Target-enriched enzymatic methyl sequencing: flexible, scalable and inexpensive hybridization capture for quantifying DNA methylation

Dustin R. Rubenstein[a,b,1] and Joseph Solomon[a]

[a] Department of Ecology, Evolution and Environmental Biology, Columbia University, New York, NY 10027, USA
[b] Center for Integrative Animal Behavior, Columbia University, New York, NY 10027, USA

[1] Corresponding Author: Phone: +1-212-854-4881; Email: dr2497@columbia.edu

Running Head: Targeted capture to study DNA methylation

Abstract

The increasing interest in studying DNA methylation to understand how traits or diseases develop requires new and flexible approaches for quantifying methylation states in a diversity of organisms. Here, we develop TEEM-Seq (target-enriched enzymatic methyl sequencing), a method that combines enzymatic methyl sequencing with a custom-designed hybridization capture bait set that can be scaled to reactions including large numbers of samples in any species for which a reference genome is available. TEEM-Seq is able to quantify DNA methylation states similarly well to the more traditional approaches of whole-genome and reduced-representation sequencing, and the downstream bioinformatic analysis is the same as for any sequence-based approach to studying DNA methylation. By maximizing the number of samples in the hybridization reaction and minimizing sequencing depth, TEEM-Seq is an inexpensive and flexible sequence-based approach for quantifying DNA methylation in species where other capture-based methods are unavailable or too expensive, particularly for non-model organisms.

Keywords: TEEM-Seq, epigenetics, DNA methylation, CpG site, hybridization capture, target enrichment

Introduction

There is increasing recognition in fields as diverse as ecology, evolutionary biology, conservation biology, neuroscience, and biomedicine that epigenetics is critical to understanding how traits, diseases, or other phenotypes develop (Waddington 1942, Jablonka and Lamb 2006, Cavalli and Heard 2019, Rey et al. 2020). One of the most widely studied epigenetic mechanisms underlying phenotypic variation is DNA methylation (Robertson and Wolffe 2000, Bossdorf et al. 2008), the addition of a methyl group to a DNA nucleotide that impacts gene expression through transcriptional silencing (Bird 1986). In mammals and many other vertebrates, DNA methylation often occurs at CpG sites, or regions of DNA where a cytosine nucleotide is followed by a guanine nucleotide. With the increasing availability of reference genomes for non-model organisms and the continually decreasing costs of high-throughput sequencing, sequencing-based approaches are becoming the preferred method for studying DNA methylation across the diversity of life because they can provide quantitative values of CpG methylation states (Harris et al. 2010). Sequence-based approaches to studying DNA methylation range from whole- or reduced-representation sequencing of genomes to sequencing one or a few candidate genes, yet the performance and cost of different sequence-based approaches to quantifying and mapping DNA methylation varies greatly (Bock et al. 2010, Consortium 2016, Crary-Dooley et al. 2017, Gouil and Keniry 2019). Whole-genome sequencing for DNA methylation profiling remains the gold standard, but is costly because of the necessary sequencing depth (Harris et al. 2010). In contrast, reduced-representation approaches are cheaper than whole-genome sequencing because they cover only a fraction of the genome, but since such approaches require restriction enzymes to target specific genomic regions (e.g., CpG islands), they inevitably miss potentially important CpG sites in genomic regions of interest (Husby 2022). Despite the differences in cost and coverage, most whole-genome and reduced-representation sequencing approaches share the chemical conversion of unmethylated cytosines to uracils with bisulfite treatment (Gu et al. 2011), a method that enables quantification of methylation states but that can fragment DNA and complicate downstream library preparation and even bias results (Tanaka and Okamato 2007, Gouil and Keniry 2019).

As the number of studies of DNA methylation grows, it becomes increasingly important to develop new methods that can quantitatively measure CpG methylation states over large regions of the genome in an efficient yet cost-effective way. Such approaches exist for humans and some model organisms in the form of bead or probe panels, but flexible methods for studying targeted genomic regions in any organism are much rarer. Although hybridization capture using biotinylated oligonucleotide baits (probes) to hybridize specific regions of interest across the genome of any species have become standard in molecular biology (Kozarewa et al. 2015), their application in epigenetics has been more limited (Lee et al. 2011, Li et al. 2015, Masser et al. 2016). Commercial approaches such as Agilent's SureSelect^{XT} Methyl-Seq Target Enrichment System and Roche's SeqCap Epi Enrichment System are neither particularly scalable nor cost-effective, and both use bisulfite conversion. Yet, with the recent development of alternative approaches to bisulfite treatment for converting DNA (Vaisvila et al. 2020), newer methods of hybridization capture may be viable for quantitatively studying DNA methylation, particularly in non-model organisms.

Here, we develop a flexible, scalable, and inexpensive method called target-enriched enzymatic methyl sequencing (TEEM-Seq) that uses hybridization capture to quantitively measure CpG methylation states. We achieve this goal by combining enzymatic methyl sequencing (EM-seq) (Vaisvila et al. 2020) with a custom-designed, species-specific hybridization capture bait set that can be scaled to reactions that include much larger numbers of samples than other currently available methods and that only requires low amounts of input DNA to capture both 5mC and 5hmC methylation. EM-seq effectively overcomes DNA fragmentation caused by bisulfite treatment, and a custom bait set allows for cost-effective and efficient targeting of (multiple) specific genomic regions. In this study, we selected genomic regions encompassing selected genes known to be associated with the hypothalamic-pituitary-adrenal axis (HPA), the hypothalamic–pituitary–gonadal axis (HPG), and DNA methyltransferase activity. We chose these specific HPA and HPG genes because of the critical roles that they play in the avian stress response and in reproduction in vertebrates (Sapolsky et al. 2000). After designing baits targeting all of the putative promoters (4 kb upstream of the transcription start site) and some of the exons of 22 of these genes, we developed the TEEM-Seq protocol using DNA extracted from the blood of the superb starling (*Lamprotornis superbus*), a passerine bird with a published reference genome (Rubenstein et al. 2021). We also tested our starling bait set performance on another passerine species, the house sparrow (*Passer domesticus*). To gauge the effectiveness and efficiency of TEEM-Seq against more traditional methods of quantifying DNA methylation, we compared data from some of the same individual starlings to data from whole-genome enzymatic methyl sequencing (WGEM-Seq) and/or reduced-representation bisulfite sequencing (RRBS). We found that necessary sequencing coverage levels were similar to RRBS, but an order of magnitude lower than for WGEM-Seq for similarly high-quality coverage of complete target regions. Ultimately, we demonstrate the utility of TEEM-Seq for quantifying DNA methylation in an efficient and cost-effective manner that should be applicable to studying any organism in a variety of tissue and sample types, including formalin-fixed, paraffin-embedded (FFPE) samples and those from degraded or ancient DNA.

Results

TEEM-Seq metrics

Our results from a pool of 48 libraries prepared using avian DNA from whole blood in a single capture reaction sequenced on an Illumina HiSeq at 110 Gb (2.3 Gb per sample) indicated that target read coverage for the superb starling samples prior to deduplication ranged from a mean of 400-4300x, with an overall mean of 1700x for both DNA strands. Moreover, background, non-target coverage was mostly scattered but evenly distributed single strand alignment, with a mean of 22x. We determined that while duplication of total reads was 90-95% according to Bismark deduplication reports (versus 15-17% duplicated reads for three WGEM-Seq starling samples, which suggests that sequencing optimization for this protocol might first focus post-capture before reducing pre-capture PCR cycles and per-sample template for the capture reaction), TEEM-Seq target region read coverage depth after deduplication ranged from an average per target of 20-270x, with an overall mean coverage depth for all targets of 173x (Figure 1). In contrast, WGEM-Seq target region coverage after deduplication ranged from an average per target of 2-41x, with an overall mean coverage of 15x (Figure 1). Figure 2 illustrates deduplicated read coverage depth for TEEM-Seq and WGEM-Seq for three representative target regions using the Integrative Genomics Viewer (IGV) v2.12.3 (Robinson et al. 2011). In

addition, the average mean depth of TEEM-Seq coverage after deduplication for just the seven starling samples also included in the RRBS study ranged from 30-374x for the lowest to highest depth promoter targets, with an overall average of 248x. For those same RRBS libraries, the average mean coverage depth per target promoter region (non-deduplication) ranged from 0-17x, with an overall mean of 7x. Although RRBS coverage is biased toward restriction sites (C/CGG for MspI) and certain targets had zero coverage, on average 120 CpGs were covered above 5x across strictly 4 kb promoter regions (Table S4). With the TEEM-Seq and WGEM-Seq libraries, the average was also coincidentally 120 CpGs above 5x in the three samples compared across methods, though the minimum number of 5x sites covered was never zero for either method (Table S2). Finally, for four samples run as duplicate or triplicate libraries, there were no significant differences in mean DNA methylation between the replicates (all paired $t_{23} < 1.63$, all $P > 0.12$; Table S1).

Assessing bait performance in another avian species

Although our bait set was designed specifically for the superb starling, we included some house sparrow samples to assess starling-derived bait performance in another passerine bird species. For four house sparrow libraries, SAMtools reported that 19 of 23 putative promoter targets had an average percent of covered bases of <50%, and 11 of 23 targets had <25% coverage. In contrast, 20 of 23 targets had an average percent of covered bases >99.5% for the 44 superb starling libraries, with *POMC* having 96% covered bases, and *NR3C1* and *VTG* having reference gaps of 488 bp and 182 bp, respectively, but still >99% coverage elsewhere in the putative promoters. The maximum average for the target with the highest coverage was 71% in the house sparrow, and the mean depth of coverage after deduplication ranged from an average across samples of 0-39.5x (versus 20-270x for the superb starling) for the lowest to highest depth promoter targets, with an overall average of 9.3x (versus 173x for the superb starling). These results suggest that superb starling bait sequence was not well conserved for most targets in house sparrows, and that species-specific baits should be designed for each study organism using its own reference genome whenever possible.

TEEM-Seq comparison and validation

To compare and validate TEEM-Seq against more traditional methods of analysis for DNA methylation, we compared many of the same starling DNA samples prepared for WGEM-Seq and RRBS. At the CpG site-level, TEEM-Seq and WGEM-Seq libraries from the same individual were highly correlated (Figure 3). Moreover, we found no major differences in mean DNA methylation between approaches. CpG methylation varied little at the site-level when absolute value mean differences in DNA methylation for sites were compared across gene bodies and putative promoter regions (Figures 4a-c), or when sites were averaged across target regions, as methylation did not differ between WGEM-Seq and TEEM-Seq data within putative promoter ($F_{1,134} = 0.0013$, $P = 0.97$; Figure 4d and Table S2) or exonic regions ($F_{1,68} = 0.00001$, $P = 0.99$; Table S3), between TEEM-Seq and RRBS data within putative promoter regions ($F_{1,249} = 0.049$, $P = 0.83$; Figure 4e and Table S4), or between WGEM-Seq and RRBS in 20 random 2.5 kb "promoter" regions ($F_{1,114} = 1.36$, $P = 0.25$; Figure 4f and Table S5). Finally, we note that TEEM-Seq data with and without deduplication were nearly identical (Figure S1) and that despite the high number of duplicates in the TEEM-Seq data, comparisons of WGEM-Seq versus TEEM-Seq were qualitatively similar when using TEEM-Seq data with and without deduplication (Tables S2 and S6). As expected, CpG site-level correlations were still higher for

171 TEEM-Seq deduplicated versus non-deduplicated data when using the same sites that were
172 highly correlated between TEEM-Seq and WGEM-Seq data (Figure S1 legend and Figure 3).
173 Overall, mean DNA methylation was comparable in all three methods, indicating that TEEM-
174 Seq is an effective approach for quantifying DNA methylation states in specific regions of the
175 genome.
176
177
178 **Discussion**
179
180 We introduce a flexible, scalable, and inexpensive method for studying DNA methylation in a
181 large number of candidate genes that combines enzymatic conversion of DNA (Vaisvila et al.
182 2020) with a custom-designed hybridization capture bait set, and uses starting input DNA as low
183 as 100 ng (though we expect lower amounts would also work). The TEEM-Seq approach
184 described here is able to quantify DNA methylation states similarly well to the more traditional
185 approaches of whole-genome and reduced-presentation sequencing (Figure 5), and the
186 downstream bioinformatic analysis is the same for all three approaches. We show that TEEM-
187 Seq is both reliable and repeatable, as duplicate libraries from the same samples did not vary
188 much. Although we demonstrate its effectiveness on DNA from birds, TEEM-Seq should work
189 on DNA from any organism, as well as from DNA extracted from tissues other than blood (Siller
190 2022). And because it employs enzymatic rather than chemical conversion of DNA (i.e., bisulfite
191 treatment), we also expect that TEEM-Seq will be useful for studying formalin-fixed, paraffin-
192 embedded (FFPE) samples and degraded or ancient DNA, particularly samples from museum
193 specimens that are likely to already be partially degraded and only damaged further by traditional
194 bisulfite conversion (Rubi et al. 2020).
195
196 One of the greatest strengths of TEEM-Seq is its scalability. Although we only scaled our
197 capture reaction to 48 libraries here, we note that in a parallel TEEM-Seq study using a similar
198 bait set, capture scaled consistently with a 96-library pool (Siller 2022). Moreover, New England
199 Biolabs currently produces 120 total unique indexes across their 24- and 96-sample EM-seq kits,
200 meaning that TEEM-Seq should be readily able to scale to a 120-library pool, and perhaps even
201 further with the development of additional unique indexes. We caution, however, that there is
202 likely to be a tradeoff in the number of baits used in the bait set and the number of samples used
203 in the hybridization capture, though assessing this was beyond the scope of the current study.
204 Since myBaits 80 nt baits are designed to tolerate >20% target divergence in conventional DNA-
205 Seq, it is likely that the eight methylation-state targets myBaits recommended for this pilot bait
206 design in addition to the unmodified targets were to some degree redundant. If this proves to be
207 the case, costs could be reduced further in bait design, though assessing the capture performance
208 and mismatch allowance of the unmodified baits and redundancy of methylation-state bait
209 permutations was beyond the scope of this study. Although we cannot directly assess bias in
210 whether and how specific baits hybridize and the affect on methylation levels, under the current
211 design parameters, no systemic bias was evident in capture products. Ultimately, TEEM-Seq is
212 highly scalable because the conversion of unmethylated cytosines to uracils is done prior to the
213 hybridization reaction. In contrast, Agilent's SureSelect^{XT} Methyl-Seq Target Enrichment
214 System performs hybridization capture before library pooling, which limits the number of
215 samples to a fraction of that possible in each TEEM-Seq capture reaction (Lee et al. 2011). And
216 while Roche's SeqCap Epi Enrichment System (Li et al. 2015) performs hybridization capture

after bisulfite conversion, it is also currently limited to a fraction of those indexes available for EM-seq, which similarly limits the number of samples in each capture reaction.

Finally, TEEM-Seq is more cost-effective and per sample data-efficient than not only other commercial capture-based approaches, but also than whole-genome sequencing. It is comparable in cost to reduced-representation sequencing. For an avian species with a genome size of roughly 1.07 Gb (Rubenstein et al. 2021), we sequenced 2.3 Gb per sample for TEEM-Seq, 1.3 Gb per sample for RRBS, and 27.5 Gb per sample for WGEM-Seq. We note, however, that even with an order of magnitude lower sequencing demand for TEEM-Seq compared to WGEM-Seq, the specificity of the bait set meant that we routinely produced 100x coverage or more of the targets, including control regions, and all target regions had average coverage of at least 20x after deduplication. Sequencing depth for TEEM-Seq could be further reduced with additional optimization and perhaps with a reduction in bait redundancy. In addition, since we were able to scale EM-seq kit reagents down to three quarter volumes, by maximizing the number of samples per total capture pool (up to 120 samples per reaction with current index availability), TEEM-Seq becomes an inexpensive method for accurately quantifying DNA methylation states across continuous regions of the genome in a large sample for a diversity of organisms. Finally, all three methods are roughly similar in terms of library preparation time, with RRBS and WGEM-Seq taking roughly two days and TEEM-Seq three to four days because of the added capture step.

Ultimately, we expect that TEEM-Seq could replace traditional approaches for studying single candidate genes (e.g., amplicon sequencing or pyrosequencing), which are time and labor intensive, and be effectively paired with other DNA methylation quantification methods (e.g., whole-genome or reduced-representation sequencing) to increase project sample sizes. Moreover, TEEM-Seq can be paired with RNA-Seq to examine how DNA methylation in promoters or other regulatory regions is related to the expression of individual genes or gene networks. As enzymatic conversion of DNA and EM-seq approaches continue to expand and replace bisulfite conversion and sequencing (e.g., Feng et al. 2020, Han et al. 2021, Liu et al. 2021, Morrison et al. 2021), we expect that TEEM-Seq can be further optimized to improve efficiency, scale to ever larger numbers of samples, and use other bait set designs. Ultimately, because of its flexibility, scalability, and cost-effectiveness, TEEM-Seq should prove useful to researchers in a number of biological disciplines for rapidly quantifying DNA methylation states in any organism of interest in order to link epigenetics to the development of phenotypes.

Methods

Sample collection

Superb starlings (N = 39) were captured using baited traps or mist nests at the Mpala Research Centre, Laikipia, Kenya (0° 17′N, 37° 52′E). Whole blood was collected in Queens lysis buffer (Seutin et al. 1991) and genomic DNA was extracted using a DNeasy Blood & Tissue Kit (Qiagen). All research was approved by the Institutional Animal Care and Use Committee at Columbia University (AC-AAAW6451), as well as the Kenyan National Commission for Science, Technology and Innovation, the Kenyan National Environmental Management Authority, the Kenya Wildlife Service, and the Mpala Research Centre.

263 ***Library preparation***
264 After DNA was quantified using a Qubit 3.0 Fluorometer (Thermo Fisher Scientific), 100-150 ng
265 was sheared to 250-275 bp on a Covaris ME220 and inspected on an Agilent 4200 TapeStation.
266 The sheared DNA was then used as a template for libraries prepared with a NEBNext Enzymatic
267 Methyl-seq Kit (EM-seq) (New England Biolabs). We note that previous attempts to develop a
268 hybridization capture protocol using bisulfite-converted DNA were unsuccessful, presumably
269 because chemical conversion led to fragmented DNA and failed hybridization reactions.
270 However, the enzymatic conversion overcame challenges associated with fragmented DNA from
271 bisulfite conversion and allowed for successful hybridization capture. Following the EM-seq Kit
272 protocol, sheared DNA was end repaired at 20°C, ligated with an EM-seq adaptor at 20°C, and
273 then used as template in a TET2 reaction at 37°C for oxidation of 5-Methylcytosines and 5-
274 Hydroxymethylcytosines. The TET2 converted DNA was denatured with formamide at 85°C,
275 underwent APOBEC cytosine deamination at 37°C, and was then PCR amplified with index
276 primers and NEBNext Q5U Master Mix at an annealing temperature of 62°C. Since we intended
277 to perform downstream in-solution hybridization capture, we increased library yield through 12
278 total PCR cycles for the enzymatic conversion amplification compared with the standard
279 maximum of eight cycles. Although we did not vary the number of PCR cycles in this study, this
280 could be optimized in the future to possibly reduce PCR duplicates. Whereas the duplicate rate
281 increased most in the hybridization reaction and amplification, it remains unknown whether EM-
282 seq duplicates were compounded in the capture process through bait bias toward more highly
283 represented duplicate molecules. We also increased all SPRI bead volumes by 50%,
284 supplementing the included sample purification beads with AMPure XP beads (Beckman
285 Coulter) at the PCR clean-up stage. For post-ligation clean-up, we used a ratio of beads to
286 sample volume of 1.76:1 instead of the NEB EM-seq standard protocol's 1.18:1; for the TET2
287 reaction cleanup, we used a ratio of 2.65:1 instead of 1.76:1; for the APOBEC deanimation
288 clean-up we used 1.5:1 instead of 1:1; and for post-amplification clean-up we used 1.35:1 instead
289 of 0.9:1. We increased the ratio of beads to sample volume because we intended the EM-seq
290 libraries for a myBaits capture pool, not for direct sequencing. In addition, we varied kit reagent
291 volumes down to half of those recommended, ultimately settling on three quarter volumes for
292 our final reactions. Importantly, results with three quarter volumes of reagents were qualitatively
293 similar to those with full volumes (data not shown). Finally, we Qubit-quantified the libraries,
294 inspected them on the TapeStation, and pooled 25 ng from each library into 1.2 µg total for the
295 pool of 48 libraries.
296
297 ***Bait design***
298 Target sequences for bait design were identified from the superb starling reference genome
299 (CU_Lasu_v1) (Rubenstein et al. 2021). Baits targeted 23 regions (*GNRHR2* had two separate
300 regions) of 4 kb in length upstream of the transcription start site (i.e., putative promoter regions)
301 and exons (excluding introns) from a subset of genes known to be associated with the HPA axis,
302 the HPG axis, and DNA methyltransferases (Table 1). Biotinylated RNA baits were then
303 prepared commercially using myBaits v4.01 Custom 1-20K 16 Reaction Kits (Daicel Arbor
304 Biosciences). We submitted 71 sequences (113,940 bp) for 80 nt bait design at 2x tiling density.
305 For each bait candidate, in addition to the unmodified original, myBaits designed eight additional
306 genome variants to reasonably cover conversion permutations (i.e., all methylated, a random
307 50% CpGs methylated, the other 50% CpGs methylated, unmethylated, and sense/antisense for
308 each version), resulting in 21,792 baits, all of which were quality-assessed and filtered by

myBaits Design based on masked repeats in the original sequences (repeatmasker [engine: cross_match, organism: Aves] resulting in 5.4% of total sequences masked), GC content, and BLAST hits on the zebra finch (*Taeniopygia guttata*), European starling (*Sturnus vulgaris*), collared flycatcher (*Ficedula albicollis*), and great tit (*Parus major*) reference genomes for the final myBaits kits (Tables S7 and S8). In addition to unmodified versions of each genome, myBaits blasted bait candidates against sense strand unmethylated (i.e., all Cs converted to Ts) and anti-sense strand unmethylated (i.e., all Gs converted to As) versions of each genome, for a total of 12 genome variants. The top 500 blast hits (by bit score) were tabulated for each bait candidate. The final bait set contained sequences with <25% repeat-masking, 10-80% GC content, and ≤10 max BLAST hits across all 12 genome variants to avoid non-specific targets.

Hybridization capture

A pool of 48 libraries contained 39 starling samples, two house sparrow samples (see below), and a total of seven replicate libraries from each species. Eight superb starling samples were used in a previously unpublished DNA methylation study that used RRBS. Three of these eight were included in duplicate libraries and one in triplicate libraries. The remaining 31 additional superb starling samples were included to assess target capture performance when scaling up to a pool of 48 libraries. Although our bait set was designed specifically for the superb starling, we also included two house sparrow samples run in duplicate to assess starling-derived bait performance in more highly conserved target regions of another avian species.

We followed the myBaits v4.01 protocol, but lowered the default hybridization and clean up temperature of 65°C to 63°C as myBaits recommends for fragmented DNA. We note, however, that a parallel TEEM-Seq study worked similarly well using a temperature of 65°C (Siller 2022). Hybridization beads bound to the pooled library-blocker mix were cleaned with three washes, and the resuspended bead-bound DNA was then amplified with 16 PCR cycles in a KAPA HiFi reaction using a 60°C annealing temperature and a one minute extension step at 72°C. The amplified capture pool was subsequently cleaned up with AMPure XP beads and sequenced at 2x150 bp reads with 5% PhiX spike-in in one lane (110 Gb) of an Illumina HiSeq 4000 at the Novogene Corporation, Sacramento, CA.

Control sequences

Bait library sequence alignments to NEB pUC19 methylated and lambda unmethylated controls were sparse in background coverage because controls were not specifically targeted in our initial bait set. Overall, coverage in control regions ranged from 0-42x without deduplication, averaging 18x for pUC19 and 25x for lambda, but methylation calls were unreliable with deduplication. Although our three EM-seq libraries that were also used for WGEM-Seq passed control checks, and the RRBS libraries used to compare CpG methylation percentages were spiked with lambda controls, we confirmed that a known low methylation (average 1.5%) 57 bp region with 12 CpGs 10 bp upstream of Arginine vasopressin receptor 1A (Rubenstein and Solomon unpublished data) also had means within 1% of this average in the seven TEEM-Seq libraries used for comparison to other methods. Nonetheless, we then specifically targeted pUC19 and lambda control sequences in a parallel TEEM-Seq study using a house sparrow-specific bait set to ensure that we would have reliable EM-seq conversion estimates for all samples (Siller 2022), in the absence of comparable libraries for these samples prepared with alternative methods. In that study, we found that lambda unmethylated and pUC19 methylated control baits were highly specific,

355 targeting the first 1 kb for pUC19 and the first 2 kb for lambda included in the bait design. For a
356 capture pool of 96 libraries, SAMtools v1.14 (Danecek et al. 2021) reported coverage for the 1
357 kb pUC19 alignments as a range of 18,582-538,322x (mean = 96,689x) for non-deduplicated
358 BAMs and a range of 4,540-86,441x (mean = 27,929x) for deduplicated BAMs. For the 2 kb
359 Lambda alignments, SAMtools reported a range of 23,445-537,148x (mean = 115,380x) for non-
360 deduplicated BAMs and a range of 4,984-88,019x (mean = 30,943x) for deduplicated BAMs. To
361 ensure requisite control alignments with an untested protocol, and because we only targeted a
362 portion of each control genome, we followed EM-seq recommendations for a higher spike-in
363 with 0.3 ng of pUC19 and 6 ng of lambda controls per 100-150 ng sample. However, given the
364 excessive depth of capture, a spike-in at normal HiSeq levels (a 1:100 dilution, or 0.001 ng
365 pUC19 and 0.02 ng lambda) should be more than sufficient when control regions are targeted.
366 Since the bait design and hybridization reaction had no negative effects on the assessment of
367 control alignments, we recommend that all capture designs for TEEM-Seq libraries follow this
368 protocol and include baits for pUC19 and lambda controls.
369
370 ***TEEM-Seq comparison and validation***
371 To compare and validate TEEM-Seq against more traditional methods of analysis for DNA
372 methylation, we compared many of the same starling DNA samples to those that underwent
373 WGEM-Seq and/or RRBS. For WGEM-Seq, extra material from three of the EM-seq libraries
374 were confirmed clear of adapter dimers on the TapeStation, pooled and sequenced at 2x150 bp
375 reads with 5% PhiX spike-in in 3/4th of a lane (82.5 Gb) on a HiSeq 4000 at Novogene with the
376 same parameters to provide a whole-genome sequence and partial methylome analysis
377 comparison to our targeted sequencing. For a previous RRBS study (Rubenstein and Solomon
378 unpublished data), we used NuGEN Ovation Methyl-Seq kits to prepare RRBS libraries (with
379 added Promega unmethylated cl857 Sam7 lambda control spike-in of 1 ng per 80-120 ng of
380 sample), which were sequenced in several pools of 16 samples per lane (20 Gb) at 1x100 bp
381 reads with 10% PhiX spike-in on a HiSeq 2500 at the Cornell Institute of Biotechnology
382 Genomics Core, Ithaca, NY. Libraries for RRBS were prepared according to kit protocols.
383 Briefly, samples were digested with MspI at 37°C, ligated with bisulfite-compatible barcoded
384 adaptors at 25°C, end repaired at 60°C, bisulfite converted with the Qiagen EpiTect Fast
385 Bisulfite Conversion Kit, PCR amplified with Nugen's standard 12 cycles and 60°C annealing,
386 and purified with Agencourt beads for sequencing. Eight of these RRBS samples were included
387 in the present study, though one failed enzymatic conversion. Data from the remaining seven
388 samples (from different sequencing pools) were used as a comparison to the TEEM-Seq and
389 WGEM-Seq derived data generated from the same individuals.
390
391 ***DNA methylation analysis***
392 Sequencing data were trimmed with the Trim Galore v0.4.2 (Krueger et al. 2021) wrapper of
393 Cutadapt v1.12 (Martin 2011) with standard parameters. As per NuGEN's recommendations,
394 RRBS reads were not trimmed with Trim Galore's RRBS option, but instead trimmed of
395 adaptors with standard parameters and with NuGEN's trimRRBSdiversityAdaptCustomers.py
396 script. Trimmed reads were aligned to the superb starling bisulfite genome reference generated
397 by Bismark v0.19.0 (Krueger and Andrews 2021). WGEM-Seq and TEEM-Seq alignments were
398 deduplicated, and CpG coverage files with methylation percentages (100 * methylated
399 cytosines/total of methylated plus unmethylated cytosines in CpG context) were extracted from
400 alignments with Bismark (non-deduplicated methylation calls were also made for TEEM-Seq

401 data to compare percentages with and without deduplication). Although there does not appear to
402 be a consensus on whether target-enriched libraries should be deduplicated (for amplicon
403 sequencing it is avoided), and because (1) the EM-seq ligation adaptor is specifically designed
404 for the EM-seq prep and (2) EM-seq currently does not support NEB Unique Molecular
405 Identifiers (which would have to be methylated for compatibility with methylation sequencing
406 workflows), our goal for bait coverage was a minimum of 20x per 4 kb promoter target region
407 and 5x for smaller promoter and exonic regions after deduplication.
408
409 To compare DNA methylation quantification methods in superb starlings, Bismark coverage files
410 from TEEM-Seq and WGEM-Seq alignments were intersected with 4kb bait "promoter" and
411 exon ranges (for a separate exon comparison) using Bedtools v2.29.2 (Quinlan and Hall 2010).
412 To confirm read and CpG site coverage, BAM alignments and Bismark coverage tracks were
413 visually inspected with Geneious v10.0.4 using an updated chromosomal-level genome assembly
414 (CU_Lasu_v2; GenBank: GCA_015883425.2) that combined Hi-C data with the previously
415 published superb starling reference genome (Rubenstein et al. 2021). SAMtools coverage was
416 used to calculate mean read coverage stats in bait target ranges. To compare WGEM-Seq and
417 RRBS data, Bedtools was used to generate 200 random 10,000 bp tracks across the starling
418 genome that were then intersected with putative 2.5 kb promoter tracks, leaving us with 27
419 random "promoter" regions. This targeted 2.5kb range was chosen instead of a whole genome
420 comparison because RRBS coverage is dense at restriction sites (e.g., in CG clusters localized
421 just prior to the transcription start site). Visual inspection confirmed that these random regions
422 appeared rich in RRBS coverage. Bismark coverage files from WGEM-Seq and RRBS
423 alignments were likewise intersected with these random regions, and then filtered for unique
424 sites (in case of promoter overlap) above 5x coverage (Ziller et al. 2015). Intersected coverage
425 files for each sample were then matched between each method for common CpG sites. That is,
426 we compared matched sites on a per sample basis within random regions between WGEM-Seq
427 versus RRBS data (resulting in 20 regions with 5x CpG site coverage for all but one sample), or
428 within target regions for WGEM-Seq versus RRBS or WGEM-Seq versus TEEM-Seq data.
429 Thus, WGEM-Seq, TEEM-Seq, and RRBS data were compared at the CpG site-level in target
430 regions, and WGEM-Seq and RRBS data were compared at the CpG site-level in 20 random
431 "promoter" regions that were not included in the bait set. We chose this approach of an
432 additional random 2.5 kb "promoter" region comparison because we expected that some RRBS
433 regions corresponding to target ranges might have sparse coverage in comparison to the high
434 coverage bait alignments. Finally, Bedtools map was used to calculate the mean of Bismark
435 methylation percentages for these regions.
436
437 ***Statistical analysis***
438 DNA methylation values target regions for libraries run in duplicate (or triplicate) were
439 compared using paired t-tests. Pearson correlations were used to compare site-level CpG data for
440 TEEM-Seq and WGEM-Seq libraries. Data for each pairwise comparison of methods of
441 quantifying DNA methylation (TEEM-Seq versus WGEM-Seq, TEEM-Seq versus RRBS, and
442 WGEM-Seq versus RRBS) were analyzed using generalized linearized mixed models
443 (GLMMS), with DNA methylation as the dependent variable, quantification method as the
444 independent variable, and bird ID included as a random effect to account for repeated measures
445 (multiple gene regions per library). All statistical tests were done in JMP Pro v16.2 or R v4.1.3
446 (R Core Team 2022).

Data Availability

Raw genetic data are available at the NCBI Sequence Read Archive (SRA) (BioProject: PRJNA861281). Summarized data are provided in the Supplemental Information in table format.

Computer Code

Code is available at https://doi.org/10.6084/m9.figshare.20634759.

Author Contributions

D.R.R. conceived of the project. D.R.R. and J.S. designed the experiment. D.R.R. collected the samples. J.S. collected and analyzed the data. D.R.R. and J.S. wrote the paper.

Competing Interests

The authors declare no competing interests.

Acknowledgements

We thank W. Watetu, G. Manyaas, and J. Mosiany for assistance collecting superb starlings and the Mpala Research Centre for logistical support in Kenya. We acknowledge A. Devault and Z. Hanf at Daicel Arbor Biosciences for help designing the bait sets. We acknowledge the Sackler Institute for Comparative Genomics at the American Museum of Natural History and the Department of Biological Sciences at Columbia University for access to equipment. We thank S. Siller for helping to organize gene sequences and, along with B. Heidinger, for supplying house sparrow DNA. A. Bendesky and J. Merritt provided constructive feedback on previous versions of this manuscript. This work was supported by the U.S. National Science Foundation grants IOS-1257530 and IOS-1656098.

476 **References**

477 Bird, A. P. CpG-rich islands and the function of DNA methylation. *Nature* **321**, 209-213 (1986).
478 Bock, C., Tomazou, E. M, Brinkman, A. B., Müller, F., Simmer, F., Gu, H. et al. Quantitative
479 comparison of genome-wide DNA methylation mapping technologies. *Nature*
480 *Biotechnology* **10**, 1106-1116 (2010).
481 Bossdorf, O., Richards, C. L. & Pigliucci, M. Epigenetics for ecologists. *Ecology Letters* **11**,
482 106-115 (2008).
483 Cavalli, G. & Heard, E. Advances in epigenetics link genetics to the environment and disease.
484 *Nature* **571**, 489-499 (2019).
485 Consortium, B. Quantitative comparison of DNA methylation assays for biomarker development
486 and clinical applications. *Nature Biotechnology* **34**, 726-37 (2016).
487 Crary-Dooley, F. K., Tam, M. E., Dunaway, K. W., Hertz-Picciotto, I., Schmidt, R. J. & LaSalle
488 J. M. Comparison of existing global DNA methylation assays to low-coverage whole-
489 genome bisulfite sequencing for epidemiological studies. *Epigenetics* **12**, 206–214
490 (2017).
491 Danecek, P., Bonfield, J. K., Liddle, J., Marshall, J., Ohan, V., Pollard, M. O. et al. Twelve years
492 of SAMtools and BCFtools. *GigaScience* **10**, 1-4 (2021).
493 Feng, S., Zhong, Z., Wang, M. & Jacobsen, S. E. Efficient and accurate determination
494 of genome-wide DNA methylation patterns in *Arabidopsis thaliana* with enzymatic
495 methyl sequencing. *Epigenetics & Chromatin* **13**, 42 (2020).
496 Gouil, Q. & Keniry, A. Latest techniques to study DNA methylation. *Essays in Biochemistry* **63**,
497 639–48 (2019).
498 Gu, H., Smith, Z. D., Bock, C., Boyle, P., Gnirke, A. & Meissner, A.Preparation of reduced
499 representation bisulfite sequencing libraries for genome-scale DNA methylation
500 profiling. *Nature Protocols* **6**, 468-481 (2011).
501 Han, Y., Zheleznyakova, G. Y., Marincevic-Zuniga, Y., Kakhki, M. P., Raine, A., Needhamsen,
502 M. & Jagodic, M. 2021. Comparison of EM-seq and PBAT methylome library methods
503 for low-input DNA. *Epigenetics* https://doi.org/10.1080/15592294.2021.1997406 (2021).
504 Harris, R. A., Wang, T., Coarfa, C., Nagarajan, R. P., Hong, C., Downey, S. L. et al. Comparison
505 of sequencing-based methods to profile DNA methylation and identification of
506 monoallelic epigenetic modifications. *Nature Biotechnology* **28**, 1097-1105 (2010).
507 Husby, A. Wild epigenetics: insights from epigenetic studies on natural populations. *Proceedings*
508 *of the Royal Society B* **289**, 20211633 (2022).
509 Jablonka, E. & Lamb, M. J. The changing concept of epigenetics. *Annals of the New York*
510 *Academy of* Sciences **981**, 82-96 (2006).
511 Kozarewa, I., Armisen, J., Gardner, A. F., Slatko, B. E. & Hendrickson, C. L. Overview of target
512 enrichment strategies. *Current Protocols in Molecular Biology* **112**, 7.21.1-7.21.23
513 (2015).
514 Krueger, F., James, F., Ewels, P., Afyounian, E. & Schuster-Boeckler B. TrimGalore v0.4.2.
515 https://github.com/FelixKrueger/TrimGalore (2021).
516 Krueger, F. & Andrews, S. R. Bismark: a flexible aligner and methylation caller for Bisulfite-
517 Seq applications. *Bioinformatics* **27**, 1571–1572 (2011).
518 Lee, E.-J., Pei, L., Srivastava, G., Joshi, T., Kushwaha, G., Choi J.-H. et al. Targeted bisulfite
519 sequencing by solution hybrid selection and massively parallel sequencing. *Nucleic Acids*
520 *Research* **39**, e127–e127 (2011).

521 Li, Q., Suzuki, M., Wendt, J., Patterson, N., Eichten, S. R., Hermanson, P. J. et al. 2015. Post-
522 conversion targeted capture of modified cytosines in mammalian and plant genomes.
523 *Nucleic Acids Research* **43**, e81 (2015).
524 Liu, Y., Hu, Z., Cheng, J., Siejka-Zielinska, P., Chen, J., Inoue, M., Ahmed, A. A. & Song, C.
525 Subtraction-free and bisulfite-free specific sequencing of 5-methylcytosine and its
526 oxidized derivatives at base resolution. *Nature Communications* **12**, 618 (2021).
527 Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads.
528 *EMBnet.journal* **17**, 10-12 (2011).
529 Masser, D. R., Stanford, D. R., Hadad, N., Giles, C. B., Wren, J. D., Sonntag, W. E. et al.
530 Bisulfite oligonucleotide-capture sequencing for targeted base- and strand-specific
531 absolute 5-methylcytosine quantitation. *AGE* **38**, 49 (2016).
532 Morrison, J., Koeman, J. M., Johnson, B. K., Foy, K. K., Beddows, I., Zhou, W. et al. Evaluation
533 of whole-genome DNA methylation sequencing library preparation protocols.
534 *Epigenetics & Chromatin* **14**, 28 (2021).
535 Quinlan, A. R. & Hall I. M. BEDTools: a flexible suite of utilities for comparing genomic
536 features. *Bioinformatics* **26**, 841–842 (2010).
537 R Core Team. R: A language and environment for statistical computing. R Foundation for
538 Statistical Computing, Vienna, Austria; https://www.R-project.org/ (2022).
539 Rey, O., Eizaguirre, C., Angers, B., Baltazar-Soares, M., Sagonas, K., Prunier, J. G. et al.
540 Linking epigenetics and biological conservation: Towards a conservation epigenetics
541 perspective. *Functional Ecology* **34**, 414-427 (2020).
542 Robertson, K. D. & Wolffe, A. P. DNA methylation in health and disease. *Nature Reviews*
543 *Genetics* **1**, 11-19 (2000).
544 Robinson, J. T., Thorvaldsdóttir, H., Winckler, W., Guttman, M., Lander, E. S., Getz, G. &
545 Mesirov, J. P. Integrative Genomics Viewer. *Nature Biotechnology* **29**, 24–26 (2011).
546 Rubenstein, D. R., Corvelo, A., MacManes, M. D., Maia, R., Narzisi, G., Rousaki, A. et al.
547 Feather gene expression elucidates the developmental basis of iridescence in African
548 starlings. *Journal of Heredity* **112**, 417-429 (2021).
549 Rubi, T. L., Knowles, L. L. & Dantzer, B. Museum epigenomics: characterizing cytosine
550 methylation in historic museum specimens. *Molecular Ecology Resources* 20, 1161-1170
551 (2020).
552 Sapolsky, R. M., Romero, L. M. & Munck, A. U. How do glucocorticoids influence stress
553 responses? Integrating permissive, suppressive, stimulatory, and preparative actions.
554 *Endocrine Reviews* **21**, 55-89 (2000).
555 Seutin, G., White, B. N. & Boag, P. T. Preservation of avian blood and tissue samples for DNA
556 analysis. *Canadian Journal of* Zoology **69**, 82–90 (1991).
557 Siller, S. J. Epigenetic modification of the hypothalamic-pituitary-adrenal axis during early life
558 of the house sparrow (*Passer domesticus*). Ph.D. Thesis, Columbia University (2022).
559 Tanaka, K. & Okamato, A. Degradation of DNA by bisulfite treatment. *Bioorganic & Medicinal*
560 *Chemistry Letters* **17**, 1912-1915 (2007).
561 Vaisvila, R., Chaithanya Ponnaluri, V. K., Sin, Z., Langhorst, B. W., Saleh, L., Guan, S. et al.
562 Enzymatic methyl sequencing detects DNA methylation at single-base resolution from
563 picograms of DNA. *Genome Research* **31**, 1280-1289 (2021).
564 Waddington, C. H. The epigenotype. *Endeavor* **81**, 18–20 (1942).

565 Ziller, M. J., Hanson, K. D., Meissner, A. & Aryee, M. J. Coverage recommendations for
566 methylation analysis by whole-genome bisulfite sequencing. *Nature Methods* **12**, 230-
567 232 (2015).

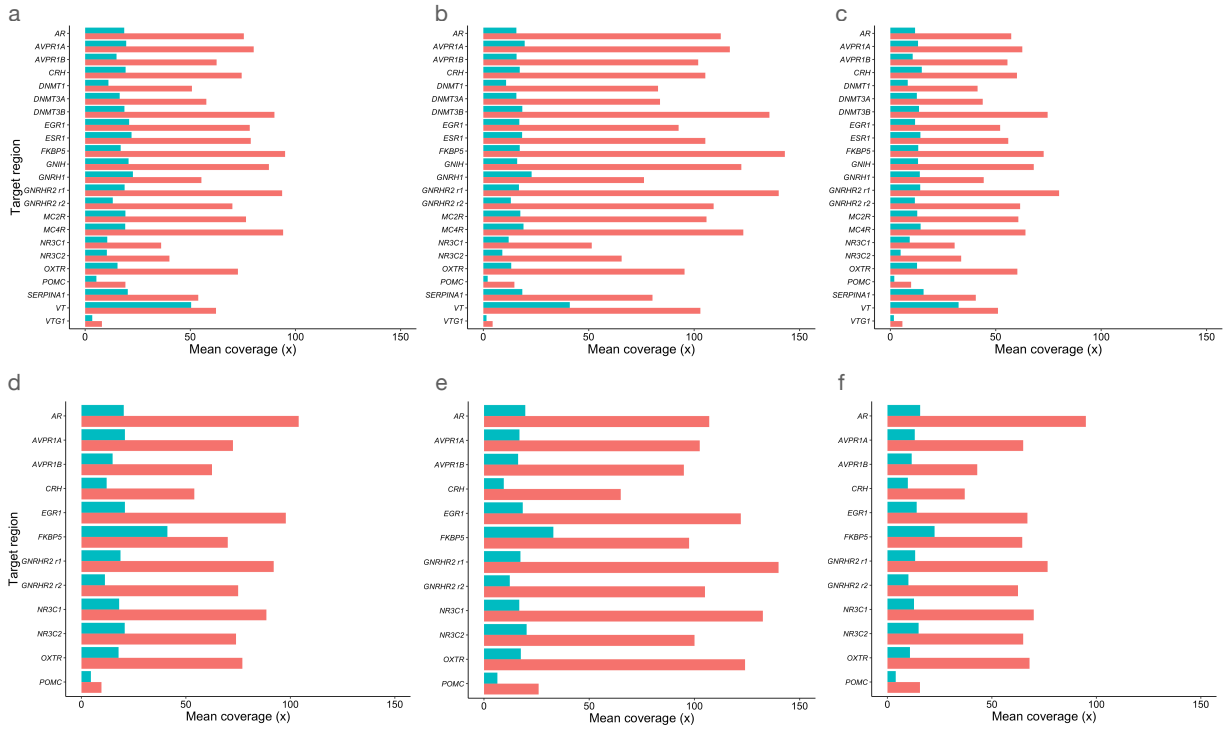


Figure 1. Comparison of mean target region coverage (after deduplication) for target-enriched enzymatic methyl sequencing (TEEM-Seq) and whole-genome enzymatic methyl sequencing (WGEM-Seq) in (a-c) putative promoter and (d-f) exonic regions. Libraries for three of the same individual superb starlings (**a** and **d**) BB-17168, (**b** and **e**) BB-17501, and (**c** and **f**) BB-14232 were sequenced using TEEM-Seq (red bars) and WGEM-Seq (blue bars).

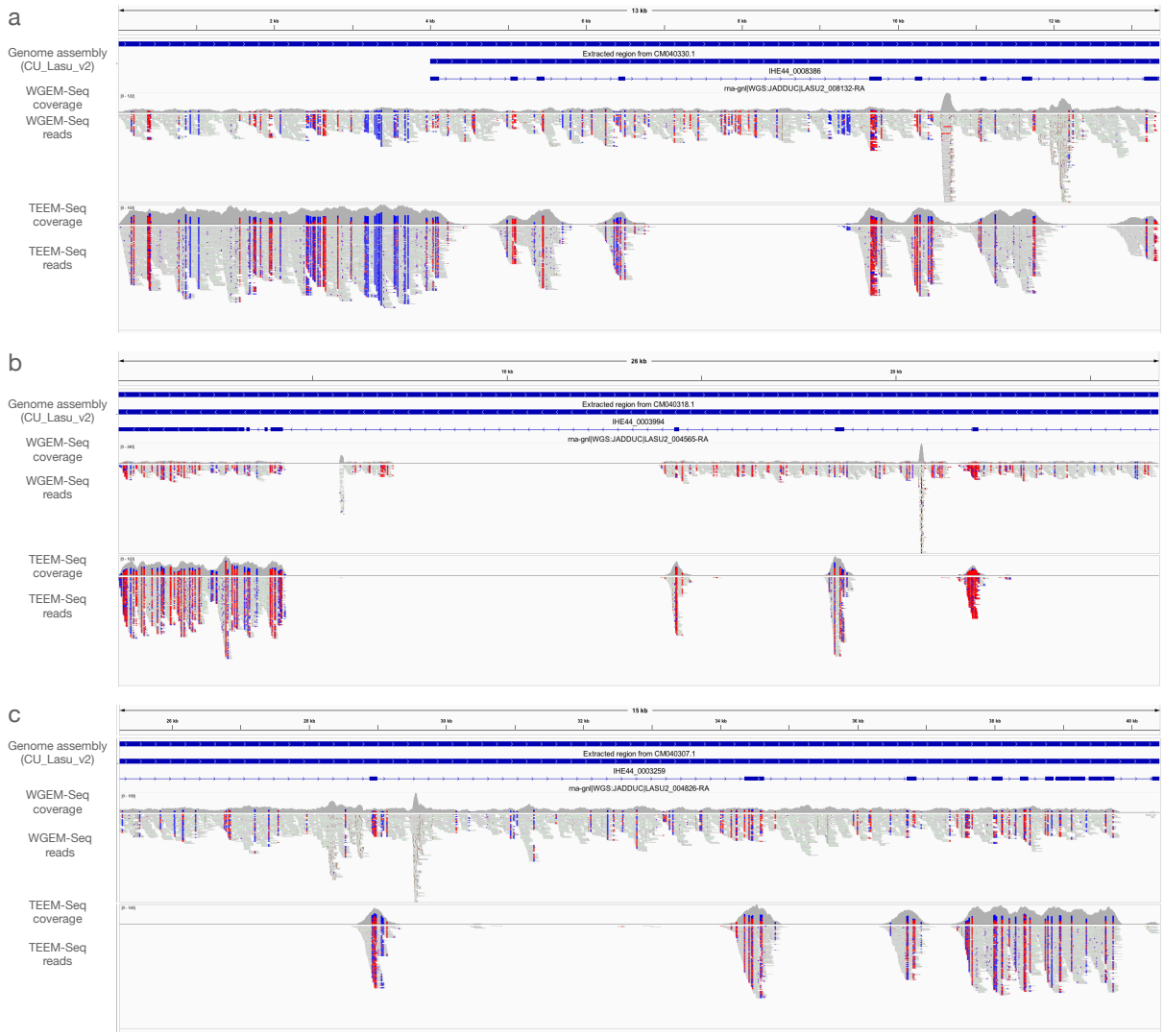
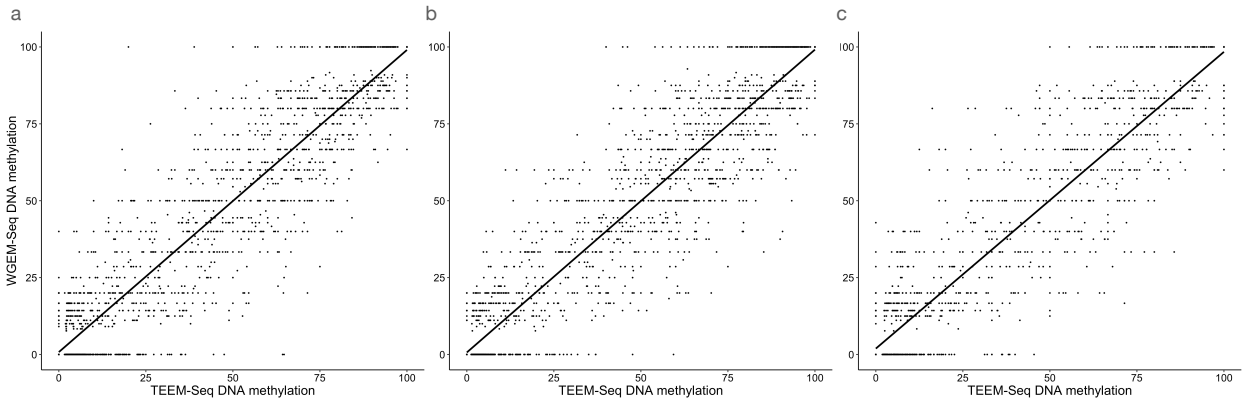


Figure 2. Representative examples of target-enriched enzymatic methyl sequencing (TEEM-Seq) versus whole-genome enzymatic methyl sequencing (WGEM-Seq) coverage. Integrative Genomics Viewer (IGV) logarithm-scale coverage and read alignment plots (for deduplicated data) of superb starling sample BB-17168 multi-exonic targets for (**a**) FK506 binding protein 5 (*FKBP5*), (**b**) the glucocorticoid receptor (*NR3C1*), and (**c**) the androgen receptor (*AR*). *FKBP5* includes the entire promoter and gene region, whereas *NR3C1* and *AR* are focused on the end of each gene region to enable IGV's display of read alignments. Alignments are colored in bisulfite mode by CG, with red indicating methylated sites (non-converted cytosines) and blue indicating unmethylated sites (converted cytosines). Target tracks for the superb starling genome assembly (CU_Lasu_v2) include GenBank sequence ID, locus tag, and protein IDs (with exonic spans also in blue).



588
589
590 **Figure 3. CpG site-level Pearson correlations between target-enriched enzymatic methyl**
591 **sequencing (TEEM-Seq) and whole-genome enzymatic methyl sequencing (WGEM-Seq)**
592 **libraries from the same individual superb starlings. (a)** BB-17168 (R = 0.95, N = 2733, P <
593 0.0001), **(b)** BB-17501 (R = 0.95, N = 2770, P < 0.0001), and **(c)** BB-14232 (R = 0.93, N =
594 1588, P < 0.0001). CpG sites at 5x coverage or above were used for comparisons.

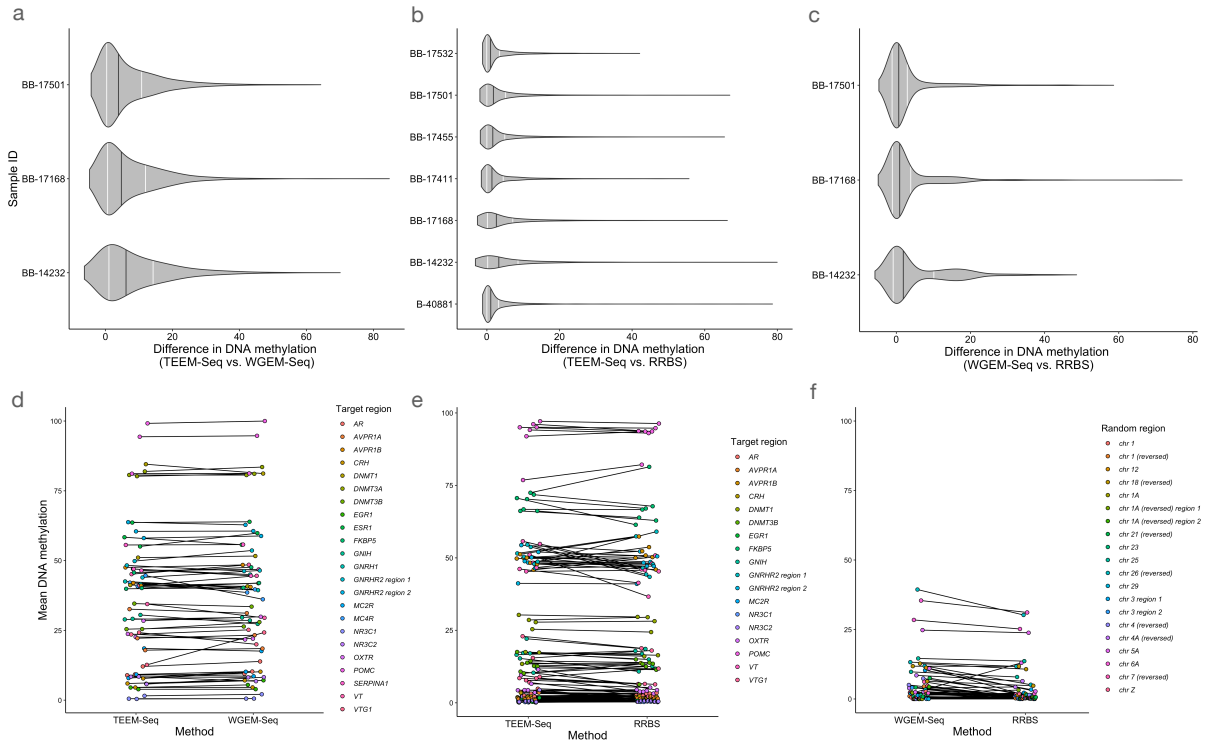


Figure 4. Pairwise comparisons of three methods for quantifying DNA methylation: target-enriched enzymatic methyl sequencing (TEEM-Seq), whole-genome enzymatic methyl sequencing (WGEM-Seq), and reduced-representation bisulfite sequencing (RRBS). (**a** and **d**) Libraries for three of the same individual superb starlings were sequenced using both TEEM-Seq and WGEM-Seq. (**b** and **e**) Libraries for seven of the same individual superb starlings were sequenced using both TEEM-Seq and RRBS. (**c** and **f**) Libraries for three of the same individual superb starlings were sequenced using both WGEM-Seq and RRBS. WGEM-Seq and RRBS data were compared in 20 random 2.5 kb "promoter" regions that were not included in the bait set, as we expected that some RRBS regions corresponding to target ranges might have sparse coverage in comparison to the high coverage bait alignments. In all cases, there were no differences in DNA methylation among quantification methods, as the absolute value mean differences in DNA methylation for promoter regions and gene bodies or random 2.5 kb "promoter" regions did not differ among pairwise comparisons (**a-c**) In each plot, grey violins show the distribution of CpG site absolute value mean differences in DNA methylation for (**a** and **b**) promoter regions and gene bodies or (**c**) random 2.5 kb "promoter" regions, black lines depict the median values, and white lines show the 1st and 3rd quartiles. Additionally, (**d** and **e**) mean DNA methylation levels for each target or (**f**) random region did not significantly differ. Target regions are indicated by different colors according to the legends on panels d-f. See Tables S2-S5 for more detailed information.

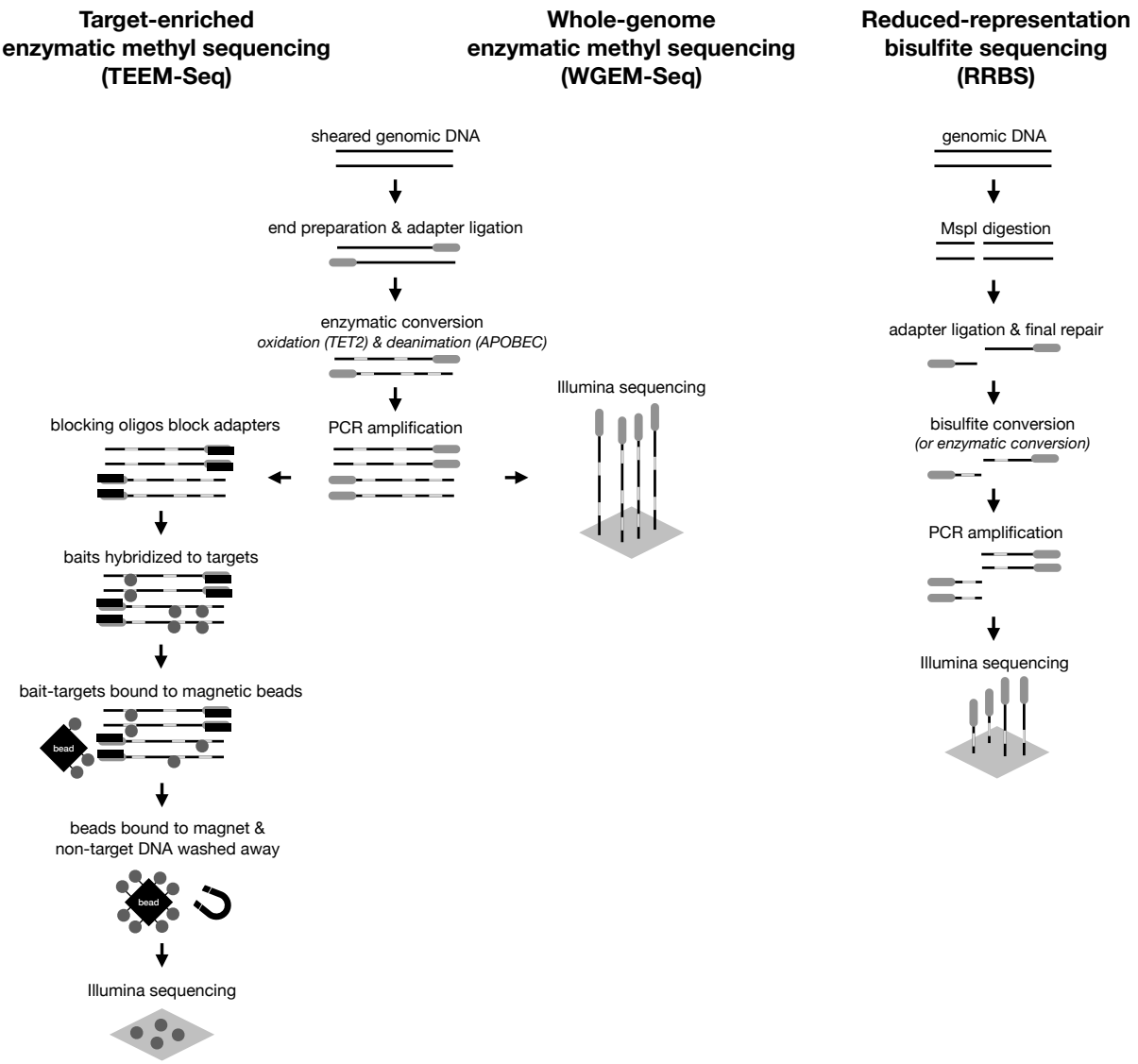


Figure 5. Diagram comparing the workflow for target-enriched enzymatic methyl sequencing (TEEM-Seq), whole-genome enzymatic methyl sequencing (WGEM-Seq), and reduced-representation bisulfite sequencing (RRBS).

621 **Table 1.** Targeted genomic regions for superb starling baits. Although we targeted putative
622 promoter regions of all of the genes (i.e., 4 kb upstream of the transcription start site where
623 annotated), we also targeted complete exonic regions (indicated by *) or gene body regions
624 (based on similarity with Ensembl genome sequences, including the collared flycatcher for
625 *GNRH1* and zebra finch for *MC4R*, indicated by **) in a subset of the genes.
626

Target region	Gene name
*AR**	Androgen receptor
*AVPR1A**	Arginine vasopressin receptor 1A
*AVPR1B**	Arginine vasopressin receptor 1B
*CRH**	Corticotropin releasing hormone
DNMT1	DNA methyltransferase 1
DNMT3a	DNA methyltransferase 3a
DNMT3b	DNA methyltransferase 3b
*EGR1**	Early growth response 1
ESR1	Estrogen receptor 1
*FKBP5**	FK506 binding protein 5
GNIH	Gonadotropin-inhibitory hormone
*GNRH1***	Gonadotropin releasing hormone 1
*GNRHR2**	Gonadotropin-releasing hormone receptor
MC2R	Melanocortin receptor 2
*MC4R***	Melanocortin receptor 4
*NR3C1**	Glucocorticoid receptor
*NR3C2**	Mineralcorticoid receptor
*OXTR**	Oxytocin receptor
*POMC**	Proopiomelanocortin
SERPINA1	Serpin family A member 1
VT	Vasotocin
VTG1	Vitellogenin-1

627