Characterizing cytosine methylation of polymorphic human transposable element insertions using human pangenome resources

Xiaoyu Zhuo[1,2], Chad Tomlinson[3], Edward A. Belter, Jr[3], Prashant Kumar Kuntala[1,2], Wesley N. Saintilnord[1,2], Tina Lindsay[1,3], Juan Macias[1,2], Human Pangenome Reference Consortium, Robert S. Fulton[1,3], Ting Wang[1,2,3]

[1]Department of Genetics, Washington University School of Medicine, St. Louis, MO, USA

[2]The Edison Family Center for Genome Sciences and Systems Biology, Washington University School of Medicine, St. Louis, MO, USA

[3]McDonnell Genome Institute, Washington University School of Medicine, St. Louis, MO, USA

13

14 **Abstract**

15 Cytosine methylation is an important epigenetic modification that plays a crucial role in
16 genomic regulation. The conventional second generation sequencing based bisulfite
17 conversion methods used to interrogate cytosine methylation require comparing
18 bisulfited-treated reads with a reference genome for methylation calling. Therefore, it
19 cannot characterize the methylation of unaligned regions. Taking advantage of the recent
20 improvement of the third generation sequencing, we investigated the methylation pattern
21 of human lymphoblastoid cell lines (LCLs) of non-reference insertions from the human
22 population, with a focus on polymorphic transposable elements. We first characterized
23 whole-genome CpG methylation using both SMRT and nanopore technology and
24 benchmarked their performance against WGBS using five human LCLs included in the draft
25 Human Pangenome Reference. Both methods are highly correlated with the conventional
26 WGBS results across the genome. The level of differences between PacBio and ONT on the
27 same sample is comparable with that of two replicates of WGBS of the same sample. Using
28 long-read data from the draft Human Pangenome Reference, we characterized CpG
29 methylation of non-reference insertions, especially polymorphic transposable elements.
30 We focused on addressing two questions: 1) do newly inserted TEs adopt the methylation
31 pattern of their genomic context? and 2) do methylation spread from new TE insertions to
32 their flanking regions? We found that most non-TE insertions exhibit DNA methylation
33 pattern consistent with their genomic context, but TE insertions are consistently
34 methylated, with a few exceptions. We also found limited methylation spreading from

35 *Alu*/L1 insertions to their flanking genomic regions. We investigated INDEL frequency in
36 both hypermethylated and hypomethylated CpG islands and found INDELs are enriched in
37 hypermethylated CpG islands. Our work demonstrated the methylation calling capability of
38 the 3rd generation sequencing and its unique advantage in characterizing epigenomic
39 features within non-reference positions.

40 **Introduction**

41 Cytosine methylation at CpG sites is prevalent in vertebrate genomes and the majority of
42 CpG sites in the human genome are methylated in most somatic tissues[1–3]. It is generally
43 believed that the CpG methylation is a repressive mark commonly associated with
44 heterochromatin and transposable elements (TEs), but is absent from promoter of active
45 genes[4]. In addition, methylation pattern is dynamic during development and disease
46 progression[5,6]. During evolution, methylation-mediated deamination converts methylated
47 cytosine to thymine, causing the general depletion of CpG sites in most vertebrate genomes.
48 The remaining regions enriched with CG motif compared with the genome background are
49 called CpG islands (CGIs), and they represent regions which remain unmethylated in germ
50 cells during evolution. Many CGIs are found in the promoters of conserved genes[7,8].

51 Cytosine methylation also responds to new genetic element insertions. For example, newly
52 inserted TEs are hypermethylated and the methylation mark can spread from the insertion
53 site to the flanking regions until blocked by other regulators[9,10], resembling the spreading
54 of heterochromatin[11]. The spreading of TE DNA methylation can affect the host genome
55 during evolution[12].

56 Given the importance of CpG methylation in genome regulation and evolution, considerable
57 efforts have been devoted to characterizing cytosine methylation. Microarray based
58 methylation chip, MeDIP-seq, MRE-seq, and RRBS-seq are all popular options to interrogate
59 methylation states of different CpGs across the genome[13–16]. Among these technologies,
60 Whole Genome Bisulfite Sequencing (WGBS) has been considered the gold standard in
61 cytosine methylation calling since its debut in the 2nd generation sequencing era[17–19].

62 Despite its many advantages, WGBS has several limitations. For example, the bisulfite
63 treatment that chemically distinguishes methylated from unmethylated cytosine damages
64 the DNA molecules and complicates the library preparation before sequencing. In addition,
65 WGBS typically relies on identifying mismatches between bisulfite-treated reads and the
66 reference genome to call methylation, making methylation calling of cytosines of non-
67 reference regions non-trivial. There are new methods developed recently to improve WGBS
68 but ultimately, they are limited by the 2nd generation sequencing[20,21].

69 The 3rd generation sequencing technologies have greatly matured in recent years[22,23]. This
70 improvement results in reads that are much longer, and the ability to distinguish
71 methylated cytosines and possibly other modified bases from unmodified ones during
72 basecalling[24,25]. Thus, methylation calling and sequencing can be done simultaneously,
73 making cytosine methylation calling much simpler and more accessible without resort to
74 specific chemical treatment to target cytosine methylation. Despite the promising outlook,
75 the signal changes caused by methylation is usually small and variable, and it is not until
76 the recent development of deep learning models that CpG methylation calling on SMRT and
77 Nanopore data has seen dramatic improvement in accuracy[26–28]. Recently, the third

78 generation sequencing were used to characterize the high resolution methylation profile of
79 repetitive elements[29,30].

80 Taking advantage of the high coverage of the Human Pangenome Reference Consortium
81 (HPRC) 3rd generation sequencing data and latest methylation calling methods, we
82 compared the whole-genome CpG methylation in five human lymphoblastoid cell lines
83 (LCLs) using Single Molecule, Real-Time (PacBio), Nanopore (ONT), and WGBS
84 technologies. After benchmarking the accuracies of methylation calling of the 3rd
85 generation sequencing, we investigated DNA methylation profiles of polymorphic TE
86 insertions, the relationship between DNA methylation patterns of the insertions and their
87 flanking genomic regions, as well as how could local methylation be affected by the new
88 insertions. Our results suggest that the 3rd generation sequencing is capable of accurately
89 characterizing CpG methylation at non-reference positions that are inaccessible with
90 conventional WGBS. Additionally, we revealed minimal methylation interaction between
91 TE insertions and their genomic context.

92 Results

93 Both PacBio HiFi and Nanopore ONT reads provide accurate DNA methylation estimation

94 The Human Pangenome Reference Consortium (HPRC) published the first human
95 pangenome reference graph representing 47 genetically diverse individuals[31]. In addition
96 to rich genetic information and various annotations, it also enabled direct methylation
97 analysis using long read raw sequencing data. Taking advantage of these resources, we
98 obtained HiFi methylation data of 32 samples (download) and performed additional CpG

99 methylation calling of ONT reads (R9.4.1) of five individuals who were part of the first draft
100 of the Human Pangenome Reference (HG00741, HG00621, HG01952, HG01978, and
101 HG03516) (Methods). In addition, we performed Whole Genome Bisulfite Sequencing
102 (WGBS) for these same five individuals (Methods). After calling CpG methylation, we
103 aligned all reads to the human reference genome hg38. Since the methylation calling of
104 both PacBio and ONT reads were obtained from raw reads directly without inferring from
105 reference sequences, we retained the per-molecule methylation estimation of both
106 reference positions and non-reference insertions. (Fig. 1a).

107 We first compared DNA methylation calls between HiFi, ONT, and WGBS for all CpG sites in
108 the reference genome hg38. We adopted a simple "pileup" method for calling methylation
109 level, which is a maximal likelihood estimate based on how many times a specific CpG is
110 called methylated or unmethylated among reads aligned to it (Methods). We also included
111 the deep learning based mathematical model (PB-model) developed by PacBio to estimate
112 methylation percentage of individual CpG sites in a given sample
113 (https://github.com/PacificBiosciences/pb-CpG-tools). We calculated mean average error
114 (MAE), 20% concordance, and Pearson correlation between WGBS, HiFi (pileup and PB-
115 model), and ONT for each sample in pairwise comparisons (HG00741 comparison in Fig. 1c
116 and the other four in Extended Data Fig. 1). We found ONT methylation calls highly
117 correlated with their matching WGBS result in all five samples (Fig. 1c and Extended Data
118 Fig. 1). Methylation percentages calculated by piling up HiFi reads did not match WGBS
119 results as well as ONT methylation calling (Fig. 1c). Upon close examination of the density
120 plot, we noticed a skewed methylation percentage distribution from HiFi pileup, with the
121 peaks of the methylation percentages located at ~20% and ~80% instead of 0% and 100%.

122 This observation suggests relatively low methylation calling accuracy per HiFi read (for
123 example, 20% accuracy a CpG site with 0% methylation will be presented as 20%
124 methylated). However, the PB-model resulted in a much better correlation/concordance
125 between the HiFi methylation percentage estimation and the WGBS (MAE: 0.1062; 20%
126 concordance: 0.846; Pearson correlation: 0.901). We examined the distribution of
127 methylation level across different genomic contexts (defined by the GM12878 chromHMM
128 states) and found strong concordance among WGBS, HiFi PB-model and ONT across all five
129 samples[32]. The HiFi pileup shared a similar trend but skewed toward intermediate
130 methylation levels (Fig. 1d). We then used the WashU Epigenome browser to illustrate the
131 comparisons between WGBS, HiFi pileup, HiFi PB-model and ONT (Fig. 1b). HiFi PB-model
132 clearly outperformed HiFi pileup and was chosen in all following analysis.

133 To gain a global view of the difference and similarity across methods, we performed
134 Principal Component Analysis (PCA) and plotted the first two Principal components in a 2D
135 plot for visualization (Fig. 1e). In all five samples, the two matching WGBS replicates
136 consistently clustered together; HiFi (PB-model) and ONT methylation results from the
137 same samples also clustered together. The PCA distance between WGBS and the long read
138 methylation results seemed to be larger than the distances between HiFi and ONT.

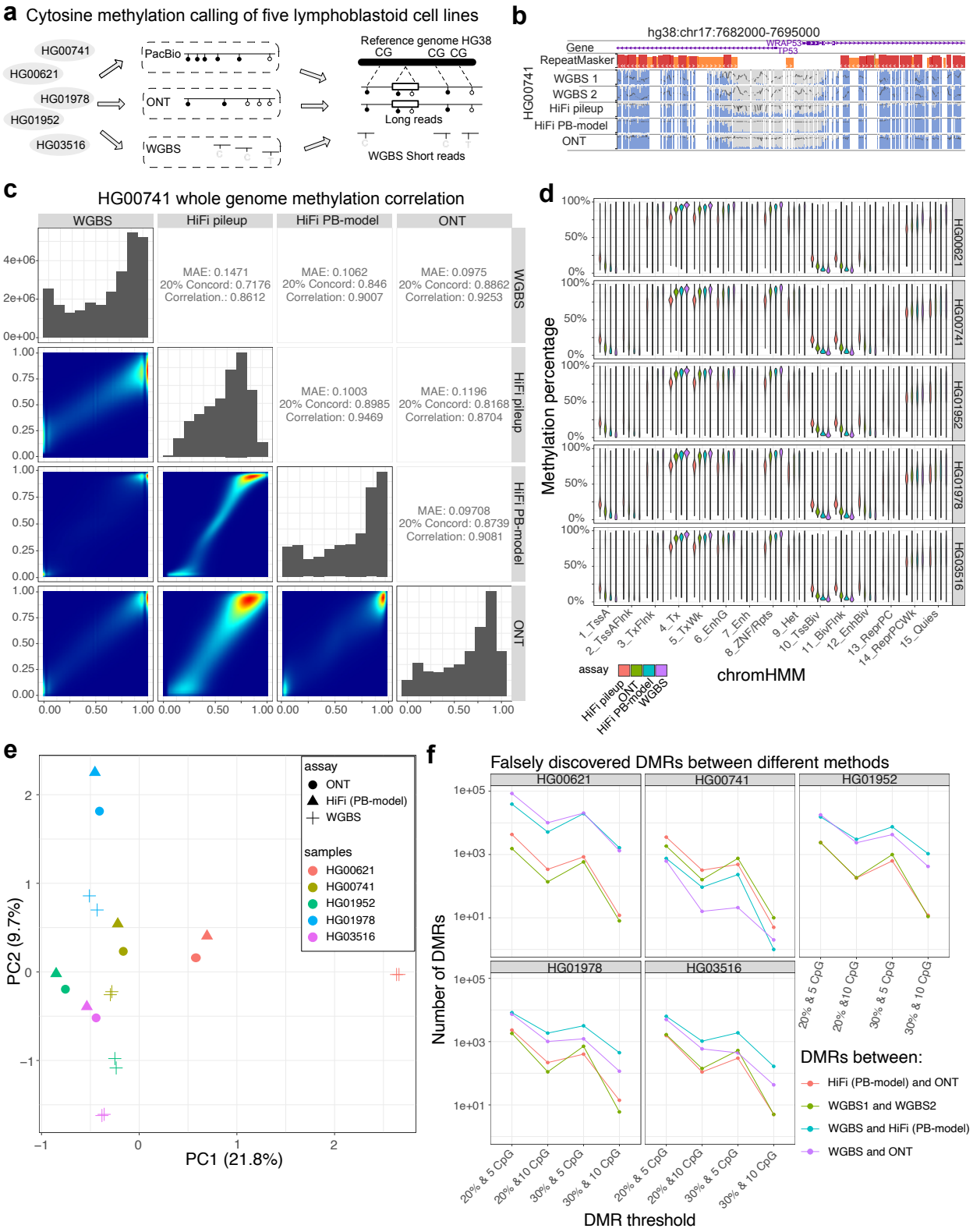
139 To better understand the scale of differences between the different methods, we defined
140 differentially methylated regions (DMRs) between HiFi (PB-model), ONT, and WGBS, as
141 well as between two WGBS replicates of the same sample[33]. Since the DMRs here were
142 defined using the same samples, they represented the false positive discoveries, and
143 allowed us to estimate an empirical false discovery rate. Using different thresholds, we

144 found that the numbers of falsely discovered DMRs between HiFi and ONT were

145 comparable with the numbers derived from the two WGBS replicates (Fig. 1f). Thus, the

146 methylation calling differences between HiFi and ONT was quite small, generally

147 comparable with the differences we observed between the two WGBS replicates.



148

149 **Figure 1.** a: Study design of methylation calling benchmarking. We performed CpG
150 methylation calling for five HPRC samples (HG00621, HG00741, HG01952, HG01978, and

151 HG03516), using PacBio, ONT, and WGBS. We mapped all reads to the human reference
152 genome hg38 to compare their difference. b: WashU Epigenome Browser screenshot of the
153 two WGBS replicates, HiFi pileup, HiFi PB-model, and ONT methylation of HG00741 on the
154 10 kb window around the TP53 gene promoter. All methylC tracks Y-axis percentage is
155 from 0% to 100%. c: whole genome methylation percentage correlation between WGBS,
156 HiFi pileup, HiFi PB-model, and ONT methylation. The whole genome methylation
157 percentage distribution is represented as a density plot. The correlation heatmap and
158 metrics (Mean Average Error, 20% concordance, and Pearson correlation) are plotted on
159 the left-bottom and right-top, respectively. d: Methylation percentage distribution of the
160 GM12878 15 chromHMM states across all five samples. The 15 chromHMM states are listed
161 on the X-axis and methylation calling results from different methods colored differently. e:
162 PCA plot of the top 2 principal components of the whole genome methylation from the five
163 HPRC samples. For each sample, we performed two replicates of WGBS in addition to HiFi
164 (PB-model) and ONT methylation calling. We use different shapes to represent different
165 methods (circle, triangle, and cross for ONT, HiFi, and WGBS) and different colors to
166 represent different samples (red, yellow, green, blue, and purple for HG00621, HG00741,
167 HG01952, HG01978, and HG03516, respectively). f: Number of falsely discovered
168 differentially methylated regions (DMRs) using different threshold cut-offs. Four different
169 thresholds (20% average methylation difference with at least 5 CpG sites, 20% average
170 methylation difference with at least 10 CpG sites, 30% average methylation difference with
171 at least 5 CpG sites, and 30% average methylation difference with at least 10 CpG sites) are
172 represented on the X-axis. The number of DMRs are represented on the Y-axis. The number
173 of DMRs between different methods under different conditions are plotted using different
174 colors.

175

176 **The methylation patterns of TE insertions and their correlation with flanking regions**

177 Because the 3rd generation methylation calling was conducted on raw reads without
178 aligning to the reference genome, it enabled methylation calling of non-reference regions
179 previously inaccessible with 2nd generation sequencing based methods, at least not
180 trivially. Thus, we investigated the methylation status of non-reference insertions with a
181 focus on polymorphic TE insertions, and their correlation with the methylation levels of
182 their flanking regions. Here we took advantage of the recent release of the draft Human
183 Pangenome Reference made available by the HPRC, which included HiFi reads of the 32
184 HPRC Year1 samples (download) as well as their variant calling results (download)[31]. We

185 further annotated insertions using the Transposable Element Insertion Annotation pipeline
186 xTEA[34]. We classified TE insertions into four categories: *Alu*, LINE1, SVA, and ERV. Non-
187 reference insertions other than TEs were classified as non-TE.

188 To help analyze and visualize non-reference insertions, we assembled chromosome level
189 individual diploid genome assemblies of the five HPRC individuals using hifiasm assemblies
190 (download) and mapped phased long reads onto maternal or paternal genomes. We also
191 developed methods to extract insertion sequences along with their methylation annotation
192 from BAM files mapped to either reference or individual genome (Fig. 2a, Methods). Using
193 these pipelines, we extracted and aligned TE insertions with methylation to their respected
194 TE consensus.

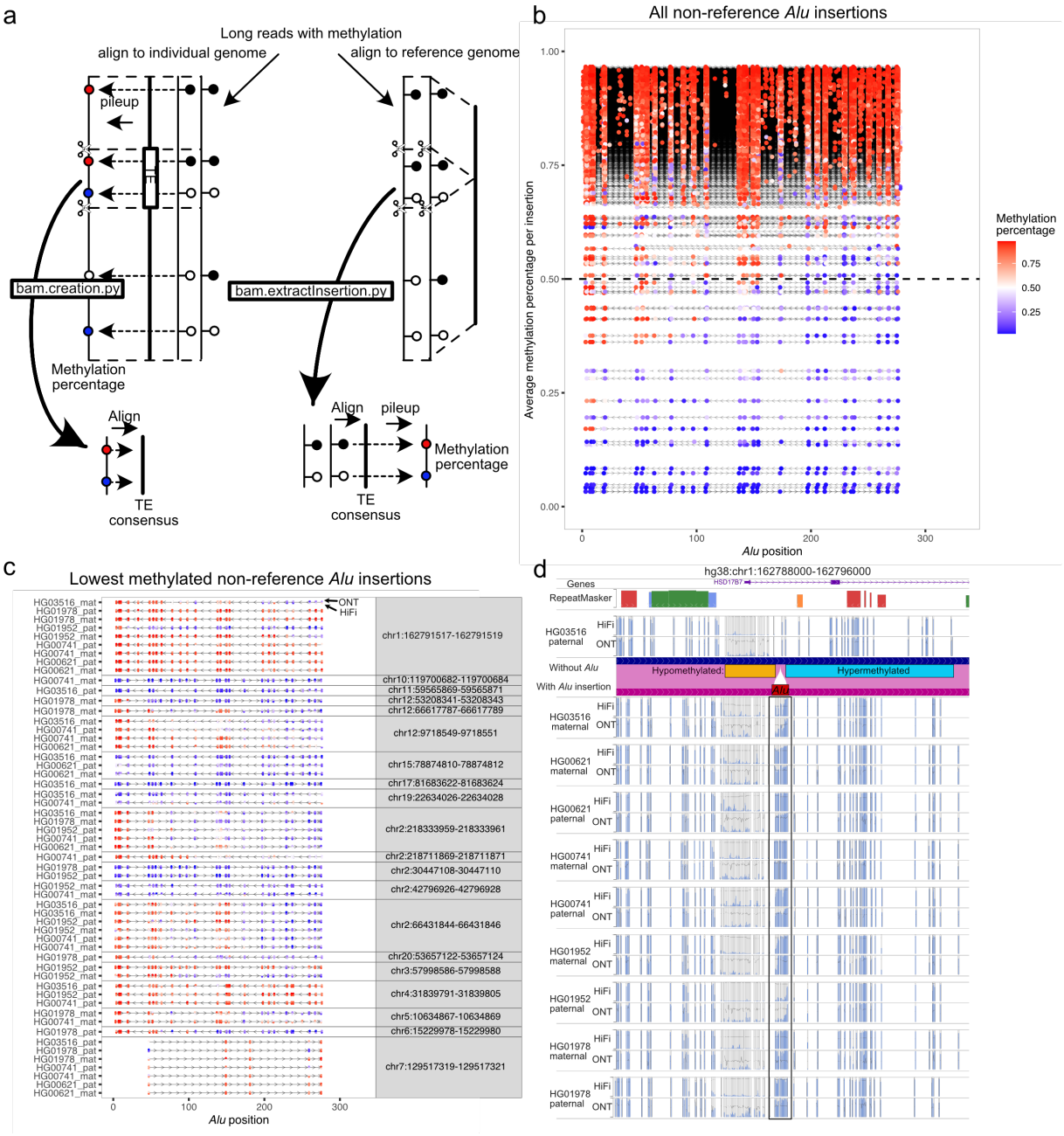


Figure 2. a: Schematic representation of our pipeline to extract non-reference insertions along with their methylation annotation. b: Per CpG HiFi methylation percentage of all non-reference *Alu* insertions. Each insertion is represented by a grey line and we used the average methylation of an insertion as Y-axis. We used the Blue-Red color scale to represent methylation status of each CpG sites (Blue: hypomethylated; Red: hypermethylated). A dotted line represents the cut-off (0.5) we used to filter lowest methylated regions to plot in c. c: Per CpG site methylation percentage of non-reference *Alu* insertions with average methylation < 0.5. Long-read methylation is colored with same color scale as Fig. 2b along *Alu* consensus (X-axis) with ONT methylation above and HiFi

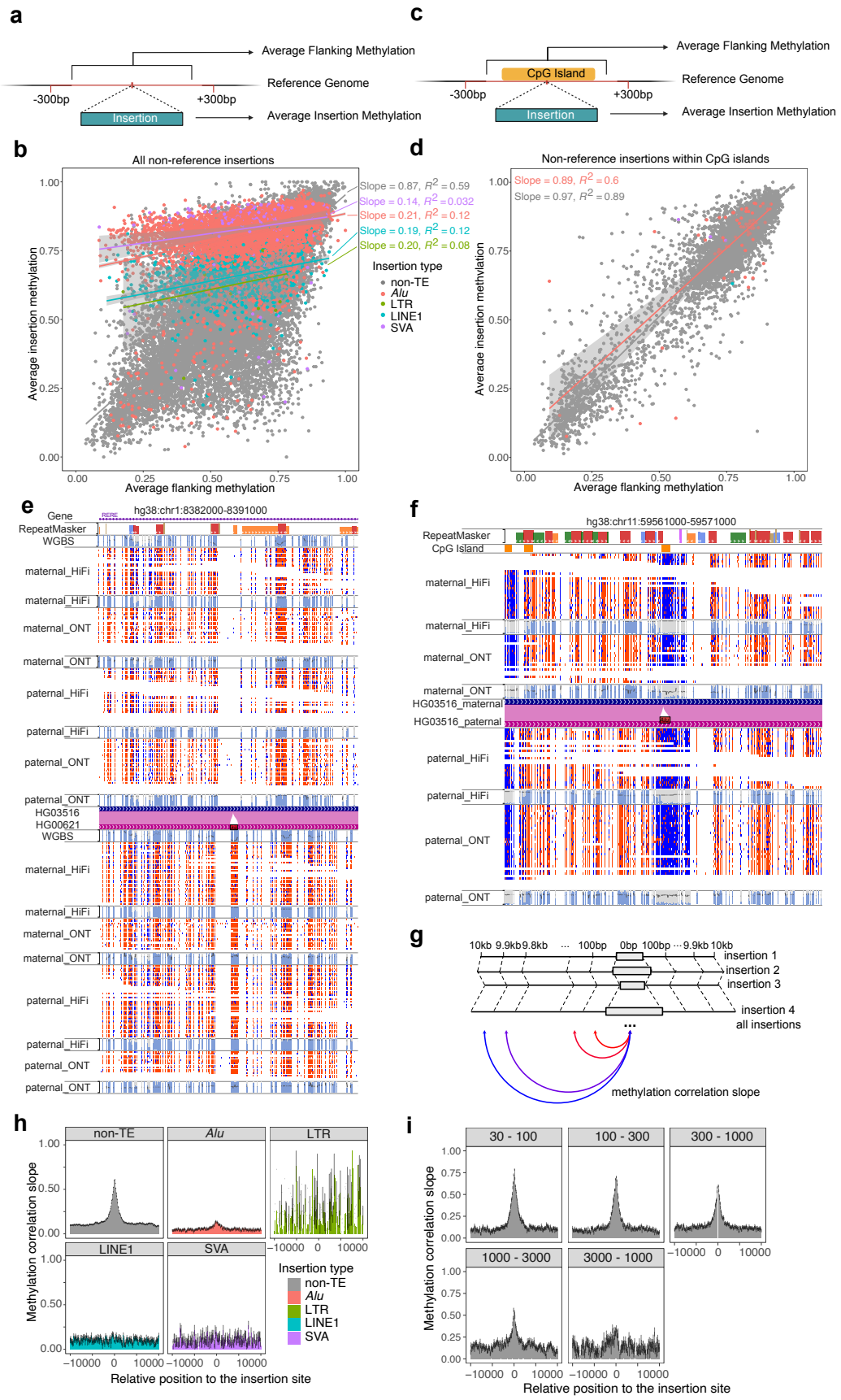
205 methylation below the grey line representing insertions. All insertions were grouped by the
206 insertion position on hg38. d: The WashU Comparative Epigenome Browser screenshot of
207 the methylation status of *Alu* insertion at hg38:chr1:162791517-162791519 in different
208 samples. Methylation percentage derived from HiFi and ONT reads were plotted from 0-
209 100% scale below gene and RepeatMasker annotation. A genome-align track separated
210 haplotypes with and without insertion. It inserted at the boundary between a
211 hypomethylated region (promoter of HSD17B7) and a hypermethylated region (gene body
212 of HSD17B7).

213

214 We displayed the methylation of all CpG sites across different TEs in Fig. 2 and Extended
215 Data Fig. 2. In Fig. 2b we showed the HiFi methylation of each CpG site of all *Alu* insertions
216 vertically separated by their average methylation, and 97% of *Alu* insertions were
217 hypermethylated (defined as average methylation > 0.7) (Fig. 2b). In Fig. 2c we plotted
218 both ONT and HiFi methylation of *Alu* insertions with average methylation < 0.5 in the five
219 phased samples grouped by the insertion site. Out of 13 *Alu* insertions presented in at least
220 two haplotypes, the methylation status of 11 of them were shared across different
221 individuals (Fig. 2c). Both exceptions (chr1:162791517-162791519 and chr3:57998586-
222 57998588) were inserted at the boundaries of methylation valleys. For example, the *Alu*
223 insertion at chr1:162791517-162791519 was hypomethylated in the HG01952 paternal
224 haplotype but hypermethylated in all other alleles (Fig. 2c). We then built a datahub using
225 the WashU Comparative Epigenome Browser[35] and its associated methylC and modbed
226 tracks[36,37] (datahub). Instead of mapping sequencing reads from different samples to a
227 single reference genome, we aligned them to their respected individual assemblies and
228 anchored them on hg38 by using genome-align tracks. This allowed us to directly visualize
229 and compare the insertions between the genome assemblies, aligned long reads, as well as
230 their DNA methylation pattern. This *Alu* element was present within the first intron of the
231 HSD17B7 gene at the boundary between the hypomethylated promoter and the

232 hypermethylated gene body (Fig. 2d). Its hypomethylation in the HG01952 paternal
233 haplotype reflects a haplotype-specific boundary shift. This example highlights the
234 variation in regional dependance of methylation of specific *Alu* elements. The methylation
235 pattern of LTR5, SVA, and L1 insertions were similar to *Alu* in general and were shown in
236 Extended Data Fig. 2. Interestingly, for L1Hs and SVAs, we found several consistently
237 hypomethylated CpG sites surrounded by hypermethylated CpG sites (position 5875 in
238 L1Hs consensus, 1116 in SVA consensus, etc.) (Extended Data Fig. 2).

239 After characterizing the methylation patterns of TE insertions, we next evaluated the
240 relationship in DNA methylation levels between insertions and their flanking regions. We
241 calculated the average methylation levels of non-reference insertions and compared them
242 with the average methylation levels of their 300bp flanking regions on both directions (Fig.
243 3a). We plotted this relationship in Fig. 3b. Consistent with our observation in Fig. 2 and
244 Extended Data Fig. 2, it is evident that TE insertions, especially *Alu* and SVA insertions,
245 were highly methylated in general, regardless of the methylation levels of their flanking
246 regions. On average, we observed weak correlations of methylation between TEs and their
247 flanking regions (linear regression slope ~0.2, $R^2 \sim 0.1$), albeit being statistically significant
248 except for LTR (p value <2.2e-16, <2.2e-16, 0.27, 5e-4 of L1, *Alu*, LTR, and SVA,
249 respectively). This pattern is sharply distinct from non-TE insertions, whose methylation
250 levels were highly correlated with their nearby flanking regions (slope = 0.87, $R^2 = 0.59$)
251 (Fig. 3b).


a
Average Flanking Methylation
Reference Genome
-300bp
+300bp
Insertion
Average Insertion Methylation
c
Average Flanking Methylation
CpG Island
Reference Genome
-300bp
+300bp
Insertion
Average Insertion Methylation
b
All non-reference insertions
Slope = 0.87, R2 = 0.59
Slope = 0.14, R2 = 0.032
Slope = 0.21, R2 = 0.12
Slope = 0.19, R2 = 0.12
Slope = 0.20, R2 = 0.08
Insertion type
non-TE
Alu
LTR
LINE1
SVA
Average insertion methylation
Average flanking methylation
d
Non-reference insertions within CpG islands
Slope = 0.89, R2 = 0.6
Slope = 0.97, R2 = 0.89
Average insertion methylation
Average flanking methylation
e
Gene
hg38:chr1:8382000-8391000
RepeatMasker
WGBS
maternal_HiFi
maternal_ONT
paternal_HiFi
paternal_ONT
HG03516
HG00621
f
hg38:chr11:59561000-59571000
RepeatMasker
CpG Island
maternal_HiFi
maternal_ONT
HG03516_maternal
HG03516_paternal
paternal_HiFi
paternal_ONT
g
10kb 9.9kb 9.8kb ··· 100bp 0bp 100bp ··· 9.9kb 10kb
insertion 1
insertion 2
insertion 3
insertion 4
all insertions
methylation correlation slope
h
non-TE
Alu
LTR
LINE1
SVA
Methylation correlation slope
Insertion type
non-TE
Alu
LTR
LINE1
SVA
Relative position to the insertion site
i
30 - 100
100 - 300
300 - 1000
1000 - 3000
3000 - 1000
Methylation correlation slope
Relative position to the insertion site

253 **Figure 3.** a: Schematic representation of average methylation calculation of all insertions
254 and associated 300bp flanking regions. b: Methylation correlation between insertions and
255 300bp flanking regions. The average methylation of insertions is represented on the Y-axis,
256 while the methylation of flanking regions on the X-axis. *Alu*, LTR, LINE1, SVA, and non-TE
257 insertions are represented by red, green, blue, and purple dots, respectively whereas non-
258 TE insertions are labeled grey, serving as a background. Linear regression lines with
259 standard errors are also plotted and labeled with same color scheme. c: Schematic
260 representation of average methylation calculation of insertions within CpG islands and
261 associated 300bp flanking regions. d: Methylation correlation between insertions within
262 CpG islands and 300bp flanking regions. The average methylation of insertions is
263 represented on the Y-axis, while the methylation of flanking regions on the X-axis. *Alu*, LTR,
264 LINE1, SVA, and non-TE insertions are represented in the same way as they are in 3b. e:
265 WashU Epigenome browser shots of a hypermethylated *Alu* insertion present in HG00621
266 but absent in HG03516 as characterized by WGBS, HiFi, and ONT methylation. Long-reads
267 are separated by their haploid type. All the methylC tracks display methylation percentage
268 from 0%-100%. f: Depiction of a hypomethylated *Alu* insertion within CpG island present in
269 HG03516 paternal but absent in HG03516 maternal. Both HiFi and ONT reads are
270 separated by their haploid type and aligned to the respective haploid assembly. A genome-
271 align track was used to illustrate the heterozygous insertion. g: Schematic representation of
272 insertion-flanking methylation regression as the function of insertion-flanking distance. For
273 each distance from 100bp to 10kb, we performed linear regression between insertion and
274 flanking average methylation and estimated the linear regression slope and standard
275 errors. h: Linear regression slope between insertion and flanking methylation as the
276 function of the insertion-flanking distance from 10kb upstream to 10kb downstream. *Alu*,
277 LTR, LINE1, SVA and non-TE insertions are plotted in separated plots with different colors.
278 Standard Error for each window were shown as black bar. i: Linear regression slope
279 between non-TE insertion and flanking methylation similar to 3h. All non-TE insertions
280 were separated by their length and plotted by the following five bins: 30bp-100bp, 100bp-
281 300bp, 300bp-1kb, 1kb-3kb.

282

283 CpG island (CGI) is an important genomic feature, and their methylation status is highly

284 relevant to gene regulation[1]. When we restricted the above analysis to non-TE insertions

285 within CGIs only (Fig. 3c), we found the insertion-flanking methylation correlation to be

286 stronger (slope = 0.97, R^2 = 0.89) (Fig. 3d). Surprisingly, this strong correlation was also

287 extended to TE insertions within CGIs (colored dots and regression line Fig. 3d). For

288 example, genome-wide *Alu* insertions are generally hypermethylated with no apparent

289 relationship with that of their flanking regions (Fig. 3b). However, *Alu* methylation states

290 are correlated with their flanking methylation when only insertions within CGIs are
291 selected (slope = 0.89, R^2 = 0.6) (Fig. 3d). Thus, methylation of TE insertions seems to
292 follow different rules between CGIs and other parts of the genome.

293 We displayed two examples using the WashU comparative Epigenome Browser. we found a
294 polymorphic *Alu* insertion at hg38:chr1:8386566 that was absent in HG03516 and present
295 in HG00621 (Fig. 3e). Since this insertion was homozygous in HG00621, we also displayed
296 the WGBS short reads mapped to the maternal assembly. Since the insertion was
297 homozygous, mapping the same WGBS reads to either paternal or maternal assembly did
298 not make a difference (Extended Data Fig. 3a). The *Alu* insertion in HG00621was
299 hypermethylated (methylation level = 95.3%, 95.8% and 96.2% based on HiFi, ONT and
300 WGBS, respectively), as illustrated by both the individual aligned HiFi reads and the
301 methylation percentage tracks. There were only two CpG sites within the 500bp flanking
302 window and they were intermediately methylated (average methylation level = 63.4%
303 combining all three methods) (Fig. 3e).

304 At hg38:chr11:59565869 the HG03516 maternal and the paternal genome were different
305 by an *Alu* insertion in the paternal genome, as illustrated by the genome-align track. The
306 *Alu* insertion was in an hypomethylated CpG island (chr11:59565807-59566091, with
307 average maternal CGI methylation = 3.9%, paternal CGI methylation = 7.9%) and was itself
308 hypomethylated (*Alu* methylation = 10.3%, 8.7%, and 4.0% based on HiFi, ONT and WGBS,
309 respectively) (Fig. 3f and Extended Data Fig. 3b).

310 Our analysis thus far has revealed a strong correlation in DNA methylation levels between
311 non-TE insertions and their 300bp flanking regions, and a much weaker correlation when

312 the insertions are TEs. We next investigated the insertion-flanking methylation correlation
313 beyond 300 bp as a function of the distance between the insertion and the nearby regions.
314 For each non-reference insertion, we calculated DNA methylation levels in its 10 kb
315 flanking regions on both directions at a resolution of 100-bp non-overlapping windows
316 (Fig. 3g). We grouped the average methylation of all 100-bp windows from different
317 insertions by their distance from the insertion sites and performed linear regression
318 between insertion and flanking regions. Since the slopes of the linear regression can reflect
319 the methylation correlation between these regions, we then displayed the slope of the
320 linear regression as a function of the distance between the flanking windows and the
321 insertion (Fig. 3h). This analysis revealed that the methylation correlation slope between
322 non-TE insertions and their flanking windows quickly faded away as the distance between
323 the flanking window and insertion site increased, with an estimated elbow point being
324 ~2.5kb away from the insertion. The slope decayed to and stayed at around 0.1 when the
325 distance reached 5kb and beyond. TEs once again displayed a strikingly distinct pattern.
326 There was no observable correlation between LINE1, SVA, and LTR insertions and their
327 flanking regions. We did however observe a weak correlation between *Alu* insertions and
328 their flanking regions but it also quickly faded away (Fig. 3h). This observation could be
329 confounded by the fact that TE insertions are longer than non-TE insertions (length
330 distribution of TE insertions in Extended Data Fig. 3c). To rule out this possibility, we
331 investigated how the size of non-TE insertion could affect the insertion-flanking
332 methylation correlation. If the methylation correlation is merely a function of the insertion
333 size, we would expect ~300bp and ~1000bp non-TE insertions behave similarly to *Alu* and
334 LTR insertions, for example. We categorized non-TE insertions by their sizes into five bins,

335 then calculated the regression slope in each bin (Fig. 3i). Despite a slightly decrease of the
336 peak methylation correlation as a function of increased non-TE insertion size, the
337 correlation was much higher than that of TE insertions of comparable length (for example,
338 the slope distribution of 300-1000bp non-TE insertions was much higher than that of *Alu*
339 and LTR insertions). This suggests that the methylation of non-TE insertions, but not TE
340 insertions, are highly correlated with that of their flanking regions.

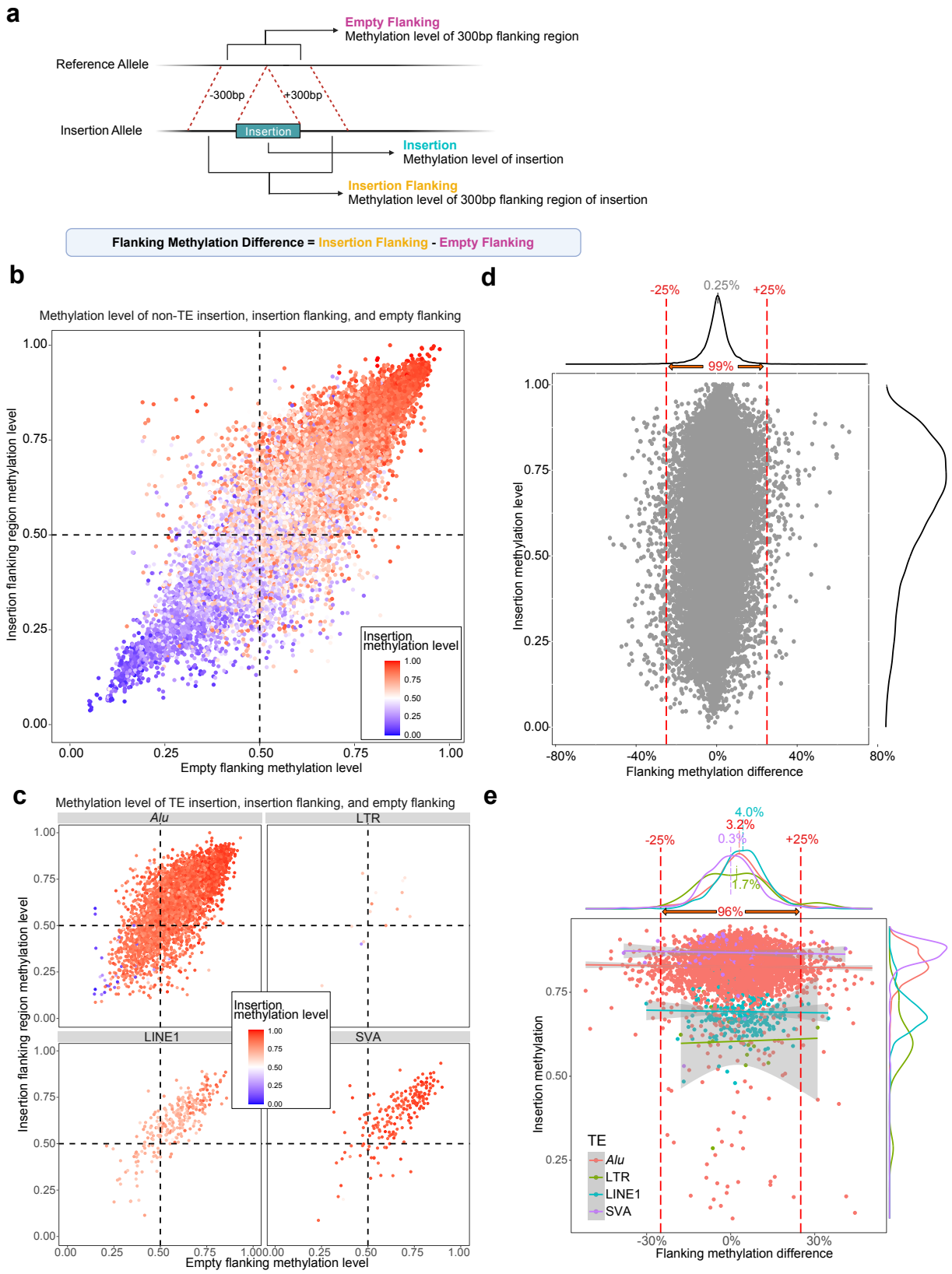
341 **Limited methylation spreading from polymorphic human L1s/*Alus* to flanking regions**

342 Previous studies have suggested that repressive chromatin marks including DNA
343 methylation deposited on newly inserted TEs can spread beyond TE/genome
344 boundary[10,11,38]. This relationship has not been examined for human polymorphic TEs at a
345 genome-wide scale, or for other types of human polymorphic insertions. Some non-
346 reference insertions could be deletion events in the human population[39]. However, TE
347 insertions are essentially homoplasy-free[40] and represent recent insertions that are
348 polymorphic in the human population. Thus, we explored how these insertions might have
349 interacted with their genomic insertion sites epigenetically, either influencing DNA
350 methylation of the insertion sites, or being influenced by the DNA methylation pattern of
351 the insertion sites. We also characterized the methylation pattern of other non-reference
352 insertions as a comparison.

353 For each polymorphic non-reference insertion relative to hg38, we phased HiFi reads from
354 all 32 HPRC Year1 samples and separated them into two groups: alleles with the insertion
355 and alleles without the insertion (Fig. 4a). We calculated the average methylation level of
356 the inserted regions (insertion), the 300bp flanking regions of the insertion from both ends

357 (insertion flanking), as well as the same flanking regions of the allele without insertion

358 (empty flanking) (Fig. 4a).


a
Empty Flanking
Methylation level of 300bp flanking region
Reference Allele
-300bp
+300bp
Insertion Allele
Insertion
Insertion
Methylation level of insertion
Insertion Flanking
Methylation level of 300bp flanking region of insertion
Flanking Methylation Difference = Insertion Flanking - Empty Flanking
b
Methylation level of non-TE insertion, insertion flanking, and empty flanking
Insertion flanking region methylation level
Empty flanking methylation level
Insertion methylation level
c
Methylation level of TE insertion, insertion flanking, and empty flanking
Alu
LTR
LINE1
SVA
Insertion flanking region methylation level
Empty flanking methylation level
Insertion methylation level
d
-25%
0.25%
+25%
99%
Insertion methylation level
Flanking methylation difference
e
4.0%
3.2%
0.3%
1.7%
-25%
+25%
96%
Insertion methylation
Flanking methylation difference
TE
Alu
LTR
LINE1
SVA

359

360 **Figure 4.** a: Schematic representation of a polymorphic insertion with aligned long reads.
361 The allele without the insertion site (top) and the allele with the insertion site (bottom) are
362 represented as reference allele and insertion allele. The average methylation level of the
363 insertion, insertion flanking regions, and empty flanking regions are calculated from reads
364 from all 32 samples. We then calculated flanking methylation difference by subtracting the
365 methylation level of empty flanking from that of insertion flanking. b: methylation level of
366 empty flanking (X-axis), insertion flanking (Y-axis), and non-TE inserted regions (color
367 scale). c: Similar to Fig. 4b, but the methylation level of different TE insertions are
368 presented and separated by the TE insertion types. d: dot plot representing the
369 relationship between methylation level of insertion (Y-axis) with that of the flanking region
370 (X-axis). Methylation percentage of insertions with flanking differences within 25% are
371 boxed and labeled. e: TE insertion flanking methylation differences plotted the sample way
372 as in Fig. 4d. Different TE insertions are distinguished by different colors. Percentage of all
373 insertions within 25% differences and the average flanking methylation differences of
374 different TEs are labeled.

375

376 To illustrate the relationship between these three regions, we also plotted the methylation
377 level of insertion flanking and empty flanking on the X-Y axis using a dot plot and colored
378 the dot using the methylation level of insertions (Fig. 4b, c). In general, the methylation
379 level of insertion, insertion flanking and empty flanking correlated with each other in case
380 of non-TE insertions (Fig. 4b). As for TE insertions, the flanking methylation level of
381 insertion and empty site showed a strong correlation despite the high methylation of the
382 insertion (Fig. 4c).

383 To better characterize the subtle changes in flanking methylation level between insertion
384 and empty sites, we defined the flanking methylation difference of each insertion by
385 subtracting the flanking methylation of empty site from the flanking methylation of
386 insertion sites (Fig. 4a). We reasoned that an allele without a TE insertion represents the
387 ancestral allele[40]. If the insertion increases flanking methylation, the difference would be
388 positive; and if it decreases methylation the difference would be negative. On the other

389 hand, if the insertion has no impact on flanking region methylation, the difference would be
390 zero.

391 We first plotted the relationship between the flanking methylation differences and the
392 methylation of non-TE insertions (Fig. 4d). The overall methylation differences between
393 orthologous flanking regions with or without insertions were heavily distributed around
394 zero, (99% (41755/42168) of the data points being within the -25% to +25% range), and
395 independent of the methylation level of the insertion itself. However, we observed a few
396 cases in which hypermethylated insertions were associated with increased flanking
397 methylation while hypomethylated insertions were associated with decreased methylation
398 (Fig. 4d). Altogether, these results strongly suggest that, albeit a few individual exceptions,
399 non-TE insertions generally do not affect flanking region methylation.

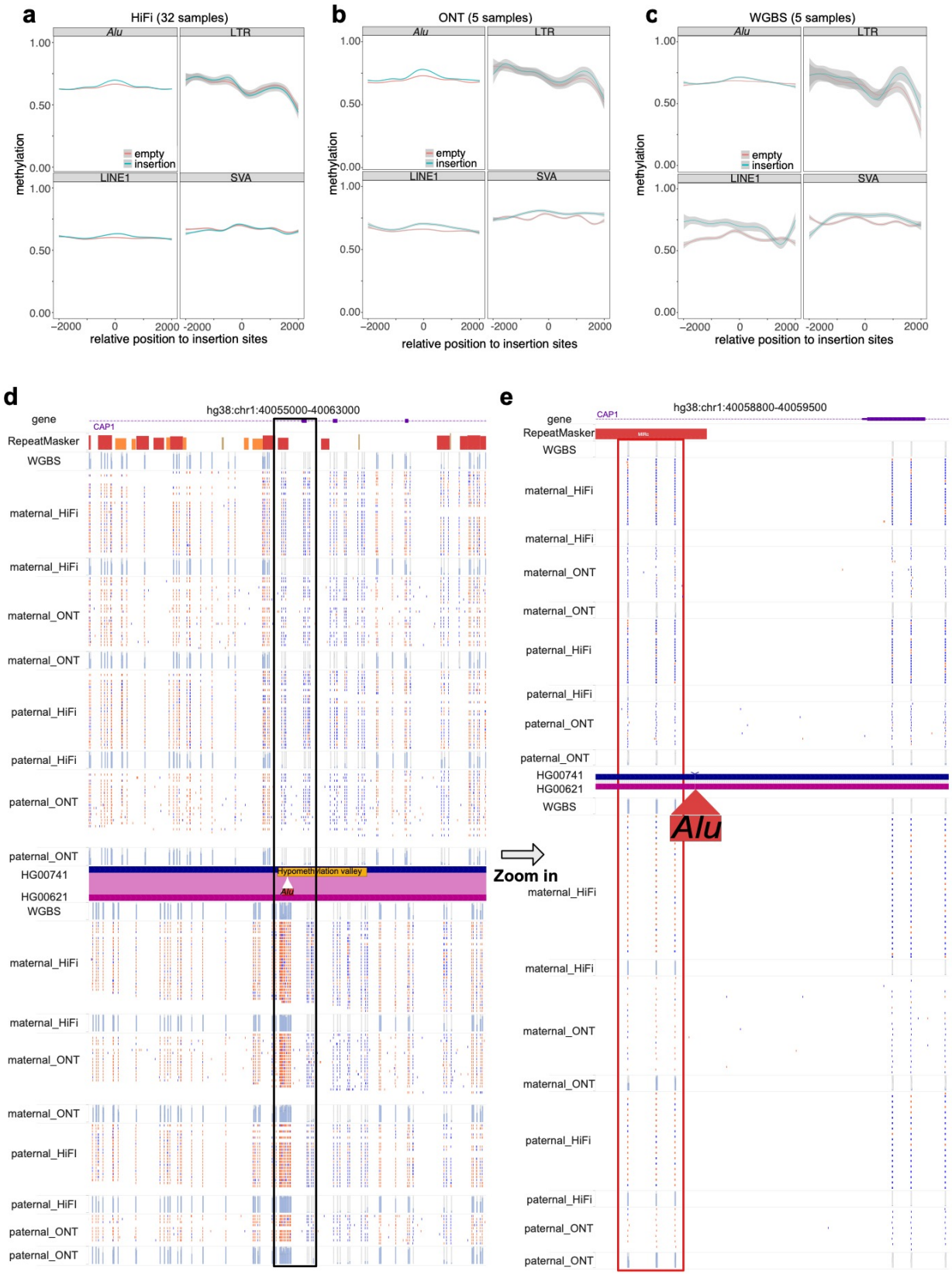
400 We then investigated the same relationship in the context of TE insertions (Fig. 4e).
401 Comparing with non-TE insertions, the range of the methylation differences was slightly
402 broader (96%, 5218/5413, within -25% and 25%), indicating increased variations in
403 flanking region methylation associated with TE insertions (F-Test to Compare Two
404 Variances: $F=2.4528$, $df1=5412$, $df2=42167$, p-value $< 2.2e-16$). On average, the flanking
405 methylation centered around zero for LTR and SVA insertions. However, *Alu* and L1
406 insertions were more highly associated with increased flanking methylation (117 and 3
407 increased by 25%, 66 and 1 decreased by 25%, respectively). On average, flanking
408 methylation increased by 3.2% ($p < 2e-16$) and 4.0% ($p = 5e-14$) for *Alu* and L1 insertions,
409 respectively. The range of methylation differences we observed was quite large. For
410 example, the mean methylation level of inserted *Alus* was 82.7%, and *Alu* insertions with

411 methylation level greater than this mean value were associated with a flanking region
412 methylation difference ranging from -42.74% to +42.66%, with 21 of them associated with
413 an increased flanking region methylation by greater than 30%. Similarly, the mean
414 methylation level of inserted L1s was 69.2%, and L1 insertions with methylation level
415 greater than this mean value were associated with a flanking region methylation difference
416 ranging from -30.1% to +27.8%, with none of them associated with an increased flanking
417 region methylation by greater than 30%. These examples are not fully consistent with the
418 hypothesis that DNA methylation deposited on newly inserted TEs could spread and
419 increase local DNA methylation level. Based on our genome-wide assessment, the
420 spreading is limited, and more drastic influence of TEs on nearby genomic methylation may
421 be an exception rather than the rule.

422 To further investigate the magnitude and the distance to which DNA methylation spreads
423 from polymorphic TE insertions, we aggregated all polymorphic insertions and plotted the
424 average methylation trendline based on their relative position in relation to the insertion
425 sites from the reference genome (Fig. 5a). We found a subtle increase of methylation
426 flanking the polymorphic L1/*Alu* insertions, with the highest increase right at the insertion
427 site (3.1% for L1 and 3.2% for *Alu*). This flanking methylation difference between insertion
428 and empty site decreased as the distance to insertion position increased, and the
429 methylation difference diminished at ~1kb away from the insertion site (Fig. 5a). The exact
430 same pattern is recapitulated by both the ONT methylation data and the WGBS data from
431 the five HPRC individuals (Fig. 5b, c), where we phased ONT reads and mapped the WGBS
432 reads of homozygous insertions to the individual's own genome assemblies (Methods) (Fig.
433 5c). Taken together, these data suggest polymorphic *Alu* and L1 insertions are associated

434 with slightly elevated DNA methylation levels at their insertion sites, and the effect is

435 localized within a thousand base pairs.


a
HiFi (32 samples)
b
ONT (5 samples)
c
WGBS (5 samples)
Alu
LTR
LINE1
SVA
methylation
empty
insertion
relative position to insertion sites
d
hg38:chr1:40055000-40063000
gene
CAP1
RepeatMasker
WGBS
maternal_HiFi
maternal_ONT
paternal_HiFi
paternal_ONT
HG00741
HG00621
Hypomethylation valley
Alu
e
hg38:chr1:40058800-40059500
Zoom in

436

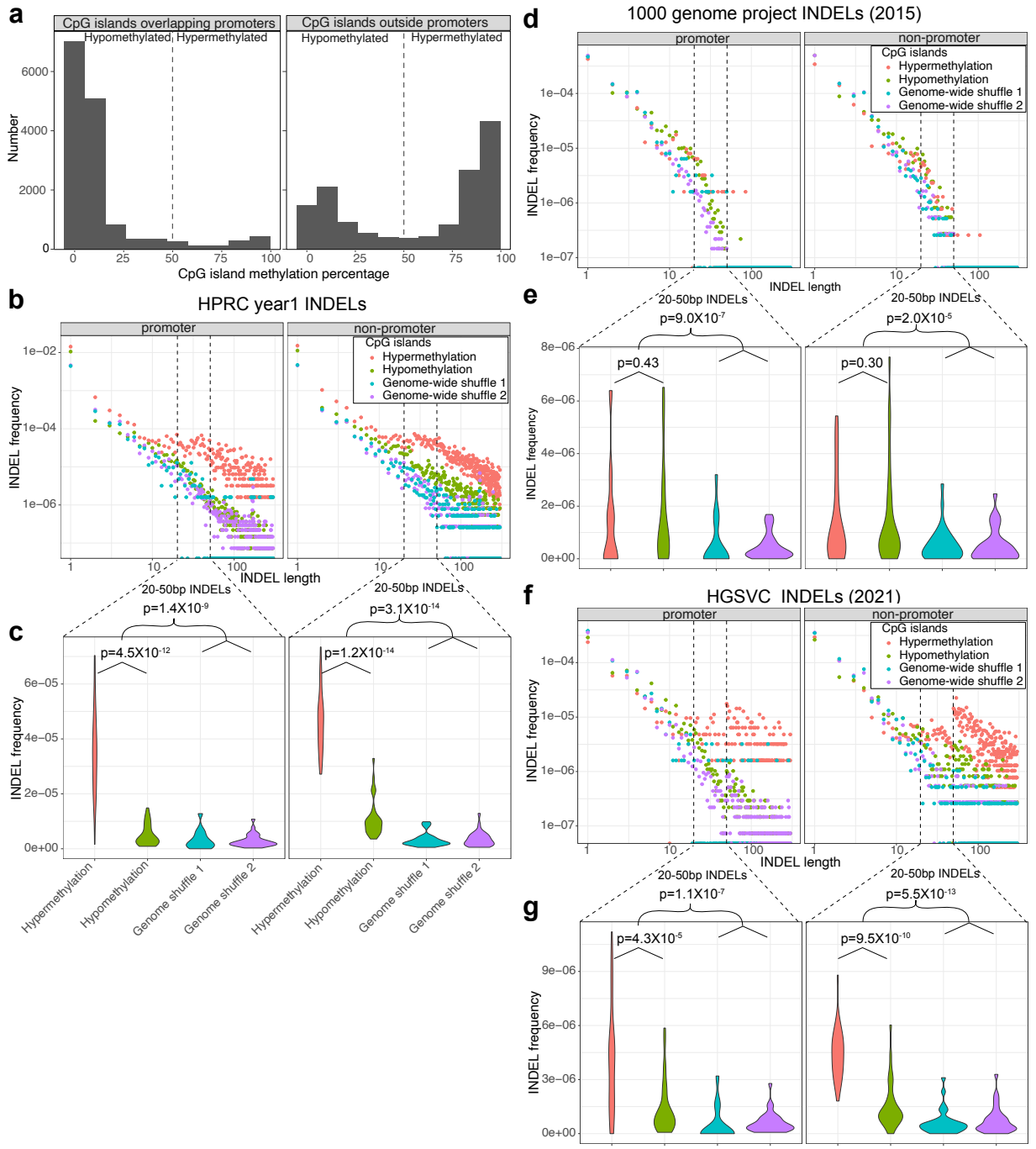
437 **Figure 5.** a: Average flanking methylation percentage trendline predicted by HiFi reads of
438 either empty or TE insertion site in the reference genome. Average methylation derived
439 from both empty and insertion alleles are represented by different color (blue represents
440 insertion flanking while red represents empty flanking). Different TE types are separated
441 by the grid. b: Average flanking methylation percentage trendline predicted by ONT reads
442 of either empty or TE insertion site in the reference genome similar to Fig. 5a. c: Average
443 flanking methylation percentage trendline predicted by WGBS of homozygous *Alu*
444 insertions and empty sites in the reference genome similar to Fig. 5a. d: WashU Epigenome
445 Browser shots of a hypermethylated *Alu* insertion within a hypomethylated valley present
446 in HG00621 but absent in HG00741 characterized by WGBS, HiFi, and ONT methylation.
447 Long-reads were separated by their haploid type, all the methylC tracks displaying
448 methylation percentage from 0%-100%. e: Zoom-in view of the methylation of the same
449 *Alu* element from Fig. 5d. The 3 CpG sites within 200bp of the insertion site went from
450 hypomethylated to hypermethylated associated with the *Alu* insertion.

451

452 Here we illustrated an example where an *Alu* insertion was associated with limited DNA
453 methylation spreading in its flanking regions (Fig. 5d, e). In Fig. 5d, we displayed a
454 homozygous *Alu* insertion present in HG00621 but absent in HG00741 (this polymorphic
455 *Alu* insertion was also found in HG03516 but absent in HG01952 and HG01978). The
456 insertion site was within a hypomethylation valley in HG00741 (hg38:chr1:40058800-
457 40060800, methylation level 4%) where the *Alu* insertion was absent. In HG00621, the *Alu*
458 insertion was hypermethylated (methylation level 85%), and methylation level of the
459 hypomethylation valley was elevated to 21% (Fig. 5d). Specifically, there were three CpG
460 sites 200bp upstream of the insertion that were lowly methylated in HG00741 (18%, 5%,
461 4%) but highly methylated in HG00621 (89%, 88%, 93%). On the other hand, multiple CpG
462 sites more than 300bp downstream from the insertion site seemed to have remained their
463 hypomethylated state (3 CpGs, 3% in HG00741, 5% in HG00621) (Fig. 5e).

464 **Elevated insertion/deletion (INDEL) rate in hypermethylated CpG Islands**

465 It is well known that the mutation rate is much higher at methylated CpG sites due to the
466 deamination of methylated cytosine[7]. However, we know little about how methylation level
467 would affect other mutation processes including insertions/deletions (INDELs). Previous
468 work comparing human and chimpanzee reference genomes as well as 1000 Genome
469 Project data revealed a surprising finding that CpG islands (CGIs) are enriched with indels
470 with moderate sizes (20 bp - 50 bp)[42]. Using the Human Pangenome resource, we set to
471 investigate whether DNA methylation of CGIs would affect their INDEL frequency. We first
472 grouped CGIs as promoter CGI and non-promoter CGI based on whether they overlapped
473 annotated promoters and compared their average methylation distribution (Fig. 6a). We
474 found that promoter CGIs were overwhelmingly unmethylated while most of the non-
475 promoter CGIs were highly methylated, consistent with previous findings[16,43] (Fig. 6a).
476 Since intermediately methylated CGIs were rare, we divided all CGIs into hypermethylated
477 and hypomethylated ones using the 50% average methylation as a cut-off (Fig. 6a). It is
478 possible that many non-promoter CGIs represent unannotated transcription start sites in
479 the genome[16]. To avoid CGI methylation status confounded by their relationship to
480 annotated promoters (Fig. 6a), we kept promoter and non-promoter CGIs separated in the
481 following analysis.



482

483 **Figure 6.** a: Histograms of the number of CpG islands with different average methylation
484 percentage. The histogram represent number of CpG islands in promoter (left), number of
485 CpG islands outside of any promoter (non-promoter, right). b: INDELs derived from HPRC
486 samples length-frequency relationship from 1bp to 150bp. INDEL frequency in
487 hypermethylated CpG islands, in hypomethylated CpG islands, and in two (shuffled
488 hypermethylated regions as shuffle 1, shuffled hypomethylated regions as shuffle 2)
489 matching genome-wide controls are represented by different colors (Red, green, blue, and

490 purple represents hypermethylated CGI, hypomethylated CGI, shuffle 1, and shuffle 2,
491 respectively). c: Violin plot representing INDEL frequency distribution from 20-50bp of the
492 same four categories of CpG islands as in Fig. 6b. Paired T-test p-values are labeled next to
493 the violin plot bars. d: INDELs derived from 1000 genome project length-frequency
494 relationship from 1bp to 150bp. INDEL frequency in hypermethylated CpG islands, in
495 hypomethylated CpG islands, and in two matching genome-wide controls are represented
496 by different colors. e: Violin plot representing INDEL frequency distribution from 20-50bp
497 of the same four categories of CpG islands as in Fig. 6d. T-test p-values are labeled next to
498 the violin plot bars. f: The same CpG island INDEL frequency vs methylation plotted as in
499 Fig. 6b and Fig. 6d but with INDELs identified by HGSVC. g: Violin plot representing INDEL
500 frequency distribution from 20-50bp of the same four categories of CpG islands as in Fig. 6f.
501 T-test p-values are labeled next to the violin plot bars.

502

503 To further evaluate the relationship between DNA methylation and INDEL frequency, we
504 plotted the relationship between INDEL size and their frequency within CGIs normalized by
505 the CGI length (Fig. 6b). We shuffled both hypermethylated and hypomethylated CGIs
506 within the genome used as background INDEL frequency, and calculated the INDEL
507 frequency the same way as background INDEL frequency. We observed a decrease of
508 INDEL frequency as a function of increase INDEL size across all comparisons, indicating a
509 negative correlation between INDEL size and frequency. In both promoter and non-
510 promoter CGIs, we recapitulated the result of enrichment of 20-50bp INDELs in CpG islands
511 compared to the shuffled genomic control (Fig. 6c, paired t-test p-value = $1.4X10^{-9}$ and
512 $3.1X10^{-14}$ in promoter and non-promoter regions, respectively)[42]. We found
513 hypermethylated CpG islands had the highest INDEL frequency among these groups (Fig.
514 6c). The difference for INDELs > 10bp was quite staggering (Fig. 6b).

515 Using the same CGI and methylation annotation, we repeated the analysis using INDELs
516 derived from the 1000 genome project[44]. We found the enrichment of 20-50bp INDELs in
517 CpG islands compared to the shuffled genomic control (Fig. 6d, e, p-value = $9X10^{-7}$ and
518 $2X10^{-5}$ in promoter and non-promoter regions, respectively)[42]. However, we did not find

519 any significant difference between hypermethylated and hypomethylated CpG islands
520 within either promoter or non-promoter CGIs (p-value = 0.43 and 0.30 in promoter and
521 non-promoter regions, respectively) (Fig. 6e). We suspect that the 2nd generation
522 sequencing, which the 1000 Genome Project relied upon, under characterized INDELs
523 greater than 10bp leading to the inability to discern the INDEL frequency difference
524 between hypomethylated and hypermethylated CGIs (Fig. 6d). To verify our findings
525 derived from HPRC long reads, we performed the same analysis but using the INDEL call
526 set derived from 3rd generation sequencing generated by the Human Genome Structural
527 Variation Consortium[45]. We confirmed the increased INDEL frequency in hypermethylated
528 CGIs from both promoter and non-promoter regions (Fig. 6f, g). Taken together, combining
529 variant calling and methylation calling using HiFi reads, we found that hypermethylated
530 CGIs had a higher frequency of INDELs greater than 10bp than both hypomethylated CGIs
531 and genomic background. Our observation here highlighted the advantage of 3rd
532 generation sequencing in INDEL calling in addition to methylation calling, and the need for
533 additional work to better understand the mechanism behind this phenomenon.

534 **Discussion**

535 Characterizing cytosine methylation is important for both evolutionary analysis and tissue-
536 specific gene regulation. However, conventional bisulfite or enzyme conversion-based
537 methods do not readily characterize methylation of repetitive regions or non-reference
538 regions. Using 3rd generation sequencing, we can call methylation during the base calling
539 step, greatly simplifying the methylation calling process and allowing new investigations
540 into the genomic methylation. Here we compared the methylation calling results between

541 HiFi, ONT, and WGBS from five HPRC LCL samples, and we found both HiFi and ONT can
542 measure CpG methylation in human with high correlation and concordance with WGBS.

543 How are genomic insertions, in particular transposable element insertions, methylated, and
544 how do they affect the methylation level of their flanking regions, has been a long-standing
545 question in the field. It has promoted several paradigms, especially in the context of
546 spreading of epigenetic modifications. For example, the hypermethylation of an inserted B2
547 element close to the Aprt promoter in mouse genome can spread to nearby CpG sites until
548 stopped by SP1 binding[10]. By examining CpG density flanking *Alu* insertions, Zhou et al. also
549 found evidence of CpG methylation spreading from old TEs to flanking regions during
550 evolution[12]. Recently, Lanciano et al. described examples of LINE1 insertion in cancer cell
551 lines affecting the methylation level of nearby regions up to 300bp upstream[46]. These
552 previous studies provided enormous insights on how newly inserted transposable
553 elements are epigenetically modified and how the modifications influence genomic
554 neighborhood. However, most represented specific examples or anecdotes, leaving the
555 question open as to what degree these examples represent the genome-scale rules or
556 exceptions.

557 Taking advantage of both SV and methylation calling using the HiFi reads from 32 samples
558 of HPRC Year1 dataset, we addressed this question by analyzing differential methylation
559 between alleles with polymorphic insertions at a genome-wide scale. Our analysis suggests
560 that the methylation of non-TE insertions is influenced by their genomic context.
561 Consistent with Lanciano et al., we did find limited methylation spreading from L1 to
562 nearby regions. However, the effect is very small (on average ~3% increase of methylation

563 within 300bp). The minimum spreading of methylation from polymorphic TE insertions to
564 their flanking regions we found suggests that methylation spreading is a slow evolutionary
565 process with limited effect after TE insertion but can potentially have large impact at
566 evolutionary timescale[12].

567 To visualize the methylation of non-reference insertions, we mapped reads to the high
568 quality individual assembly and displayed them on the WashU Epigenome Browser using
569 the genome comparison feature[35,47]. However, this approach is limited by the underlying
570 pairwise alignment and more crucially, the availability of high quality individual
571 assemblies. We hope the development of pangenome based tools would facilitate data
572 display of non-reference sequences in the future.

573 Using the HPRC long read methylation data, we also found hypermethylated CpG islands
574 enriched with INDELs, further clarifying the initial discovery from 2020[42]. We envision the
575 publicly available long read methylation data can be used to address other questions in the
576 biomedical field, and we invite other scientists to explore the rich, high-fidelity HPRC
577 dataset[31].

578 **Methods**

579 **Genome assembly and scaffolding**

580 The genome assemblies used in this study were produced for the draft human pangenome,
581 combining PacBio HiFi long-read and Illumina short-read sequencing for de novo assembly
582 with parental short-read data used to for phasing of these assemblies. Phased contig
583 assembly was performed using hifiasm[48]. Post-assembly, the RagTag toolset was employed

584 for reference-guided scaffolding, organizing contigs into structured, phased chromosomal
585 assemblies[49]. Validation was performed using BioNano optical mapping, in order to detect
586 and amend potential assembly inaccuracies[50]. For each of the 10 assemblies derived from
587 the five diploid samples, we created a whole-genome pairewise alignment BAM file with
588 hg38 using minimap2[51].

589 **Whole Genome Bisulfite Sequencing (WGBS)**

590 The cytosine methylation was estimated by mapping to reference genome hg38 or the
591 individual assemblies using the standard WGBS pipeline Bismark as described by Lee et
592 al[52,53]. WGBS QC data including bisulfite conversion rate and mapping ratios is available in
593 Supplementary Table 1.

594 **PacBio HiFi Methylation**

595 PacBio single-molecule high fidelity (HiFi) Circular Consensus Sequencing (CCS) reads with
596 kinetics tags can be generated using pbccs (https://github.com/nlhepler/pbccs). We then
597 predicted cytosine methylation probability per molecule in all CpG context by applying
598 primrose to PacBio HiFi reads (https://github.com/PacificBiosciences/jasmine). The CpG
599 methylation probability was saved as the ML and MM tags in the BAM file following
600 standard BAM format convention (http://samtools.github.io/hts-specs/SAMtags.pdf).
601 Unaligned HiFi BAM files were aligned to human reference genome hg38 using PacBio
602 pbmm2 (https://github.com/PacificBiosciences/pbmm2), and CpG methylation percentage
603 on the reference genome were calculated using PacBio pb-CpG-tools
604 (https://github.com/PacificBiosciences/pb-CpG-tools). We ran the methylation percentage

605 calculation twice for each alignment using both “count” and “model” as pileup (–pileup-
606 mode count and –pileup-mode model).

607 **Nanopore ONT methylation**

608 We downloaded raw sequencing signal of ONT Nanopore (FAST5 files) of the same five
609 samples and performed methylation calling using basecaller guppy with methylation model
610 remora (https://github.com/nanoporetech/remora). CpG methylation probability of ONT
611 reads were estimated using Nanopore basecaller Guppy v6 with remora methylation model
612 (v9.4.1) (https://community.nanoporetech.com). The CpG methylation probability were
613 saved as the standard ML and MM tags in the BAM format (http://samtools.github.io/hts-
614 specs/SAMtags.pdf). ONT unaligned BAM files were aligned to human reference genome
615 hg38 using minimap2 (https://github.com/lh3/minimap2), and methylation percentage
616 were calculated using modbam2bed (https://github.com/epi2me-labs/modbam2bed).

617 **Methylation comparison**

618 With genome-wide methylation percentage of all three methods calculated, we made
619 pairwise correlation heatmap, PCA plot and calculated the mean average error, 20%
620 concordance, and person correlation using R[54,55]. We also applied DSS to define pairwise
621 differential methylation regions (DMRs)[33,56].

622 **Long reads phasing**

623 We used HiPhase for HiFi reads phasing[57]. To determine in each phase block which
624 haplotype is paternal, and which is maternal, we downloaded HiFi reads used for
625 constructing paternal and maternal assemblies from hifiasm output gfa files (HG00621,

626 HG00741, HG01952, HG01978, HG03516). We labeled the haplotype paternal if the
627 number of reads labeled with it from the same phase block used for paternal assembly is
628 >10 times higher than that used for maternal assembly, otherwise maternal. Once we
629 separated HiFi reads to paternal and maternal with whatshap[58], we aligned them to the
630 individual paternal and maternal assemblies, accordingly using pbmm2
631 (https://github.com/PacificBiosciences/pbmm2). At last, we calculated the genome-wide
632 methylation percentage on individual assemblies using pb-CpG-tools
633 (https://github.com/PacificBiosciences/pb-CpG-tools).

634 To phase ONT reads according to the individual assemblies, we created a diploid VCF file
635 for each sample using dipcall (https://github.com/lh3/dipcall)[59] and haptagged all reads
636 using haplotype information from the VCF file. We then separated paternal and maternal
637 reads using whatshap[58] and aligned them to paternal and maternal assembly using
638 minimap2[60], respectively. At last, we converted the bam file to methylbed files with
639 modbam2bed (https://github.com/epi2me-labs/modbam2bed).

640 **Individual genome datahub**

641 For each pairwise genome alignment between hg38 and individual genome, we created a
642 genome-align track file using the pairwise genome alignment BAM file[35]. We built
643 individual genome assembly of both maternal and paternal chromosomes of the five
644 samples (10 assemblies in total) and aligned each of them to the hg38 to create genome-
645 align tracks. We also phased HiFi reads using HiPhase using alignment to hg38 and
646 assigned haplotypes to paternal/maternal for each phasing block based on hifiasm
647 assembly construction graph[48,57]. We then aligned maternal and paternal HiFi reads to the

648 maternal/paternal assembly separately for each of the five samples. Applying the genome-
649 align track we anchored gene, repeat, and CpG island annotations from hg38 onto the
650 individual genomes. We colored each CpG site by their methylation prediction (red:
651 methylated; blue: unmethylated) per HiFi read using the modbed tracks and displayed
652 piled up methylation level using the methylC track. Both modbed and methylC tracks were
653 mapped to the individual genome. Combining genome-align, methylC, and modbed track
654 files, we created three datahubs on the WashU Epigenome browser representing HiFi, ONT,
655 and WGBS data
656 (https://epigenomegateway.wustl.edu/browser/?genome=hg38&hub=https://wangcluste
657 r.wustl.edu/~xzhuo/hifi_methylation/hifi.all.json,
658 https://epigenomegateway.wustl.edu/browser/?genome=hg38&hub=https://wangcluster
659 .wustl.edu/~xzhuo/hifi_methylation/ont.all.json,
660 <https://wangcluster.wustl.edu/~xzhuo/hifi_methylation/wgbs_individual/wgbs.all.json)[3
661 6,37,47].

662 At last, we created a browser session for each panel in screenshot (Fig. 2d:
663 https://epigenomegateway.wustl.edu/browser/?sessionFile=https://wangcluster.wustl.ed
664 u/~xzhuo/hifi_methylation/Fig2D.session.json; Fig. 3e:
665 https://epigenomegateway.wustl.edu/browser/?sessionFile=https://wangcluster.wustl.ed
666 u/~xzhuo/hifi_methylation/Fig3E.session.json; Fig3. f:
667 https://epigenomegateway.wustl.edu/browser/?sessionFile=https://wangcluster.wustl.ed
668 u/~xzhuo/hifi_methylation/Fig3F.session.json; Fig. 5d-e:
669 https://epigenomegateway.wustl.edu/browser/?sessionFile=https://wangcluster.wustl.ed
670 u/~xzhuo/hifi_methylation/Alu.spread.session.json).

671 **Structural variation and TE annotation**

672 We obtained the structural variations of each sample from HPRC consortium (Download[31].
673 To further annotate transposable element insertions among all structural variations, we
674 ran xTEA on the same PacBio HiFi library mapped to hg38[34]. We then intersected the
675 coordinates of TE insertions with SV and annotated which SVs are TE insertions (Supp.
676 Table 2).

677 With TE insertions annotated in all samples, we developed two methods to calculate
678 methylation percentage per CpG site of all TE insertions. With individual assembly
679 available, we extracted TE sequences and their methylation percentage based on the
680 insertion coordinates after aligning long-reads with methylation to the individual genomes
681 and calculating methylation percentage using pb-CpG-tools or modkit pileup. We then
682 constructed a new unmapped BAM file of inserted TE copies and attached methylation
683 percentage using modified base tags (MM and ML) (bam.creation.py). At last, we mapped
684 the new BAM file to TE consensus using Minimap2 to align CpG methylation percentage to
685 the TE consensus.

686 We also made another script for samples mapped to reference genome. Using a BED file of
687 annotated insertion sites, it can extract inserted and soft-clipped sequences from raw reads
688 and generate a new unmapped BAM file of inserted sequences with base modification tags
689 (MM and ML) (bam.extractInsertion.py). We can then align the unmapped BAM file to TE
690 consensus to create CpG methylation pileup results.

691 To calculate average methylation of certain regions, we extracted methylation prediction
692 value of each CpG site of each read from bam files with methylation tags with a python

693 script (python bam.Mmtag.regions.oo.py -b <bam> -r <insertion.bed> -o
694 <ins.methylC.reldist.txt> -f <bed>) available in Github
695 repository(https://github.com/xzhuo/modbamUtil)

696 To evaluate the relationship between CpG methylation and indels, we aggregated
697 methylation prediction of all 32 samples and calculated the average CpG methylation of
698 each CpG Island (CGI)[61].

699 **Insertion and flanking methylation calculation**

700 We selected non-reference insertions from annotated SVs and used a custom python script
701 in the Github repository(https://github.com/xzhuo/modbamUtil) to calculate average
702 methylation of the insertion and flanking regions (python read.methylation.mean.py -i
703 <input> -o <output> -s <start> -l <length> -c). Two-sided F-test were performed to
704 compare methylation variation.

705 **CpG island methylation and INDEL frequency calculation**

706 Unmasked CpG island annotation file was downloaded from UCSC
707 (https://hgdownload.soe.ucsc.edu/goldenPath/hg38/database/cpgIslandExtUnmasked.txt
708 .gz). For each CpG site on hg38, we calculated the median methylation percentage of the 32
709 HiFi methylation. We then calculated the average methylation percentage of each CpG
710 island using the median methylation.

711 After separating CpG islands to hypermethylated and hypomethylated ones using 50%
712 methylation as cut-off, we calculated the INDEL frequency of different sizes within CpG
713 islands. To create the background INDEL frequency in the genome, we shuffled CpG islands

714 in the genome and repeated the calculation. INDEL frequency difference were compared
715 using two-sided t-test.

716 We also repeated the same calculation using same methylation data and INDEL callset from
717 1000 genome project and HGSVC project[44,45].

718 **Ethics and Inclusion statement**

719 The research is conceptualized and conducted by the colleagues from Washington
720 University and HPRC consortium using both publicly available and newly generated data.
721 Being a study about overall human population, it is not locally relevant. All results and raw
722 data would be open to public upon publication.

723 **Data availability statement**

724 HiFi reads with methylation are available at https://s3-us-west-2.amazonaws.com/human-
725 pangenomics/index.html?prefix=submissions/560047a6-6d16-4b0c-aac9-7d0c83e2188e-
726 -HIFI-METHYLATION-READS/.

727 ONT reads are available at https://s3-us-west-2.amazonaws.com/human-
728 pangenomics/index.html?prefix=submissions/CBC1E2D6-3925-4CD0-9682-
729 555C64F81B02--UCSC_HPRC_ONT_Y1_WTOPUP_GUPPY6/.

730 Variants calls of the same individuals were available at https://s3-us-west-
731 2.amazonaws.com/human-pangenomics/index.html?prefix=submissions/B581EBA7-
732 8BDE-4C7C-9DEA-78B99A051155--Yale_HPP_Year1_Variant_Calls/.

733 WGBS illumina reads will be make available at NIH GEO dataset GSE261315.

734 Processed data mapped to either hg38 or individual assemblies are available in datahub
735 provided above in the "individual genome hub" session.

736 **Code availability**

737 All custom scripts used in the manuscript are available at
738 https://github.com/xzhuo/modbamUtil.

739

740 **Reference**

741 1. Bird, A. P. CpG-rich islands and the function of DNA methylation. *Nature* **321**, 209–
742 213 (1986).

743 2. The ENCODE Project Consortium. An integrated encyclopedia of DNA elements in
744 the human genome. *Nature* **489**, 57–74 (2012).

745 3. Roadmap Epigenomics Consortium *et al.* Integrative analysis of 111 reference
746 human epigenomes. *Nature* **518**, 317–330 (2015).

747 4. Moore, L. D., Le, T. & Fan, G. DNA Methylation and Its Basic Function.
748 *Neuropsychopharmacol* **38**, 23–38 (2013).

749 5. Jones, P. A. Functions of DNA methylation: islands, start sites, gene bodies and
750 beyond. *Nat Rev Genet* **13**, 484–492 (2012).

751 6. Smith, Z. D., Hetzel, S. & Meissner, A. DNA methylation in mammalian development
752 and disease. *Nat Rev Genet* (2024) doi:10.1038/s41576-024-00760-8.

753 7. Sved, J. & Bird, A. The expected equilibrium of the CpG dinucleotide in vertebrate
754 genomes under a mutation model. *Proceedings of the National Academy of Sciences of the*
755 *United States of America* **87**, 4692–4696 (1990).

756 8. Han, L., Su, B., Li, W.-H. & Zhao, Z. CpG island density and its correlations with
757 genomic features in mammalian genomes. *Genome Biology* **9**, R79 (2008).

758 9. Turker, M. S. The establishment and maintenance of DNA methylation patterns in
759 mouse somatic cells. *Seminars in Cancer Biology* **9**, 329–337 (1999).

760 10. Turker, M. S. Gene silencing in mammalian cells and the spread of DNA methylation.
761 *Oncogene* **21**, 5388–5393 (2002).

762 11. Sentmanat, M. F. & Elgin, S. C. R. Ectopic assembly of heterochromatin in Drosophila
763 melanogaster triggered by transposable elements. *Proceedings of the National Academy of*
764 *Sciences* **109**, 14104–14109 (2012).

765 12. Zhou, W., Liang, G., Molloy, P. L. & Jones, P. A. DNA methylation enables transposable
766 element-driven genome expansion. *PNAS* **117**, 19359–19366 (2020).

767 13. Hodges, E. *et al.* High definition profiling of mammalian DNA methylation by array
768 capture and single molecule bisulfite sequencing. *Genome Res.* **19**, 1593–1605 (2009).

769 14. Weber, M. *et al.* Chromosome-wide and promoter-specific analyses identify sites of
770 differential DNA methylation in normal and transformed human cells. *Nat Genet* **37**, 853–
771 862 (2005).

772 15. Meissner, A. *et al.* Genome-scale DNA methylation maps of pluripotent and
773 differentiated cells. *Nature* **454**, 766–770 (2008).

774 16. Maunakea, A. K. *et al.* Conserved role of intragenic DNA methylation in regulating
775 alternative promoters. *Nature* **466**, 253–257 (2010).

776 17. Lister, R. *et al.* Human DNA methylomes at base resolution show widespread
777 epigenomic differences. *Nature* **462**, 315–322 (2009).

778 18. Cokus, S. J. *et al.* Shotgun bisulphite sequencing of the Arabidopsis genome reveals
779 DNA methylation patterning. *Nature* **452**, 215–219 (2008).

780 19. Laurent, L. *et al.* Dynamic changes in the human methylome during differentiation.
781 *Genome Res.* **20**, 320–331 (2010).

782 20. Vaisvila, R. *et al.* Enzymatic methyl sequencing detects DNA methylation at single-
783 base resolution from picograms of DNA. *Genome Res.* **31**, 1280–1289 (2021).

784 21. Nunn, A., Otto, C., Fasold, M., Stadler, P. F. & Langenberger, D. Manipulating base
785 quality scores enables variant calling from bisulfite sequencing alignments using
786 conventional bayesian approaches. *BMC Genomics* **23**, 477 (2022).

787 22. van Dijk, E. L., Jaszczyszyn, Y., Naquin, D. & Thermes, C. The Third Revolution in
788 Sequencing Technology. *Trends in Genetics* **34**, 666–681 (2018).

789 23. Sigurpalsdottir, B. D. *et al.* A comparison of methods for detecting DNA methylation
790 from long-read sequencing of human genomes. *Genome Biol* **25**, 69 (2024).

791 24. Flusberg, B. A. *et al.* Direct detection of DNA methylation during single-molecule,
792 real-time sequencing. *Nat Methods* **7**, 461–465 (2010).

793 25. Schatz, M. C. Nanopore sequencing meets epigenetics. *Nat Methods* **14**, 347–348
794 (2017).

795 26. Tse, O. Y. O. *et al.* Genome-wide detection of cytosine methylation by single molecule
796 real-time sequencing. *PNAS* **118**, (2021).

797 27. Ni, P. *et al.* DNA 5-methylcytosine detection and methylation phasing using PacBio
798 circular consensus sequencing. 2022.02.26.482074 Preprint at
799 https://doi.org/10.1101/2022.02.26.482074 (2022).

800 28. Liu, Y. *et al.* DNA methylation-calling tools for Oxford Nanopore sequencing: a
801 survey and human epigenome-wide evaluation. *Genome Biology* **22**, 295 (2021).

802 29. Gerdes, P. *et al.* Locus-resolution analysis of L1 regulation and retrotransposition
803 potential in mouse embryonic development. *Genome Res.* (2023)
804 doi:10.1101/gr.278003.123.

805 30. Bodea, G. O. *et al.* LINE-1 retrotransposons contribute to mouse PV interneuron
806 development. *Nat Neurosci* 1–11 (2024) doi:10.1038/s41593-024-01650-2.

807 31. Liao, W.-W. *et al.* A draft human pangenome reference. *Nature* **617**, 312–324 (2023).

808 32. Ernst, J. & Kellis, M. ChromHMM: automating chromatin-state discovery and
809 characterization. *Nature methods* **9**, 215–216 (2012).

810 33. Wu, H. *et al.* Detection of differentially methylated regions from whole-genome
811 bisulfite sequencing data without replicates. *Nucleic Acids Res* **43**, e141 (2015).

812 34. Chu, C. *et al.* Comprehensive identification of transposable element insertions using
813 multiple sequencing technologies. *Nat Commun* **12**, 3836 (2021).

814 35. Zhuo, X. *et al.* Comparing genomic and epigenomic features across species using the
815 WashU Comparative Epigenome Browser. *Genome Res.* **33**, 824–835 (2023).

816 36. Zhou, X., Li, D., Lowdon, R. F., Costello, J. F. & Wang, T. methylC Track: visual
817 integration of single-base resolution DNA methylation data on the WashU EpiGenome
818 Browser. *Bioinformatics* **30**, 2206–2207 (2014).

819 37. Li, D., Zhuo, X., Harrison, J. K., Liu, S. & Wang, T. Modbed track: Visualization of
820 modified bases in single-molecule sequencing. *Cell Genomics* **3**, 100455 (2023).

821 38. Zhang, Y. *et al.* Repetitive elements and enforced transcriptional repression co-
822 operate to enhance DNA methylation spreading into a promoter CpG-island. *Nucleic Acids*
823 *Research* **40**, 7257–7268 (2012).

824 39. Lee, Y., Lee, J., Kim, J. & Kim, Y.-J. Insertion variants missing in the human reference
825 genome are widespread among human populations. *BMC Biology* **18**, 167 (2020).

826 40. Batzer, M. A. & Deininger, P. L. Alu repeats and human genomic diversity. *Nature*
827 *Reviews Genetics* **3**, 370–379 (2002).

828 41. Coulondre, C., Miller, J. H., Farabaugh, P. J. & Gilbert, W. Molecular basis of base
829 substitution hotspots in Escherichia coli. *Nature* **274**, 775–780 (1978).

830 42. Zhuo, X., Du, A. Y., Pehrsson, E. C., Li, D. & Wang, T. Epigenomic differences in the
831 human and chimpanzee genomes are associated with structural variation. *Genome Res.* **31**,
832 279–290 (2021).

833 43. Deaton, A. M. & Bird, A. CpG islands and the regulation of transcription. *Genes &*
834 *development* **25**, 1010–1022 (2011).

835 44. The 1000 Genomes Project Consortium. A global reference for human genetic
836 variation. *Nature* **526**, 68–74 (2015).

837 45. Ebert, P. *et al.* Haplotype-resolved diverse human genomes and integrated analysis
838 of structural variation. *Science* **372**, eabf7117 (2021).

839 46. Lanciano, S. *et al.* Locus-level L1 DNA methylation profiling reveals the epigenetic
840 and transcriptional interplay between L1s and their integration sites. *Cell Genomics* **4**,
841 100498 (2024).

842 47. Li, D. *et al.* WashU Epigenome Browser update 2022. *Nucleic Acids Research* **50**,
843 W774–W781 (2022).

844 48. Cheng, H., Concepcion, G. T., Feng, X., Zhang, H. & Li, H. Haplotype-resolved de novo
845 assembly using phased assembly graphs with hifiasm. *Nat Methods* **18**, 170–175 (2021).

846 49. Alonge, M. *et al.* Automated assembly scaffolding using RagTag elevates a new
847 tomato system for high-throughput genome editing. *Genome Biology* **23**, 258 (2022).

848 50. Lam, E. T. *et al.* Genome mapping on nanochannel arrays for structural variation
849 analysis and sequence assembly. *Nat Biotechnol* **30**, 771–776 (2012).

850 51. Li, H. New strategies to improve minimap2 alignment accuracy. *Bioinformatics* **37**,
851 4572–4574 (2021).

852 52. Krueger, F. & Andrews, S. R. Bismark: a flexible aligner and methylation caller for
853 Bisulfite-Seq applications. *Bioinformatics* **27**, 1571–1572 (2011).

854 53. Lee, H. J. *et al.* Regenerating zebrafish fin epigenome is characterized by stable
855 lineage-specific DNA methylation and dynamic chromatin accessibility. *Genome Biology* **21**,
856 52 (2020).

857 54. Wickham, H. *ggplot2*. (Springer Science & Business Media, New York, NY, 2009).
858 doi:10.1007/978-0-387-98141-3.

859 55. Emerson, J. W. *et al.* The Generalized Pairs Plot. *Journal of Computational and*
860 *Graphical Statistics* **22**, 79–91 (2013).

861 56. Feng, H., Conneely, K. N. & Wu, H. A Bayesian hierarchical model to detect
862 differentially methylated loci from single nucleotide resolution sequencing data. *Nucleic*
863 *Acids Res* **42**, e69 (2014).

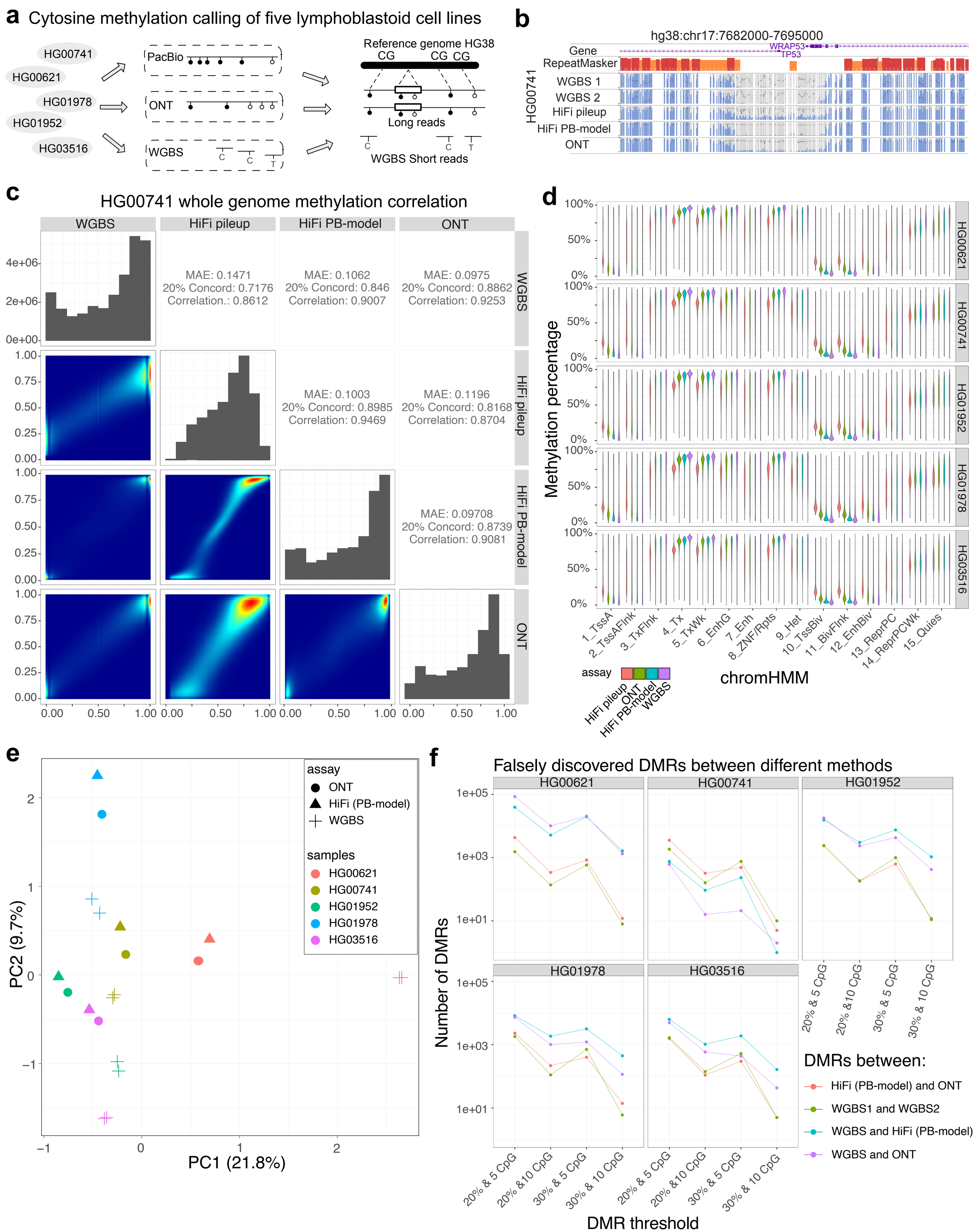
864 57. Holt, J. M. *et al.* HiPhase: jointly phasing small, structural, and tandem repeat
865 variants from HiFi sequencing. *Bioinformatics* **40**, btae042 (2024).

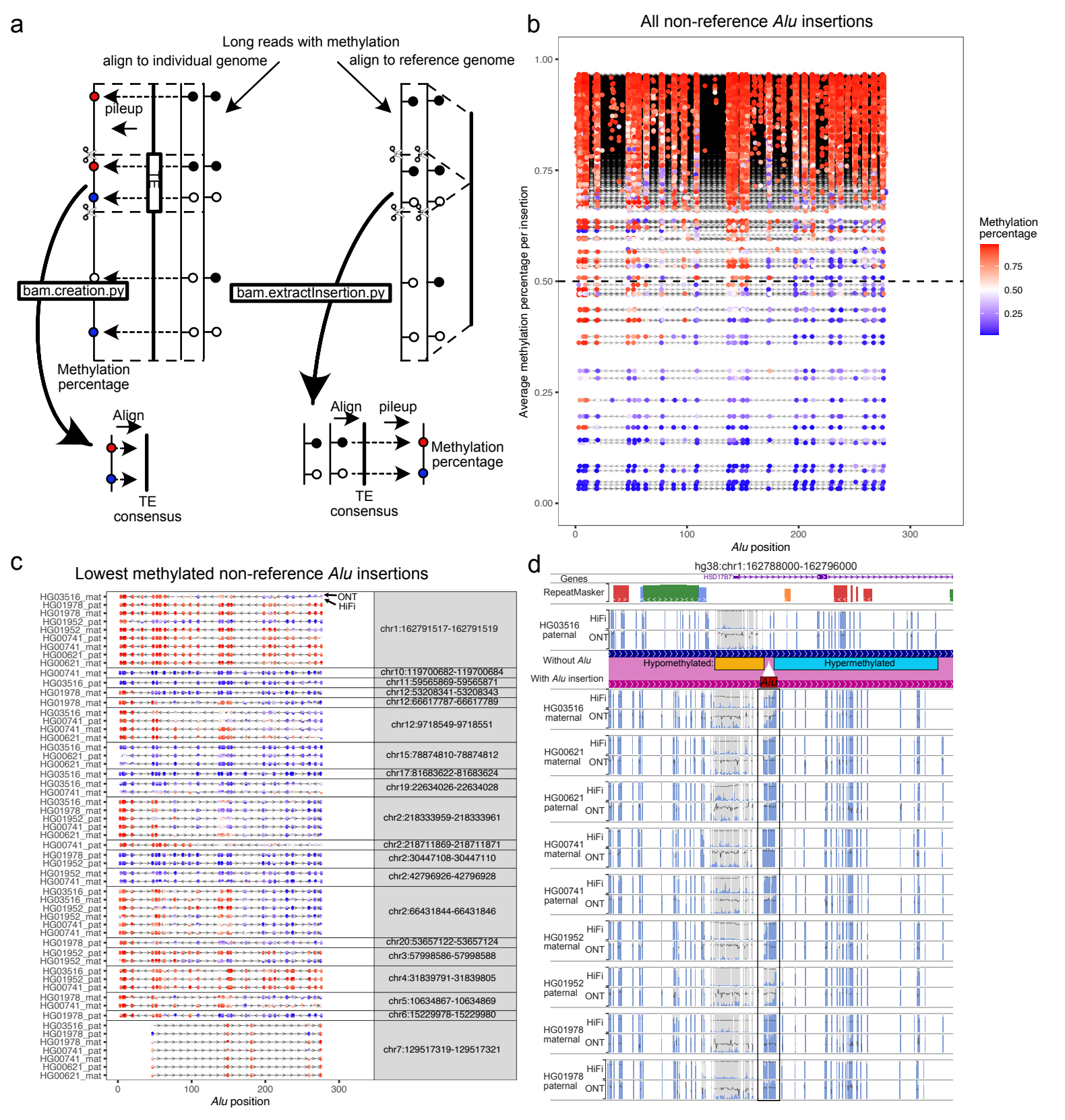
866 58. Patterson, M. *et al.* WhatsHap: Weighted Haplotype Assembly for Future-Generation
867 Sequencing Reads. *Journal of Computational Biology* **22**, 498–509 (2015).

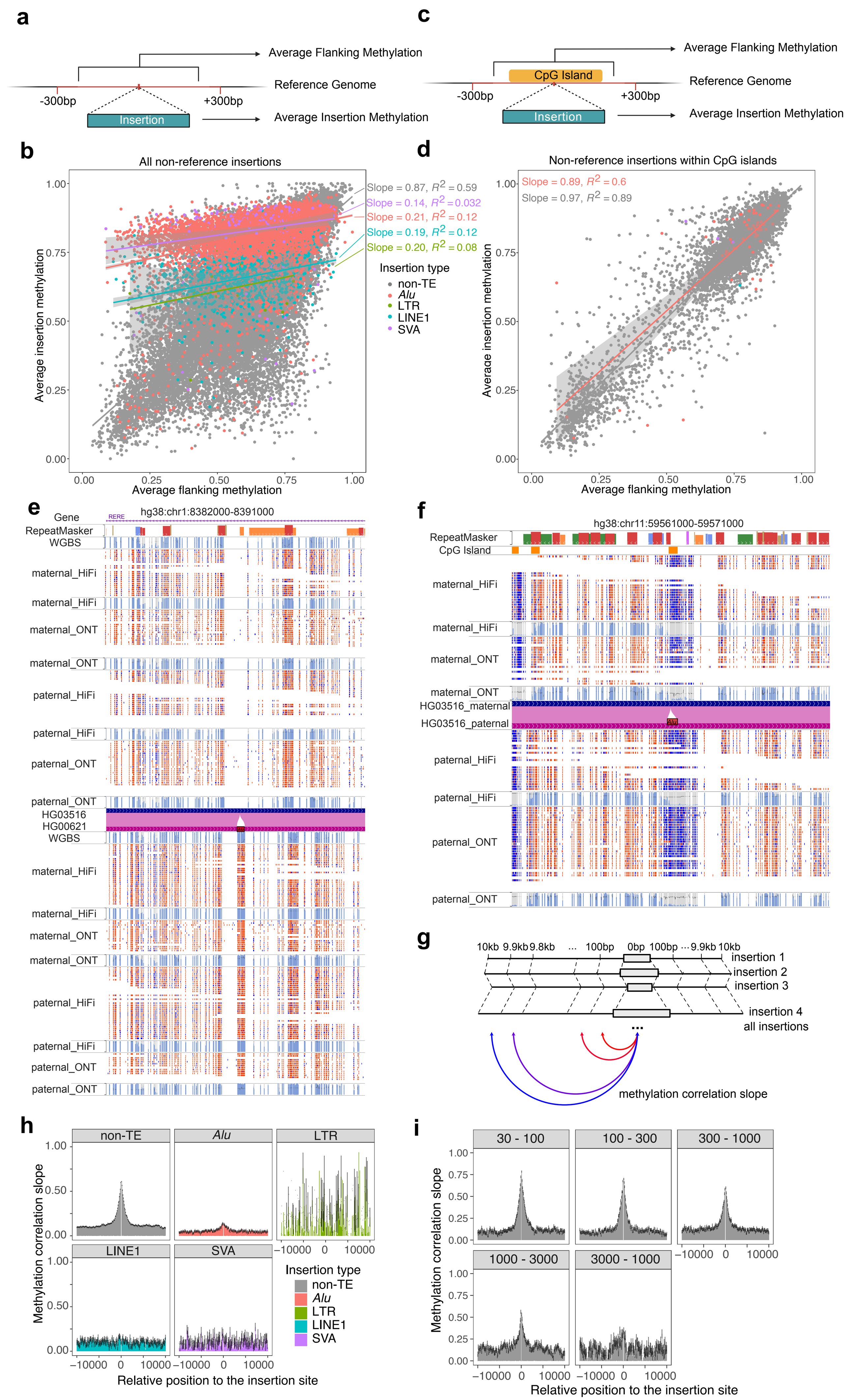
868 59. Li, H. *et al.* A synthetic-diploid benchmark for accurate variant-calling evaluation.
869 *Nat Methods* **15**, 595–597 (2018).

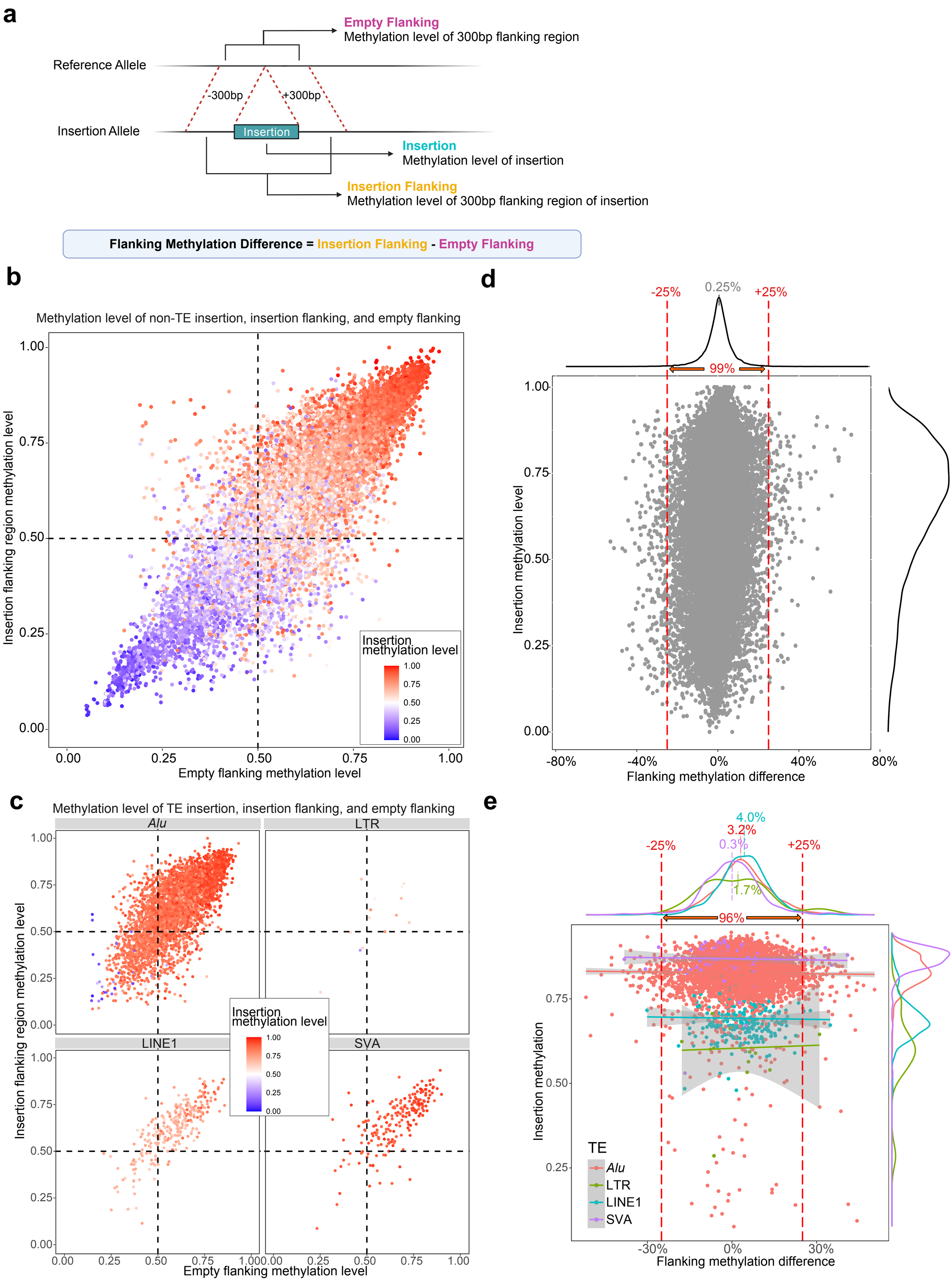
870 60. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**,
871 3094–3100 (2018).

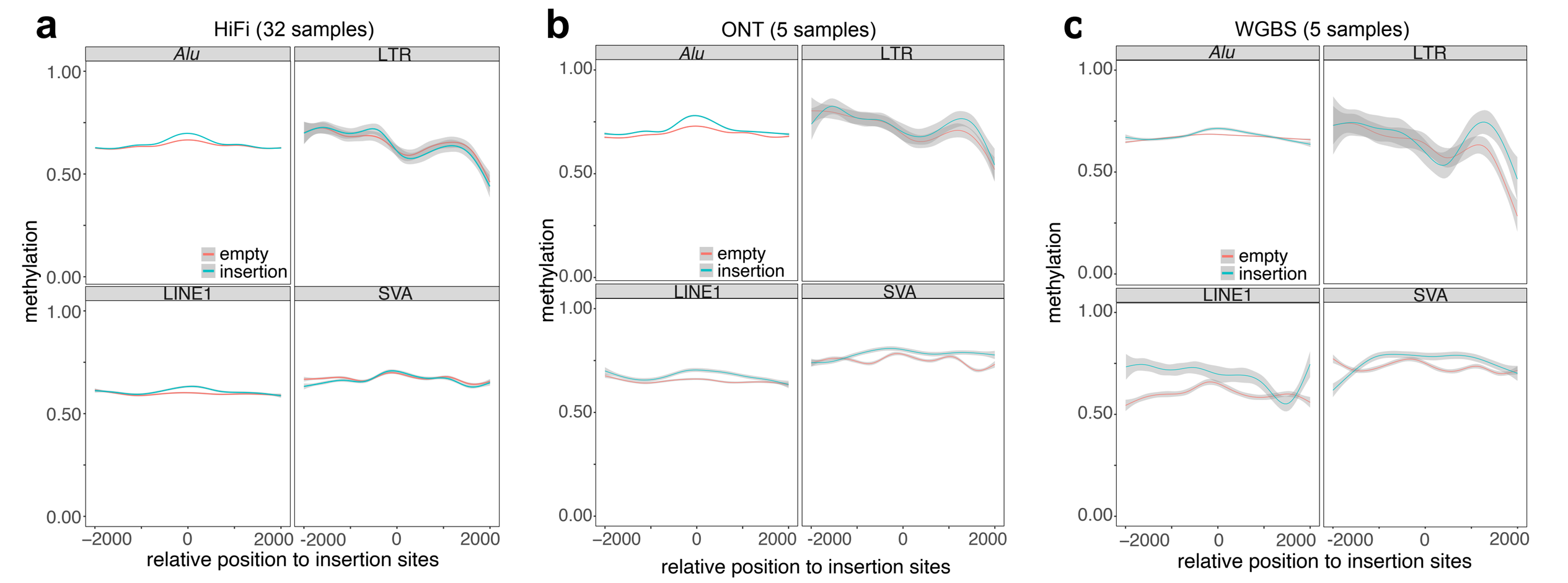
872 61. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing
873 genomic features. *Bioinformatics* **26**, 841–842 (2010).

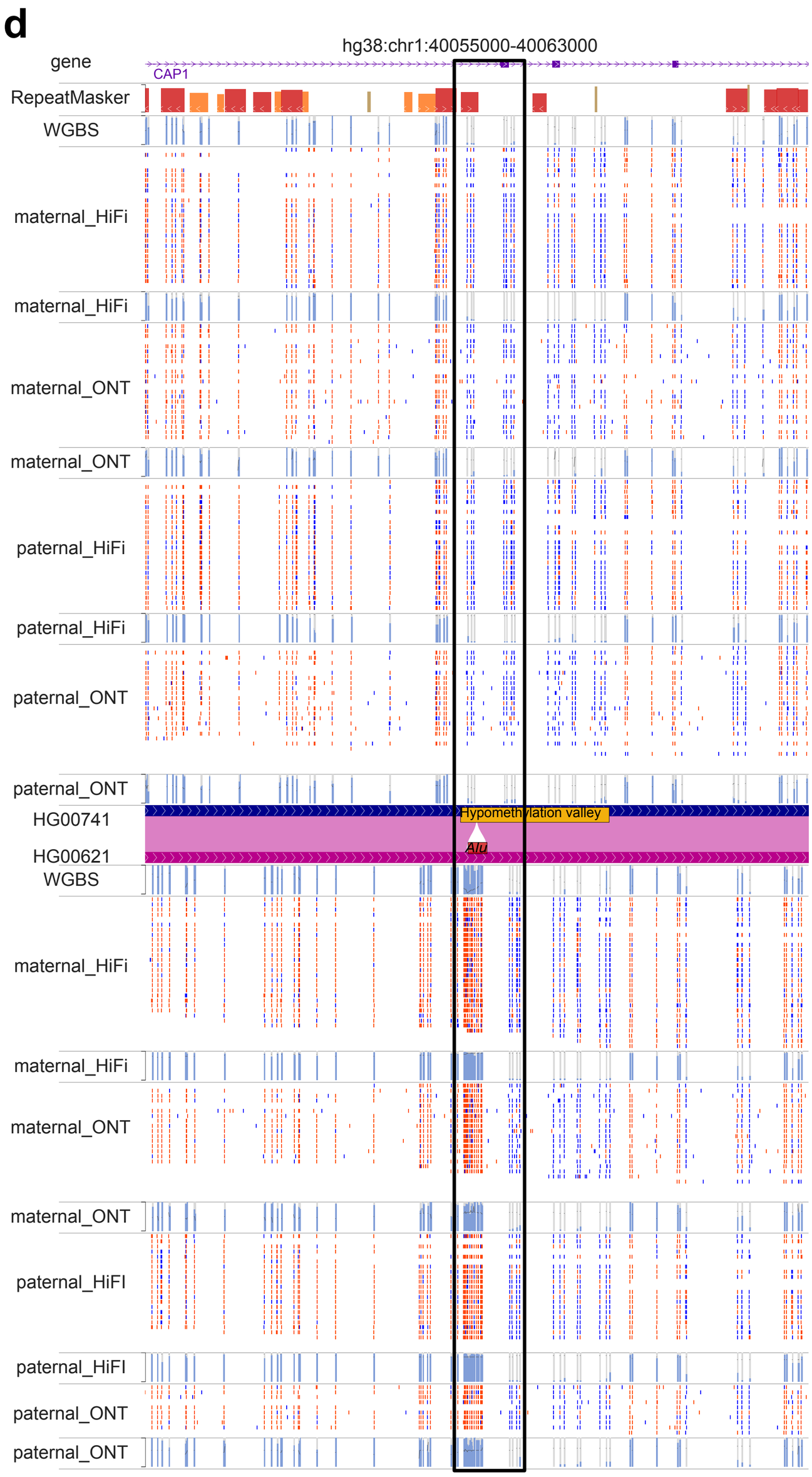

a
Cytosine methylation calling of five lymphoblastoid cell lines
HG00741
HG00621
HG01978
HG01952
HG03516
PacBio
ONT
WGBS
Reference genome HG38
CG
CG CG
Long reads
WGBS Short reads
b
hg38:chr17:7682000-7695000
Gene
WRAP53
TP53
RepeatMasker
HG00741
WGBS 1
WGBS 2
HiFi pileup
HiFi PB-model
ONT
c
HG00741 whole genome methylation correlation
WGBS
HiFi pileup
HiFi PB-model
ONT
MAE: 0.1471
20% Concord: 0.7176
Correlation.: 0.8612
MAE: 0.1062
20% Concord: 0.846
Correlation: 0.9007
MAE: 0.0975
20% Concord: 0.8862
Correlation: 0.9253
MAE: 0.1003
20% Concord: 0.8985
Correlation: 0.9469
MAE: 0.1196
20% Concord: 0.8168
Correlation: 0.8704
MAE: 0.09708
20% Concord: 0.8739
Correlation: 0.9081
d
Methylation percentage
HG00621
HG00741
HG01952
HG01978
HG03516
1_TssA
2_TssAFlnk
3_TxFlnk
4_Tx
5_TxWk
6_EnhG
7_Enh
8_ZNF/Rpts
9_Het
10_TssBiv
11_BivFlnk
12_EnhBiv
13_ReprPC
14_ReprPCWk
15_Quies
chromHMM
assay
HiFi pileup
ONT
HiFi PB-model
WGBS
e
assay
ONT
HiFi (PB-model)
WGBS
samples
HG00621
HG00741
HG01952
HG01978
HG03516
PC2 (9.7%)
PC1 (21.8%)
f
Falsely discovered DMRs between different methods
HG00621
HG00741
HG01952
HG01978
HG03516
Number of DMRs
20% & 5 CpG
20% &10 CpG
30% & 5 CpG
30% & 10 CpG
DMR threshold
DMRs between:
HiFi (PB-model) and ONT
WGBS1 and WGBS2
WGBS and HiFi (PB-model)
WGBS and ONT

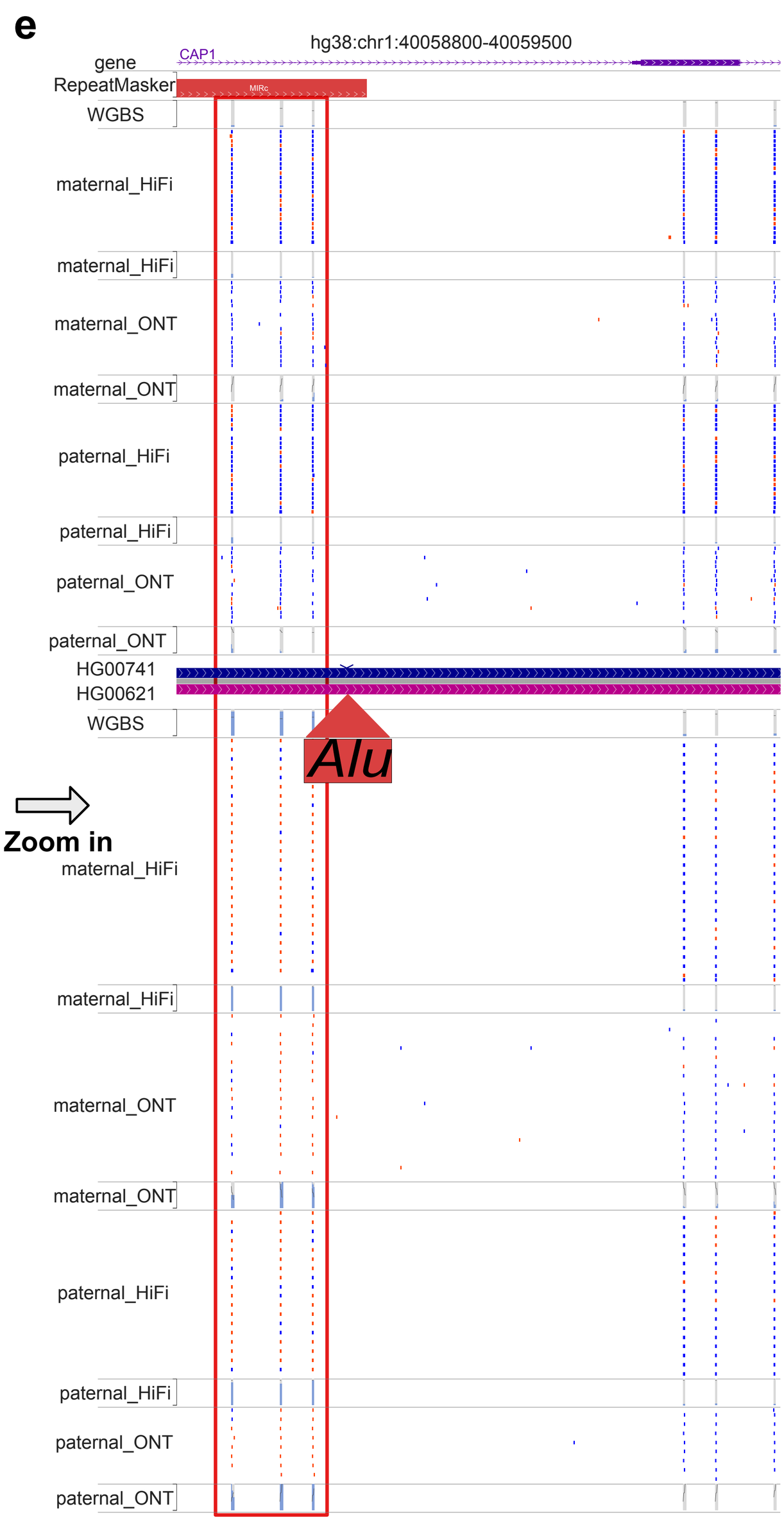

a
Long reads with methylation
align to individual genome
align to reference genome
pileup
TE
bam.creation.py
bam.extractInsertion.py
Methylation percentage
Align
TE consensus
Align
pileup
Methylation percentage
TE consensus
b
All non-reference Alu insertions
Average methylation percentage per insertion
1.00
0.75
0.50
0.25
0.00
0
100
200
300
Alu position
Methylation percentage
0.75
0.50
0.25
c
Lowest methylated non-reference Alu insertions
ONT
HiFi
HG03516_mat
HG01978_pat
HG01978_mat
HG01952_pat
HG01952_mat
HG00741_pat
HG00741_mat
HG00621_pat
HG00621_mat
chr1:162791517-162791519
chr10:119700682-119700684
chr11:59565869-59565871
chr12:53208341-53208343
chr12:66617787-66617789
chr12:9718549-9718551
chr15:78874810-78874812
chr17:81683622-81683624
chr19:22634026-22634028
chr2:218333959-218333961
chr2:218711869-218711871
chr2:30447108-30447110
chr2:42796926-42796928
chr2:66431844-66431846
chr20:53657122-53657124
chr3:57998586-57998588
chr4:31839791-31839805
chr5:10634867-10634869
chr6:15229978-15229980
chr7:129517319-129517321
0
100
200
300
Alu position
d
hg38:chr1:162788000-162796000
Genes
RepeatMasker
HG03516 paternal
HG03516 maternal
HG00621 maternal
HG00621 paternal
HG00741 maternal
HG00741 paternal
HG01952 maternal
HG01952 paternal
HG01978 maternal
HG01978 paternal
HiFi
ONT
Without Alu
Hypomethylated
Hypermethylated
With Alu insertion
Alu

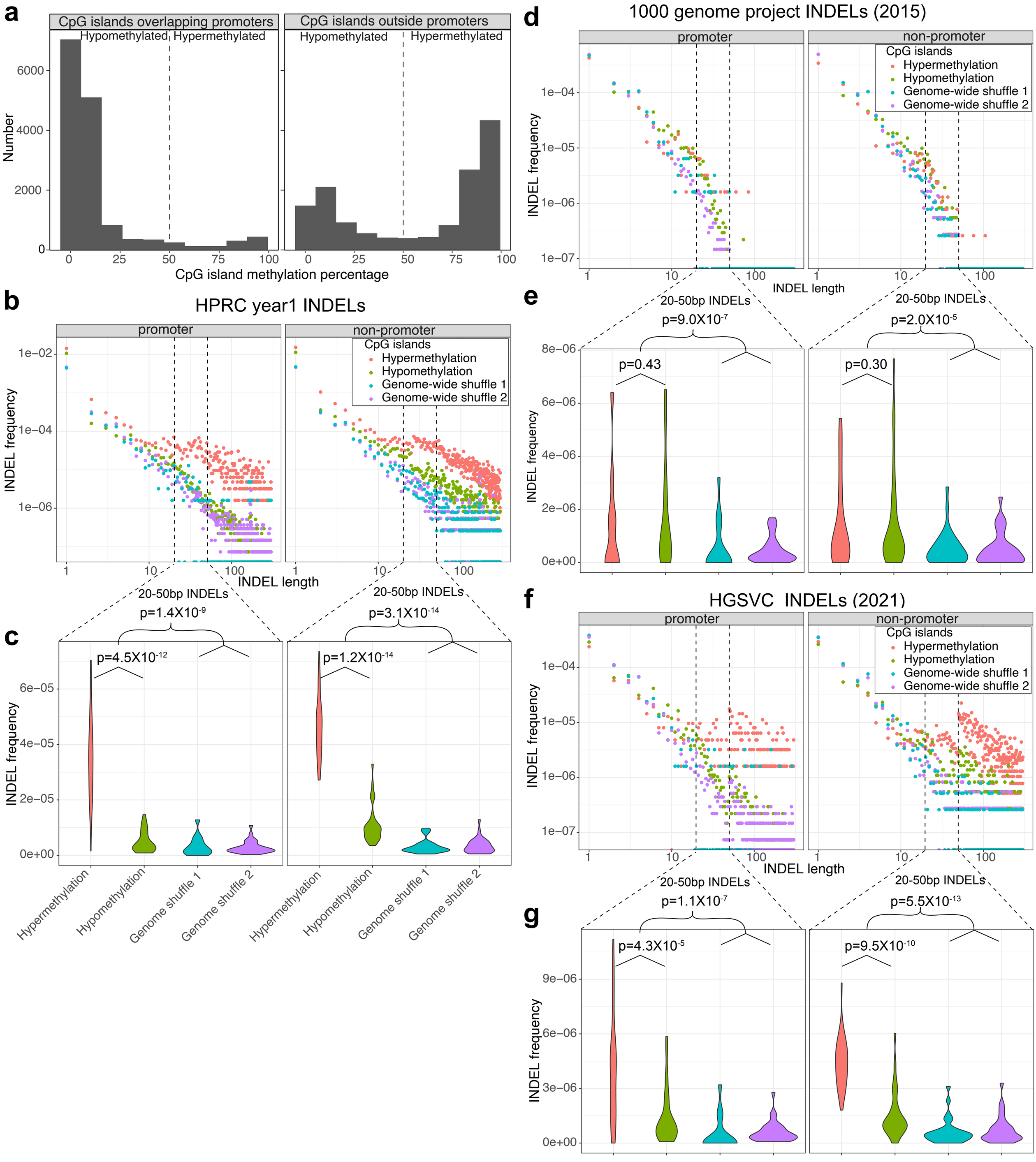

a
Average Flanking Methylation
Reference Genome
-300bp
+300bp
Insertion
Average Insertion Methylation
c
CpG Island
b
All non-reference insertions
Slope = 0.87, R2 = 0.59
Slope = 0.14, R2 = 0.032
Slope = 0.21, R2 = 0.12
Slope = 0.19, R2 = 0.12
Slope = 0.20, R2 = 0.08
Insertion type
non-TE
Alu
LTR
LINE1
SVA
Average insertion methylation
Average flanking methylation
d
Non-reference insertions within CpG islands
Slope = 0.89, R2 = 0.6
Slope = 0.97, R2 = 0.89
e
hg38:chr1:8382000-8391000
Gene
RERE
RepeatMasker
WGBS
maternal_HiFi
maternal_ONT
paternal_HiFi
paternal_ONT
HG03516
HG00621
Alu
f
hg38:chr11:59561000-59571000
CpG Island
HG03516_maternal
HG03516_paternal
g
10kb 9.9kb 9.8kb ... 100bp 0bp 100bp ... 9.9kb 10kb
insertion 1
insertion 2
insertion 3
insertion 4
all insertions
methylation correlation slope
h
non-TE
Alu
LTR
LINE1
SVA
Methylation correlation slope
Relative position to the insertion site
i
30 - 100
100 - 300
300 - 1000
1000 - 3000
3000 - 1000


a
Empty Flanking
Methylation level of 300bp flanking region
Reference Allele
-300bp
+300bp
Insertion Allele
Insertion
Insertion
Methylation level of insertion
Insertion Flanking
Methylation level of 300bp flanking region of insertion
Flanking Methylation Difference = Insertion Flanking - Empty Flanking
b
Methylation level of non-TE insertion, insertion flanking, and empty flanking
Insertion flanking region methylation level
Empty flanking methylation level
Insertion methylation level
1.00
0.75
0.50
0.25
0.00
d
-25%
0.25%
+25%
99%
Insertion methylation level
Flanking methylation difference
-80%
-40%
0%
40%
80%
c
Methylation level of TE insertion, insertion flanking, and empty flanking
Alu
LTR
LINE1
SVA
Insertion flanking region methylation level
Empty flanking methylation level
Insertion methylation level
e
4.0%
3.2%
0.3%
1.7%
-25%
+25%
96%
Insertion methylation
Flanking methylation difference
-30%
0%
30%
TE
Alu
LTR
LINE1
SVA


a
HiFi (32 samples)
b
ONT (5 samples)
c
WGBS (5 samples)
Alu
LTR
LINE1
SVA
methylation
empty
insertion
relative position to insertion sites


d
hg38:chr1:40055000-40063000
gene
CAP1
RepeatMasker
WGBS
maternal_HiFi
maternal_ONT
paternal_HiFi
paternal_ONT
HG00741
HG00621
Hypomethylation valley
Alu


e
hg38:chr1:40058800-40059500
gene
CAP1
RepeatMasker
MIRc
WGBS
maternal_HiFi
maternal_ONT
paternal_HiFi
paternal_ONT
HG00741
HG00621
Alu
Zoom in


a
CpG islands overlapping promoters
CpG islands outside promoters
Hypomethylated
Hypermethylated
Number
CpG island methylation percentage
b
HPRC year1 INDELs
promoter
non-promoter
CpG islands
Hypermethylation
Hypomethylation
Genome-wide shuffle 1
Genome-wide shuffle 2
INDEL frequency
INDEL length
20-50bp INDELs
p=1.4X10-9
p=3.1X10-14
c
p=4.5X10-12
p=1.2X10-14
Hypermethylation
Hypomethylation
Genome shuffle 1
Genome shuffle 2
d
1000 genome project INDELs (2015)
e
p=9.0X10-7
p=2.0X10-5
p=0.43
p=0.30
f
HGSVC INDELs (2021)
p=1.1X10-7
p=5.5X10-13
g
p=4.3X10-5
p=9.5X10-10