Chapter 3

Estimating Global Methylation and Erasure Using Low-Coverage Whole-Genome Bisulfite Sequencing (WGBS)

Oscar Ortega-Recalde, Julian R. Peat, Donna M. Bond, and Timothy A. Hore

Abstract

Whole-genome bisulfite sequencing (WGBS) is a popular method for characterizing cytosine methylation because it is fully quantitative and has base-pair resolution. While WGBS is prohibitively expensive for experiments involving many samples, low-coverage WGBS can accurately determine global methylation and erasure at similar cost to high-performance liquid chromatography (HPLC) or enzyme-linked immunosorbent assays (ELISA). Moreover, low-coverage WGBS has the capacity to distinguish between methylation in different cytosine contexts (e.g., CG, CHH, and CHG), can tolerate low-input material (<100 cells), and can detect the presence of overrepresented DNA originating from mitochondria or amplified ribosomal DNA. In addition to describing a WGBS library construction and quantitation approach, here we detail computational methods to predict the accuracy of low-coverage WGBS using empirical bootstrap samplers and theoretical estimators similar to those used in election polling. Using examples, we further demonstrate how non-independent sampling of cytosines can alter the precision of error calculation and provide methods to improve this.

Key words Low-coverage bisulfite sequencing, DNA methylation, Epigenetics, Methylation erasure, Whole-genome bisulfite sequencing, Post-bisulfite adaptor tagging, Bootstrap sampling, Asymptotic estimator

1 Introduction

Cytosine methylation is a well-described epigenetic modification that is essential for vertebrate development [1, 2], and when in the CG dinucleotide context, can transmit molecular memory post-replication. This unique ability is due to the symmetry of CG dinucleotides; methylation signals on the template strand are recognized by maintenance methylation enzymes and copied to

The original version of this chapter was revised. The correction to this chapter is available at https://doi.org/10.1007/978-1-0716-1294-1_18

Ozren Bogdanovic and Michiel Vermeulen (eds.), *TET Proteins and DNA Demethylation: Methods and Protocols*, Methods in Molecular Biology, vol. 2272, https://doi.org/10.1007/978-1-0716-1294-1_3,
© The Author(s) 2021, Corrected Publication 2021

cognate (unmethylated) cytosines on the new strand [3]. In plants, DNA methylation is also prominent, with both CG dinucleotides and CHG trinucleotides (where H is any base other than C) having the capacity to harbor epigenetic memory. Considerable methylation does exist on cytosines outside of the CG context in mammals, and the CG/CHG context in plants; however, this methylation is unable to maintain epigenetic memory post-replication and has as yet unknown function in mammals [4].

During mammalian germline development, the genome undergoes two waves of global DNA demethylation [5], effectively erasing the majority of epigenetic memory encoded by this mark. It has been hypothesized that this iconic process facilitates dramatic changes in cell identity (e.g., enhanced levels of developmental potency germline function, as in primordial germ cells) or achieves enhanced levels of developmental potency (e.g., the post-fertilization zygote). Global loss in DNA methylation is a feature of many cancers [6], which are also associated with enhanced cellular plasticity, and loss in cellular identity. The mechanisms by which global DNA demethylation occurs in mammals are yet to be fully understood, and its existence outside of eutherians is largely untested. In order to understand the prevalence and nature of global DNA methylation erasure, cheap and effective methods for quantifying global DNA methylation are required.

Several methods allow DNA methylation quantitation across the whole genome, with the most cost-effective involving high-performance liquid chromatography (HPLC) or enzyme-linked immunosorbent assay (ELISA) [7]. While these techniques can accurately measure DNA methylation globally, they are naïve to genomic position and cannot distinguish between memory-capable CG/CHG methylation and DNA methylation in other contexts. Conversely, “gold-standard” sequencing-based techniques such as whole-genome bisulfite sequencing (WGBS) provide fully quantitative information with single base-pair resolution [8–10]. Yet, even with rapidly diminishing sequencing costs, WGBS is prohibitively expensive for large genomes; consequently, most experiments feature a limited number of samples (usually <100).

A simple method to assay DNA methylation at low cost, but retain sequence context information, is to perform WGBS at low coverage [11–15]. In this way, global DNA methylation is effectively surveyed as if one was undertaking polling to predict the outcome of political election from a large population of voters. Applications for low-coverage WGBS range from screening hundreds of cancer samples for aberrant global methylation patterns to identifying unique epigenetic regulation in divergent taxa [15]; however, it is particularly useful for tracking developmental transitions and changes in cell identity that involve global epigenetic reprogramming [12–14]. An additional benefit of this approach is that once samples of interest are identified by low-coverage WGBS, a full methylome with base-pair resolution

Fig. 1 Low-coverage WGBS strategy. Low-coverage WGBS is a cost-effective technique to measure global levels of methylation and can provide a useful screening tool for selection of interesting libraries for deep-coverage WGBS. Nevertheless, there are important considerations when using the low-coverage approach, such as the level of data dependency and alternative cytosine modifications

can be acquired simply by performing deeper sequencing; additional library construction is not required (Fig. 1). While any form of WGBS library preparation can be used, we find that post-bisulfite adaptor tagging (PBAT) is one of the most versatile and least expensive [16]. This method also has the benefit of being successful in experiments where input material is limited (<100 cells) [17].

Despite the benefits of low-coverage WGBS, some aspects of its use are not well developed in the literature. For example, as with election polling, it is important to understand the margin of error (ϵ) associated with WGBS where a given number of cytosine sites are sampled. This protocol outlines two methods (one theoretical, one empirical) to calculate the margin of error (ϵ) associated with estimating global methylation level. In doing so, this informs the experimentalist the number of cytosine calls, and therefore how much sequencing is required in order to accurately predict global DNA methylation levels within a given confidence interval. For the sake of completeness, we have also included a post-bisulfite adapter tagging (PBAT) protocol outlining WGBS library construction; however, any WGBS protocol can be used to predict global methylation using our theoretical and empirical sampling estimation methods.

2 Materials

2.1 WGBS Library Construction Using PBAT

1. 0.5 mL DNA LoBind tubes (e.g., Eppendorf).
2. Low-binding pipette tips (e.g., MultiMax).
3. 0.2 mL PCR tubes with individual caps.
4. Magnetic racks for 0.2 mL PCR tubes and 0.5 mL tubes.
5. PCR thermocycler.
6. Cell lysis buffer: 10 mM Tris-HCl pH 8.0, 1 mM EDTA, 0.6% SDS.
7. EZ Methylation Direct Mag Prep kit (Zymo Research).
8. PCR-Grade dNTPs.
9. NEB Buffer 2 (New England Biolabs).
10. Klenow fragment 3' → 5' exo- (50 U/μL).
11. Exonuclease I (20 U/μL).
12. H₂O (molecular biology grade).
13. 100% Ethanol.
14. PEG/NaCl-diluted beads: 0.1% magnetic beads, 18% PEG-8000, 1 M NaCl, 10 mM Tris-HCl pH 8.0, 1 mM EDTA, 0.05% Tween 20. Carboxylate-Modified Sera-Mag Magnetic beads (Merck) or any other magnetic beads can be used to prepare this solution, *see* **Note 1**.
15. TE buffer (1×): 10 mM Tris-HCl pH 8.0, 1 mM EDTA.
16. M-280 Streptavidin Dynabeads (Thermo Fisher).
17. Binding and washing (B&W) buffer (2×): 10 mM Tris-HCl, 1 mM EDTA, 2 M NaCl, pH 7.5.
18. 0.1 M NaOH (make fresh).
19. HiFi HotStart Uracil+ Mix (KAPA Biosystems).
20. BioP5N7 oligos, 5' biotin- ACACCTCTTCCCTACAC GACGCTCTTCCGATCTNNNNNNN (where N are random nucleotides).
21. P7N7 oligos, 5' GTGACTGGAGTTCAGACGTGTGCTCT TCCGATCTNNNNNNN.
22. Forward primer index (FPI): Illumina TruSeq-type forward indexes, e.g., AATGATACGGCGACCACCGAGATCTA CACNNNNNNNACACTCTTCCCTACAC GACGCTCTTCCGATCT, where NNNNNN is unique index.
23. Reverse primer index (RPI): Illumina TruSeq-type reverse indexes.
CAAGCAGAAGACGGCATACGAGATNNNNNNNNNG
TGACTGGAGTTCAGACGTGTGCTCTTCCGATC.

2.2 Computing Requirements

1. Operating system: This protocol uses a Unix-like operating system (e.g., Linux or Mac), with a bash shell.
2. The R statistical computing environment (available at <https://www.r-project.org/>).
3. Packages ggplot2 and dplyr (available as libraries for R at https://cran.r-project.org/web/packages/available_packages_by_name.html).
4. SRA toolkit (available at https://ncbi.github.io/sra-tools/install_config.html).
5. fastq-tools (available at <https://homes.cs.washington.edu/~dcjones/fastq-tools/>).
6. Trim galore! (available at https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/).
7. Bismark (available at <https://www.bioinformatics.babraham.ac.uk/projects/bismark/>).
8. SeqMonk (available at <https://www.bioinformatics.babraham.ac.uk/projects/seqmonk/>), *see* **Note 2**.
9. Scripts and the full data used for this example is available at https://github.com/OscarOrt/Methods_Low_coverage_WGBS.

3 Methods

This protocol is divided into two main sections. First, we describe the preparation of WGBS libraries using the PBAT approach which is both amendable to high-throughput analysis (i.e., no sonication required) and is optimized for low-input material (<100 cells) (*see* Subheading 3.1). Second, we describe a method for predicting the accuracy of low-coverage WGBS (*see* Subheadings 3.2 and 3.3). As mentioned, margin of error calculations are not restricted to PBAT but can be used along with other WGBS protocols [9, 16, 18].

3.1 WGBS Library Preparation by PBAT

PBAT enables high quality and base-pair specific DNA methylation information using bisulfite conversion and high-throughput sequencing. Negative controls, where no template DNA is added during library preparation, are strongly recommended. This is especially important when low-input material is used as even low-level contamination may confound the results. Low-DNA binding plasticware is also recommended to maximize yield from low-input material.

3.1.1 Cell Lysis and DNA Purification

While commercial kits can be used to perform this step, open-source magnetic bead based protocols are a feasible and cost-effective method for DNA purification [19]. Alternatively, when

very low inputs are processed (<500 cells) and a strong cell lysis method is not required for DNA extraction, cell lysates can be used directly for bisulfite conversion.

Option 1: BOMB Genomic DNA Extraction

Perform the gDNA extraction using the BOMB (Bio-On-Magnetic-Beads) method. This method is based on paramagnetic beads which can be acquired commercially or synthesized in the laboratory [19]. Protocols for magnetic bead preparation and DNA extraction are provided in the BOMB.bio webpage (<https://bomb.bio/protocols/>). While approaches using magnetic beads are the most suitable for high-throughput experiments, any robust DNA extraction protocol can be used in its place.

Option 2: Using Cell Lysate Directly for Bisulfite Conversion

1. Lyse cells in 10 mM Tris-HCl pH 8.0, 1 mM EDTA, 0.6% SDS buffer to give a final volume of 25 μ L. The volume of cell lysis buffer should not be lower than 75% the final volume. This method is particularly useful for low-cell number PBAT [17].

3.1.2 Bisulfite Conversion

Perform bisulfite treatment on 20 μ L of purified DNA or cell lysate using the EZ Methylation Direct Mag Prep kit (Zymo Research) following manufacturer's instructions. Converted DNA can be used immediately or stored at -20°C for short-term use.

3.1.3 First Strand Synthesis

1. Add 20 μ L of bisulfite-converted DNA (Subheading 3.1.2) to a 0.2 mL PCR tube and add 1 μ L of dNTPs (5 mM, 1.25 mM each), 0.5 μ L of BioP5N7 (10 μ M), and 2.5 μ L of 10 $\times$ NEB2 buffer per sample.
2. Incubate samples at 65°C in thermocycler for 5 min, then at 4°C for 5 min and place the samples on ice immediately. Hold thermocycler at 4°C .
3. Add 0.5 μ L of Klenow fragment $3' \rightarrow 5'$ exo- and mix gently by pipetting.
4. Incubate at 4°C for 15 min, then raise to 37°C at a rate of 6°C every minute, and hold temperature at 37°C for 90 min. Pause at 4°C .

3.1.4 Exonuclease Treatment and Purification

1. Add 0.5 μ L of Exonuclease I directly into the first strand synthesis reaction (subheading 3.1.3).
2. Incubate at 37°C for 60 min and hold at 4°C .
3. Add 20 μ L of PEG/NaCl-diluted magnetic beads to each ExoI-treated first strand synthesis reaction (25 μ L from previous steps—sample:bead ratio = 0.8 $\times$) (see **Note 1**).
4. Mix gently by pipetting and incubate the mixture for 5 min at RT.
5. Place the tubes on a magnetic rack, wait for 1–2 min until the solution becomes clear and remove supernatant.

6. Wash beads twice with 100 μ L of 90% ethanol, keeping tubes on magnet; remove the supernatant.
7. Dry the beads for approximately 5–10 min.
8. Add 100 μ L of 1 $\times$ TE buffer and resuspend beads; heat the elution at 55 $^{\circ}$ C for 4 min, place tubes on magnetic rack, wait for 1–2 min until the solution becomes clear and transfer supernatant to a new 0.5 mL tube.

3.1.5 Biotin Capture

1. Take 10 μ L of M-280 Streptavidin Dynabeads per sample to a new 0.2 mL PCR tube, place tubes on magnetic rack, wait for 1–2 min until the solution becomes clear and remove supernatant.
2. Wash the Dynabeads twice with 2 $\times$ B&W buffer according to manufacturer's protocol.
3. Resuspend Dynabeads in 100 μ L per sample of 2 $\times$ B&W buffer.
4. Add 100 μ L of purified ExoI-treated first strand synthesis reaction (subheading 3.1.4) to 100 μ L of resuspended Dynabeads (subheading 3.1.5, step 3).
5. Mix gently by pipetting and incubate for 30 min at RT with continuous rotation.
6. Place the tubes on a magnetic rack, wait for 1–2 min until the solution becomes clear and remove supernatant.
7. Wash the beads twice with 100 μ L of freshly prepared NaOH (0.1 M). Briefly vortex the tubes, place tubes in magnetic rack, wait for solution to clarify, remove supernatant. Perform second wash while keeping the tubes on the magnetic rack.
8. Wash the beads twice with 100 μ L of Tris (10 mM), pH 8.0. Briefly vortex the tubes, place tubes in magnetic rack, wait for solution to clarify, remove supernatant. Perform second wash while keeping the tubes on the magnetic rack.

3.1.6 Second Strand Synthesis

1. Remove liquid remnants and resuspend washed beads (subheading 3.1.5) with the following reaction mix, per sample: 2 μ L of dNTPs (5 mM), 2 μ L of P7N7 (10 μ M), 5 μ L of 10 $\times$ NEB2 buffer, and 40 μ L of H₂O.
2. Transfer mix to new 0.2 mL PCR tubes.
3. Incubate at 95 $^{\circ}$ C in a thermocycler for 45 s. Place the samples on ice immediately. Hold thermocycler at 4 $^{\circ}$ C.
4. Add 1 μ L of Klenow fragment 3' $\rightarrow$ 5' exo- and mix gently by pipetting.
5. Incubate at 4 $^{\circ}$ C for 5 min, then raise to 37 $^{\circ}$ C at a rate of 6 $^{\circ}$ C every minute, and hold temperature at 37 $^{\circ}$ C for 90 min. Pause at 4 $^{\circ}$ C.

3.1.7 PCR Amplification and Analysis

1. Place the tubes on a magnetic rack, wait for 1–2 min until the solution becomes clear and remove supernatant.
2. Wash the beads twice with 50 μL of $1\times$ TE buffer, keeping tubes on the magnetic rack.
3. Resuspend washed beads in the following PCR amplification mix, per sample: 6 μL of $2\times$ KAPA HiFi HotStart Uracil+ Mix, 5 μL of H_2O , 0.5 μL of FPI (10 μM), and 0.5 μL of RPI (10 μM). If several samples will be pooled, combine indexes appropriately.
4. Incubate the reaction in a thermocycler at 98 $^\circ\text{C}$ for 5 min, followed by 13 cycles of 98 $^\circ\text{C}$ for 80 s, 65 $^\circ\text{C}$ for 30 s, 72 $^\circ\text{C}$ for 30 s, and a final elongation of 72 $^\circ\text{C}$ for 5 min. Hold at 10 $^\circ\text{C}$. The maximum number of cycles recommended is 15.
5. Run 3 μL of PCR product on 1% agarose gel to check library quality/quantity relative to 3 μL of 100-bp ladder (e.g., Bionline).

3.1.8 Pooling and Library Purification

1. Pool PCR products based on quantity of DNA in each library (generally 2 μL).
2. Add $0.8\times$ volume of PEG/NaCl-diluted beads to pooled PCR products (i.e., 40 μL diluted beads for 20 μL pooled PCR product). Alternatively, the pooled PCR products (~300–600 bp in size) can be gel extracted from a 1.2% agarose gel using a commercial gel extraction system (e.g., QiaQuick Gel Extraction kit) following the manufacturer's instructions.
3. Mix gently by pipetting and incubate the mixture for 5 min at RT.
4. Place the tubes on a magnetic rack, wait for 1–2 min until the solution becomes clear and remove supernatant.
5. Keeping tubes on magnetic rack, wash beads twice with 100 μL of 70% ethanol. Remove the supernatant.
6. Dry the beads for approximately 5–10 min.
7. Resuspend beads in H_2O , typically 50% of the initial PCR pool volume (i.e., 10 μL H_2O for 20 μL pooled PCR product).
8. Heat the elution at 55 $^\circ\text{C}$ for 5 min and transfer to a new 0.5 mL tube.
9. Samples are ready to be quantified and sequenced.

The experimental strategy and limitations for this approach are discussed in **Note 3**. Although rare, overrepresented reads may bias the method accuracy, therefore assessing mapping homogeneity is recommended (*see Note 2*).

3.2 Theoretical Prediction of Margin of Error

The margin of error (ϵ) associated with particular levels of methylation determined by low-coverage WGBS can be estimated through the use of a confidence interval known as a “one-sample proportion in the Z -interval” according to the formula below:

$$\epsilon = z^* \sqrt{\frac{p(1-p)}{n}}$$

where p is the sample proportion, n is the sample size, and the value of the z^* multiplier depends on the level of confidence.

To simplify calculation of ϵ , one can use the statistical programming language, R. The following code uses variables from low-coverage WGBS from bovine blastocysts as an example:

```
zstar = qnorm(.995) #This defines 99.5% confidence level
p = 0.25 #The approximate methylation level from this example
n = 10000 #Number of cytosine calls assayed
zstar*sqrt((p*(1-p))/n) #Margin of error calculation given
above variables
```

Alternatively, one may wish to calculate the number of calls required to achieve a given margin of error.

```
zstar = qnorm(.995) #This defines 99.5% confidence level
p = 0.25 #The approximate methylation level from this example
E = 0.005 #Margin of error
zstar^2 * p * (1-p) / E^2 # Calculates cytosine calls required
to achieve given margin of error
```

By plotting the predicted margin of error with variable numbers of cytosine calls, one can see in this example that additional sequencing beyond 20,000 CG calls does not significantly improve the accuracy of methylation prediction (Fig. 2a).

```
zstar = qnorm(.995) #This defines 99.5% confidence level
p = 0.25 #The approximate methylation level from this example
n = seq(100,50000,by=100) #Numbers of cytosine calls assayed
E = zstar*sqrt((p*(1-p))/n) #Margin of error calculation given
above variables
plot(x=n,y=E,pch=20,xlab = " CG calls", ylab = "Margin of
error") #Plot X and Y axes
abline(h=0.01) # Margin of error 1%
```

Given useful read numbers even in the low thousands, with library multiplexing, it is possible to predict CG methylation levels for hundreds of samples simultaneously using only low-output sequencing machines such as the Illumina® MiSeq. We used this approach to estimate methylation levels in bovine blastocysts

Fig. 2 Theoretical model of margin of error for low-coverage WGBS. **(a)** The approximate margin of error (percentage points, p.p.) can be calculated using a theoretical estimator of population proportions. **(b)** CG methylation levels in bovine blastocysts under various culture conditions and somatic control samples (SNAZY, LJ801, adult skin fibroblast cell lines). Error bars indicate the theoretical error associated with sampling according to 99.5% confidence

(Fig. 2b); a result indicating that irrespective of culture media, blastocysts in bovine have dramatically reduced CG methylation compared to somatic cell controls.

3.3 Empirical Bootstrap Sampling

If one has access to deep-sequencing datasets that were constructed in a similar fashion to the dataset generated by low-coverage sequencing, it is possible to calculate the margin of error associated with methylation sampling using random sampling with replacement, or “bootstrapping” [20]. With sufficient sampling repetitions (e.g., 1000), smooth distributions of methylation estimates from which the margin of error over a given confidence interval can be calculated in an empirical manner.

Below we have provided an example of the empirical bootstrap method for WGBS performed on the gonads of bluehead wrasse (*Thalassoma bifasciatum*). This species is a protogynous sequential hermaphrodite—dominant “initial phase” adult females can dramatically transition to a “terminal phase” male phenotype and assume control over a harem of females with 1–2 weeks. During this process the gonads transform from ovaries to testes, and with this transformation comes a distinctive increase in global CG methylation within gonadal DNA [14].

The scripts below are written for Unix-based machines and for the sake of demonstration, examine only a fraction of the sampling reads from bluehead wrasse ovary (SRR8137404).

1. Download WGBS dataset from the NCBI short read archive SRA database using SRA-Toolkit.

```
# Download SRA dataset
wget ftp://ftp-trace.ncbi.nih.gov/sra/sra-
instant/reads/ByRun/sra/SRR/SRR813/SRR8137404/SRR8137404.sra

# Dump SRA file into a fastq file (file saved in 01_sample folder)
fastq-dump --outdir 01_sample --split-files SRR8137404.sra
```

2. Bootstrap sampling with replacement using fastq-tools.

```
# Sampling (files save in 02_sampling folder)
for i in 100 500 1000 5000 10000 # Number of reads by sample
(i.e. 100, 500, 1000, 5000, 10000).
do
    for j in {1..1000} # Number of replicates (i.e. 1000)
    do
        seed=$((i+j)) # Variable seed
        fastq-sample -r -n $i -o 02_sampling/sampling\_$_i\_$_j -s $seed
01_sample/SRR8137404_1.fastq
        echo $j # Progress number of replicates
    done
    echo $i # Progress number of reads by sample
done
```

3. Read trimming with Trim galore!

```
# Trimming (files saved in 03_trimmed folder)
trim_galore -q 20 --clip_R1 10 -o 03_trimmed 02_sampling/*.
fastq # Two steps trimming. First, adaptors are removed and
10bp are hard-trimmed from the 5' end of all reads, finally
low-quality base calls (Phred score<20) are removed.
```

4. Rename reads using a bash script. This step is required because files with duplicated read names are not allowed in the Bismark mapping program.

```
# Rename (files save in 04_renamed folder)
for i in 100 500 1000 5000 10000 # Number of reads by sample
(i.e. 100, 500, 1000, 5000, 10000).
do
    for j in {1..1000} # Number of replicates (i.e. 1000).
    do
        awk '{print (NR%4 == 1) ? "@1_" ++i : $0}' <
03_trimmed/sampling\_$_i\_$_j\_$_trimmed.fq >
04_renamed/sampling\_$_i\_$_j\_$_trimmed.fq.temp
        awk '{print (NR%4 == 3) ? "+1_" ++i : $0}' <
```

```
04_renamed/sampling\_i\_j\_trimmed.fq.temp >
04_renamed/sampling\_i\_j\_trimmed.fq
rm 04_renamed/sampling\_i\_j\_trimmed.fq.temp
done
done
```

5. Map to reference genome using Bismark (Reference genome used for this example is available at the GitHub repository).

```
# Variables (Bismark prepared genome folder)
GENOME=Thalassoma_bifasciatum_v1

# Run Bismark (files saved in 05_bismark)
bismark --pbat $GENOME --output 05_bismark 04_renamed/*.fq # Bismark
mapping
for PBAT libraries.
```

6. Summarize data using a bash script (Variables to extract from Bismark output file could change according to the data).

```
# Make summary (file saved in 06_summary)
echo -e "Sample\tReads_number\tUnique_best_hit\tMapping_effi-
ciency\tNo_alignments\tTotal_Cs\tTotal_CsM_CpG\tTotal_CsM_CHG
\tTotal_CsM_CHH\tTotal_CsM_UNKNOWN\tTotal_CsU_CpG\tCsM_CpG
\tCsM_CHG\tCsM_CHH\tCsM_UNKNOWN" > 06_summary/summary_fema-
le_sample.txt # This is the first line of the summary

FILES=05_bismark/*.txt
for f in $FILES
do
    SAMPLE=$(grep -oP "(?<=renamed\\)[^ ]+" $f)
    READS_NUMBER=$(grep -oP "(?<=Sequences analysed in total:
\\t)[^ ]+" $f)
    UNIQUE_BEST_HIT=$(grep -oP "(?<=Number of alignments with
a unique best hit from the different alignments:\\t)[^ ]+" $f)
    MAPPING EFFICIENCY=$(grep -oP "(?<=Mapping efficiency:\\t)
[^ ]+" $f)
    NO_ALIGNMENTS=$(grep -oP "(?<=Sequences with no alignments
under any condition:\\t)[^ ]+" $f)
    Total_Cs=$(grep -oP "(?<=Total number of C's analysed:\\t)
[^ ]+" $f)
    Total_CsM_CpG=$(grep -oP "(?<=Total methylated C's in CpG
context:\\t)[^ ]+" $f)
    Total_CsM_CHG=$(grep -oP "(?<=Total methylated C's in CHG
context:\\t)[^ ]+" $f)
    Total_CsM_CHH=$(grep -oP "(?<=Total methylated C's in CHH
context:\\t)[^ ]+" $f)
    Total_CsM_UNKNOWN=$(grep -oP "(?<=Total methylated C's in
```

```

Unknown context:\t)[^ ]+" $f)
    Total_CsU_CpG=$(grep -oP "(?<=Total unmethylated C's in
CpG context:\t)[^ ]+" $f)
    CSM_CpG=$(grep -oP "(?<=C methylated in CpG context:\t)[^ ]
+" $f)
    CSM_CHG=$(grep -oP "(?<=C methylated in CHG context:\t)[^ ]
+" $f)
    CSM_CHH=$(grep -oP "(?<=C methylated in CHH context:\t)[^ ]
+" $f)
    CSM_UNKNOWN=$(grep -oP "(?<=C methylated in Unknown
context \ (CN or CHN\):\t)[^ ]+" $f)
    echo -e "$SAMPLE\t$READS_NUMBER\t$UNIQUE_BEST_HIT\t$MAP-
PING_EFFICIENCY\t$NO_ ALIGMENTS\t$Total_Cs\t$Total_CSM_CpG\t
$Total_CSM_CHG\t$Total_CSM_CHH\t$Total_CSM_UNKNOWN\t$Total_C-
sU_CpG\t$CSM_CpG\t$CSM_CHG\t$CSM_CHH\t$CSM_UNKNOWN" >>
06_summary/summary_female_sample.txt # This loop summarizes
each file into a single output file.
done

```

As expected from the theoretical estimator, greater numbers of reads produce more consistent predictions of global methylation levels in both male and female gonads of the bluehead wrasse (Fig. 3a). Moreover, once sequencing surpasses ~10,000 reads, there is diminishing improvement in the margin of error. Interestingly, the margin of error from bootstrap sampling appears to be greater than that predicted by the theoretical estimator (Fig. 3b, red and blue dots). This is likely due to data dependency—a caveat which is discussed in **Note 4**. Processed data and scripts used for the downstream analysis and plotting are available at https://github.com/OscarOrt/Methods_Low_coverage_WGBS.

4 Notes

1. It is recommended once PEG/NaCl-diluted magnetic beads are prepared to test the size selection protocol with 100 bp ladder and different volumes of PEG/NaCl-diluted beads (i.e., 0.4×, 0.6×, 0.8×, 1×, 1.2×). The PEG/NaCl bead solution can be aliquoted and stored at 2–8 °C for until 3 months.
2. In contrast to HPLC and mass spectrometry, the ability to map WGBS data to a genome makes possible to identify regions which are overrepresented for technical and biological reasons. A striking example of this is seen in the oocytes and zygotes of fish and amphibians, where mitochondrial DNA and ribosomal DNA is massively overrepresented due to increased numbers of mitochondria and rolling circle amplification of ribosomal DNA [5, 13]. Because these overrepresented sequences are

Fig. 3 Empirical bootstrap sampling of high-coverage gonad bluehead wrasse methylation datasets. (a) Percentage of methylation per number of CGs sampled for 1000 replicates at each sample size. (b) Error associated with different sample sizes (99.5% confidence level, $n = 1000$). Error was calculated for all cytosines sampled in females (red) and males (blue) and just for the first CG dinucleotide (CG) of each CG-containing read (orange for female and green for male)

highly demethylated, it is impossible to accurately calculate global methylation of the nuclear genome without first knowing sequence identity. Therefore, assessing the homogeneity of reads mapped along the genome is recommended. Appropriate software for this purpose includes SeqMonk (available at <https://www.bioinformatics.babraham.ac.uk/projects/seqmonk/>) and other high-throughput sequencing visualization tools.

- Sequencing coverage of 5–15 $\times$ is required to identify accurately differentially methylated regions (DMRs) from WGBS [21]. It should be highlighted that low-coverage WGBS is not a technique to quantify DMRs at single copy loci, but a method to survey global levels of methylation in the entire genome, or for a large class of common sequences (e.g., common k-mers, transposable elements, and metagenes). Because the technique requires just a fraction of the WGBS library, once good quality and interesting samples have been identified, deep-coverage sequencing can be performed without the need to create additional sequencing libraries (Fig. 1).

Another consideration when using low-coverage WGBS methods is discrimination of methylated cytosine (mC) from alternative cytosine modifications such as hydroxymethylcytosine (hmC), formylcytosine (fmC), and carboxylcytosine (caC). mC and hmC cannot be distinguished using standard bisulfite treatment (because they are both resistant to it), and fmC and caC are both susceptible to deamination to uracil in the same way as unmethylated cytosine. Discrimination of mC and hmC is possible using modified bisulfite sequencing protocols such as TET-assisted bisulfite sequencing [22], and oxidative

bisulfite sequencing [23]; and estimating the margin of error associated with predicting their levels by sampling will follow the same principles as outlined here.

4. When predicting population proportions using samples, one of the key assumptions is that the sampled data is independent. However, multiple cytosines on the same sequencing read are not independent; DNA methylation (or lack thereof) often occurs in localized patches such as transposable elements or CpG islands in vertebrates [24]. Accordingly, cytosines that are closely linked to each other often show the same methylation value. This might be comparable to election polling by phone survey, and instead of just asking the person who picks up the phone who they will vote for, the surveyor would also ask who other people in the household are going to vote for (i.e., often the same candidate). Because the independence of data in the survey is violated, power is effectively reduced. Indeed, from the bluehead wrasse example, when all CGs on a single read are examined, the empirical bootstrap sampling method predicts a higher number of mapped CGs are required to achieve a given margin of error compared to the theoretical model (Fig. 3b, red and blue dots). However, if only the first CG is counted per CG-containing read, non-independent data are removed, and the theoretical estimator and empirical model predict a very similar margin of error (Fig. 3b, orange and green dots).

Acknowledgements

The authors would like to thank all members of the Hore laboratory for their contribution to refining the low-coverage WGBS protocol, and Andrew Gray for providing statistical advice. We are indebted to Bjorn Oback and Pavla Turner for providing the bovine samples. The Hore laboratory is supported by funding from the University of Otago (111899.01.R.LA) and the Marsden Fund (18-UOO-200) and the Ministry for Business, Innovation and Enterprise (UOOX1909).

References

1. Li E, Bestor TH, Jaenisch R (1992) Targeted mutation of the DNA methyltransferase gene results in embryonic lethality. *Cell* 69:915–926. 0092-8674(92)90611-F
2. Okano M, Bell DW, Haber DA et al (1999) DNA methyltransferases Dnmt3a and Dnmt3b are essential for de novo methylation and mammalian development. *Cell* 99:247–257. S0092-8674(00)81656-6
3. Goll MG, Bestor TH (2005) Eukaryotic cytosine methyltransferases. *Annu Rev Biochem* 74:481–514. <https://doi.org/10.1146/annurev.biochem.74.010904.153721>
4. Jang HS, Shin WJ, Lee JE et al (2017) CpG and non-CpG methylation in epigenetic gene regulation and brain function. *Genes (Basel)* 8:148. <https://doi.org/10.3390/genes8060148>
5. Ortega-Recalde O, Hore TA (2019) DNA methylation in the vertebrate germline: balancing memory and erasure. *Essays Biochem.* <https://doi.org/10.1042/EBC20190038>

6. Ehrlich M (2009) DNA hypomethylation in cancer cells. *Epigenomics* 1:239–259. <https://doi.org/10.2217/epi.09.33>
7. Kurdyukov S, Bullock M (2016) DNA methylation analysis: choosing the right method. *Biology* (Basel) 5. <https://doi.org/10.3390/biology5010003>
8. Lister R, O'Malley RC, Tonti-Filippini J et al (2008) Highly integrated single-base resolution maps of the epigenome in Arabidopsis. *Cell* 133:523–536. [j.cell.2008.03.029](https://doi.org/10.1016/j.cell.2008.03.029)
9. Lister R, Pelizzola M, Dowen RH et al (2009) Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature* 462:315–322. <https://doi.org/10.1038/nature08514>
10. Li N, Ye M, Li Y et al (2010) Whole genome DNA methylation analysis based on high throughput sequencing technology. *Methods* 52:203–212. [j.jymeth.2010.04.009](https://doi.org/10.1016/j.jymeth.2010.04.009)
11. Feng S, Cokus SJ, Zhang X et al (2010) Conservation and divergence of methylation patterning in plants and animals. *Proc Natl Acad Sci U S A* 107:8689–8694. <https://doi.org/10.1073/pnas.1002720107>
12. Popp C, Dean W, Feng S et al (2010) Genome-wide erasure of DNA methylation in mouse primordial germ cells is affected by AID deficiency. *Nature* 463:1101–1105. <https://doi.org/10.1038/nature08829>
13. Ortega-Recalde O, Day RC, Gemmell NJ et al (2019) Zebrafish preserve global germline DNA methylation while sex-linked rDNA is amplified and demethylated during feminisation. *Nat Commun* 10:3053. <https://doi.org/10.1038/s41467-019-10894-7>
14. Todd EV, Ortega-Recalde O, Liu H et al (2019) Stress, novel sex genes, and epigenetic reprogramming orchestrate socially controlled sex change. *Sci Adv* 5:eaaw7006. <https://doi.org/10.1126/sciadv.aaw7006>
15. Bewick AJ, Hofmeister BT, Lee K et al (2015) FASTmC: a suite of predictive models for nonreference-based estimations of DNA methylation. *G3* (Bethesda) 6:447–452. <https://doi.org/10.1534/g3.115.025668>
16. Miura F, Enomoto Y, Dairiki R et al (2012) Amplification-free whole-genome bisulfite sequencing by post-bisulfite adaptor tagging. *Nucleic Acids Res* 40:e136. <https://doi.org/10.1093/nar/gks454>
17. Peat JR, Smallwood SA (2018) Low input whole-genome bisulfite sequencing using a post-bisulfite adaptor tagging approach. *Methods Mol Biol* 1708:161–169. https://doi.org/10.1007/978-1-4939-7481-8_9
18. Miura F, Ito T (2015) Highly sensitive targeted methylome sequencing by post-bisulfite adaptor tagging. *DNA Res* 22:13–18. <https://doi.org/10.1093/dnares/dsu034>
19. Oberacker P, Stepper P, Bond DM et al (2019) Bio-On-Magnetic-Beads (BOMB): open platform for high-throughput nucleic acid extraction and manipulation. *PLoS Biol* 17: e3000107. <https://doi.org/10.1371/journal.pbio.3000107>
20. Efron B (1992) Bootstrap methods: another look at the jackknife. In: Kotz S, Johnson NL (eds) *Breakthroughs in statistics: methodology and distribution*. Springer New York, New York, NY, pp 569–593
21. Ziller MJ, Hansen KD, Meissner A et al (2015) Coverage recommendations for methylation analysis by whole-genome bisulfite sequencing. *Nat Methods* 12:230–232, 1 p following 232. <https://doi.org/10.1038/nmeth.3152>
22. Yu M, Hon GC, Szulwach KE et al (2012) Base-resolution analysis of 5-hydroxymethylcytosine in the mammalian genome. *Cell* 149:1368–1380. [j.cell.2012.04.027](https://doi.org/10.1016/j.cell.2012.04.027)
23. Booth MJ, Ost TWB, Beraldi D et al (2013) Oxidative bisulfite sequencing of 5-methylcytosine and 5-hydroxymethylcytosine. *Nat Protoc* 8:1841–1851. <https://doi.org/10.1038/nprot.2013.115>
24. Schmitz RJ, Lewis ZA, Goll MG (2019) DNA methylation: shared and divergent features across eukaryotes. *Trends Genet. j. tig.* 2019.07.007

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.