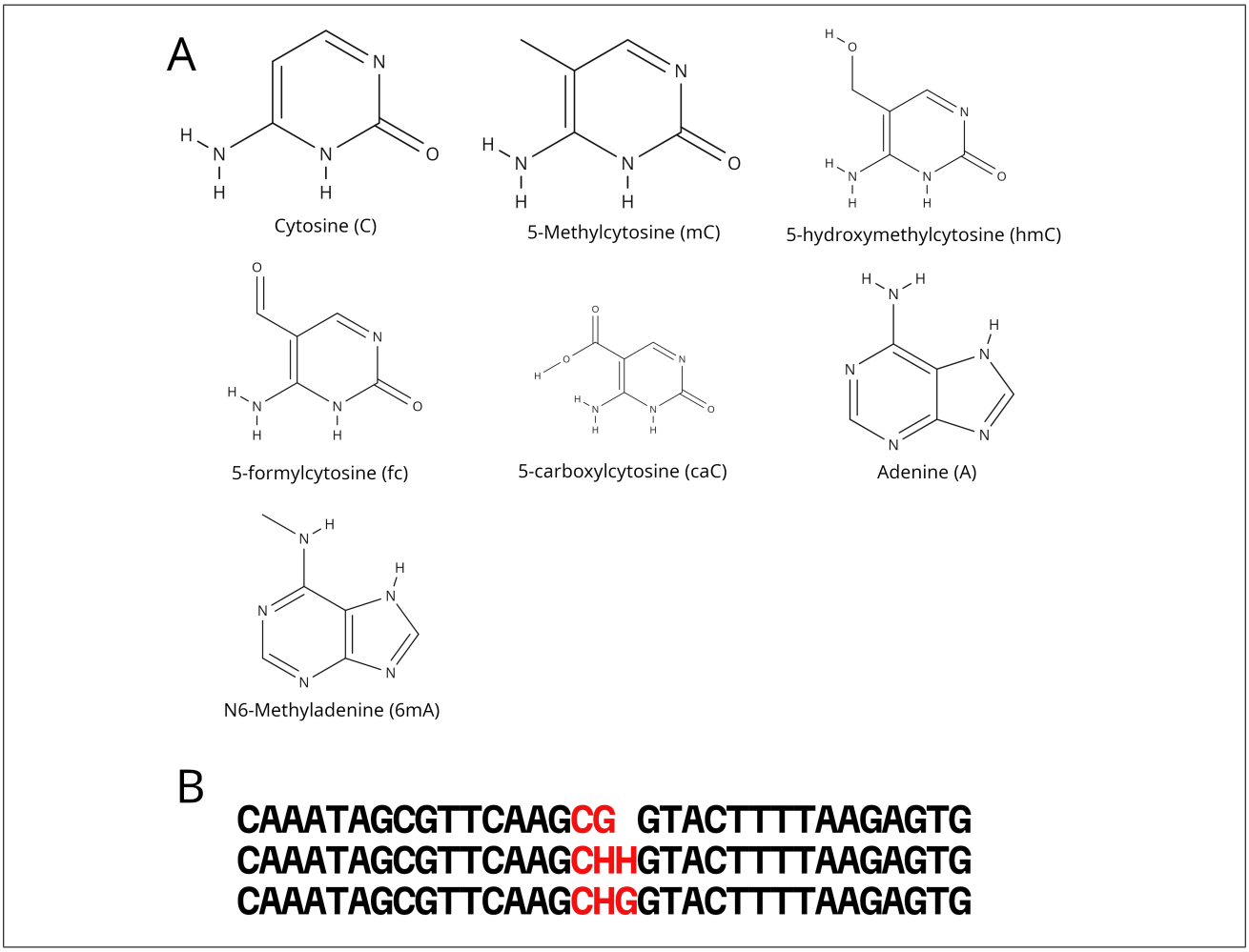


Figure 1: Overview of methylation types and contexts. **a**) Chemical structures of cytosine, adenine, and types of modifications. **b**) The three cytosine contexts highlighted in a DNA sequence.

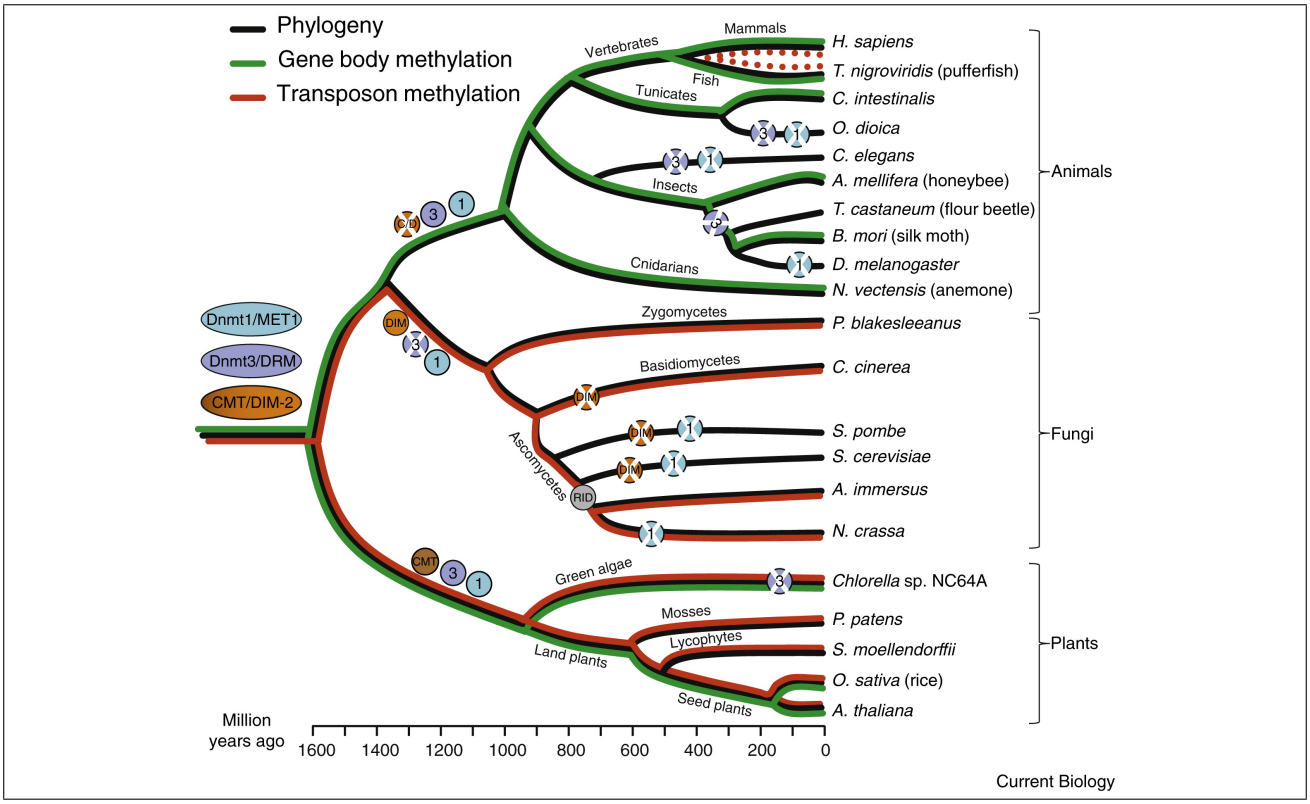


Figure 2: Phylogenetic tree showing the presence, absence, and location of methyltransferases and methylation on the genomes of several animals, fungi, and plants. Figure from [Zemach and Zilberman, 2010].

lution, bias in enrichment, and loss of information due to the loss of the non-methylated DNA [Oakeley, 1999].

Due to the advent of whole-genome sequencing, we have been able to investigate and determine the biological functions of DNA and RNA at genetic-, transcriptomic- and epigenetic- layers [Zhong et al., 2021]. An application of this in epigenetics is whole-genome bisulfite sequencing (Fig. 3). In this method, bisulfite converts cytosine residues to uracil in DNA, while leaving methylated cytosines unaffected. This results in information on single-nucleotide resolution on the 5m-cytosine methylation status of a segment of DNA in all contexts. Most of the knowledge about DNA methylation has been discovered through this method. The drawbacks, however, are that it is limited to read lengths typically shorter than 500 base pairs and can only detect 5mC methylation.

Third-generation sequencing techniques like PacBio and Oxford Nanopore sequencing enable, long-read sequencing with single-base, single-molecule sensitivity, allowing for the distinction between modified bases and unmodified ones (Fig. 5) [Rang et al., 2018] [Rhoads and Au, 2015]. PacBio uses a polymerase enzyme to add nucleotides with a fluorescent label to a DNA strand. By capturing the flashes that occur upon the addition of these nucleotides to the DNA strand, the sequence can be resolved (Fig. 4) [Rhoads and Au, 2015]. PacBio distinguishes modified bases from unmodified ones by the differences in time lengths between base incorporations, which are indicated by these flashes [Rhoads and Au, 2015].

In contrast, Nanopore can directly detect DNA or RNA by pulling nucleotides of a sequence through a pore. When a nucleotide is pulled through, a signal is produced and recorded, which can be translated into a base. The differences in these "squiggles" are used to distinguish bases from each other [Liu et al., 2021]. Relating this electric current back to a base is not

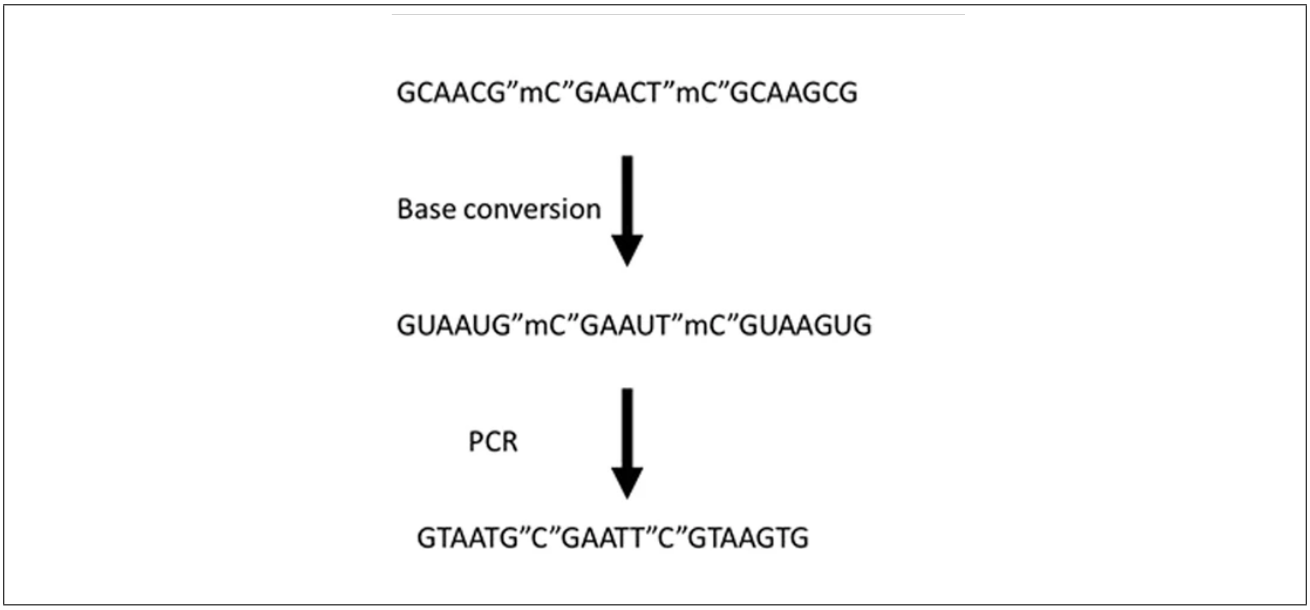


Figure 3: Schematic of the bisulfite conversion process. Unmodified cytosines in the DNA strand are converted to uracils, while leaving modified cytosines the same. After PCR, the uracils are now thymines in the copied strand, regular cytosines in the copied strand now indicate methylated ones. Figure adapted from [Xu and Seki, 2019].

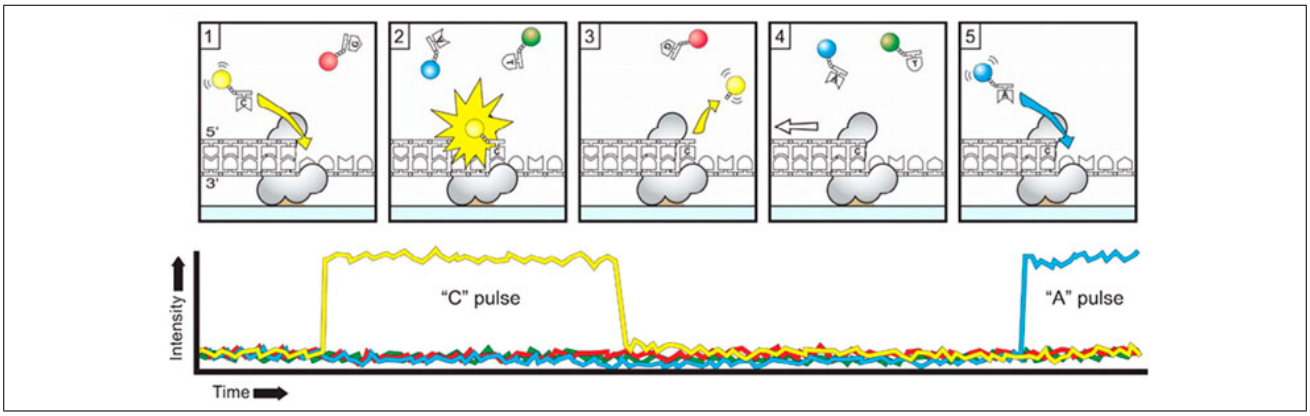


Figure 4: Schematic of the PacBio sequencing process. The addition of “tagged” nucleotides to the DNA strand causes a flash of light. Through the duration of pauses between these flashes of light, the sequence is determined. Figure adapted from [Rhoads and Au, 2015].

trivial, as there is more signal than there are bases in the sequence that is pulled through the pore. Additionally, we do not know how many bases are present in the pore at a given time. There are currently two pore types, R9 and R10. The R10 type is a more recent iteration that features two sites for recording base signal, rather than the R9 pore type's single site. Complex Neural Networks are employed to translate squiggles to sequence, but this process of basecalling is error-prone. It has been shown that Nanopore sequencers are capable of discriminating between 5mC, 5hmC, 6mA, 5caC, and 5fC [Wescoe et al., 2014]. However, the ability to detect and discriminate between these types of methylation are completely depended on the basecaller and the neural network model used to decode the signal. Models capable of detecting certain types of methylation or contexts are often not available for each pore type, or available at all. Additionally, models are typically tailored for human methylation analysis, potentially causing incompatibility in organisms that exhibit different methylation patterns. Thus, it is currently unknown to which extent these models can be applied to non-model systems such as fungi.

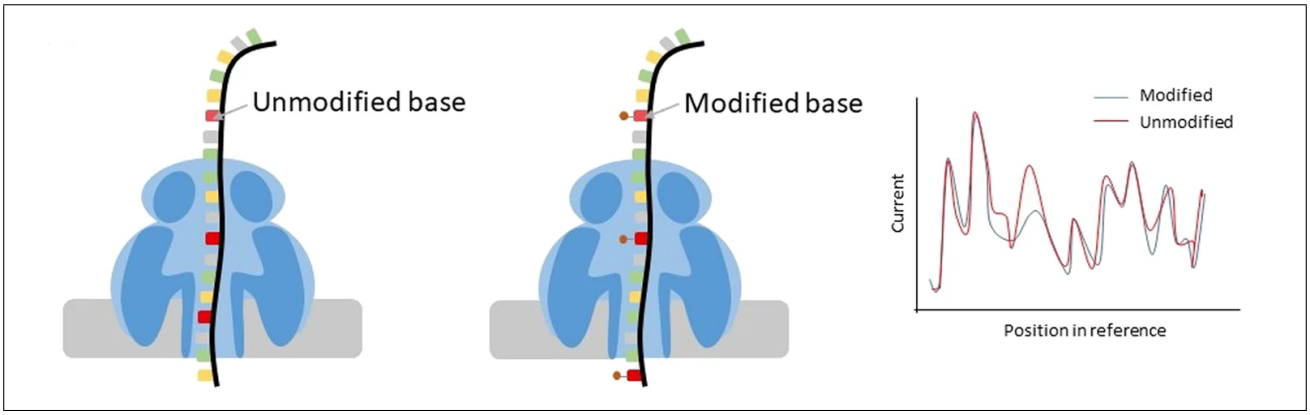


Figure 5: Two strands of DNA with modified and unmodified bases being pulled through two Nanopores, producing squiggles. Squiggles from modified bases can be distinguished from squiggles from unmodified bases. Figure adapted from [Xu and Seki, 2019].

3.5 Aim of this project

Verticillium dahliae is a fungal plant pathogen that lives in the soil, and it is responsible for causing Verticillium wilt in plants [Berlanger and Powelson, 2000]. It is presumed to be asexual and targets a broad spectrum of over 400 plant species [Usami et al., 2009] [Song et al., 2020] [Berlanger and Powelson, 2000]. Our understanding of its genome is that it is characterized by a scarcity of methylation, TE-rich regions such as centromeres, and lineage-specific regions that contain TEs marked by H3K9me, which can be found across different *Verticillium dahliae* strains [Cook et al., 2020] [Seidl et al., 2020] [Kramer et al., 2021]. A complex of DIM2 and HP1 are responsible for all methylation on its genome. This methylation predominantly occurs in highly repetitive genomic regions and TEs. Our understanding of its methylation originates from gene knockout experiments and the analysis of bisulfite data, providing valuable insights [Kramer et al., 2021]. We are going to test Nanopore methylation detection in *Verticillium dahliae* with the goal of assessing its applicability in fungi. We will achieve this by 1) establishing a baseline with bisulfite data from previous studies, 2) detecting methylation with several Nanopore tools and comparing them against each other, and 3) comparing the Nanopore and bisulfite results.

4 Results

4.1 *De novo* genome assembly of *V. dahliae* JR2 as a reference for DNA methylation studies

The current *Verticillium dahliae* reference genome assembly was created using PacBio technology eight years ago [Faino et al., 2015]. To be able to test Nanopore methylation detection, we generated Nanopore data of the *Verticillium dahliae* JR2 strain with the R9 pore type. Even though our new Nanopore data originates from the same JR2 strain, there is a potential for differences between the two datasets due to technological advancements and inherent variations in biological samples. Such disparities could introduce unexpected variability into our analyses. Thus, we created a new assembly of the *Verticillium dahliae* JR2 genome using Oxford Nanopore R9 reads.

The raw data was basecalled with Guppy [Oxford Nanopore PLC., 2023b], the Nanopore data consists of 301,479 reads with 2,605,587,933 bases, with a median read quality of 14.6 and a N50 of 14,368 bp. After internal quality checks and corrections by Canu [Koren et al., 2017],

2,720,69 reads consisting of 2,588,578,378 bases with genome-wide coverage of 51.77 reads, were used for the assembly. The *de novo* genome assembly was assembled with Canu using an estimated genome size of 50mb, which generated 21 contigs with a total length of 36,637,128 bp (Fig S1). Of the 21 contigs, 11 contigs smaller than 50kb were discarded. Contig 15 was attached to 16 and contig 14 attached to contig 3, placing 100 N bases between them, resulting in 8 chromosomes. The contigs were combined based on manual curation with Tapestry and dot plots, which revealed that these contigs aligned and that each contig contained a single telomere (Fig. 6a) [Davey et al., 2020]. Chromosomes were reverse complemented where necessary and renamed according to the reference genome assembly. The new assembly was corrected with short Illumina reads of the same JR2 strain, using six iterations of Pilon [Walker et al., 2014] to compensate for the higher error rates found in Nanopore reads. In total, 19,703 changes were made to the genome, with the last iteration only making 6 changes. The corrections range from single base insertions to insertions of entire reads. A separate copy of the same assembly was corrected with Nanopolish [Loman et al., 2015] using the same R9 Nanopore reads. Nanopolish made corrections per chromosome, with a total of 19,727 changes. The identities of the assemblies were determined with Mummer [Marçais et al., 2018], the identity of the non-corrected assembly to the public reference genome was 99.8989%, the identity of the Nanopolish corrected assembly was 99.9301%, and the identity of the Pilon corrected assembly was 99.9345%. This indicates the high similarity of all the assemblies to the reference, with the Pilon corrected one having a slight edge. The TE annotations and gene annotations of the publicly available reference genome assembly were transferred to the new assembly (non-corrected, Nanopolish corrected, and Pilon corrected) using Liftoff [Shumate and Salzberg, 2021] and polished after. In the Pilon corrected assembly, 1 gene "VDAG_JR2_Chr3g00010a" failed to transfer, while no genes failed to transfer in the other assemblies. To check for changes in the genes, protein sequences of the original gene annotation were compared with the protein sequences of the transferred genes. This resulted in 8,284 identical proteins and 3,120 changed proteins in the non-corrected assembly, 8,099 identical proteins and 3,305 changed proteins in the Nanopolish correction, and 11,117 identical proteins and 286 changes, with 1 not being transferred in the Pilon correction. To give an example, of the 286 changed sequences, 81 have a different length and 205 had at least one amino acid change. Thus, we concluded that the Pilon corrected assembly was the best, and we decided to move forward with the Pilon corrected assembly for use in the downstream analysis (Fig. 6b).

4.2 DIM2 and HP1 knockouts lead to reduced methylation in repetitive sequences and TEs

As a framework for the assessment of DNA methylation patterns in *V. dahliae* JR2 we first utilized a publicly available bisulfite dataset of the *V. dahliae* JR2 strain [Kramer et al., 2021], serving as a positive control. The dataset comprises six samples, including the wild type, as well as knockouts for RID1, HP1, DNMT5, DIM2, and a double knockout DIM2DNMT5 [Kramer et al., 2021]. The six bisulfite samples were all processed twice, once with Bismark [Krueger and Andrews, 2011] and once with BSmapz [Greg Zynda, 2020]. The number of aligned reads for all samples hovered around 14 million, which was reduced to 12 million after deduplication. The median coverage for all samples ranged from 31 to 34. The per site methylation of the wild type was plotted as a histogram, revealing an unimodal distribution, where most of the cytosines have no methylation (Fig. 7). Because the fraction modified is so low, we cannot perform a bimodal test to determine the methylation status of an individual cytosine, which is the standard in organisms with higher methylation levels [Schultz et al., 2012]. Instead, we considered cytosines to be methylated if it has minimum coverage of four and a minimum of three reads that indicate methylation.

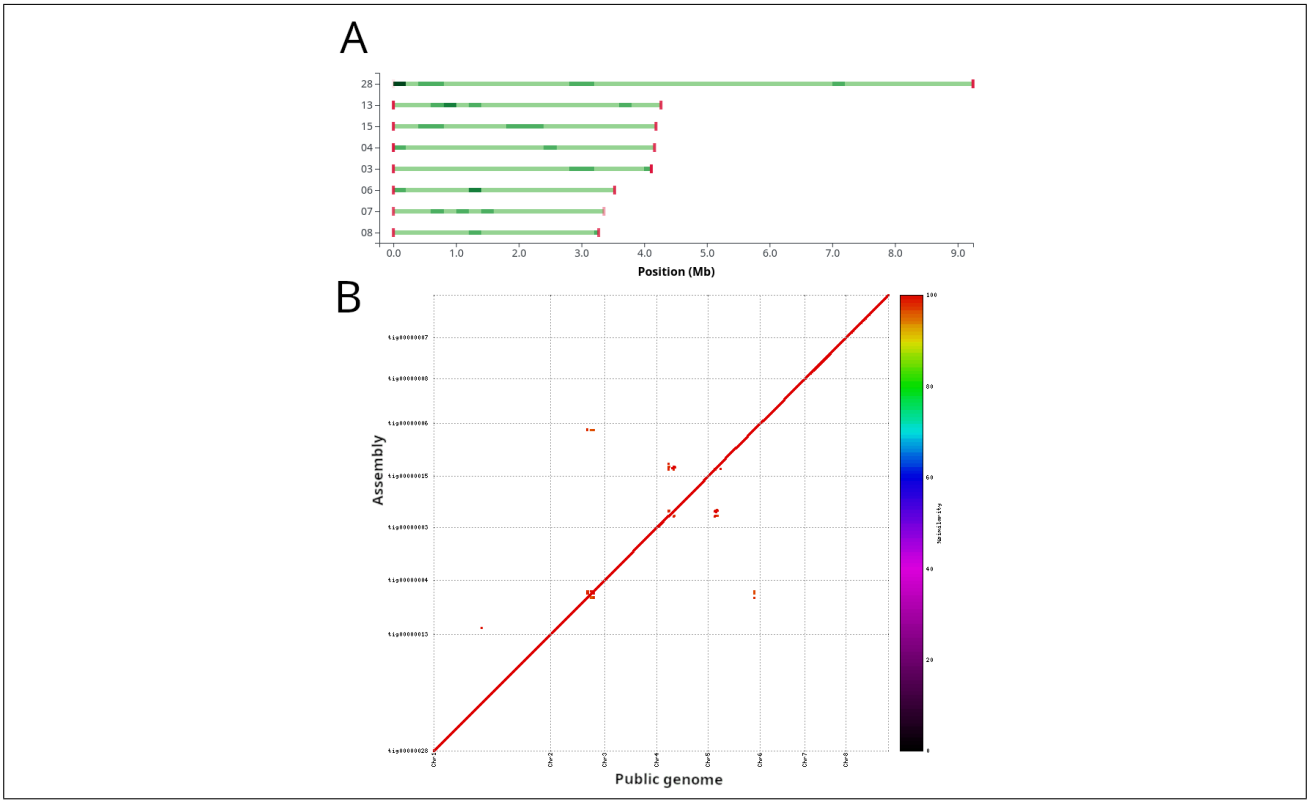


Figure 6: *De novo* genome assembly of the *Verticillium dahliae* genome with Nanopore R9 data. **a**) Contigs shown in Tapestry after combining and correcting with six iterations of Pilon. **b**) Dot plot of the final *de novo* assembly aligned against the reference genome assembly [Faino et al., 2015], indicating high similarity.

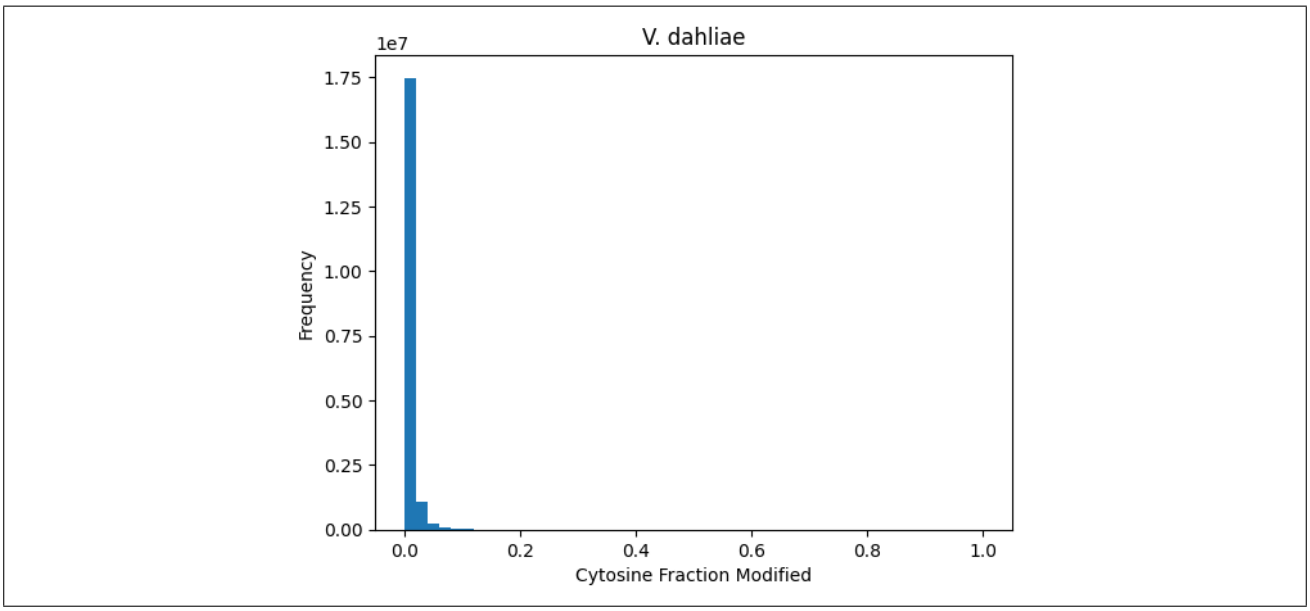


Figure 7: Per site methylation distribution of the *Verticillium dahliae*. The distribution is unimodal and visibly ranges from 0.0% to 0.2%, indicating low levels of methylation.

The wild type sample with Bismark detected 18,758,324 cytosines (10,099,428 CHH, 4,904,882 CG, 3,754,014 CHG) of which, 11,860 (4,978 CHH, 4,843 CG, 2,039 CHG) are methylated. The wild type sample with BSmapz detected 19,307,970 cytosines (10,413,535 CHH, 5,034,482 CG, 3,859,953 CHG) of which, 9,307 (3,224 CHH, 4,321 CG, 1,762 CHG) are methylated. The total fraction of methylated cytosines is 0.06% in Bismark and 0.05% in BSmapz. The methylated fraction per total context is similar across the two tools, with the most considerable difference in CHH context (0.05% vs 0.03%). Overall, the tools called comparable numbers of cytosines.

The weighted methylation score was calculated for each sample by dividing the genome into windows of 10kb, and subsequently the score was calculated by dividing the sum of methylated reads in the window through the total number of reads in that window. Only reads that were covering a cytosine in the given window were considered in the calculation. Hypo-methylated windows were determined by comparing the wild type windows to the windows of the other samples, respectively. The weighted methylation scores were then plotted over the chromosomes alongside the TE and gene annotations (Fig. 8). The overall weighted methylation scores across all samples range from 0 to 0.03, these low scores are as expected in fungi [Bewick et al., 2019] [He et al., 2020] [Kramer et al., 2021]. As expected, we observed that the wild type shows peaks at areas of the chromosome where TEs and repeats are located [Kramer et al., 2021]. The score over areas where genes are located is zero or close to zero. The DNMT5 and RID1 knockouts follow the same methylation pattern as the wild type, indicating that there was no significant loss of methylation in these mutants [Kramer et al., 2021]. In contrast, the DIM2, DIM2DNMT5, and HP1 knockout samples show a significant loss of methylation over TE and repeat areas of the chromosome, as indicated by the hypo-methylated windows (Fig. 8), confirming that DIM2 and HP1 are essential for methylation. Next, taking a look at the total number of methylated cytosines per sample, we can see that the DIM2, DIM2DNMT5, and HP1 knockouts show a decreased number of methylated cytosines which are in line with the results from the weighted methylation scores (Fig. 9a). Lastly, we used Deeptools [Ramírez et al., 2016] to collapse the annotated genome features into a length of 3000 bp and plotted the per site methylation score over this area for each sample. The per site methylation score goes as high as 0.8% over the TE and repeat features in the wild type, DNMT5, and RID samples. While the scores over the genes do not go higher than 0.3% in these same samples (Fig. 9b). In the DIM2, DIM2DNMT5, and HP1 knockouts the per site methylation over TE features was reduced to the same level as the genes at 0.3% (Fig. 9b). Thus, further reinforcing that DIM2 and HP1 are essential for methylation.

4.3 Dorado and Guppy basecalls are highly similar

There are many basecallers, developed both by ONT and third parties, such as Megalodon, Bonito, Causacall, and Halcyon. We selected two basecallers called Dorado [Oxford Nanopore PLC., 2023a] and Guppy because these basecallers are directly developed and recommended by Oxford Nanopore Technologies. They can perform both basecalling and modified basecalling having various models available for 5mC methylation detection and other types of methylation. Additionally, DeepSignalPlant, which is a methylation calling tool trained for plant methylation detection [Ni et al., 2021], recommends Guppy for its basecalling. This tool will be used in the downstream analysis.

To ensure a fair comparison when evaluating Nanopore methylation calls from different tools, we benchmarked the basecallers with Nanopore_benchmark [Pagès-Gallego and de Ridder, 2023]. This benchmarking method considers the reference genome as the truth, evaluating how effectively the basecallers generate reads that align to the genome. We used the *V. dahliae*

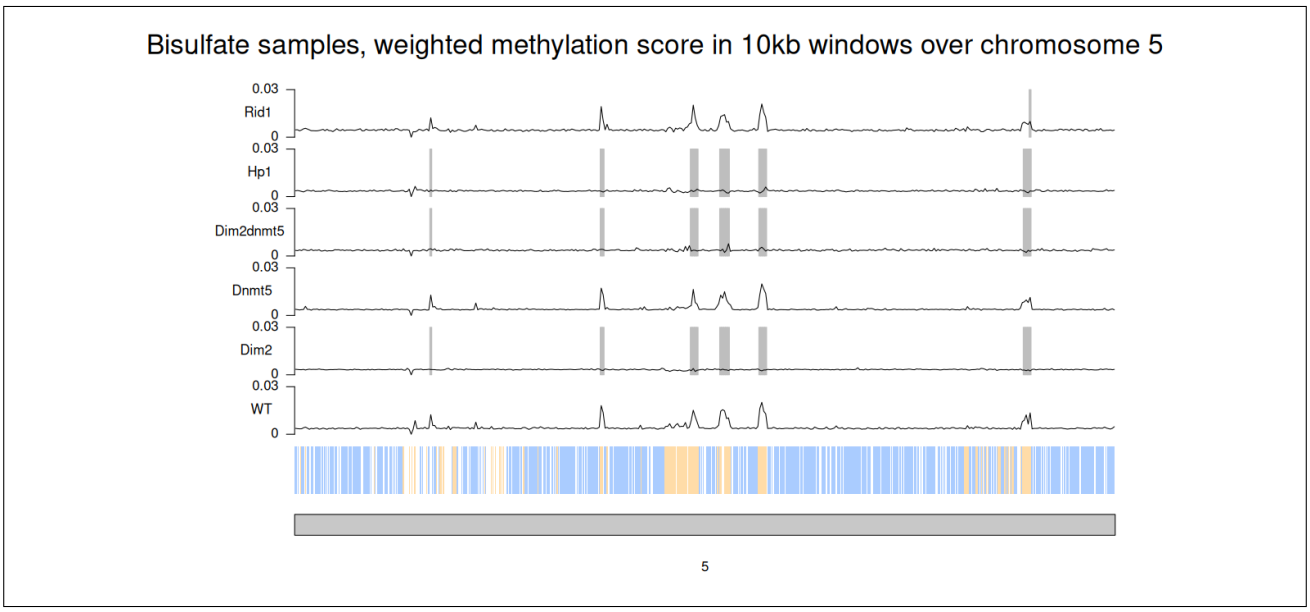


Figure 8: Global view of methylation across chromosome five of the wild type and knockouts of the bisulfite data. The weighted methylation score is shown in 10kb windows over chromosome 5. Grey squares highlight hypo-methylated windows. At the bottom, genes are represented in blue, while transposable elements and repeats are highlighted in yellow.

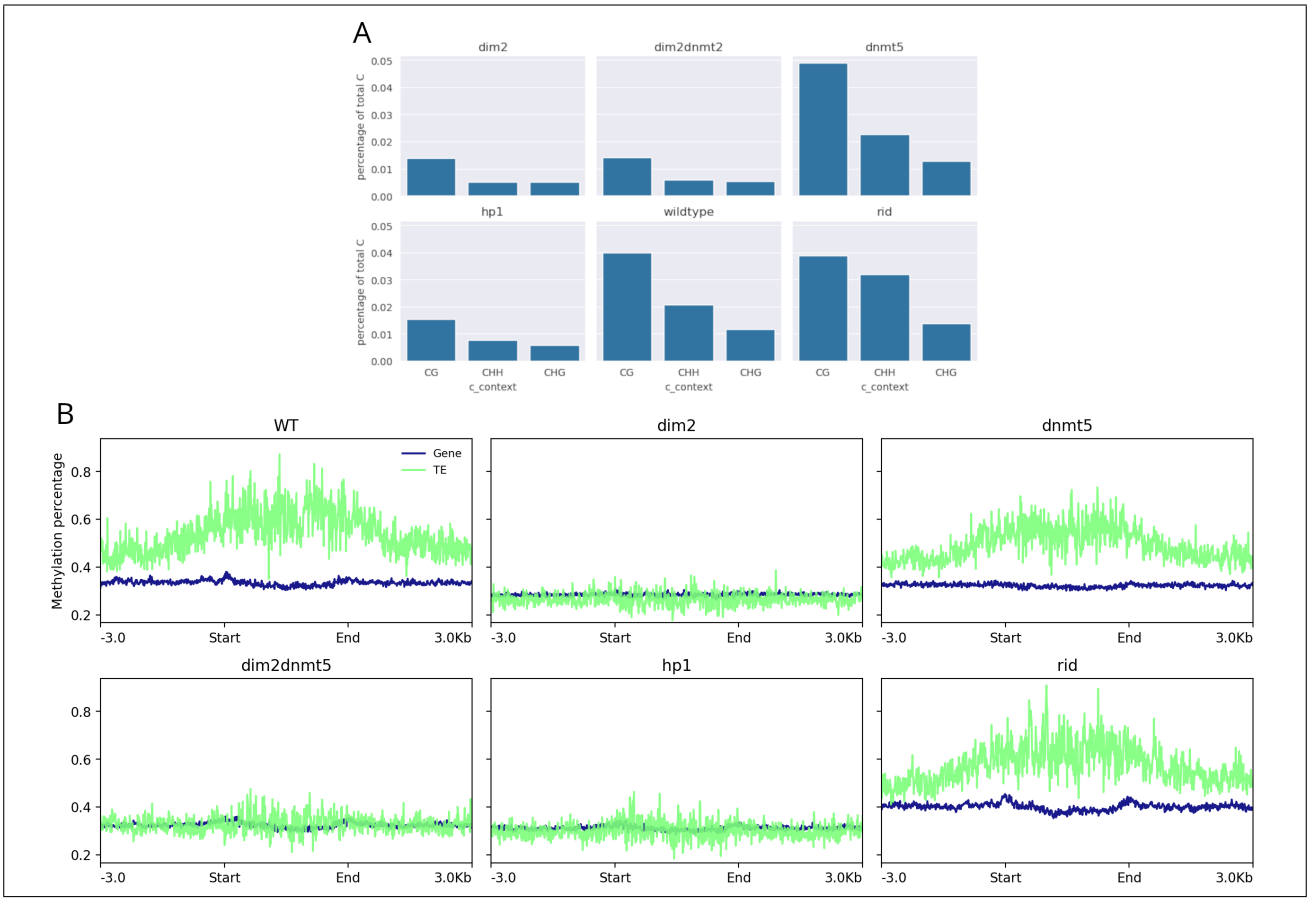


Figure 9: Counts and percent methylation shown across the different bisulfite samples. **a**) Percentage of cytosines methylated on chromosome 5 per mutant and cytosine context. **b**) Methylation percentage plotted over genome features with Deeptools.

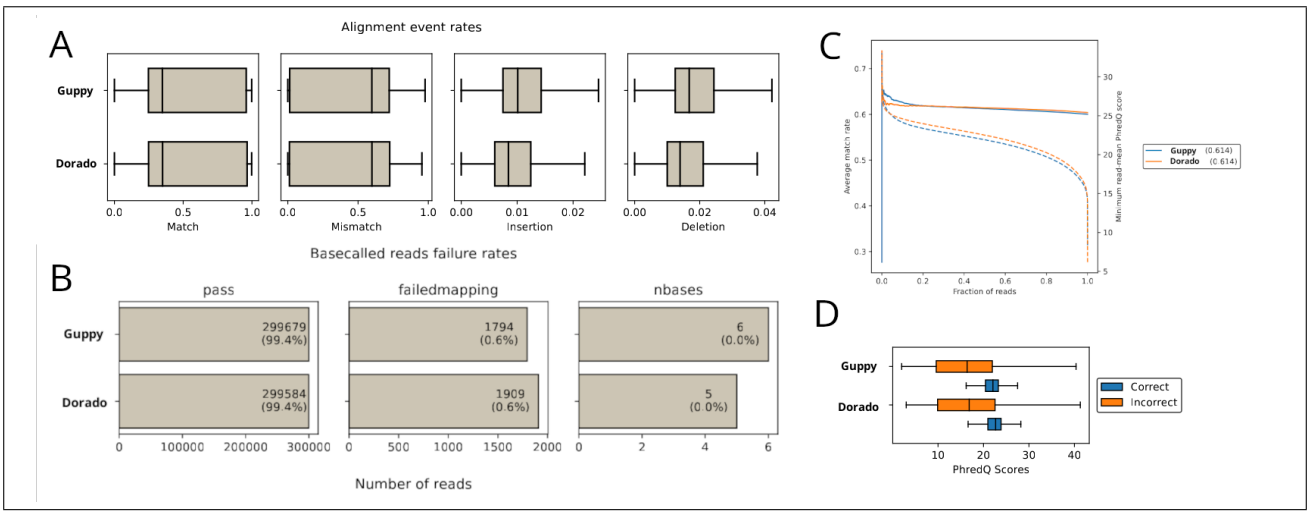


Figure 10: Benchmarking results of Dorado & Guppy basecallers. **a**) Comparison of alignment event rates. **b**) Passed reads, alignment failure rates and number of bases aligned to N bases in the genome. **c**) AUC scores of the match rate, sorted by average read Q-score **d**) Q-score of correctly and incorrectly predicted bases.

Nanopore R9 reads for this benchmark. The median coverage resulting from both tools is 33. The alignment event rates for both basecallers show highly similar results, with box plot medians indicating a rate of 0.4. The median mismatch rates are between 0.5 and 0.6. The insertion median rates are at 0.01 while the deletions are between 0.01 and 0.02. In Guppy, 2,996,79 reads passed the quality check, while 2,995,84 reads passed in Dorado. In Dorado, 200 more reads failed to map than the Guppy reads. The AUC scores of the match rate are the same, at 0.614. The mean Q-scores for incorrectly predicted bases in both tools ranges from 1 to 40 with the median at 16, while the Q-score for the correctly predicted bases ranges from 16 to 28 with a mean of 23. Overall, the basecalls by Guppy and Dorado demonstrate a high degree of similarity (Fig. 10).

4.4 Nanopore captures methylation patterns identified through bisulfite treatment on a global level, displaying higher levels of methylation.

To evaluate methylation calls in all contexts and a model trained for plant methylation, we also performed methylation calling with DeepSignalPlant. DeepSignalPlant is a tool for methylation calling that uses a combination of raw ONT signal and basecalls from Guppy to determine methylation status in all contexts [Ni et al., 2021]. The tool was trained for methylation calling in plants, where methylation mostly occurs in TEs, which is more similar to methylation patterns in fungi [Ni et al., 2021] [Zemach et al., 2010].

Before comparing methylated cytosines, the exact positions of basecalled cytosines were compared, resulting in an overlap of 5,054,454 cytosines with 1,100 positions being unique to Dorado and 1,276 positions unique to Guppy. Comparing positions from Dorado and Guppy to the CG positions of DeepSignalPlant resulted in overlaps of 5,055,421 (combined remainder: 2134) and 5,055,665 (combined remainder: 1822) respectively. Having established that Dorado and Guppy basecalls are similar, we examined their methylation calling capabilities. Guppy detected 5,055,730 cytosines (after filtering) of which 31,452 are methylated, Dorado detected 5,055,554 cytosines (after filtering) of which 34,888 are methylated. As the R9 models for Guppy and Dorado are only capable of detecting 5mCG, all cytosines are in CG context. DeepSignalPlant detected 19,497,234 cytosines (10,562,997 CHH, 5,057,422, CG, 3,876,815 CHG) of

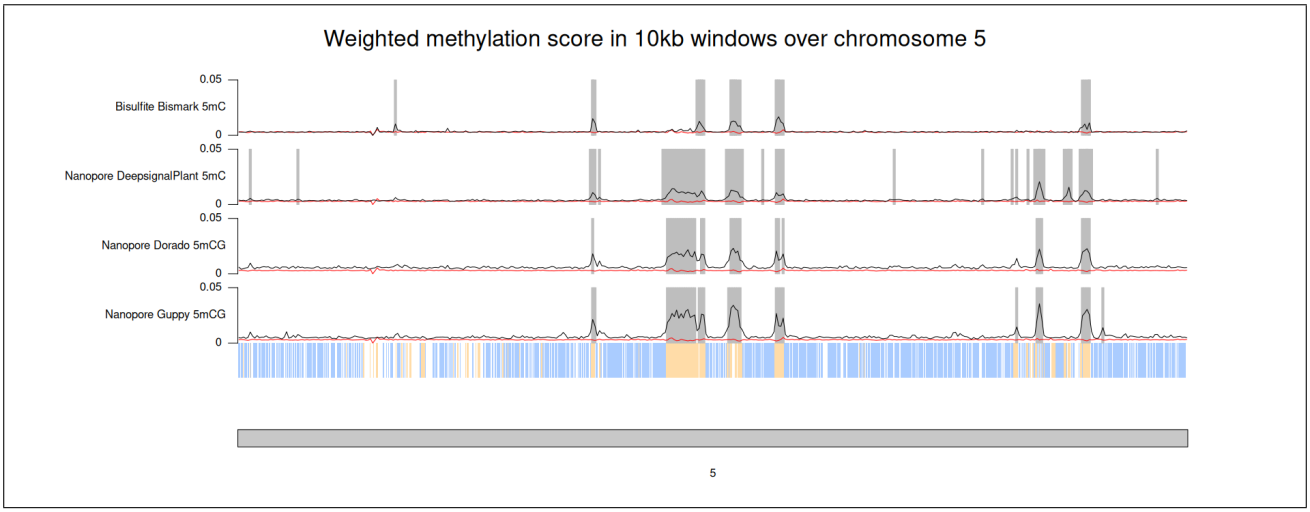


Figure 11: Global methylation levels for bisulfite data and multiple Nanopore methylation callers. The red lines depict the baseline noise level established by the bisulfite knockout of HP1. At the bottom, genes are represented in blue, while transposable elements and repeats are highlighted in yellow. Grey boxes highlight hyper-methylated windows.

which 150,729 (44,463 CHH, 94,436 CG, 11,830 CHG) are methylated.

To compare the Nanopore tools to each other and to the bisulfite data on a global level, the genomes were split into windows of 10kb and weighted methylation scores were calculated for all tools. In this instance, the hyper-methylated windows were calculated instead of hypo-methylated ones. Which was done by comparing the scores of the windows from the bisulfite HP1 knockout to all other windows. On the global level, the three Nanopore tools agree with each other, the weighted methylation scores peak over the same areas where transposable elements are located (Fig. 11). The height of the peaks differ between the tools, with the highest score being 0.03 in Dorado, 0.05 in Guppy, and 0.024 in DeepSignalPlant. DeepSignalPlant shows that a couple of windows at the beginning and end of the chromosome are hyper-methylated, but these windows are not shown as such in Dorado or Guppy.

Having compared the Nanopore tools, we now make a comparison to the bisulfite data to see if Nanopore sequencing detects similar methylation patterns and levels. The three Nanopore tools highlight all the peaks present in the bisulfite track. Additionally, the Nanopore tools show peaks at the end of the chromosome not present in the bisulfite track, over an area where genes and TEs occur alternately (Fig. 11). Nanopore peaks are higher and not interrupted over highly repeating areas, as opposed to the bisulfite track, which has a peak score of 0.022. These trends occurred over all eight chromosomes of the genome (Fig. S2-S9). This suggests that Nanopore is capable of detecting methylation identified by bisulfite conversion and that it might have an advantage over it, thanks to its longer reads.

4.5 Methylation calls disagree on exact methylated cytosine locations

Since we globally observed that Nanopore can detect the same patterns as bisulfite conversion, possibly having an advantage. We now sought to establish a similar pattern on the level of exact cytosines. Zooming in on individual cytosines called by all tools, the overlap of methylated cytosines between Dorado and Guppy is 16.4%. The overlap between Dorado and DeepsignalPlant 9.9% (CG only), and the overlap between Guppy and DeepSignalPlant is 10.4% (CG only) (Fig. 12b). The overlap between DeepSignalPlant and the bisulfite wild type resulted in

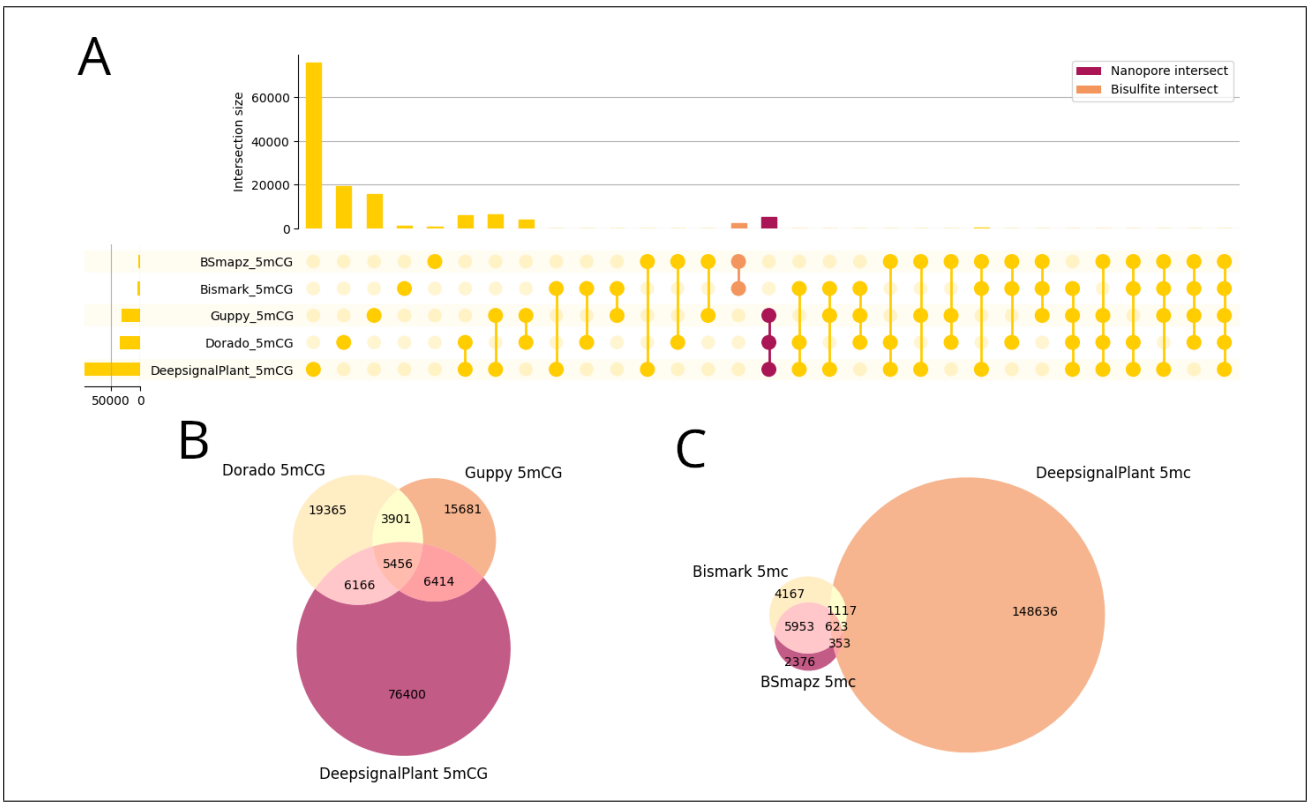


Figure 12: Overlap of methylated cytosine positions across different tools. **a**) Upset plot of all tools in 5mCG context. **b**) Venn diagram of Dorado, Guppy, and DeepSignalPlant in 5mCG context. **c**) Venn diagram of Bismark, BSmapz, and DeepSignalPlant in 5mC all contexts.

Table 1: Scores of methylated cytosine location overlap between tools

	Precision	Recall	F1	AUC_ROC
Bismark vs DeepSignalPlant	0.0162	0.290	0.0307	0.644
DeepsignalPlant 5mCG vs Dorado	0.345	0.123	0.181	0.561
Guppy vs Dorado	0.277	0.304	0.290	0.651

a 1.1% overlap with Bismark (all context), and a 0.62% overlap with BSmapz (all context) (Fig. 12c). To put a more precise number on the performance of the methylation calling, we used machine learning scoring metrics (Table 1) (Figs S10-S14). The precision, recall, and F1 scores are extremely low across the board. To ensure that there was no cut-off issue in terms of defining methylated cytosines, parameter sweeping was done, but scores virtually stayed the same. These results show that exact methylated cytosines positions disagree among Nanopore tools and when compared to the bisulfite data. Which suggests that the Nanopore tools are incapable of capturing the methylation indicated by bisulfite data at the exact cytosine level.

4.6 Guppy and Dorado have better local overlap in human data

On the exact methylated cytosine level, we saw that the overlap between Dorado and Guppy was very low at 16.4% and inconsistent with the global methylation view. To narrow down possible causes of this discrepancy, we basecalled a low coverage public human R9 Nanopore dataset with Dorado and Guppy. The human data set has 747,475,084 bases of which 709,920,532 aligned, a total of 126,897 reads, N50 of 8,184, and a median Q-score of 14. The coverage over the genome ranged from 1 to 32 with a mean of 1. After filtering, Dorado called 23,324 cytosines of which, 12,127 are methylated and Guppy called 20,197 cytosines with 9,884 methylated cytosines. The

per site methylation was plotted, which shows a bimodal distribution. Methylated cytosines are either fully methylated or not methylated (Fig. 13), there is a small peak at 50% most likely due to heterozygous sites. The overlap of Dorado and Guppy for the human dataset is 56.1% compared to the 16.4% overlap in the fungal data (Fig. 14). This suggests that the discrepancy between methylated cytosine overlap in the fungal data is caused by the models not being generalised enough for methylation calling in fungi. As, the human is a model organism with high levels of methylation.

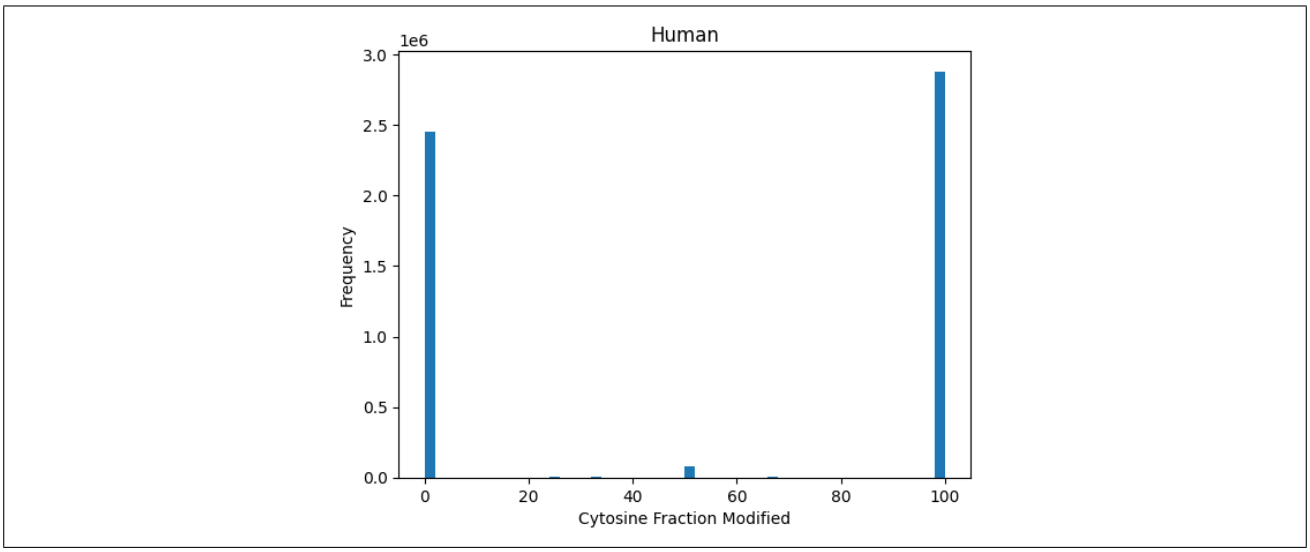


Figure 13: Per site methylation distribution of the human. Peaks can be seen at 0% and 100%, which indicates that positions are often completely methylated or not methylated at all. The small peak at 50% possibly originates from heterozygous sites.

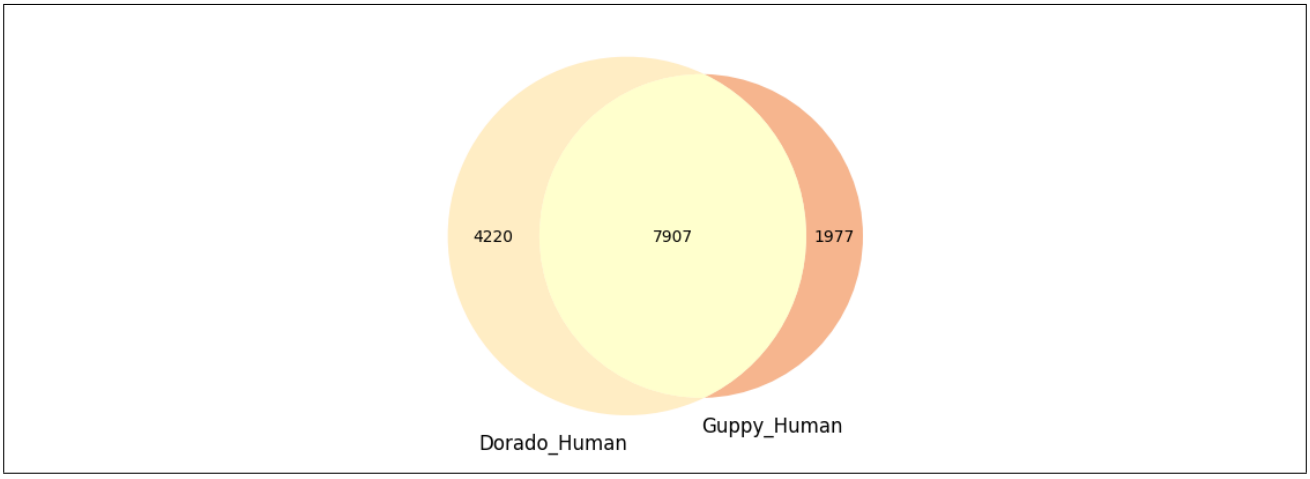


Figure 14: Methylated cytosine overlap of Guppy and Dorado with human data. Of the total, 14104 cytosines indicated as methylated between Dorado and Guppy, 56% are overlapping. Dorado has more methylated cytosines of which 34% are not overlapping.

5 Discussion

We tested Nanopore methylation detection in *Verticillium dahliae* with the goal of assessing its applicability in fungi. To achieve this we completed three objectives 1) establishing a baseline with bisulfite data from previous studies, 2) detecting methylation with several Nanopore

tools and comparing them against each other, and 3) comparing the Nanopore and bisulfite results.

Using the bisulfite data, we set a baseline, allowing us to evaluate the Nanopore methylation calls. Our results showed low levels of methylation in the *V. dahliae*, which peaked at TE and repeat-rich regions in the wild type. Using the gene knockouts, we reproduced results from [Kramer et al., 2021] that showed that an absence of DIM2 and HP1 leads to reduced levels of methylation over these areas. This demonstrates that DIM2 and HP1 are essential for methylation in *V. dahliae*, while also showing that the biological signal occurs over TEs and repeats. The DNMT5 and RID1 knockouts mirrored the wild type methylation pattern, indicating their limited impact on overall methylation. These findings were found on the global level, the per site methylation level, and at the total number of methylated cytosines which shows that the findings are consistent across different metrics. Additionally, the HP1 knockout established a baseline noise level. These characterizations of the methylation pattern in *Verticillium dahliae* set an overall expectation, which allowed us to test and compare the Nanopore data to it.

Nanopore basecalling is a quickly evolving field, with many basecalling/methylation calling tools developed by first and third parties alike. By selecting multiple Nanopore basecallers, we were able to compare performance between Nanopore tools, giving us a better idea of what to expect when comparing Nanopore to bisulfite data. The comparison of Dorado and Guppy revealed a high degree of similarity in basecalling performance. There was a substantial overlap of exact cytosine positions between Dorado and Guppy, and the number of cytosines and methylated cytosines were all comparable. Unexpectedly, the methylation calling results at the local level had little to no overlap. Checking the overlapping positions between Dorado, Guppy, and DeepSignalPlant revealed that not all overlapping positions necessarily occurred over TEs and repeats. The unique calls for each tool also partially overlapped TE and repeat positions. Meaning that the unique methylation calls in each tool cannot be disregarded as noise, and overlapping calls cannot be assumed to be the truth.

We were only able to detect methylation in CG context with Dorado and Guppy, highlighting a disadvantage of the R9 pore type. In the case of DeepSignalPlant, there were more methylated cytosines, but this did not result in an increased overlap. Similarly, to the Nanopore tools, little overlap was found when comparing to the local view of the bisulfite tools.

The small (16.4%) overlap between exact methylated cytosine positions in Dorado and Guppy, despite basecalling similarities, prompted us to test the tools with human data. The human sample showed higher overlap at 56%, while you would expect a higher overlap of such similar basecallers, it was still significantly higher than the 16.4% in *V. dahliae*. This suggests that the models used in Nanopore methylation calling are not generalized well enough for application in fungi. Perhaps if a bigger human sample was used with more coverage, the overlap could have been higher in the human test. For example, in the study from [Liu et al., 2021], they performed a benchmark of many Nanopore methylation callers in human where they showed that Guppy has a high Pearson Correlation on the per site methylation status with both bisulfite data (0.901) and other Nanopore methylation callers such as DeepSignal (not to be confused with DeepSignalPlant) (0.938) and Megalodon (0.947).

The global methylation analysis, by use of weighted methylation score, demonstrated overall agreement among the three Nanopore tools. They exhibited peak methylation scores over TEs rather than genes. Still, small differences in peak heights were observed, with DeepSignalPlant having the lowest and Guppy the highest scores. The peaks in the Nanopore tools aligned with the peaks in the bisulfite tools. However, the peaks in the Nanopore tools displayed heightened methylation levels, often extending over entire TEs where bisulfite did not. This increase can be attributed to the advantage of long reads. Aligners often face challenges when attempting

to align short reads in densely repeating areas, resulting in sparser methylation levels. Starting this project, we had assumed that the bisulfite data would be our positive control. However, as highly repeating areas are precisely the regions where methylation is expected to occur in fungi, assuming that the bisulfite data is the ground truth here would be incorrect. Which means that bisulfite data is not the ground truth and a comparison of bisulfite and Nanopore at exact cytosine positions may lead to many false negatives.

By adding DeepSignalPlant to the comparison, we hoped to gain insight into the differences between more specialized methylation callers and generalized methylation callers such as Dorado and Guppy. However, no such advantage was apparent, even when considering that DSP calls cytosines in all contexts. This is not conclusive evidence, however. While plant methylation is more similar to fungi than mammals, the level of methylation in fungi is still significantly lower than in plants, possibly affecting DSP's performance in *V. dahliae*.

Nanopore technology has demonstrated its capability to distinguish various types of methylation, a distinguishing feature from bisulfite conversion. Although the potential for discriminating between different methylation types has been demonstrated in general, our specific study faced limitations due to the absence of applicable models for the older Nanopore variants utilized in our analysis.

While bisulfite conversion cannot detect 5hmC, limiting our knowledge of the presence of this modification, our use of ONT enabled the detection of its presence. Specifically, the 5hmCG_5mCG model exhibited significantly higher scores across the entire chromosome (Fig. S15). Upon isolating the 5mCG and 5hmCG tracks from the 5hmCG_5mCG model, it became evident that the majority of the signal originated from 5hmCG modifications (Fig. S15). Nevertheless, the inability of bisulfite to detect 5hmC prevented the comparison of these results.

Nanopore sequencing has recently been used in different studies in fungi such as the studies by [Li et al., 2023], [Dvorianinova et al., 2021], [Sperschneider et al., 2021], and [Gusa et al., 2023]. They used Nanopore for methylation analysis in *Beauveria bassiana*, *Fusarium oxysporum*, *Puccinia graminis*, and *Cryptococcus deneoformans* respectively. These studies did not report benchmarks of the methylation calling tools they used. Their findings, on a global level, are similar to ours, with low methylation levels that occur mainly over centromeres and TEs. In [Li et al., 2023] basecalling was done with Guppy and methylation analysis was done with DeepSignalPlant. The other studies used Guppy and Nanopolish. We did not test Nanopolish which remains a popular tool among researchers for methylation calling. It uses a Hidden Markov Model that is limited to only detecting 5mCG methylation in groups [Simpson et al., 2017]. Its inability to call CG sites individually was the reason for us to not include it in the tested Nanopore tools. As Nanopolish calls CG sites in groups, no fair comparison of individual cytosines can be made to other tools. This raises the question if the tools used in these studies are accurate enough for use in fungi, as our results suggest that individual methylated cytosine calling with Nanopore in fungi is not accurate enough.

6 Conclusion

Our results demonstrate that various Nanopore tools and bisulfite data align when observed from a global perspective. However, discrepancies arise among Nanopore tools themselves and when compared to bisulfite data at the level of specific cytosines. In the global context, longer reads from Nanopore exhibit a distinct advantage over the short reads from bisulfite. Testing Dorado and Guppy with human data revealed a significantly higher overlap of methylated cytosines, suggesting that Nanopore models may not be sufficiently generalized for precise

methylation position detection in fungi. Revealing a need for tools designed specifically for methylation detection in fungi.

7 Methods

7.1 *De novo* genome assembly

Nanopore Fast5 files were basecalled with Guppy v6.5.7 using the dna_r9.4.1_450bps_hac model. Resulting files were concatenated and passed to Canu v2.2. From the resulting contigs, all contigs under 50kb were discarded and four contigs were combined into two separate chromosomes, resulting in eight chromosomes total. The assembled genome was corrected with six iterations of Pilon v1.24 using short Illumina reads of the JR2 strain. Annotations were transferred to the new *de novo* genome using Liftoff v1.6.3. Analysis of the protein sequences from transferred genes was done by changing nucleotide sequences to protein sequences using Gffread v0.12.7. Subsequently, the protein sequences were compared to each other using the "compare_proteins.py" script.

7.2 Bisulfite analysis

The genome was converted using Bismark v0.24.1, then the paired reads were aligned using Bowtie2 v2.5.1. After alignment, the resulting BAM files were deduplicated and the methylation was extracted from the BAM files per context. The three files were first converted to BedGraph format and then to a cytosine coverage report, resulting in the final methylation BED file.

BSmapz v1.1.3 was run for every bisulfite sample in paired read alignment mode. The resulting BAM files were deduplicated and the methylation extracted with the methratio.py script from BSmapz.

Minimum coverage of four was used to filter the reported cytosines from both tools. The weighted methylation scores were calculated in windows of 10kb using the "weighted_methylation_score.py" script with the following formula: $\sum_{i=1}^{n} C_i / \sum_{i=1}^{n} C_i + T_i$.

The calculated scores were then passed to MethylKit v1.12.0 where hypo/hyper-methylated windows were determined by use of a fast fisher test with a q value of 0.01 and then extracted to separate BED files. All the resulting data was then imported into the "Plot_scores.Rmd" script where the data was plotted over all chromosomes using karyoploteR v3.18.

Additional statistics such as the per site distribution were calculated with the "Bismark", "BSmapz" Jupyter notebooks. Lastly, Deeptools v3.5.4 was used to plot the collapsed feature plots using calculate matrix and plotProfile.

7.3 Nanopore analysis

7.3.1 Guppy

Guppy v6.5.7 was run with the dna_r9.4.1_450bps_hac model with methylation detection set to 5mCG. The BAM files that passed the internal quality check were concatenated and then passed to Modkit [Oxford Nanopore PLC., 2023c] v0.2.2. Using Modkit with the pileup –CpG –reference options, only CG cytosines that matched the reference genome were extracted.

7.3.2 Dorado

Dorado v0.4.2 was run with the dna_r9.4.1_e8_sup@v3.3 basecalling model with methylation calling set to 5mCG. The resulting BAM file was then aligned to the genome using Minimap2 v2.26. The resulting BAM file was sorted and indexed, after which it was passed to Modkit. Using Modkit only CG cytosines matching the reference genome were extracted.

7.3.3 DeepSignalPlant

The fungal Fast5 files were first converted to single Fast5 files. The single Fast5 files were then basecalled with Guppy v6.5.7 using the dna_r9.4.1_450bps_hac model. The FASTQ files that passed the quality check were concatenated together and passed to Tombo v1.5.1 alongside the Fast5 files. Tombo v1.5.1 was used to combine the basecalls from the FASTQ files with the raw signal from the Fast5 files. Then the signal in the Fast5 files were re-squiggled with Tombo v1.5.1 by referencing the genome. Modifications were called afterwards with DeepSignalPlant v0.1.6 using the "model.dp2.CNN.arabnrice2-1_120m_R9.4plus_tem.bn13_sn16.both_bilstm.epoch6.ckpt" model.

7.3.4 Downstream

The resulting BED files were all passed to "weighted_methylation_score.py" where cytosines were filtered on minimum coverage of four and the weighted methylation scores were calculated like the bisulfite data. The hyper/hypo-methylated windows were also calculated in the same manner.

The exact position overlap of regular cytosines and machine learning metrics for methylation overlap were all calculated with the "check_c_overlap.py" script. The Venn diagrams and upset plot were created with the "methylated_c_overlap.py" script. Additional statistics such as the per site distribution were calculated with the "Dorado", "Guppy", "DeepSignalPlant" Jupyter notebooks.

7.4 Automation

The steps involving Bismark, Dorado and DeepSignalPlant were automated with bash scripts (supplementary Fig. S16). The downstream analysis for calculating and plotting the weighted methylation score was automated using a Snakemake pipeline (Fig. S17).

7.5 Code availability

The scripts and pipelines developed for this project are publicly available at https://github.com/bayraktar1/Fungi_methylation_calling

References

[Adhvaryu et al., 2015] Adhvaryu, K. K., Gessaman, J. D., Honda, S., Lewis, Z. A., Grisafi, P. L., and Selker, E. U. (2015). The cullin-4 complex dcdc does not require e3 ubiquitin ligase elements to control heterochromatin in *Neurospora crassa*. *Eukaryotic Cell*, 14(1):25–28.

[Berlanger and Powelson, 2000] Berlanger, I. and Powelson, M. L. (2000). *Verticillium* wilt. *The Plant Health Instructor*.

[Bewick et al., 2019] Bewick, A. J., Hofmeister, B. T., Powers, R. A., Mondo, S. J., Grigoriev, I. V., James, T. Y., Stajich, J. E., and Schmitz, R. J. (2019). Diversity of cytosine methylation across the fungal tree of life. *Nature Ecology & Evolution*, 3(3):479–490.

[Bostick et al., 2007] Bostick, M., Kim, J. K., Esteve, P.-O., Clark, A., Pradhan, S., and Jacobsen, S. E. (2007). Uhrf1 plays a role in maintaining DNA methylation in mammalian cells. *Science*, 317(5845):1760–1764.

[Cambareri et al., 1989] Cambareri, E. B., Jensen, B. C., Schabtach, E., and Selker, E. U. (1989). Repeat-induced G-C to A-T mutations in *Neurospora*. *Science*, 244(4912):1571–1575.

[Catania et al., 2020] Catania, S., Dumesic, P. A., Pimentel, H., Nasif, A., Stoddard, C. I., Burke, J. E., Diedrich, J. K., Cooke, S., Shea, T., Gienger, E., Lintner, R., Yates, J. R., Hajkova, P., Narlikar, G. J., Cuomo, C. A., Pritchard, J. K., and Madhani, H. D. (2020). Evolutionary persistence of DNA methylation for millions of years after ancient loss of a de novo methyltransferase. *Cell*, 180(2):263–277.e20.

[Cook et al., 2020] Cook, D. E., Kramer, H. M., Torres, D. E., Seidl, M. F., and Thomma, B. P. H. J. (2020). A unique chromatin profile defines adaptive genomic regions in a fungal plant pathogen. *eLife*, 9.

[Davey et al., 2020] Davey, J. W., Davis, S. J., Mottram, J. C., and Ashton, P. D. (2020). Tapestry: validate and edit small eukaryotic genome assemblies with long reads. *bioRxiv*.

[Deniz et al., 2019] Deniz, O., Frost, J. M., and Branco, M. R. (2019). Regulation of transposable elements by DNA modifications. *Nature Reviews Genetics*, 20(7):417–431.

[Dvorianinova et al., 2021] Dvorianinova, E. M., Pushkova, E. N., Novakovskiy, R. O., Povkhova, L. V., Bolsheva, N. L., Kudryavtseva, L. P., Rozhmina, T. A., Melnikova, N. V., and Dmitriev, A. A. (2021). Nanopore and illumina genome sequencing of *Fusarium oxysporum* f. sp. lini strains of different virulence. *Frontiers in Genetics*, 12.

[Faino et al., 2015] Faino, L., Seidl, M. F., Datema, E., van den Berg, G. C. M., Janssen, A., Wittenberg, A. H. J., and Thomma, B. P. H. J. (2015). Single-molecule real-time sequencing combined with optical mapping yields completely finished fungal genome. *mBio*, 6(4).

[Feng et al., 2010] Feng, S., Cokus, S. J., Zhang, X., Chen, P.-Y., Bostick, M., Goll, M. G., Hetzel, J., Jain, J., Strauss, S. H., Halpern, M. E., Ukomadu, C., Sadler, K. C., Pradhan, S., Pellegrini, M., and Jacobsen, S. E. (2010). Conservation and divergence of methylation patterning in plants and animals. *Proceedings of the National Academy of Sciences*, 107(19):8689–8694.

[Freitag et al., 2004] Freitag, M., Hickey, P. C., Khlafallah, T. K., Read, N. D., and Selker, E. U. (2004). HP1 is essential for DNA methylation in *Neurospora*. *Molecular Cell*, 13(3):427–434.

[Gessaman and Selker, 2017] Gessaman, J. D. and Selker, E. U. (2017). Induction of h3k9me3 and DNA methylation by tethered heterochromatin factors in *Neurospora crassa*. *Proceedings of the National Academy of Sciences*, 114(45).

[Goll and Bestor, 2005] Goll, M. G. and Bestor, T. H. (2005). Eukaryotic cytosine methyltransferases. *Annual Review of Biochemistry*, 74(1):481–514.

[Greg Zynda, 2020] Greg Zynda (2020). BSMAPz v1.1.3. https://github.com/zyndagj/BSMAPz.

[Gusa et al., 2023] Gusa, A., Yadav, V., Roth, C., Williams, J. D., Shouse, E. M., Magwene, P., Heitman, J., and Jinks-Robertson, S. (2023). Genome-wide analysis of heat stress-stimulated transposon mobility in the human fungal pathogen *Cryptococcus deneoformans*. *Proceedings of the National Academy of Sciences*, 120(4).

[Han et al., 2008] Han, L., Su, B., Li, W.-H., and Zhao, Z. (2008). Cpg island density and its correlations with genomic features in mammalian genomes. *Genome Biology*, 9(5):R79.

[He et al., 2020] He, C., Zhang, Z., Li, B., and Tian, S. (2020). The pattern and function of DNA methylation in fungal plant pathogens. *Microorganisms*, 8(2):227.

[Hernando-Herraez et al., 2015] Hernando-Herraez, I., Garcia-Perez, R., Sharp, A. J., and Marques-Bonet, T. (2015). DNA methylation: Insights into human evolution. *PLOS Genetics*, 11(12).

[Honda et al., 2016] Honda, S., Bicocca, V. T., Gessaman, J. D., Rountree, M. R., Yokoyama, A., Yu, E. Y., Selker, J. M. L., and Selker, E. U. (2016). Dual chromatin recognition by the histone deacetylase complex hchc is required for proper DNA methylation in *Neurospora crassa*. *Proceedings of the National Academy of Sciences*, 113(41).

[Jacinto et al., 2008] Jacinto, F. V., Ballestar, E., and Esteller, M. (2008). Methyl-DNA immunoprecipitation (MeDIP): Hunting down the DNA methylome. *BioTechniques*, 44(1):35–43.

[Jin et al., 2011] Jin, B., Li, Y., and Robertson, K. D. (2011). DNA methylation: Superior or subordinate in the epigenetic hierarchy? *Genes & Cancer*, 2(6):607–617.

[Jjingo et al., 2012] Jjingo, D., Conley, A. B., Yi, S. V., Lunyak, V. V., and Jordan, I. K. (2012). On the presence and role of human gene-body DNA methylation. *Oncotarget*, 3(4):462–474.

[Jones, 2012] Jones, P. A. (2012). Functions of DNA methylation: islands, start sites, gene bodies and beyond. *Nature Reviews Genetics*, 13(7):484–492.

[Jones and Takai, 2001] Jones, P. A. and Takai, D. (2001). The role of DNA methylation in mammalian epigenetics. *Science*, 293(5532):1068–1070.

[Koren et al., 2017] Koren, S., Walenz, B. P., Berlin, K., Miller, J. R., Bergman, N. H., and Phillippy, A. M. (2017). Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome research*, 27(5):722–736.

[Kouzminova, 2001] Kouzminova, E. (2001). dim-2 encodes a DNA methyltransferase responsible for all known cytosine methylation in *Neurospora*. *The EMBO Journal*, 20(15):4309–4323.

[Kramer et al., 2021] Kramer, H. M., Cook, D. E., van den Berg, G. C. M., Seidl, M. F., and Thomma, B. P. H. J. (2021). Three putative DNA methyltransferases of *Verticillium dahliae* differentially contribute to DNA methylation that is dispensable for growth, development and virulence. *Epigenetics & Chromatin*, 14(1).

[Krueger and Andrews, 2011] Krueger, F. and Andrews, S. R. (2011). Bismark: a flexible aligner and methylation caller for bisulfite-seq applications. *Bioinformatics*, 27(11):1571–1572.

[Laurent et al., 2010] Laurent, L., Wong, E., Li, G., Huynh, T., Tsirigos, A., Ong, C. T., Low, H. M., Kin Sung, K. W., Rigoutsos, I., Loring, J., and Wei, C.-L. (2010). Dynamic changes in the human methylome during differentiation. *Genome Research*, 20(3):320–331.

[Law and Jacobsen, 2010] Law, J. A. and Jacobsen, S. E. (2010). Establishing, maintaining and modifying DNA methylation patterns in plants and animals. *Nature Reviews Genetics*, 11(3):204–220.

[Lewis et al., 2010] Lewis, Z. A., Adhvaryu, K. K., Honda, S., Shiver, A. L., Knip, M., Sack, R., and Selker, E. U. (2010). DNA methylation and normal chromosome behavior in *Neurospora* depend on five components of a histone methyltransferase complex, DCDC. *PLoS Genetics*, 6(11):e1001196.

[Lewis et al., 2008] Lewis, Z. A., Honda, S., Khlafallah, T. K., Jeffress, J. K., Freitag, M., Mohn, F., Schübeler, D., and Selker, E. U. (2008). Relics of repeat-induced point mutation direct heterochromatin formation in *Neurospora crassa*. *Genome Research*, 19(3):427–437.

[Li et al., 2023] Li, Y.-H., Chang, J.-C., Yen, M.-R., Huang, Y.-F., Chen, T.-H., Chen, L.-H., and Nai, Y.-S. (2023). Whole-genome DNA methylome analysis of different developmental stages of the entomopathogenic fungus *Beauveria bassiana* nchu-157 by nanopore sequencing. *Frontiers in Genetics*, 14.

[Lippman et al., 2005] Lippman, Z., Gendrel, A.-V., Colot, V., and Martienssen, R. (2005). Profiling DNA methylation patterns using genomic tiling microarrays. *Nature Methods*, 2(3):219–224.

[Liu et al., 2021] Liu, Y., Rosikiewicz, W., Pan, Z., Jillette, N., Wang, P., Taghbalout, A., Foox, J., Mason, C., Carroll, M., Cheng, A., and Li, S. (2021). Dna methylation-calling tools for oxford nanopore sequencing: a survey and human epigenome-wide evaluation. *Genome Biology*, 22(1).

[Loman et al., 2015] Loman, N. J., Quick, J., and Simpson, J. T. (2015). A complete bacterial genome assembled *de novo* using only nanopore sequencing data. *Nature Methods*, 12(8):733–735.

[Marçais et al., 2018] Marçais, G., Delcher, A. L., Phillippy, A. M., Coston, R., Salzberg, S. L., and Zimin, A. (2018). MUMmer4: A fast and versatile genome alignment system. *PLOS Computational Biology*, 14(1).

[Ming et al., 2021] Ming, X., Zhu, B., and Li, Y. (2021). Mitotic inheritance of DNA methylation: more than just copy and paste. *Journal of Genetics and Genomics*, 48(1):1–13.

[Moarefi and Chédin, 2011] Moarefi, A. H. and Chédin, F. (2011). ICF syndrome mutations cause a broad spectrum of biochemical defects in DNMT3b-mediated de novo DNA methylation. *Journal of Molecular Biology*, 409(5):758–772.

[Moen et al., 2014] Moen, E. L., Mariani, C. J., Zullow, H., Jeff-Eke, M., Litwin, E., Nikitas, J. N., and Godley, L. A. (2014). New themes in the biological functions of 5-methylcytosine and 5-hydroxymethylcytosine. *Immunological Reviews*, 263(1):36–49.

[Nai et al., 2021] Nai, Y.-S., Huang, Y.-C., Yen, M.-R., and Chen, P.-Y. (2021). Diversity of fungal DNA methyltransferases and their association with DNA methylation patterns. *Frontiers in Microbiology*, 11.

[Neri et al., 2015] Neri, F., Incarnato, D., Krepelova, A., Rapelli, S., Anselmi, F., Parlato, C., Medana, C., Dal Bello, F., and Oliviero, S. (2015). Single-base resolution analysis of 5-formyl and 5-carboxyl cytosine reveals promoter DNA methylation dynamics. *Cell Reports*, 10(5):674–683.

[Ni et al., 2021] Ni, P., Huang, N., Nie, F., Zhang, J., Zhang, Z., Wu, B., Bai, L., Liu, W., Xiao, C.-L., Luo, F., et al. (2021). Genome-wide detection of cytosine methylations in plant from nanopore data using deep learning. *Nature communications*, 12(1):5976.

[Oakeley, 1999] Oakeley, E. J. (1999). DNA methylation analysis: a review of current methodologies. *Pharmacology & Therapeutics*, 84(3):389–400.

[Oxford Nanopore PLC., 2023a] Oxford Nanopore PLC. (2023a). Dorado v0.4.2. https://github.com/nanoporetech/dorado.

[Oxford Nanopore PLC., 2023b] Oxford Nanopore PLC. (2023b). Guppy v6.5.7. https://community.nanoporetech.com/docs/prepare/library_prep_protocols/Guppy-protocol/v/gpb_2003_v1_revax_14dec2018/linux-guppy.

[Oxford Nanopore PLC., 2023c] Oxford Nanopore PLC. (2023c). Modkit v0.2.2. https://github.com/nanoporetech/modkit.

[Pagès-Gallego and de Ridder, 2023] Pagès-Gallego, M. and de Ridder, J. (2023). Comprehensive benchmark and architectural analysis of deep learning models for nanopore sequencing basecalling. *Genome Biology*, 24(1).

[Ramírez et al., 2016] Ramírez, F., Ryan, D. P., Grüning, B., Bhardwaj, V., Kilpert, F., Richter, A. S., Heyne, S., Dündar, F., and Manke, T. (2016). deepTools2: a next generation web server for deep-sequencing data analysis. *Nucleic Acids Research*, 44(W1):W160–W165.

[Rang et al., 2018] Rang, F. J., Kloosterman, W. P., and de Ridder, J. (2018). From squiggle to basepair: computational approaches for improving nanopore sequencing read accuracy. *Genome Biology*, 19(1).

[Rhoads and Au, 2015] Rhoads, A. and Au, K. F. (2015). PacBio sequencing and its applications. *Genomics, Proteomics, and Bioinformatics*, 13(5):278–289.

[Scelfo and Fachinetti, 2019] Scelfo and Fachinetti (2019). Keeping the centromere under control: A promising role for DNA methylation. *Cells*, 8(8):912.

[Schultz et al., 2012] Schultz, M. D., Schmitz, R. J., and Ecker, J. R. (2012). 'leveling' the playing field for analyses of single-base resolution DNA methylomes. *Trends in Genetics*, 28(12):583–585.

[Seidl et al., 2020] Seidl, M. F., Kramer, H. M., Cook, D. E., Fiorin, G. L., van den Berg, G. C. M., Faino, L., and Thomma, B. P. H. J. (2020). Repetitive elements contribute to the diversity and evolution of centromeres in the fungal genus *Verticillium*. *mBio*, 11(5).

[Shumate and Salzberg, 2021] Shumate, A. and Salzberg, S. L. (2021). Liftoff: accurate mapping of gene annotations. *Bioinformatics*, 37(12):1639–1643.

[Simpson et al., 2017] Simpson, J. T., Workman, R. E., Zuzarte, P. C., David, M., Dursi, L. J., and Timp, W. (2017). Detecting DNA cytosine methylation using nanopore sequencing. *Nature Methods*, 14(4):407–410.

[Song et al., 2020] Song, R., Li, J., Xie, C., Jian, W., and Yang, X. (2020). An overview of the molecular genetics of plant resistance to the *Verticillium* wilt pathogen *Verticillium dahliae*. *International Journal of Molecular Sciences*, 21(3):1120.

[Sperschneider et al., 2021] Sperschneider, J., Jones, A. W., Nasim, J., Xu, B., Jacques, S., Zhong, C., Upadhyaya, N. M., Mago, R., Hu, Y., Figueroa, M., Singh, K. B., Stone, E. A., Schwessinger, B., Wang, M.-B., Taylor, J. M., and Dodds, P. N. (2021). The stem rust fungus *Puccinia graminis* f. sp. tritici induces centromeric small rnas during late infection that are associated with genome-wide DNA methylation. *BMC Biology*, 19(1).

[Suzuki and Bird, 2008] Suzuki, M. M. and Bird, A. (2008). DNA methylation landscapes: provocative insights from epigenomics. *Nature Reviews Genetics*, 9(6):465–476.

[Tamaru and Selker, 2001] Tamaru, H. and Selker, E. U. (2001). A histone h3 methyltransferase controls DNA methylation in *Neurospora crassa*. *Nature*, 414(6861):277–283.

[Usami et al., 2009] Usami, T., Itoh, M., and Amemiya, Y. (2009). Asexual fungus *Verticillium dahliae* is potentially heterothallic. *Journal of General Plant Pathology*, 75(6):422–427.

[Walker et al., 2014] Walker, B. J., Abeel, T., Shea, T., Priest, M., Abouelliel, A., Sakthikumar, S., Cuomo, C. A., Zeng, Q., Wortman, J., Young, S. K., and Earl, A. M. (2014). Pilon: An integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS ONE*, 9(11):e112963.

[Wescoe et al., 2014] Wescoe, Z. L., Schreiber, J., and Akeson, M. (2014). Nanopores discriminate among five C5-cytosine variants in DNA. *Journal of the American Chemical Society*, 136(47):16582–16587.

[Wu et al., 2019] Wu, B., Hussain, M., Zhang, W., Stadler, M., Liu, X., and Xiang, M. (2019). Current insights into fungal species diversity and perspective on naming the environmental DNA sequences of fungi. *Mycology*, 10(3):127–140.

[Xie et al., 2023] Xie, L., Zhang, X., Xie, J., Xu, Y., Li, X.-J., and Lin, L. (2023). Emerging roles for DNA 6ma and rna m6a methylation in mammalian genome. *International Journal of Molecular Sciences*, 24(18):13897.

[Xu and Seki, 2019] Xu, L. and Seki, M. (2019). Recent advances in the detection of base modifications using the nanopore sequencer. *Journal of Human Genetics*, 65(1):25–33.

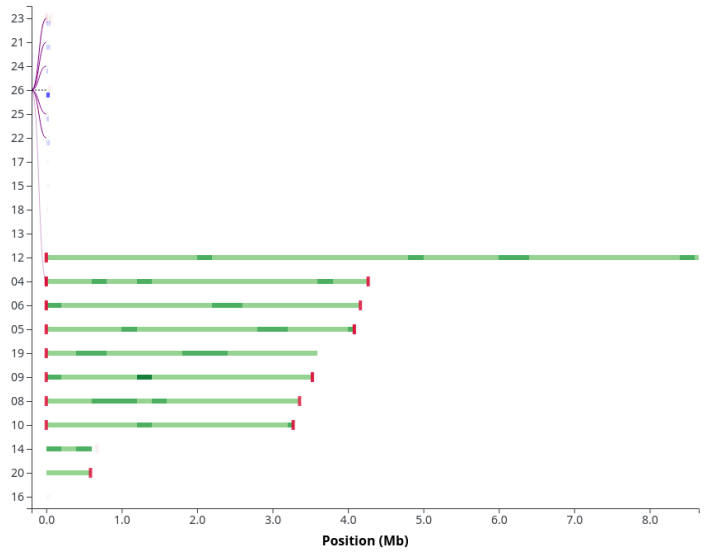
[Zemach et al., 2010] Zemach, A., McDaniel, I. E., Silva, P., and Zilberman, D. (2010). Genome-wide evolutionary analysis of eukaryotic DNA methylation. *Science*, 328(5980):916–919.

[Zemach and Zilberman, 2010] Zemach, A. and Zilberman, D. (2010). Evolution of eukaryotic DNA methylation and the pursuit of safer sex. *Current Biology*, 20(17):R780–R785.

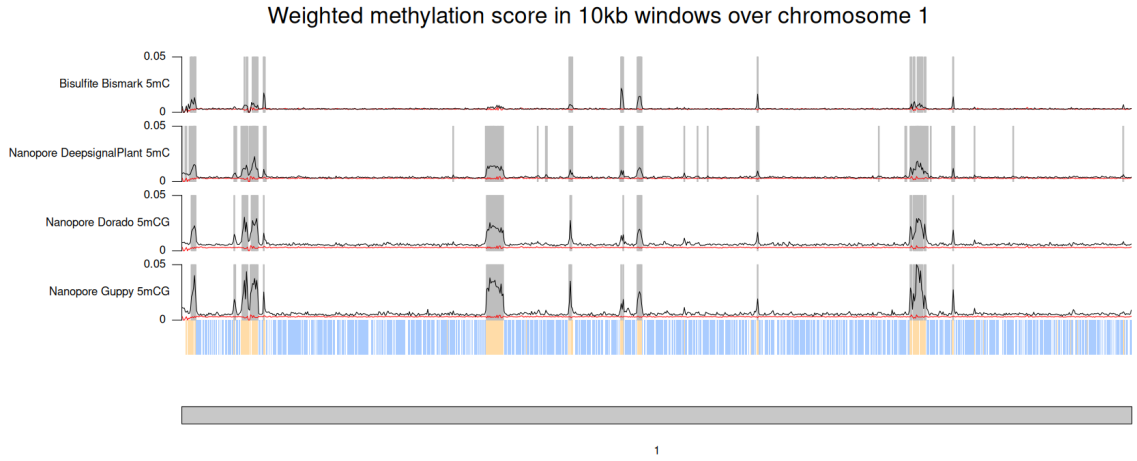
[Zhong et al., 2021] Zhong, Y., Xu, F., Wu, J., Schubert, J., and Li, M. M. (2021). Application of next generation sequencing in laboratory medicine. *Annals of Laboratory Medicine*, 41(1):25–43.

[Željko M. Svedružić, 2011] Željko M. Svedružić (2011). Chapter 6 - dnmt1: Structure and function. In Cheng, X. and Blumenthal, R. M., editors, *Modifications of Nuclear DNA and its Regulatory Proteins*, volume 101 of *Progress in Molecular Biology and Translational Science*, pages 221–254. Academic Press.

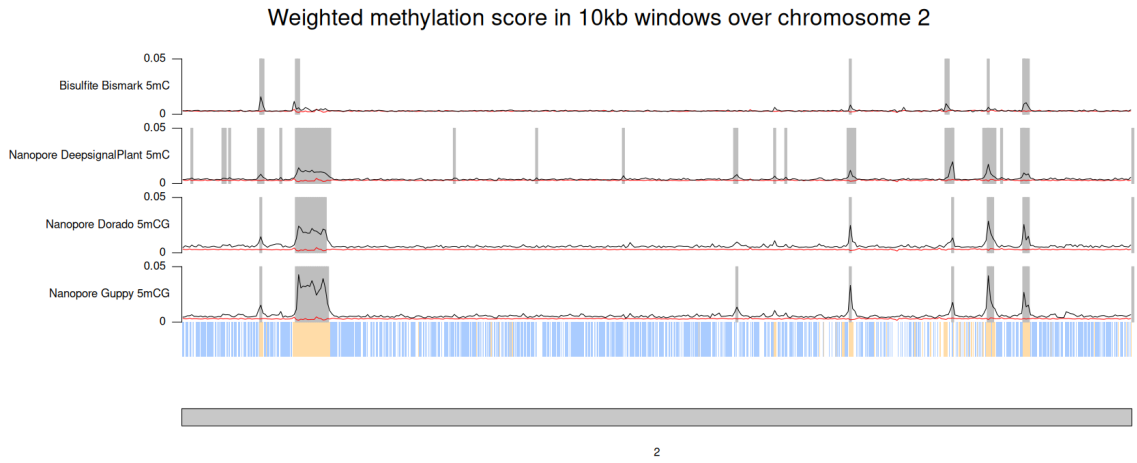
Supplementary figures



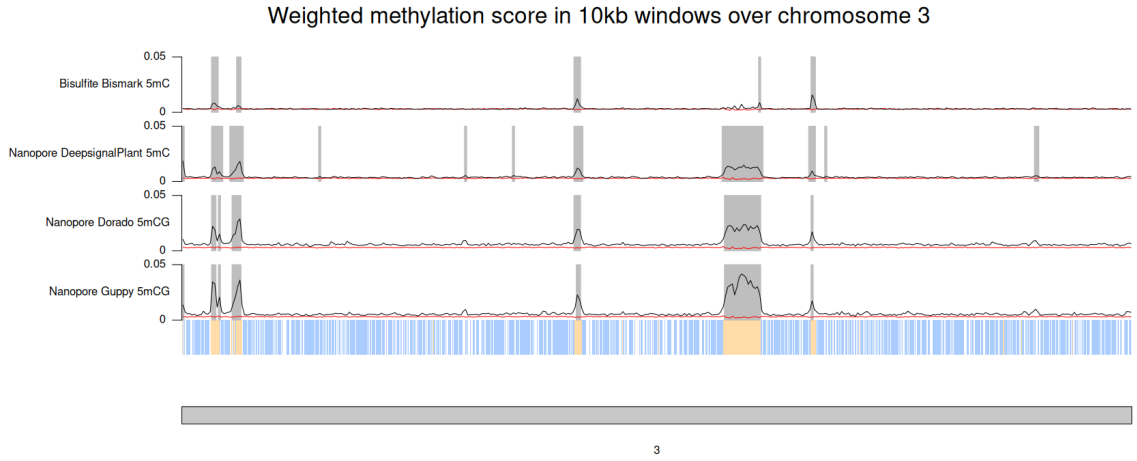
Supplementary Figure 1: All contigs assembled by the *de novo* genome assembly.



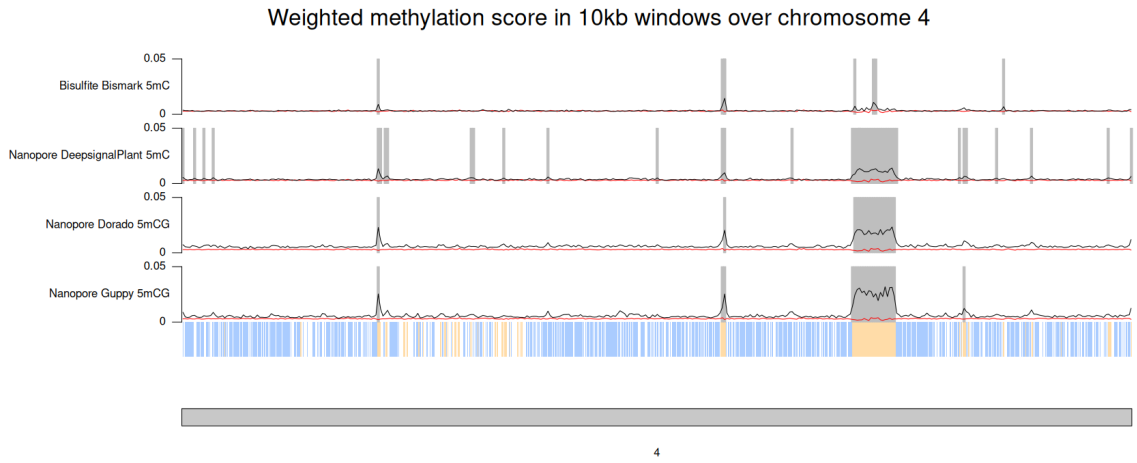
Supplementary Figure 2: Global methylation levels for bisulfite data and multiple Nanopore methylation callers over chromosome 1. The red lines depict the baseline noise level established by the bisulfite knockout of HP1. At the bottom, genes are represented in blue, while transposable elements and repeats are highlighted in yellow. Grey boxes highlight hyper-methylated windows.



Supplementary Figure 3: Global methylation levels for bisulfite data and multiple Nanopore methylation callers over chromosome 2. The red lines depict the baseline noise level established by the bisulfite knockout of HP1. At the bottom, genes are represented in blue, while transposable elements and repeats are highlighted in yellow. Grey boxes highlight hyper-methylated windows.

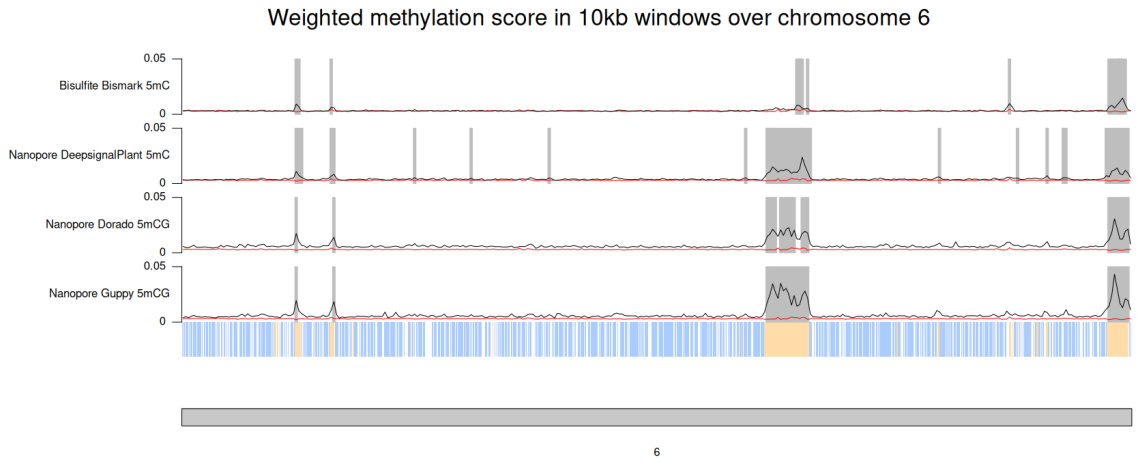


Supplementary Figure 4: Global methylation levels for bisulfite data and multiple Nanopore methylation callers over chromosome 3. The red lines depict the baseline noise level established by the bisulfite knockout of HP1. At the bottom, genes are represented in blue, while transposable elements and repeats are highlighted in yellow. Grey boxes highlight hyper-methylated windows.

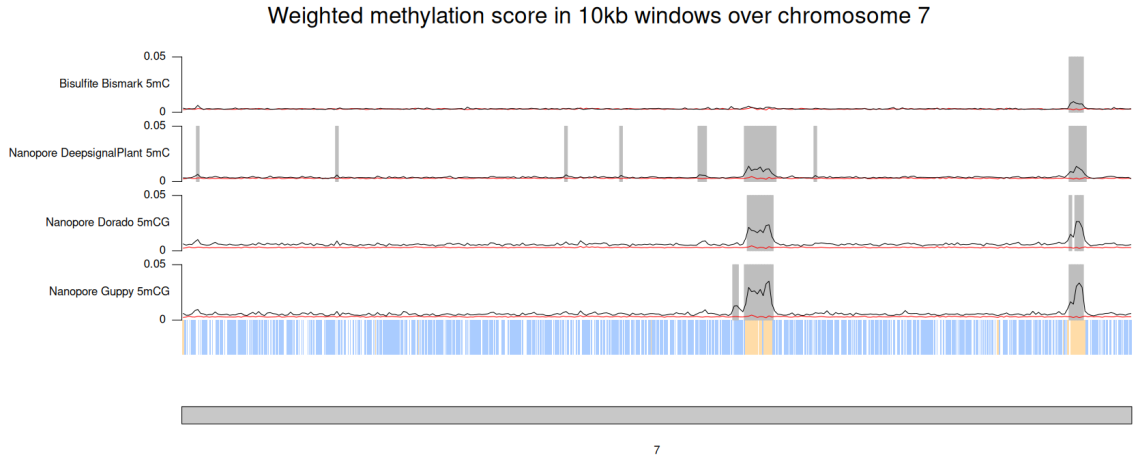


Supplementary Figure 5: Global methylation levels for bisulfite data and multiple Nanopore methylation callers over chromosome 4. The red lines depict the baseline noise level established by the bisulfite knockout of HP1. At the bottom, genes are represented in blue, while transposable elements and repeats are highlighted in yellow. Grey boxes highlight hyper-methylated windows.

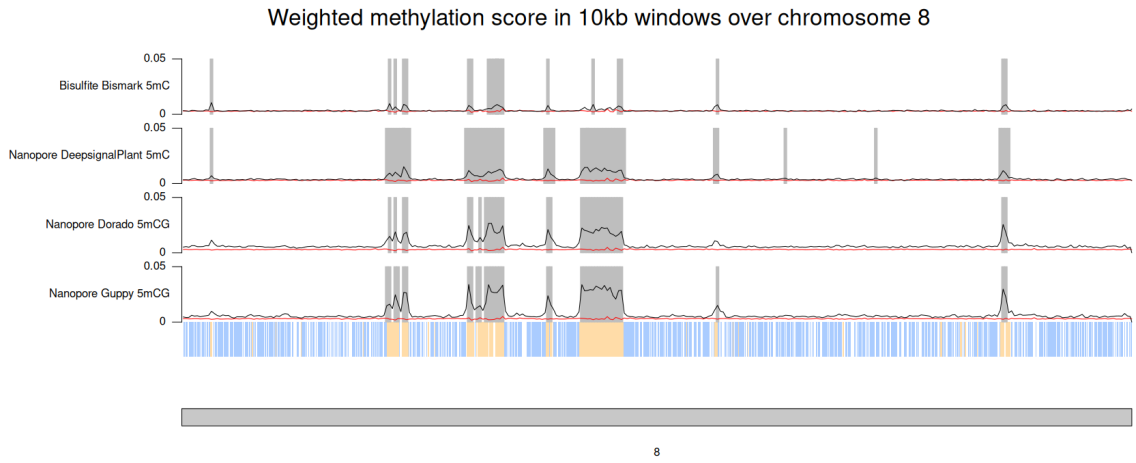
Supplementary Figure 6: Global methylation levels for bisulfite data and multiple Nanopore methylation callers over chromosome 5. The red lines depict the baseline noise level established by the bisulfite knockout of HP1. At the bottom, genes are represented in blue, while transposable elements and repeats are highlighted in yellow. Grey boxes highlight hyper-methylated windows.



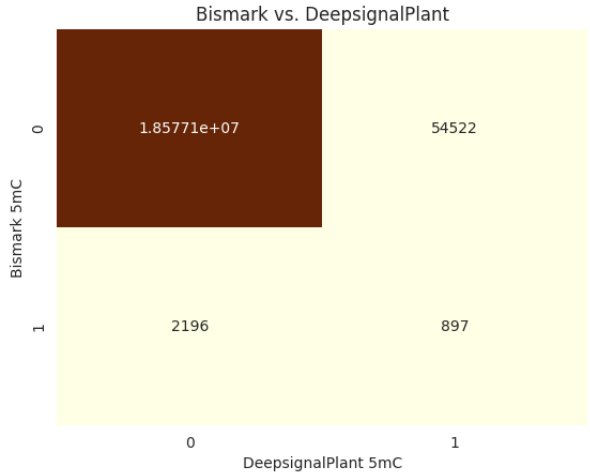
Supplementary Figure 7: Global methylation levels for bisulfite data and multiple Nanopore methylation callers over chromosome 6. The red lines depict the baseline noise level established by the bisulfite knockout of HP1. At the bottom, genes are represented in blue, while transposable elements and repeats are highlighted in yellow. Grey boxes highlight hyper-methylated windows.



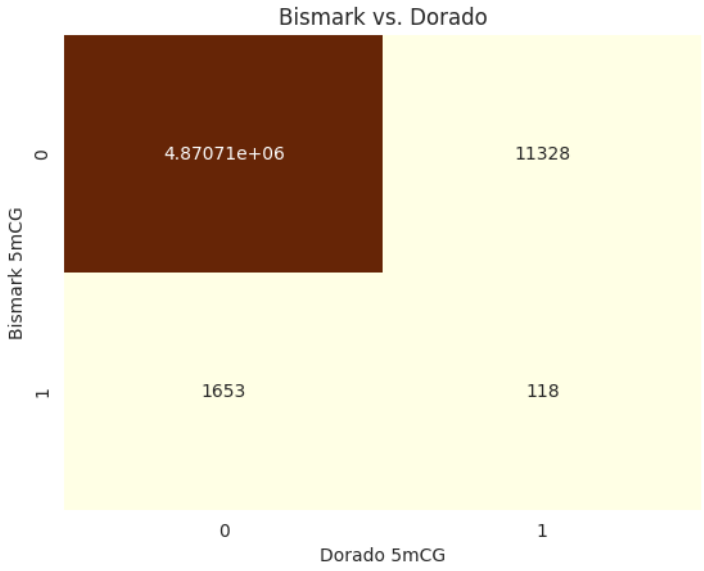
Supplementary Figure 8: Global methylation levels for bisulfite data and multiple Nanopore methylation callers over chromosome 7. The red lines depict the baseline noise level established by the bisulfite knockout of HP1. At the bottom, genes are represented in blue, while transposable elements and repeats are highlighted in yellow. Grey boxes highlight hyper-methylated windows.



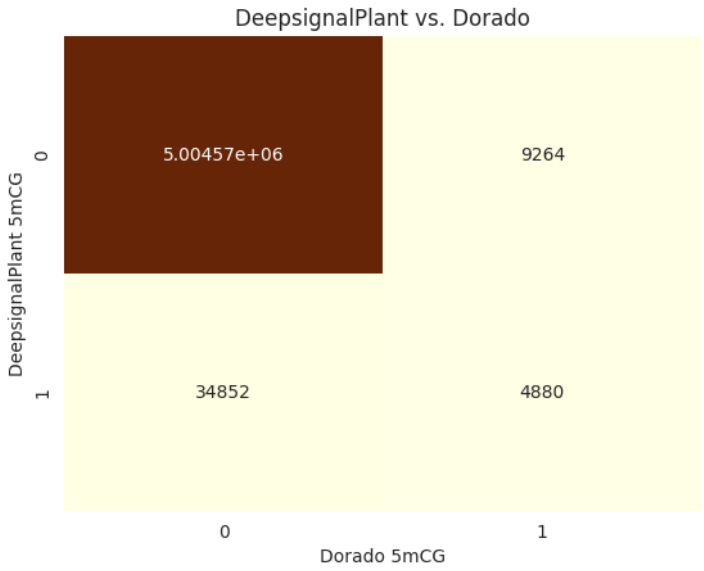
Supplementary Figure 9: Global methylation levels for bisulfite data and multiple Nanopore methylation callers over chromosome 8. The red lines depict the baseline noise level established by the bisulfite knockout of HP1. At the bottom, genes are represented in blue, while transposable elements and repeats are highlighted in yellow. Grey boxes highlight hyper-methylated windows.



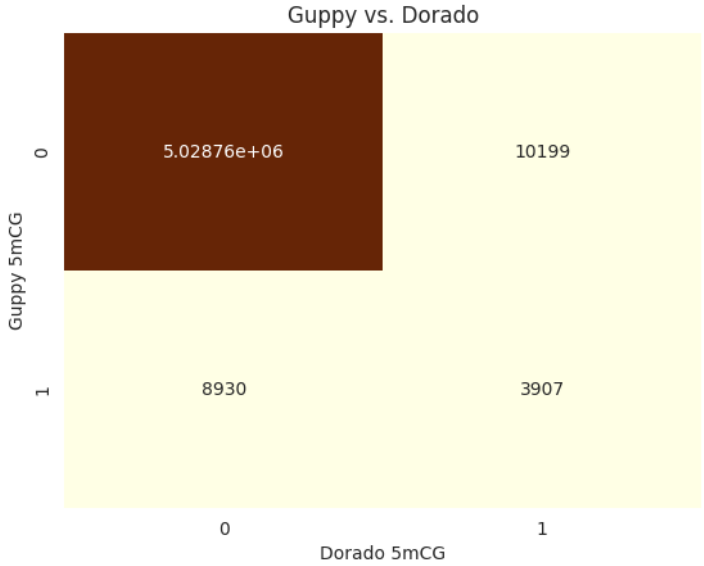
Supplementary Figure 10: Confusion matrix of Bismark vs DeepSignalPlant methylated cytosine positions.



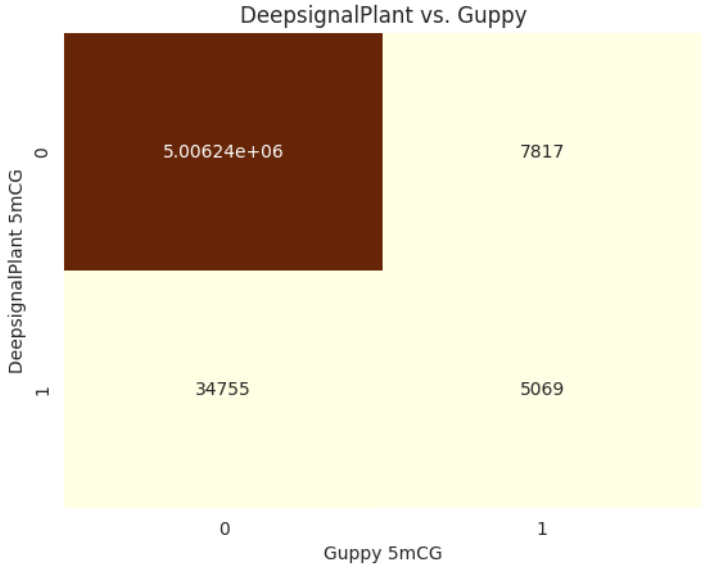
Supplementary Figure 11: Confusion matrix of Bismark vs Dorado methylated cytosine positions.



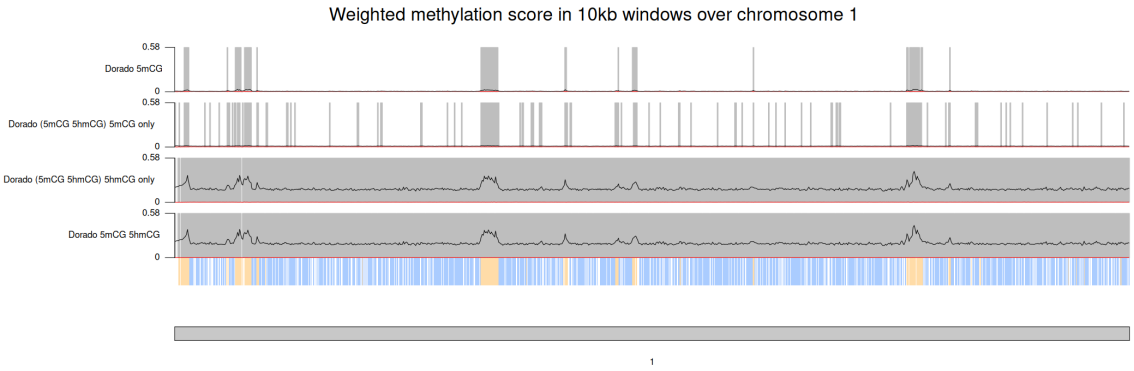
Supplementary Figure 12: Confusion matrix of DeepSignalPlant vs Dorado methylated cytosine positions.



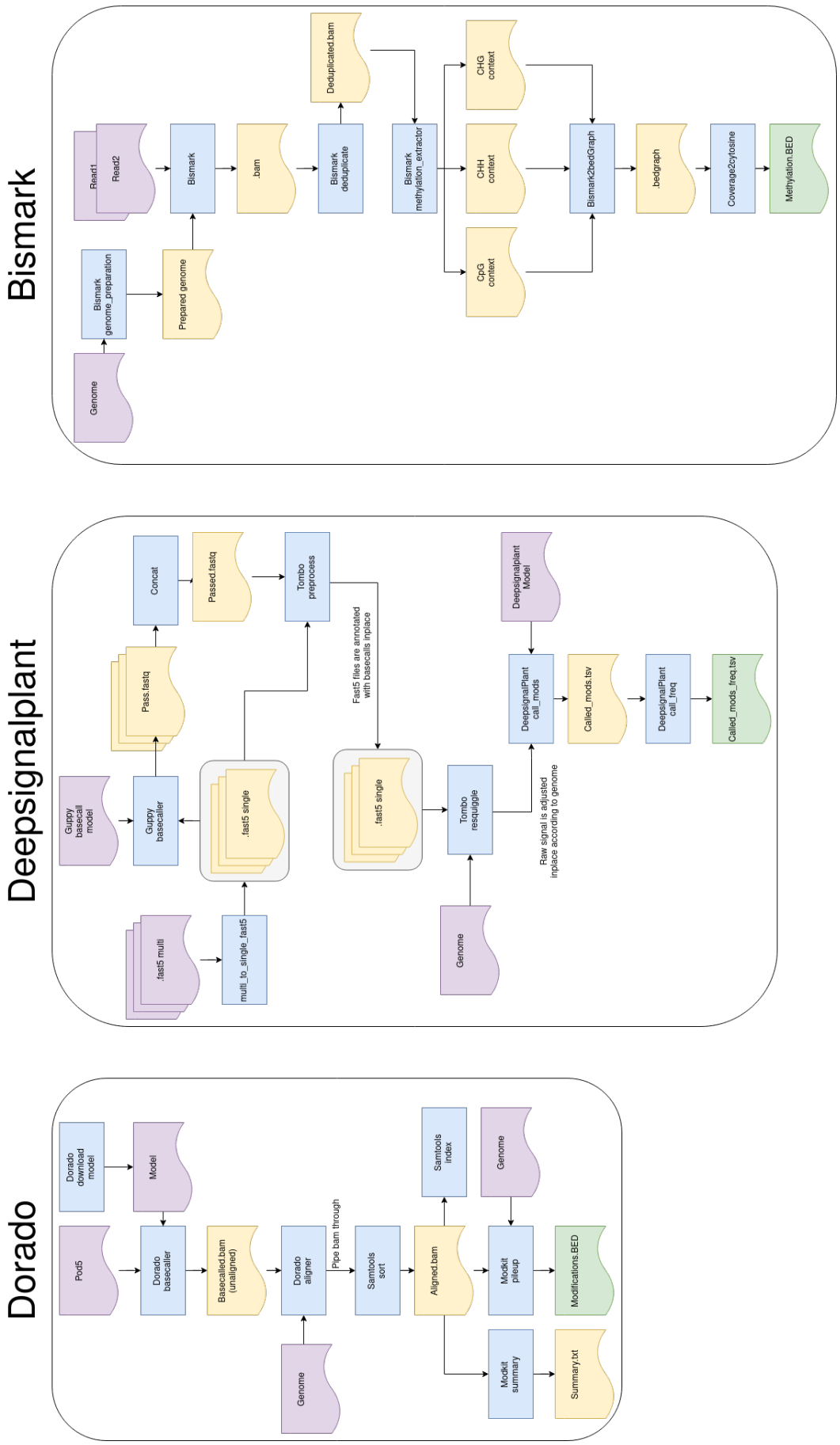
Supplementary Figure 13: Confusion matrix of Guppy vs Dorado methylated cytosine positions.



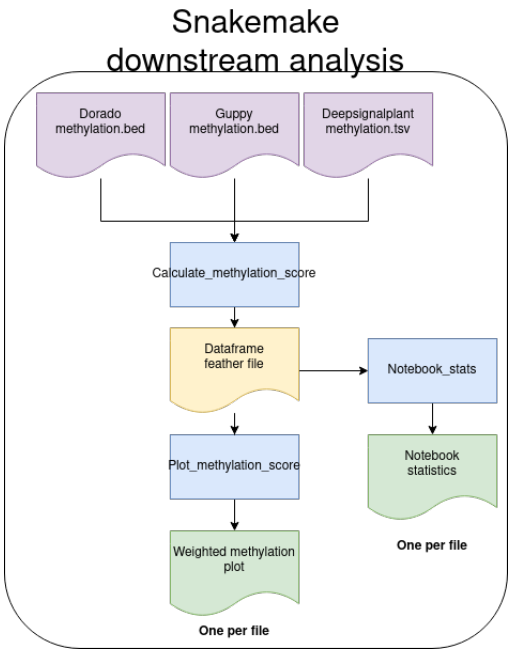
Supplementary Figure 14: Confusion matrix of DeepsSignalPlant vs Guppy methylated cytosine positions.



Supplementary Figure 15: Weighted methylation scores of Dorado 5mCG and 5hmCG models. The red lines depict the baseline noise level established by the bisulfite knockout of HP1. At the bottom, genes are represented in blue, while transposable elements and repeats are highlighted in yellow. Grey boxes highlight hyper-methylated windows.



Supplementary Figure 16: Flowchart of the Dorado, DeepSignalPlant, and Bismark automated pipelines.



Supplementary Figure 17: Flowchart of the Snakemake pipeline for calculating and plotting the weighted methylation score.