

1 **TITLE**

2 *FAST^mC*: a suite of predictive models for non-reference-based estimations of
3 DNA methylation

4

5 **AUTHORS**

6 Adam J. Bewick¹, Brigitte T. Hofmesiter², Kevin Lee¹, Xiaoyu Zhang³, Dave W.
7 Hall¹, Robert J. Schmitz¹

8

9 **AFFILIATIONS**

10 ¹Department of Genetics, University of Georgia, Athens, GA 30602, USA

11 ²Institute of Bioinformatics, University of Georgia, Athens, GA 30602, USA

12 ³Department of Plant Biology, University of Georgia, Athens, GA 30602, USA

13

14 **CORRESPONDENCE**

15 Robert J. Schmitz, schmitz@uga.edu

16

17

18

19

20 **ABSTRACT**

21 We describe a suite of predictive models, coined *FAST^mC*, for non-reference,
22 cost-effective exploration and comparative analysis of context-specific DNA
23 methylation levels. Accurate estimations of true DNA methylation levels can be
24 obtained from as few as several thousand short-reads generated from whole
25 genome bisulfite sequencing. These models make high-resolution time course or
26 developmental, and large diversity studies practical regardless of species,
27 genome size and availability of a reference genome.

28

29 **KEYWORDS**

30 Epigenetics, DNA methylation, Whole-genome bisulfite sequencing, Methylome,
31 Modeling

32

33 **BACKGROUND**

34 Advances in high-throughput sequencing has allowed for single-base resolution
35 analysis of DNA methylation at cytosines across an entire genome. This was first
36 applied to the model plant *Arabidopsis thaliana* [1],[2] and, since then, has been
37 applied to numerous species, including protists, fungi, insects, anthozoa,
38 tunicates, fish, and mammals [3]-[5]. Currently, DNA methylation is profiled
39 genome-wide by deep, whole-genome bisulfite sequencing (WGBS). The use of
40 a reference genome is essential to inform the methylation status at each

41 cytosine reference position, where a thymine *in lieu* of cytosine indicates an
42 unmethylated cytosine [6]. Thus, absence of a reference genome has
43 prevented rapid, genome-wide analysis of DNA methylation for the majority of
44 known species, and is cost-prohibitive for high-resolution developmental or time-
45 course studies in species with large genomes. To date, several methods exist to
46 accommodate the challenges associated with non-reference based analysis of
47 DNA methylation, but lack cytosine context sequence specificity [7]-[9].

48 Here we present *FAST^mC*, a suite of predictive models that can be used
49 to estimate genome-wide DNA methylation levels at all cytosine sequence
50 contexts without the use of a reference genome. These models assumed a
51 relationship between DNA methylation levels calculated from alignment of
52 WGBS reads to a reference genome (target; m) and from direct assessment
53 from raw WGBS reads (i.e., no alignment to a reference genome) (estimator; \hat{F}).
54 Methylation levels are calculated as the proportion of methylated cytosines to the
55 total number of possible methylated cytosines. The difference between the two
56 variables exists at unmethylated cytosines; the estimator value includes
57 unmethylated cytosines and true thymines when calculating the DNA methylation
58 level. Estimator DNA methylation levels were compared to target levels to
59 determine a relationship, and the strength of which, to confidently

60 predict/extrapolate genome-wide DNA methylation levels for any sample
61 regardless of the availability of a reference genome.

62 Using publicly available data, for species with reference genomes, target
63 and estimator DNA methylation levels for 44 species were used to construct
64 models capable of predicting genome-wide levels of DNA methylation for
65 species without a sequenced genome. Using additional publicly available data
66 from mutants and cell-types known to be different from wild-type samples, we
67 discuss the sensitivity, robustness and utility of the models in terms of CpG DNA
68 methylation, followed by plant- (CHG and CHH) and mammal-specific (CH) DNA
69 methylation.

70

71 **RESULTS AND DISCUSSION**

72 *FAST^mC* is able to detect intraspecific differences in DNA methylation (Fig. 1). In
73 the plant *A. thaliana*, mutants exist that are defective for enzymes that are
74 required for maintenance of CpG DNA methylation – *met1*, *met1+cmt3*, and
75 *vim1+vim2+vim3* – as they have reduced CpG methylation levels compared to
76 wild type [10]. Also, several mutant genotypes for *met1* show different degrees of
77 loss of CpG DNA methylation compared to each other: (i) An original *met1*
78 mutant genotype (high loss); (ii) A *met1* heterozygous mutant genotype (*met1 +/-*
79) (intermediate loss); and (iii) A recovered genotype (*MET1 +/+*) from a *MET1*

80 *+/+* and *met1 +/-* backcross. The recovered *MET1 +/+* is wild-type for MET1
81 function but has lost CpG methylation in some regions of the genome (low loss).
82 *FAST^mC* is able to capture the differences between these maintenance
83 methyltransferases (Fig. 1A). Additionally, the slight (~3%) difference between
84 *MET1 +/+* and the *met1 +/-* mutant can be distinguished, demonstrating the
85 sensitivity of *FAST^mC* (Fig. 1A).

86 In mammals, epigenetic reprogramming, including CpG demethylation, is
87 required to erase DNA methylation imprints and epimutations established in the
88 previous generation [11]. Following demethylation, DNA methylation patterns are
89 re-established at imprinted loci and transposable elements (TEs) during
90 gametogenesis by the *de novo* methyltransferases DNMT3A and a non-catalytic
91 paralogue, DNMT3-like (DNMT3L) (reviewed by [12]). The reductions in CpG
92 DNA methylation caused by epigenetic reprogramming in primordial germ cells
93 (PGCs) or by mutations in DNMT3L (*dnmt3L*) compared to somatic tissues are
94 captured by *FAST^mC* (Fig. 1B) [13]-[15]. Additionally, increased levels of CpG
95 DNA methylation in the brain (e.g., *NeuN+* and *glia* cells) [16] can be
96 differentiated from other somatic tissues (Fig. 1B; Suppl. Table 1) [17]. Overall,
97 as demonstrated in *A. thaliana* and *M. musculus*, *FAST^mC* can be used to
98 accurately detect intraspecific differences of DNA methylation levels at CpG sites
99 (Fig. 1A and B).

100 We determined natural interspecific variation of DNA methylation at CpG
101 sites across 44 different species (Fig. 2A). However, unlike intraspecific
102 comparisons between mutants or cell-types, nucleotide biases, such as genomic
103 GC content differences, can over- or underestimate the estimator value for the
104 CpG sequence contexts. The estimator (equation 2 of Methods) is estimating the
105 product of the methylation frequency of CpG sites and the GC content of the
106 genome, and are thus confounded. This bias can be overcome in all species
107 investigated but mammals (*H. sapiens*, *M. musculus*, and *C. l. familiaris*) by
108 dividing the estimator value by an average GC content of the genome, which
109 corrects the relationship between target and estimator to ~1:1. GC content can
110 be approximately estimated from WGBS reads (see Methods) or additional
111 genomic sequence data – 10,000, 50 base pairs (bp) reads (500,000bp) – can
112 be used to directly estimate GC content (Suppl. Table 1).

113 Nucleotide biases in genomes – such as the depletion of CpG
114 dinucleotides to localized “CpG islands” in mammalian genomes – may interfere
115 when estimating \hat{F} . CpG dinucleotides can be directly measured from 10,000, 50
116 bp genomic sequencing reads (Suppl. Table 1), and this can then be used to
117 directly calculate the proportion of target sites that are methylated, m , using the
118 frequency of intact target sites, e.g., CpG, that remain in the bisulfite sequencing
119 data. These are sites that were methylated and thus escaped C to T conversion.

120 Accommodating for nucleotide biases in mammalian genomes does not improve
121 assessment of DNA methylation levels by *FAST^mC* (Suppl. Table 1). However,
122 treating mammals separately from other species with CpG DNA methylation (i.e.,
123 phylogenetic correction) produces an improved, mammal-specific model with
124 similar accuracy – measured as the Mean Absolute Percentage Error (MAPE) –
125 to the remaining species (Suppl. Table 1). Additionally, only a modest increase in
126 model improvement was observed for non-mammalian species (Suppl. Table 1).
127 Overall, GC content correction (\hat{F}/\hat{p}) and treating mammalian species
128 separately improves model accuracy without introducing additional genomic
129 sequencing data.

130 *FAST^mC* also tolerates high contamination and error rates associated with
131 sodium bisulfite conversion. We used *A. thaliana met1* mutants generated by
132 [10], which show minor (~3%) to large (~14%) differences in CpG DNA
133 methylation compared to the wild-type *A. thaliana*. By artificially introducing un-
134 methylated chloroplast reads to 10,000 reads to *met1* and *met1 +/-* mutant
135 genotypes, and *MET1 +/+* and *A. thaliana* wild-type genotypes, we were able to
136 demonstrate that a ~3% difference in DNA methylation can still be detected with
137 <10% chloroplast contamination, and a difference of 13-14% with 40-50%
138 chloroplast contamination (Suppl. Table 1). Similarly, nonconversion rates >3%
139 still allow for detection of differences between samples (Suppl. Table 1). It

140 should be noted that the *met1* mutants and *A. thaliana* samples had
141 nonconversion rates of 0.50%, 0.82%, 1.86%, and 0.56% for *met1*, *met1 +/-*,
142 *MET1 +/-*, and wild-type *A. thaliana*, respectively. The artificially introduced error
143 rates are extremely high, but possible. For example, <1% of reads typically map
144 to the chloroplast genome, and nonconversion rates are typically <2% (data not
145 shown). However, it is recommended that Lambda DNA be sequenced for each
146 batch of WGBS libraries prepared to estimate the rate of sodium bisulfite non-
147 conversion. Reducing technical error is especially important for identifying
148 differences between species with small amounts of or no DNA methylation like
149 insects (Suppl. Table 1). Regardless, the *FAST^mC* method is robust as it is able
150 to tolerate technical and biological contamination.

151 The number of short reads (≥ 30 bp) required to make accurate
152 estimations is low, and we have determined that a few thousand reads produce
153 high-confidence estimates of genome-wide methylation levels (Suppl. Fig. 1).
154 Hence, these models can be used to accurately, and cost-effectively, identify
155 differences of DNA methylation levels for any species regardless of the
156 availability of a reference genome assembly.

157 Non-CpG DNA methylation can also be confidently predicted within and
158 between species using *FAST^mC*. In *A. thaliana*, the majority of DNA methylation
159 at CHG sites is maintained by chromomethylase CMT3 through a reinforcing

160 loop with H3K9me2 methylation catalyzed by the KRYPTONITE (KYP)/SUVH4
161 protein [18]-[20]. Similarly to MET1, mutations in CMT3 causes reductions in
162 CHG DNA methylation [10], which are accurately detected by *FAST^mC* (Fig. 1C).
163 Also, in *A. thaliana*, cell-type specific levels of CHH DNA methylation in the
164 sperm cell (SC) (i.e., hypo-CHH DNA methylation) and vegetative nucleus (VN)
165 (i.e., hyper-CHH DNA methylation), and depletion of CHH DNA methylation in
166 mutants in the *de novo* DNA methylation pathway (e.g., the DNA-dependent
167 RNA polymerase, NRPD1) were recapitulated (Fig. 1D) [21],[10].

168 In mammals, non-CpG DNA methylation can be found at CH sites. Work
169 by [16] has demonstrated the overall increase of CH DNA methylation during
170 brain development in *M. musculus* and *Homo sapiens*. *FAST^mC* was able to
171 capture the overall trend of increasing CH methylation through brain
172 development in *H. sapiens* (Fig. 1E). Furthermore, despite only small differences
173 in brain CH methylation in the intervals from 2 years to 5 years (0.068%), and
174 from 55 years to 64 years (0.062%) of age, the *FAST^mC* model accurately
175 detected these changes (Fig. 1E) [16].

176

177 **CONCLUSIONS**

178 We propose several models, which capture the variation of, and can accurately
179 predict, genome-wide DNA methylation levels between species to represent

180 *FAST^mC* and can be found at <http://fastmc.genetics.uga.edu>. Additionally, the
181 web-based interface makes *FAST^mC* universally accessible, and models will be
182 continuously updated when new whole genome and methylome data is
183 analyzed and becomes available. Although genome content biases interfere with
184 the accuracy of *FAST^mC*, treating mammalian species separately for CpG DNA
185 methylation overcame this obstacle. *FAST^mC* makes practical previously
186 intractable studies (e.g. high-resolution time course, developmental, and large
187 diversity panels) regardless of species, genome size and availability of a
188 reference genome. Furthermore, these models will greatly contribute to high-
189 resolution screening of either developmental- or environmental-induced
190 epigenomic reprogramming events. *FAST^mC* is a suite of powerful models that
191 can aid researchers to make better investments in more comprehensive, fruitful
192 studies.

193

194 **METHODS**

195 Whole genome bisulfite sequencing (WGBS) data was downloaded from the
196 Short Read Archive (SRA)/Gene Expression Omnibus (GEO) or sequenced in-
197 house (Suppl. Table 1). WGBS data was aligned using methods described in
198 [22] to generate “allC” files. The allC files were used to determine target DNA
199 methylation levels, and can be downloaded from GEO under accession number

200 GSE72155. Prior to estimation of predictor DNA methylation levels, WGBS data
201 was trimmed of adaptor sequences using Cutadapt v1.9 [23], end-trimmed using
202 Trimmomatic [24], and quality filtered using FASTX-toolkit
203 (http://hannonlab.cshl.edu/fastx_toolkit/). Reads of at least 30 base pairs (bp) in
204 length with $\geq 20\%$ of nucleotides having a quality score $\geq 75\%$ were retained.
205 Random sampling without replacement was performed with increasing fold-
206 change from $1-10^5$ reads using the program fastq-tools
207 (<http://homes.cs.washington.edu/~dcjones/fastq-tools/>). Custom Perl scripts
208 were used to sum the number of C^m and $C^?$ sites for each randomly sampled
209 read, and subsequently to estimate the predictor DNA methylation level at CpG,
210 CHG, CHH, and CH sites (Suppl. Table 1).

211 Predictive modeling is used to find the mathematical relation between a
212 target, (dependent variable) and various estimators (independent variables);
213 subsequent values of an estimator(s) are used to predict the target variable
214 using the established mathematical relationship between them. The goal of the
215 *FAST^mC* models were to predict reference-based (target) from non-reference-
216 based (estimator) DNA methylation levels. These models assume that in
217 MethylC-Seq data [6]: (i) all cytosines at CpG, CHG, CHH, and CH sites are
218 methylated. (ii) all thymines at TpG, THG, THH, and TH sites are converted
219 unmethylated cytosines or true thymines, and (iii) all nucleotides are randomly

220 distributed in the genome. Our goal is to estimate the proportion of Cs in
221 potential target sites that are in fact methylated, m , which is

222

$$223 \quad m = \frac{\sum C^m}{\sum (C^m + C^u)}, \quad (1)$$

224

225 where $\sum C^m$ and $\sum C^u$ are the total number of methylated and unmethylated target
226 sites in the genome, respectively. Since m is unknown, we use an estimator, \hat{F} ,
227 which is obtained from the bisulfite sequencing data:

228

$$229 \quad \hat{F} = \frac{\sum_s C^m}{\sum_s (C^m + C^?)}, \quad (2)$$

230

231 where $\sum_s C^m$ is the total number of methylated target sites in the sample and
232 $\sum_s C^?$ is the sum of unmethylated target sites plus sites that are equivalent to
233 unmethylated target sites after bisulfite sequencing in the sample, e.g. all TG
234 dinucleotides in the case of CpG methylation. With our assumptions, it is
235 straightforward to show that for CpG methylation, the expected value of \hat{F} is mp .
236 Thus, \hat{F} divided by the estimated genomic GC content, \hat{p} , is an estimate of m .
237 We estimate GC content from the frequencies of G nucleotides in the sample
238 because these sites are unaffected from bisulfite treatment. Estimates of GC

239 content from WGBS reads are on average within $4.56\% \pm 3.52\%$ standard
240 deviations of the true GC content. For the other three targets of methylation (CH,
241 CHH and CHG), it can be easily shown that $\frac{\hat{F}}{\hat{p}}$ is also equal to m . *FAST^mC*
242 calculates $\frac{\hat{F}}{\hat{p}}$ from a whole genome bisulfite sample and uses it to estimate m , the
243 fraction of Cs that are methylated.

244 Violation of the assumptions can cause inaccuracies in estimating \hat{F} . We
245 discuss some of these violations in the results section. In addition, we note that
246 when additional genomic short read data ($\geq 500,000$ bp) is available, the
247 frequency of the target site in the genome, e.g., the GC content and frequency of
248 CpG dinucleotides, can be directly measured. This can then be used to directly
249 calculate the proportion of target sites that are methylated, m , using the
250 frequency of intact target sites, e.g., CpG, that remain in the bisulfite genome
251 data. These are sites that were methylated and thus escaped C to T conversion.

252

253 **AVAILABILITY OF SUPPORTING DATA**

254 All data used in this study can be found on the Short Read Archive (SRA)/Gene
255 Expression Omnibus (GEO) webpages. Accession identifiers can be found in
256 Suppl. Table 1.

257

258 **LIST OF ABBREVIATIONS**

259 Whole-Genome Bisulfite Sequencing (WGBS), Primordial Germ Cells (PGCs),
260 Base Pairs (bp), Short Read Archive (SRA), Gene Expression Omnibus (GEO),
261 Mean Absolute Percentage Error (MAPE), Sperm Cell (SC), Vegetative Nucleus
262 (VN)

263

264 **COMPETING INTERESTS**

265 The author's declare no competing interests.

266

267 **AUTHORS' CONTRIBUTIONS**

268 All authors contributed equally to this work.

269

270 **ACKNOWLEDGEMENTS**

271 We would like to thank Nathan Springer for critical comments on this manuscript.

272 Also, we would like to thank David Brown for webpage setup. The study was

273 funded by grants from the National Science Foundation (MCB-1339194) and the

274 National Institutes of Health (R00GM100000) to RJS.

275

276 **REFERENCES**

- 277 1. Cokus, S. J., Feng, S., Zhang, X., Chen, Z., Merriman, B., Haudenschild, C.
278 D., Pradhan, S., Nelson, S. F., Pellegrini, M., and Jacobsen, S. E.
279 Shotgun bisulfite sequencing of the *Arabidopsis* genome reveals DNA
280 methylation patterning. *Nature* 2008;452:215–219.
- 281 2. Lister, R., O'Malley, R. C., Tonti-Filippini, J., Gregory, B. D., Berry, C. C.,
282 Millar, A. H., and Ecker, J. R. Highly integrated single-base resolution
283 maps of the epigenome in *Arabidopsis*. *Cell* 2008;133:523–536.
- 284 3. Lister, R., Pelizzola, M., Downen, R. H., Hawkins, R. D., Hon, G., Tonti-Filippini,
285 J., Nery, J. R., Lee, L., Ye, Z., Ngo, Q.-M., Edsall, L., Antosiewicz-
286 Bourget, J., Stewart, R., Ruotti, V., Millar, A. H., Thomson, J. A., Ren, B.,
287 and Ecker, J. R. Human DNA methylomes at base resolution show
288 widespread epigenomic differences. *Nature* 2009;462:315–322.
- 289 4. Feng, S., Cokus, S. J., Zhang, X., Chen, P.-Y., Bostick, M., Goll, M. G.,
290 Hetzel, J., Jain, J., Strauss, S. H., Halpern, M. E., Ukomadu, C., Sadler,
291 K. C., Pradhan, S., Pellegrini, M., and Jacobsen, S. E. Conservation and
292 divergence of methylation patterning in plants and animals. *PNAS*
293 2010;107:8689–8694.

- 294 5. Zemach, A., McDaniel, I. E., Silva, P., and Zilberman, D. Genome-wide
295 evolutionary analysts of eukaryotic DNA methylation. *Science*
296 2010;328:916–919.
- 297 6. Urich, M. A., Nery, J. R., Lister, R., Schmitz, R. J., and Ecker, J. R. Methyl-
298 seq library preparation for base-resolution whole-genome bisulfite
299 sequencing. *Nature Protocols* 2015;10:475–483.
- 300 7. Kuo, K. C., McCune, R. A., Gehrke, C. W., Midgett, R., Ehrlich, M.
301 Quantitative reversed-phase high performance liquid chromatographic
302 determination of major and modified deoxyribonucleosides in DNA.
303 *Nucleic Acids Research* 1980;8:4763–4776.
- 304 8. Fraga, M. F., Uriol, E., Borja, D. L., Berdasco, M., Esteller, M., Cañal, M. J.,
305 Rodríguez, R. High-performance capillary electrophoretic method for the
306 quantification of 5-methyl 2'-deoxycytidine in genomic DNA: application to
307 plant, animal and human cancer tissues. *Electrophoresis* 2002;23:1677–
308 1681.
- 309 9. Karimi M., Johansson S., Stach D., Corcoran M., Grander D. LUMA
310 (LUMinometric Methylation Assay)-a high throughput method to the
311 analysis of genomic DNA methylation. *Experimental Cell Research*
312 2006;312:1989–1995.

- 313 10. Stroud, H., Greenberg, M. V. C., Feng, S., Bernatavichute, Y. V., and
314 Jacobsen, S. E. Comprehensive analysis of silencing mutants reveals
315 complex regulation comprehensive analysis of silencing mutants reveals
316 complex regulation of the *Arabidopsis* methylome. *Cell* 2013;152:352–
317 364.
- 318 11. Reik, W., Dean, W., and Walter, J. Epigenetic reprogramming in mammalian
319 development. *Science* 2001;293:1089–1093.
- 320 12. Law, J. A. and Jacobsen, S. E. Establishing, maintaining and modifying DNA
321 methylation patterns in plants and animals. *Nature Reviews Genetics*
322 2010;11:204–220
- 323 13. Popp, C., Dean, W., Feng, S., Cokus, S. J., Andrews, S., Pellegrini, M.,
324 Jacobsen, S. E., and Reik, W. Genome-wide erasure of DNA methylation
325 in mouse primordial germ cells is affected by AID deficiency. *Nature*
326 2010;463:1101–1105.
- 327 14. Kobayashi, H., Sakurai, T., Imai, M., Takahashi, N., Fukuda, A., Yayoi, O.,
328 Sato, S., Nakabayashi, K., Hata, K., Sotomaru, Y., Suzuki, Y., and Kono,
329 T. Contribution of intragenic DNA methylation in mouse gametic DNA
330 methylomes to establish oocyte-specific heritable marks. *PLoS Genetics*
331 2012;8:e1002440

- 332 15. Seisenberger, S., Andrews, S., Krueger, F., Arand, J., Walter, J., Santos, F.,
333 Popp, C., Thienpont, B., Dean, W., Reik, W. The dynamics of genome-
334 wide DNA methylation reprogramming in mouse primordial germ cells.
335 *Molecular Cell* 2012;48:849–862.
- 336 16. Lister, R., Mukamel, E. A., Nery, J. R., Urich, M., Puddifoot, C. A., Johnson,
337 N. D., Lucero, J., Huang, Y., Dwork, A. J., Schultz, M. D., Yu, M., Tonti-
338 Filippini, J., Heyn, H., Hu, S., Wu, J. C., Rao, A., Esteller, M., He, C.,
339 Haghghi, F. G., Sejnowski, T. J., Behrens, M. M., and Ecker, J. R. Global
340 epigenomic reconfiguration during mammalian brain development.
341 *Science* 2013;341:1237905.
- 342 17. Hon, G. C., Rajagopal, N., Shen, Y., McCleary, D. F., Yue, F., Dang, M. Y.,
343 and Ren, B. Adult tissue methylomes harbor epigenetic memory at
344 embryonic enhancers. *Nature Genetics* 2013;45:1198–1206.
- 345 18. Jackson, J. P., Lindroth, A. M., Cao X., and Jacobsen, S. E. Control of
346 CpNpG DNA methylation by the KRYPTONITE histone H3
347 methyltransferase. *Nature* 2002;416:556-560.
- 348 19. Du, J., Zhong, X., Bernatavichute, Y.V., Stroud, H., Feng, S., Caro, E.,
349 Vashisht, A.A., Terragni, J., Chin, H.G., Tu, A., Hetzel, J., Wohlschlegel,
350 J. A., Pradhan, S., Patel, D. J., and Jacobsen, S. E. Dual binding of

- 351 chromomethylase domains to H3K9me2-containing nucleosomes directs
352 DNA methylation in plants. *Cell* 2012;151:167–180.
- 353 20. Du, J., Johnson, L. M., Groth, M., Feng, S., Hale, C. J., Li, S., Vashisht, A.
354 A., Gallego-Bartolome, J., Wohlschlegel, J. A., Patel, D. J., and Jacobsen,
355 S. E. Mechanism of DNA methylation-directed histone methylation by
356 KRYPTONITE. *Molecular Cell* 2014;55:495-504.
- 357 21. Calarco, J. P., Borges, F., Donoghue, M. T., Van Ex, F., Jullien, P. E., Lopes,
358 T., Gardner, R., Berger, F., Feijo, J.A , Becker, J. D., and Martienssen, R.
359 A. Reprogramming of DNA methylation in pollen guides epigenetic
360 inheritance via small RNA. *Cell* 2012;151:194-205.
- 361 22. Schultz, M. D. , He, Y., Whitaker, J. W. , Hariharan, M., Mukamel, E. A.,
362 Leung, D., Rajagopal, N., Nery, J. R., Urich, M. A., Chen, H., Lin, S., Lin,
363 Y., Jung, I., Schmitt, A. D., Selvaraj, S., Ren, B., Sejnowski, T. J., Wang.
364 W., and Ecker, J. R. Human body epigenome maps reveal noncanonical
365 DNA methylation variation. *Nature* 2015;523:212-216.
- 366 23. Martin, M. Cutadapt removes adapter sequences from high-throughput
367 sequencing reads. *EMBnet* 2011;17:10–12.
- 368 24. Bolger, A. M., Lohse, M., and Usadel, B. Trimmomatic: a flexible trimmer for
369 illumina sequence data. *Bioinformatics* 2014;30:2114–2120.

Figure 1. Detection of intraspecific DNA methylation levels by *FAST^mC*.

Generalized linear models (GLMs) for estimator (\hat{F}) versus target (m) CpG, CHG, CHH, and CH DNA methylation levels using 10,000 reads corrected for estimated GC content (\hat{p}) (A-E). Differences between *A. thaliana* mutants can be detected (A, C, and D). Also, differences of CpG DNA methylation between mutants, cell-types and tissues in *M. musculus* can be differentiated by *FAST^mC* (B). Finally, increasing CH methylation through brain development is captured with *FAST^mC* (E). Shaded area represents the 95% confidence interval.

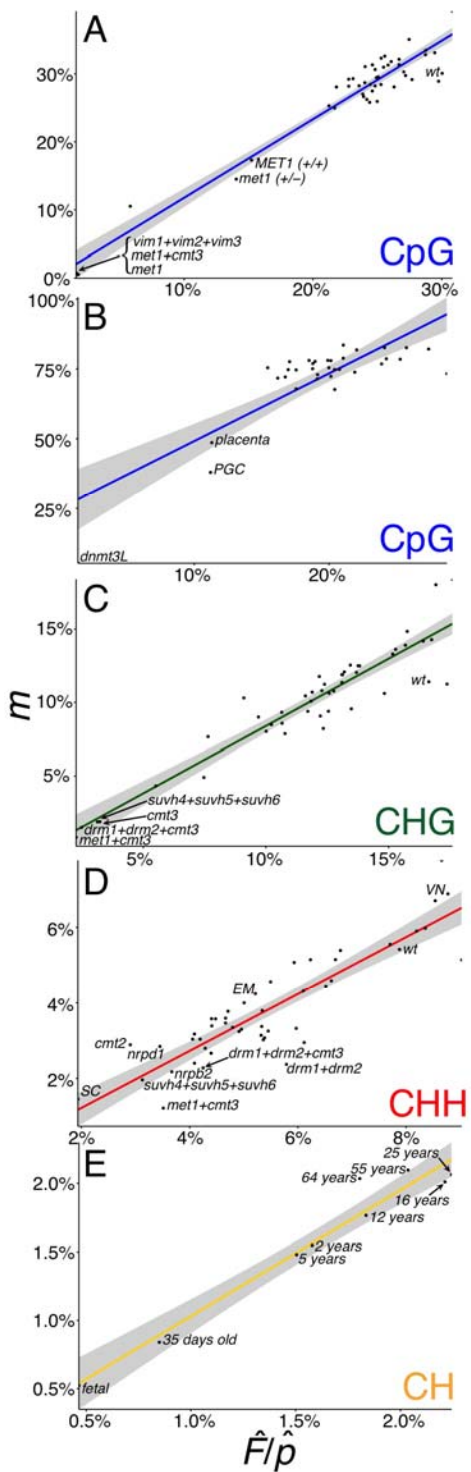


Figure 2. Detection of interspecific DNA methylation levels by *FAST^mC*.

Generalized linear models (GLMs) for estimator (\hat{F}) versus target (m) CpG (A-B), CHG (C), CHH (D), and CH (E) DNA methylation levels using 10,000 reads corrected for estimated GC content (\hat{p}). Species included in each plot can be found in Suppl. Table 1. Shaded area represents the 95% confidence interval.

