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], Juan Jiang[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

ABSTRACT

Cytosine methylation, a crucial epigenetic modification, plays a vital role in genomic regulation. Leveraging the advancements in third-generation sequencing, we investigated the methylation patterns of non-reference insertions of human lymphoblastoid cell lines (LCLs), particularly polymorphic transposable elements (TEs). We validated the high concordance between long-read methylation calls and conventional whole genome bisulfite sequencing (WGBS) method. By characterizing thousands of polymorphic TE insertions genome-wide using long reads from the draft Human Pangenome Reference, we aimed to establish general rules of TE methylation by addressing three key questions: 1) what is the methylation profile of each insertion? 2) do newly inserted TEs adopt the methylation pattern of their genomic context? and 3) do new TE insertions affect the methylation of their flanking regions? While most non-TE insertions exhibit DNA methylation patterns consistent with their genomic context, TE insertions are generally highly methylated, exhibiting distinct, class-specific patterns, and with profound variation within TE bodies. A small percentage of *Alu* insertions are hypomethylated, particularly those inserted within hypomethylated CpG islands. By comparing DNA methylation of flanking regions of TE insertions between individuals with and without the TE insertions, we revealed that majority of TEs exhibited minimal impact on nearby regions, although numerous exceptions exist where the methylation status of both L1 and *Alu* insertions "leak" into nearby regions, leading to either methylation spreading or hypomethylation sloping shores. In conclusion, we demonstrated the methylation calling capability of third-generation sequencing and its unique advantage in characterizing epigenomic features

41 within non-reference positions. While TE insertions primarily exhibit methylation patterns
42 restricted within their boundaries, some TEs are able to engage in context-dependent
43 complex interactions with genomic neighborhood.

44 INTRODUCTION

45 Cytosine methylation at CpG sites is prevalent in vertebrate genomes. The majority of CpG
46 sites in the human genome are methylated in most somatic tissues (Bird 1986; The
47 ENCODE Project Consortium 2012; Roadmap Epigenomics Consortium et al. 2015). It is
48 generally believed that the CpG methylation is a repressive mark commonly associated
49 with heterochromatin and transposable elements (TEs) silencing, but is absent from
50 promoter of active genes (Moore et al. 2013). In addition, methylation pattern is dynamic
51 during development and disease progression (Jones 2012; Smith et al. 2024). During
52 evolution, methylation-mediated deamination converts methylated cytosine to thymine,
53 causing the general depletion of CpG sites in most vertebrate genomes (Bird 2002).
54 However, there are CpG islands (CGIs) in the genome that are relatively enriched with GC
55 content and CG motif (generally defined by sequences > 200bp with >=50% GC content and
56 the observed/expected CpG ratio >=0.6) (Fazzari and Greally 2004). It is believed that they
57 represent conserved unmethylated regions in germ cells during development (Jones 2012;
58 Smallwood et al. 2011) and many were found in the promoters of conserved genes (Sved
59 and Bird 1990; Han et al. 2008).

60 Given the importance of CpG methylation in genome regulation and evolution, considerable
61 efforts have been devoted to characterizing cytosine methylation. Microarray based
62 methylation chip, MeDIP-seq, MRE-seq, and RRBS-seq are all popular options to interrogate

63 methylation states of different CpGs across the genome (Hodges et al. 2009; Weber et al.
64 2005; Meissner et al. 2008; Maunakea et al. 2010). Because of its universal coverage across
65 the genome, Whole Genome Bisulfite Sequencing (WGBS) has been considered the gold
66 standard in cytosine methylation calling since its debut in the 2nd generation sequencing
67 (shotgun based sequencing of several hundred base pairs) era (Lister et al. 2009; Cokus et
68 al. 2008; Laurent et al. 2010).

69 Despite its many advantages, WGBS has several limitations. For example, the bisulfite
70 treatment that chemically distinguishes methylated from unmethylated cytosine damages
71 the DNA molecules and complicates the library preparation before sequencing. In addition,
72 WGBS typically relies on identifying mismatches between bisulfite-treated reads and the
73 reference genome to call methylation, making methylation calling of cytosines of non-
74 reference regions non-trivial. There are new methods developed recently to improve WGBS
75 but ultimately, they are limited by the 2nd generation sequencing (Vaisvila et al. 2021;
76 Nunn et al. 2022).

77 The 3rd generation sequencing technologies, pioneered by PacBio and ONT, have greatly
78 matured in recent years (van Dijk et al. 2018; Sigurpalsdottir et al. 2024). This
79 improvement results in much longer reads and the ability to distinguish modified bases,
80 included methylated cytosine, during basecalling (Flusberg et al. 2010; Schatz 2017;
81 Stergachis et al. 2020; Xu and Seki 2020). Thus, methylation calling and sequencing can be
82 done simultaneously, making cytosine methylation calling much simpler and more
83 accessible without resort to specific chemical treatment to target cytosine methylation.
84 Despite the promising outlook, the signal changes caused by methylation is usually small

85 and variable, and it is not until the recent development of deep learning models that CpG
86 methylation calling on SMRT and Nanopore data has seen dramatic improvement in
87 accuracy (Simpson et al. 2017; Tse et al. 2021; Ni et al. 2023; Liu et al. 2021).

88 More than just providing better mapping and coverage to highly repetitive regions, the 3rd
89 generation sequencing enabled us to identify non-reference insertions along with their
90 methylation status at individual level including generally methylated TEs. However, there
91 are many outstanding questions regarding the epigenetic response to newly inserted TEs.
92 Which of them are methylated, which of them are not? Is their methylation being
93 influencing by genomic regions into which they insert? Can they affect the methylation
94 status of nearby genomic regions? If they can, how far away would their influence
95 continue? The advance in 3rd generation sequencing makes it much easier to directly
96 address these questions at scale. Recently, third generation sequencing was used to
97 characterize the methylation profile of repetitive elements with high resolution (Ewing et
98 al. 2020; Gerdes et al. 2023; Bodea et al. 2024).

99 It has been reported in mice that a newly inserted hypermethylated TE can spread
100 methylation from the insertion site to the flanking regions until blocked by other regulators
101 (Turker 1999, 2002), resembling the spreading of heterochromatin (Sentmanat and Elgin
102 2012). The spreading of TE methylation can further influence the host genome during
103 evolution by altering mutation pattern (Zhou et al. 2020). Not only hypermethylation can
104 spread out of TEs, tissue-specific hypomethylated CGIs introduced by new L1 insertions
105 can also create so-called “sloping shores” in otherwise hypermethylated regions (Grandi et
106 al. 2015). There were also studies reporting the methylation of new TE insertions have

107 limited effect on nearby gene expression pattern in plants (Choi and Purugganan 2018;
108 Wyler et al. 2020). Recently, Lanciano et al. investigated the local impact of L1 insertion
109 using bsATLAS-seq and ONT reads in multiple cell lines and found regions 300bp upstream
110 of L1 insertions can be influenced by the methylation status of L1 (Lanciano et al. 2024).

111 Taking advantage of the high coverage of the Human Pangenome Reference Consortium
112 (HPRC) 3rd generation sequencing data and latest methylation calling methods, we
113 compared the whole-genome CpG methylation in five human lymphoblastoid cell lines
114 (LCLs) using Single Molecule, Real-Time (PacBio), Nanopore (ONT), and WGBS
115 technologies. After benchmarking the accuracies of methylation calling of the 3rd
116 generation sequencing, we investigated DNA methylation profiles of polymorphic TE
117 insertions, the relationship between DNA methylation patterns of the insertions and their
118 flanking genomic regions, as well as how local methylation could be affected by the new
119 insertions. We found that the 3rd generation sequencing is capable of accurately
120 characterizing CpG methylation at non-reference positions that are inaccessible with
121 conventional WGBS and determined DNA methylation of polymorphic TE insertions and
122 how they interact with their flanking regions.

123 RESULTS

124 Both PacBio HiFi and Nanopore ONT reads provide accurate DNA methylation 125 estimation

126 The Human Pangenome Reference Consortium (HPRC) published the first human
127 pangenome reference graph representing 47 genetically diverse individuals (Liao et al.

128 2023). Out of 47, the DNA methylation status of 32 were determined using PacBio HiFi
129 reads (download). In addition to the rich genetic information and various annotations, the
130 HPRC resources also enabled direct methylation analysis using long read sequencing data.
131 Taking advantage of these resources, we performed additional CpG methylation calling
132 using ONT reads (R9.4.1) from five individuals who were part of the first draft of the
133 Human Pangenome Reference (HG00741, HG00621, HG01952, HG01978, and HG03516)
134 and Whole Genome Bisulfite Sequencing (WGBS) of these same five individuals (Methods).
135 After calling CpG methylation, we aligned all reads to the human reference genome hg38
136 (Fig. 1a).

137 We first compared DNA methylation calls between HiFi, ONT, and WGBS for all CpG sites in
138 the reference genome hg38. For HiFi reads, we adopted the deep learning based
139 mathematical model (PB-model) developed and suggested by PacBio to estimate
140 methylation percentage of individual CpG sites (https://github.com/PacificBiosciences/pb-
141 CpG-tools). We calculated mean average error (MAE), 20% concordance, and Pearson
142 correlation between WGBS, HiFi, and ONT for each sample in pairwise comparisons
143 (HG00741 comparison in Fig. 1c and the other four in Supplemental Fig. S1a-d). We found
144 both HiFi and ONT methylation calls highly correlated with their matching WGBS results
145 (MAE: 0.1062 and 0.0975; 20% concordance: 0.846 and 0.886; Pearson correlation: 0.901
146 and 0.925 for HiFi and ONT, respectively for HG00741) in all five samples (Fig. 1c and
147 Supplemental Fig. S1a-d). We examined the distribution of methylation level across
148 different genomic contexts (defined by the GM12878 chromHMM states) and found strong
149 concordance among WGBS, HiFi and ONT across all five samples (Ernst and Kellis 2012)

150 (Fig. 1d). We then used the WashU Epigenome Browser to illustrate the comparisons
151 between WGBS, HiFi and ONT (Fig. 1b).

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

158 To better understand the scale of differences between the different methods, we defined
159 both differentially methylated loci (DMLs) and differentially methylated regions (DMRs)
160 between HiFi, ONT, and WGBS, as well as between two WGBS replicates of the same sample
161 (Wu et al. 2015). Since the DMRs here were defined using the same samples, they
162 represented the false positive discoveries, and allowed us to estimate an empirical false
163 discovery rate. Using different thresholds, we found that the numbers of falsely discovered
164 DMRs between HiFi and ONT were comparable with the numbers derived from the two
165 WGBS replicates (Fig. 1f). Thus, the methylation calling differences between HiFi and ONT
166 was quite small, generally comparable with the differences we observed between the two
167 WGBS replicates.

168 To further understand the methylation differences between long-read and short-read
169 methods, we examined DMLs and DMRs called between the two methods. We found that
170 the DMLs between long and short reads that exhibiting a >50% methylation difference
171 were predominantly contributed by CpG sites with abnormally high short-read coverage

172 (>50X) (Supplemental Fig. S1e), suggesting short-read mapping in these regions were more
173 challenging than others. In contrast, long-read sequencing coverage or GC content did not
174 correlate with falsely discovered DNA methylation differences (Supplemental Fig. S1f,g).
175 We also found that low complexity and simple repeats contributed to a higher fraction of
176 DMRs with a methylation difference >40% (Supplemental Fig. S1h). Since these regions
177 also suffer more from imprecise mapping of short reads, we reasoned that the small
178 discrepancy between long-read and short-read based methods in assessing DNA
179 methylation was systematic and likely a result of reference mapping bias (Pollard et al.
180 2018). A systematic benchmarking analysis with a focus on alignment bias will be required
181 in the future to fully understand this phenomenon.

182 The methylation patterns of TE insertions

183 Because the 3rd generation methylation calling was conducted on raw reads without
184 aligning to the reference genome, it enabled methylation calling of non-reference regions
185 previously inaccessible with 2nd generation sequencing based methods (at least not
186 trivially). Thus, we investigated the methylation status of non-reference insertions with a
187 focus on polymorphic TE insertions and their correlation with the methylation levels of
188 their flanking regions. Here we took advantage of the recent release of the draft Human
189 Pangenome Reference made available by the HPRC, which included HiFi reads with
190 methylation calling from 32 HPRC Year1 samples (download) as well as their genetic
191 variations (download) (Liao et al. 2023). We further annotated insertions combining the
192 Transposable Element Insertion Annotation pipeline MELT-LRA, PALMER2 and xTEA
193 (Methods) (Gardner et al. 2017; Zhou et al. 2024; Chu et al. 2021). We classified TE

194 insertions into four categories: *Alu*, L1, SVA, and HERVK. In total, we identified 6,752 *Alu*,
195 410 L1, 334 SVA, 16 solo LTR5 and 9 HERVK insertions from the 32 HPRC samples
196 (Supplemental Table S1). For convenience, we termed these TE insertions non-reference
197 TEs, and TEs annotated in hg38 reference TEs.

198 Using custom pipelines (Methods), we extracted non-reference TEs and the HiFi reads
199 derived from these regions (Fig. 2a). We first assessed if we had sufficient long-read
200 coverage on non-reference TEs. We calculated the relative CpG site coverage of each TE
201 insertion compared with the genome-wide average coverage based on HiFi data and
202 observed the two expected peaks for each TE class, representing coverages of heterozygous
203 and homozygous insertions (Supplemental Fig. S2a). The average genome-wide coverage
204 was 45.9X, thus we believe that the coverage on TE insertions was sufficient for the
205 purpose of our analysis. We also extracted HiFi methylation of reference TEs for
206 comparison (Fig. 2a).

207 For the five HPRC individuals with WGBS, HiFi and ONT methylation results, we phased
208 both HiFi and ONT reads to maternal and paternal assemblies and mapped them onto
209 maternal or paternal haplotypes, respectively (download) (Patterson et al. 2015; Holt et al.
210 2024). Using the reference-assembly alignment, we created comparative epigenome
211 browser datahubs to interactively examine the methylation of non-reference insertions
212 (HiFi datahub:
213 https://epigenomegateway.wustl.edu/browser2022/?genome=hg38&hub=https://wangcl
214 uster.wustl.edu/~xzhuo/hifi_methylation/hifi.all.json; ONT datahub:

215 https://epigenomegateway.wustl.edu/browser2022/?genome=hg38&hub=https://wangcl
216 uster.wustl.edu/~xzhuo/hifi_methylation/ont.all.json).

217 We first compared DNA methylation profiles of full-length non-reference TE insertions
218 with their counterparts in the reference genome. We extracted *52,802 Alu*Y, 2,746 *Alu*Ya5,
219 36 *Alu*Ya8, 2,111 *Alu*Yb8, and 301 *Alu*Yb9 from hg38 using repeatMasker annotation.
220 Similarly, we extracted 124 SVA_E, 331 SVA_F, and 481 LTR5 as reference TEs. We also
221 downloaded 146 active L1 on hg38 from L1 Base 2 (Penzkofer et al. 2017). Active L1s were
222 further classified into L1Ta and L1Pre-Ta using the defining variant at position 5932 (A as
223 L1Ta and G as L1pre-Ta) based on L1base 2 (Boissinot et al. 2000). By anchoring
224 methylation to a consensus-based alignment, we significantly reduced the noise in
225 methylation signals compared to previous studies (Ewing et al. 2020; Gerdes et al. 2023).
226 This allowed us to observe sharp methylation fluctuations between adjacent CpG sites
227 within transposable elements (TEs) (Fig. 2b-d). Our findings suggested that certain CpG
228 sites, such as position 5875 in L1HS, were consistently lowly methylated. Overall, DNA
229 methylation profile of reference and non-reference TE insertions showed high methylation
230 levels and were similar in all four classes (*Alu*, L1, and SVA profile in Fig. 2b-d and LTR5
231 profile in Supplemental Fig. S2b). Interestingly, a single CpG site in each of *Alu*, L1 and SVA
232 (at position 213, 5875, 1166 respectively) exhibited statistically significant differences
233 between non-reference insertions and reference copies. Also profoundly, the internal CG
234 rich, variable number of tandem repeats (VNTR) region of SVA was methylated lower in the
235 non-reference insertions compared with their counterpart in the reference copies (Fig. 2d).
236 The overall methylation plots for all but *Alu* insertions (which were too numerous to
237 display individually) were available in Supplemental Fig. S2c-e.

238 To our knowledge, the relatively low methylation of VNTR within SVA elements have never
239 been reported before. Ewing et al. found that the VNTRs are methylated higher than its
240 flanking regions in SVA (Ewing et al. 2020) and we confirmed their observation within
241 reference SVAs (Supplemental Fig. S2d). However, the methylation of VNTR of non-
242 reference SVAs in our LCL samples were noticeably lower than both their flanking regions
243 and their counterparts in the reference genome (Fig. 2f), and it resembled the methylation
244 of mouse L1 5'UTRs with different length of monomers (Gerdes et al. 2023). Aligning full-
245 length SVAs, often >2kb long, to the 1.3kb consensus may create some alignment issue,
246 especially in the VNTR region. However, we do not believe this could explain our
247 observation: the long read methylation calling was done during basecalling process,
248 independent of alignment, and the presence of many lowly methylated CpGs was visible
249 across many non-reference SVAs (Supplemental Fig. S2f).

250 The transcription factor YY1 is an important regulator of L1 activity in mammals (Saha et
251 al. 2024). It was reported that L1 insertions missing the YY1 binding site, located at
252 positions from 12 to 20 in the L1HS, could escape 5' methylation in a tissue-specific
253 manner (Sanchez-Luque et al. 2019). It was also proposed that YY1 binding affects the
254 methylation of L1 in a cell type-specific manner (Lanciano et al. 2024). To investigate
255 whether the YY1 binding site could affect polymorphic L1 methylation in LCL cells, we
256 examined the 5' end methylation of 10 L1 insertions whose first 20-100bp were truncated
257 (Fig. 2e).

258 The YY1 motif is located at position 12-20 in L1 consensus and none of the 10 L1s here
259 contained the YY1 binding site. Despite that, we found five loci, missing 20 bp to 28 bp from

260 the 5'end, were hypermethylated at the L1 promoter. In contrast, three out of the five loci
261 with >32 bp truncated at the 5' end were hypomethylated. Two loci, missing 39 and 73 bp
262 respectively, remained hypermethylated. Our data thus suggest that missing the YY1
263 binding site is not sufficient for the demethylation of L1 promoter, at least not in LCL.
264 Interestingly, we found individual DNA methylation differences of the same L1 insertions.
265 The reference L1 insertion with the first 31 bp at the 5'end truncated at chr13:29641706-
266 29647706, denoted as Chr13delta31 in Sanchez-Luque et al. (Sanchez-Luque et al. 2019),
267 displayed variation among our 32 samples: its promoter was hypomethylated in most of
268 the samples but hypermethylated in HG02717 (Supplemental Fig. S2g).

269 Low methylation correlation between TE insertions and their flanking regions

270 We next investigated the relationship between DNA methylation of the insertions and their
271 flanking regions. We compared the average methylation levels of non-reference insertions,
272 categorized based on their TE annotation, with the methylation levels of their 300bp
273 flanking regions on both 5' end and 3' end (Fig. 3a). TE insertions, particularly *Alu* and SVA
274 insertions, exhibited high methylation levels regardless of the methylation levels of their
275 flanking regions (Fig. 3b). On average, we observed weak correlations between TEs and
276 their flanking regions (linear regression slope < 0.2, $R^2 \sim 0.1$). We did not calculate the
277 correlation of HERVK insertions because of the insufficient number of non-reference
278 HERVK insertions. This pattern was in sharp contrast to non-TE insertions, whose
279 methylation levels were highly correlated with their nearby flanking regions (slope = 0.88,
280 $R^2 = 0.61$) (Fig. 3b). The pattern for non-TE insertions did not change much when we
281 separated them based on whether they contained tandem repeats, whether they were in

282 CpG islands, or their CpG density (Supplemental Fig. S3a). Reference TE insertions
283 displayed a similar pattern to that of non-reference TEs (Supplemental Fig. S3b). Among
284 reference TE insertions, LTR5 exhibited the highest linear correlation slope of
285 approximately 0.5 (Supplemental Fig. S3b).

286 CpG island (CGI) is an important genomic feature, and its methylation status plays a
287 significant role in gene regulation[1]. When we restricted the above analysis to non-TE
288 insertions within CGIs, we observed a stronger correlation between the insertions and
289 their flanking methylation (slope = 0.97, R^2 = 0.90) (Fig. 3c). Remarkably, this correlation
290 extended even to *Alu* insertions within CGIs. In contrast to the general hypermethylation of
291 *Alu* insertions genome-wide, the methylation of *Alu* insertions within CGIs correlated
292 highly with their flanking methylation (slope = 1.16, R^2 = 0.82) (Fig. 3c).

293 Despite the majority of *Alus* being hypermethylated, a small percentage were
294 hypomethylated. For instance, at chr11:59565869, a polymorphic *Alu* insertion within a
295 hypomethylated CpG island was hypomethylated. We plotted the methylation of this *Alu*
296 along with its 300bp flanking regions across all 32 samples (Supplemental Fig. S3c). The
297 insertion site was hypomethylated in all samples, and the *Alu* insertion itself appeared to
298 have adopted the hypomethylation status in all five samples in which it was present. This
299 region can be further examined in sample HG03516, whose *Alu* insertion was heterozygous
300 (Supplemental Fig. S3d). The regional hypomethylation, presence of the *Alu* insertion, were
301 supported by HiFi, ONT, and WGBS analysis (Supplemental Fig. S3d,e).

302 Several *Alu* elements exhibited a methylation slope that connected hypomethylated and
303 hypermethylated regions. We illustrated four such cases in Supplemental Fig. S3f. At

304 chr1:162791517, chr2:218333959, and chr9:71308789, the 300bp upstream of the
305 insertion was hypermethylated, while the downstream region was hypomethylated. In
306 contrast, at chr4:163587802, the upstream region was hypomethylated, while the
307 downstream region was hypermethylated. In each of these four cases, the *Alu* insertion
308 displayed a gradual transition between hypo- and hypermethylation, with some variation
309 observed among samples. We displayed the antisense *Alu* insertion at chr1:162791517 in
310 our comparative browser by phasing reads to paternal and maternal alleles (Fig. 3f). The
311 ONT and HiFi data showed high agreement with each other on methylation level of this
312 non-reference insertion. Downstream of the *Alu* was hypomethylated (left side), while
313 upstream of it (right side) was hypermethylated in all samples. Interestingly, there were
314 variations among individuals and between haplotypes at this locus. While most instances of
315 this insertion were highly methylated throughout the *Alu* element, the methylation level in
316 the HG03516 maternal allele exhibited a transition from low to high in the middle of the
317 element. In contrast, this insertion was hypomethylated in the HG01952 paternal allele. No
318 other sequence feature seemed to be associated with the methylation pattern difference
319 across individual alleles around chr1:162791517.

320 Our analysis thus far has revealed a strong correlation in DNA methylation levels between
321 non-TE insertions and their 300bp flanking regions, and a much weaker correlation when
322 the insertions were TEs. We next investigated the insertion-flanking methylation
323 correlation beyond 300 bp as a function of the distance between the insertion and the
324 nearby regions. For each non-reference insertion, we calculated DNA methylation levels in
325 its 10 kb flanking regions on both directions at a resolution of 100-bp non-overlapping
326 windows (Fig. 3d). We grouped the average methylation of all 100-bp windows from

327 different insertions by their distance from the insertion sites and performed linear
328 regression between insertion and flanking regions. Since the slopes of the linear regression
329 can reflect the methylation correlation between these regions, we displayed the slope of
330 the linear regression as a function of the distance between the flanking windows and the
331 insertion (Fig. 3e). This analysis revealed that the methylation correlation slope between
332 non-TE insertions and their flanking windows quickly faded away as the distance between
333 the flanking window and insertion site increased, with an estimated elbow point being
334 ~2.5kb away from the insertion. The slope decayed to and stayed at around 0.1 when the
335 distance reached 5kb and beyond. TEs once again displayed a strikingly distinct pattern.
336 There was no observable correlation between L1, SVA, and HERV insertions and their
337 flanking regions. We did however observe a weak correlation between *Alu* insertions and
338 their flanking regions but it also quickly faded away (Fig. 3e). This observation could be
339 confounded if TE insertions were longer than non-TE insertions. To rule out this
340 possibility, we investigated how the size of non-TE insertion could affect the insertion-
341 flanking methylation correlation. If the methylation correlation is merely a function of the
342 insertion size, we would expect ~300bp and ~1000bp non-TE insertions behave similarly
343 to *Alu* and LTR insertions, for example. We categorized non-TE insertions by their sizes into
344 five bins, then calculated the regression slope in each bin (Supplemental Fig. S3g). Despite a
345 slightly decrease of the peak methylation correlation as a function of increased non-TE
346 insertion size, the correlation was much higher than that of TE insertions of comparable
347 length (for example, the slope distribution of 300-1000bp non-TE insertions was much
348 higher than that of *Alu* and LTR insertions). This result suggests that the methylation of

349 non-TE insertions, but not TE insertions, are highly correlated with that of their flanking
350 regions.

351 Most TE insertions do not change the methylation level of their flanking 352 regions

353 Some non-reference insertions could be deletion events in the human population (Lee et al.
354 2020b). However, TE insertions are essentially homoplasy-free (Batzer and Deininger
355 2002) and represent recent polymorphic insertions in the human population. Therefore,
356 they are excellent candidates to investigate how new insertions might have influenced the
357 methylation of their inserted regions.

358 Previous studies have suggested that repressive chromatin marks including DNA
359 methylation deposited on newly inserted TEs can spread beyond TE/genome boundary
360 (Turker 2002; Sentmanat and Elgin 2012; Zhang et al. 2012). It was also reported that
361 hypomethylated L1 could drag down the methylation at its 5' end regions and create
362 "sloping shores" (Grandi et al. 2015). Recently, Lanciano et al. found L1 insertions can have
363 an impact on the methylation of CpG sites within 300bp upstream based on both bs-ATLAS-
364 seq and ONT sequencing of multiple cancer cell lines (Lanciano et al. 2024).

365 The Human Pangenome Reference Consortium (HPRC) resources provided a unique
366 opportunity to investigate the impact of TE insertions on nearby regions. For each
367 polymorphic non-reference insertion relative to hg38, we separated HiFi reads from the 32
368 HPRC Year1 samples into two groups: alleles with the insertion and alleles without the
369 insertion (Fig. 4a). We then calculated the average methylation level of the inserted regions

370 (insertion), the 300bp flanking regions of the insertion from both ends (insertion flanking),
371 and the same flanking regions of the allele without insertion (empty flanking) (Fig. 4a).

372 To illustrate the relationship between these three regions, we plotted the methylation
373 levels of insertion flanking and empty flanking on the X-Y axis using a dot plot. We colored
374 the dots based on the methylation level of insertions. We separated non-TE insertions by
375 whether they were annotated as tandem repeat or CGIs, and classified TE insertions by TE
376 classes. In general, the methylation levels of insertion, insertion flanking, and empty
377 flanking correlated with each other for non-TE insertions (Fig. 4b). However, for TE
378 insertions, the flanking methylation levels of insertion and empty sites showed a strong
379 correlation independent of the methylation of the insertion (Fig. 4c).

380 To better understand the subtle changes in flanking methylation levels between insertion
381 and empty sites, we defined the flanking methylation difference for each insertion by
382 subtracting the flanking methylation of the empty site from the flanking methylation of the
383 insertion site (Fig. 4a). We reasoned that an allele without a TE insertion represents the
384 ancestral allele[56]. If the insertion increases flanking methylation, the difference would be
385 positive; if it decreases methylation, the difference would be negative. Conversely, if the
386 insertion has no impact on flanking region methylation, the difference would be zero.

387 We first plotted the correlation between flanking methylation variations and the
388 methylation of non-TE insertions (Fig. 4d). The overall methylation differences between
389 orthologous flanking regions with or without insertions were predominantly centered
390 around zero (population average was 0.25%) – 99% of the data points (28344/28621) fell
391 within the -25% to +25% range. Notably, these variations were independent of the

392 methylation level of the insertion itself. However, there indeed existed multiple instances
393 where hypermethylated insertions were associated with increased flanking methylation,
394 and hypomethylated insertions were linked to decreased methylation (Fig. 4d).
395 Collectively, these findings strongly indicate that, although there are noted exceptions, non-
396 TE insertions generally do not significantly impact flanking region methylation.

397 We then investigated the same relationship in the context of TE insertions (Fig. 4e).
398 Compared to non-TE insertions, the range of methylation differences was slightly wider
399 (96%, 5075/5263, within -25% and 25%), indicating significantly increased variations in
400 flanking region methylation associated with TE insertions (F-Test to Compare Two
401 Variances: p-value < 2.2e-16). On average, flanking methylation differences centered
402 around zero for LTR and SVA insertions. However, the average flanking methylation
403 differences for *Alu* and L1 insertions were significantly larger than zero (3.2% and 3.9%,
404 respectively). Furthermore, there were more *Alu* and L1 insertions with methylation
405 increase more than 25% than decrease more than 25% (112 vs 62, and 4 vs 1,
406 respectively). The range of methylation differences we observed was quite large. For
407 example, the mean methylation level of inserted *Alus* was 82.7%, and *Alu* insertions with
408 methylation level greater than this mean value were associated with a flanking region
409 methylation difference ranging from -42.74% to +42.66%, with 21 of them associated with
410 an increased flanking region methylation by greater than 30%. Similarly, the mean
411 methylation level of inserted L1s was 69.2%, and L1 insertions with methylation level
412 greater than this mean value were associated with a flanking region methylation difference
413 ranging from -30.1% to +27.8%, but none of them was associated with an increased
414 flanking region methylation by greater than 30%. These examples only weakly support the

415 hypothesis that DNA methylation deposited on newly inserted TEs could spread and
416 increase local DNA methylation level. Contrary to this conventional view, our genome-wide
417 assessment suggests that the spreading is limited, and those more drastic influence of TEs
418 on nearby genomic methylation may be an exception rather than the rule.

419 It has been reported that L1 promoter methylation can influence the nearby upstream
420 methylation (Lanciano et al. 2024). To test this hypothesis using our data, we created a
421 dotplot similar to Fig. 4c, but using the 5'end 500bp of L1/SVA insertions along with the
422 300bp upstream region genomic sequences (Fig. 4f). Notably, despite the existence of a few
423 exceptions, the upstream methylation of empty sites consistently matched those with
424 insertions, indicating that the vast majority of L1/SVA insertions did not affect their
425 flanking upstream methylation.

426 To further validate and generalize our findings, we aggregated all polymorphic insertions
427 and plotted the average methylation trendline based on their relative position in
428 relationship to the insertion sites from the reference genome (Supplemental Fig. S4a). We
429 found a subtle increase of methylation flanking the polymorphic L1/*Alu* insertions, with the
430 highest increase right at the insertion site (3.1% for L1 and 3.2% for *Alu*). This flanking
431 methylation difference between insertion and empty site decreased as the distance to
432 insertion position increased, and the methylation difference diminished at ~1kb away from
433 the insertion site (Supplemental Fig. S4a). The exact same pattern was recapitulated by
434 both the ONT methylation data (phased ONT reads) and the WGBS data (WGBS reads of
435 homozygous insertions to the individual's own genome assemblies) from the five HPRC
436 individuals (Supplemental Fig. S4b,c) (Methods). Taken together, these results suggest that

437 polymorphic *Alu* and L1 insertions are associated with slightly elevated DNA methylation
438 levels at their insertion sites with highly localized and small effect size.

439 The exceptions: TE insertions associated with flanking region methylation 440 change

441 Despite most TE insertions not affecting the methylation of their insertion sites, we also
442 found numerous TE insertions associated with flanking methylation changes. For instance,
443 the hypermethylated L1 insertion at chr3:85527420 (average promoter methylation level
444 88%) was associated with an increase in methylation of the CpG site upstream of the
445 insertion. We plotted the methylation of each CpG site of the 5'end 1000 bp of the L1 along
446 with both flanking regions across all 32 samples in Supplemental Fig. S5a and summarized
447 them in Fig. 5a. In samples without this insertion, the upstream CpG site was mostly
448 hypomethylated (average methylation 24.3%), whereas in samples with the insertion, it
449 was generally intermediately methylated (average methylation 42.9%) (Fig. 5a).
450 Conversely, the L1 insertion with a hypomethylated promoter (average methylation level
451 46.3%) at chr2:108834214 was associated with a reduction in methylation of its upstream
452 region (from 94.2% to 74.7%) (Fig. 5b and Supplemental Fig. S5b).

453 The impact of TE insertions on nearby genomic sequence can also be appreciated by
454 examining heterozygous insertion in the same individual, for example, the L1 insertion at
455 chr3:85527420 in both HG00621 and HG00741 (present only in the paternal haplotype).
456 Both long-read methods identified the CpG site 46bp upstream of the L1 insertion was
457 hypermethylated (average 81%) but remained hypomethylated in the remaining eight
458 haplotypes without the insertion (average 21%) (Supplemental Fig. S5h). Interestingly, this

459 specific polymorphic L1 was reported by Lanciano et al. to increase upstream CpG
460 methylation in the MCF7 cell line, but was hypomethylated and had reduced proximal
461 methylation in the 2102Ep cell line (Lanciano et al. 2024). Taken together, these results
462 suggest that this specific polymorphic L1 insertion adopts cell type-specific DNA
463 methylation which may spread to its immediate region upstream.

464 We confirmed the impact of multiple L1 loci, including the insertion at chr3:85527420 we
465 described above. However, there were also instances where the relationship between
466 methylation differences and L1 insertions was less evident. For instance, the methylation of
467 the empty allele at CpG site chr14:24523788, upstream of the L1 insertion at
468 chr14:24523704, exhibited significant variability across the HPRC samples. Meanwhile, the
469 same CpG site was intermediately methylated among samples with the insertion allele, and
470 the difference was not statistically significant (p=0.42) (Supplemental Fig. S5c,d). At CpG
471 site chr5:137679380, upstream of the L1 insertion at chr5:137679083, while the
472 methylation difference between the insertion and empty alleles was significant (p=0.0056),
473 the high methylation variation of the empty allele made it difficult to conclude that the
474 insertion allele methylation was influenced by the L1 insertion (Supplemental Fig. S5e,f).

475 Our study identified thousands of *Alu* insertions and most of them were hypermethylated.
476 In some cases, their methylation appeared to be influenced by their genomic context (Fig.
477 3). Can *Alu* insertions influence the methylation of its nearby regions? We found that while
478 the large majority of *Alu* insertions seemed to have no impact on nearby DNA methylation,
479 exceptions did exist. For example, polymorphic *Alu* insertion at chr1:40059013 was
480 hypermethylated (average methylation 94%), and so were 3 downstream CpG sites within

481 200bp (average methylation 58.1%). In samples without the *Alu* insertion, these three CpG
482 sites were hypomethylated (average methylation 12.2%) (Fig. 5c and Supplemental Fig.
483 S5g). A closeup view revealed that this *Alu* insertion was homozygous in HG00621 but
484 absent in HG00741. This *Alu* was inserted within a hypomethylation valley
485 (chr1:40058800-40060800, methylation level 4%). In HG00621, the *Alu* insertion was
486 hypermethylated (methylation level 85%), and methylation level of the hypomethylation
487 valley was elevated to 21% (Supplemental Fig. S5h) (Li et al. 2022; Zhuo et al. 2023). The
488 three CpG sites 200bp 3' to the *Alu* insertion site were lowly methylated in HG00741 (18%,
489 5%, 4%) but highly methylated in HG00621 (89%, 88%, 93%). On the other hand, multiple
490 CpG sites more than 300bp downstream from the insertion site seemed to have remained
491 their hypomethylated state (3 CpGs, 3% in HG00741, 5% in HG00621) (Supplemental Fig.
492 S5h).

493 DISCUSSION

494 Characterizing cytosine methylation is important for both evolutionary analysis and tissue-
495 specific gene regulation. However, conventional bisulfite or enzyme conversion-based
496 methods do not readily characterize methylation of repetitive regions or non-reference
497 regions. Using 3rd generation sequencing, we can call methylation during the base calling
498 step, greatly simplifying the methylation calling process and allowing new investigations
499 into genomic DNA methylation. Here we compared the methylation calling results between
500 HiFi, ONT, and WGBS from five HPRC LCL samples, and we found both HiFi and ONT can
501 measure CpG methylation in human genome with high correlation and concordance in
502 comparison with WGBS.

503 We characterized the methylation profile of non-reference TE insertions. Surprisingly, we
504 found that the VNTR regions of non-reference SVA insertions were methylated at a lower
505 level compared to their counterparts in the reference genome. Given the variable length of
506 VNTR, a more specific analysis is necessary to determine their methylation levels
507 accurately.

508 How are genomic insertions, in particular transposable element insertions, methylated, and
509 how do they affect the methylation level of their flanking regions, have been a long-
510 standing question in the field. It has promoted several paradigms, especially in the context
511 of spreading of epigenetic modifications. For example, the hypermethylation of a B2
512 element close to the Aprt promoter in mouse genome can spread to nearby CpG sites until
513 stopped by SP1 binding (Turker 2002). By examining CpG density flanking *Alu* insertions,
514 Zhou et al. found evidence of CpG methylation spreading from old TEs to flanking regions
515 during evolution (Zhou et al. 2020). Recently, Lanciano et al. described examples of L1
516 insertion in cancer cell lines affecting the methylation level of nearby regions up to 300bp
517 upstream (Lanciano et al. 2024). It was also reported that hypomethylated L1 insertions
518 can create hypomethylated "slopping shore" (Grandi et al. 2015) . These previous studies
519 provided enormous insights on how newly inserted transposable elements are
520 epigenetically modified and how the modifications influence genomic neighborhood.
521 However, most studies represented specific examples or anecdotes, leaving the question
522 open as to what degree these examples represent the genome-wide rules or exceptions.

523 Taking advantage of both SV and methylation calling using the HiFi reads from 32 samples
524 of HPRC Release1 dataset, we addressed this question by analyzing differential methylation

525 between alleles with polymorphic insertions at a genome-wide scale. Our analysis suggests
526 that the methylation of non-TE insertions is strongly associated with their genomic context.
527 Consistent with Lanciano et al., we find some methylation spreading from L1 to nearby
528 regions. However, the effect is very small (on average ~3% increase of methylation within
529 300bp). The minimum spreading of methylation from polymorphic TE insertions to their
530 flanking regions we found suggests that methylation spreading is a slow evolutionary
531 process with limited effect after TE insertion but can potentially have large impact at an
532 evolutionary timescale (Zhou et al. 2020).

533 However, our analysis was conducted on EBV-derived lymphoblastoid cell lines, which are
534 commonly used for genome sequencing. It is crucial to note that the genomic methylation
535 status observed in this cell line does not reflect the diverse epigenomic characteristics of
536 various human cell types (Roadmap Epigenomics Consortium et al. 2015). Therefore,
537 careful characterization of other cell types is still necessary to validate our findings in this
538 cell line and to gain a deeper understanding of the dynamic methylation patterns in
539 humans.

540 Because our strategy was conceptually similar to that from Lanciano et al., we directly
541 examined the potentially polymorphic L1 insertions characterized by their study, which
542 were reported to exhibit short-range epivariation in almost half of the cases. We found 21
543 out of the 87 L1 insertions from their study had sufficient HiFi coverage of both empty and
544 insertion alleles in the HPRC dataset. The 300bp upstream methylation of empty and
545 insertion alleles were highly correlated with $R^2 = 0.97$ if they were classified as "not
546 influenced"; with $R^2 = 0.86$ if they were classified as "inconclusive"; and with $R^2 = 0.61$ if

547 they were classified as “influenced” (Supplemental Fig. S6), as defined in Lanciano et al.
548 (Lanciano et al. 2024). Thus, those L1 insertions that were determined by Lanciano et al. to
549 exhibit short-range epivariation were indeed associated with slightly increased differences
550 between alleles with the insertion and the empty alleles, replicated in our own dataset.
551 However, the empty allele and insertion allele from the same locus were still strongly
552 correlated with each other, indicating the effect size of L1 influence was small. In contrast,
553 our HPRC dataset contained a much larger sample size, and we determined that the
554 flanking methylation difference between the insertion allele and the empty allele was less
555 than 25% for most cases (287 out of 292 L1 insertions and 4519 out of 4693 *Alu*
556 insertions) (Fig. 4). Thus, we concluded that the majority of polymorphic TE insertions do
557 not change their flanking region DNA methylation, while there exist numerous exceptions
558 with a limited short-range influence.

559 To visualize the methylation of non-reference insertions, we mapped reads to the high-
560 quality individual assemblies and displayed them on the WashU Epigenome Browser using
561 the genome comparison feature (Li et al. 2022; Zhuo et al. 2023). However, this approach is
562 limited by the underlying pairwise alignment and more crucially, the availability of high-
563 quality individual assemblies. We hope the development of pangenome based tools would
564 facilitate data display of non-reference sequences in the future.

565 With the ongoing effort and continuing release of high-quality data from HPRC, we envision
566 the publicly available long read methylation data can be used to address other questions in
567 the biomedical field, and we invite other scientists to explore the rich, high-fidelity HPRC
568 dataset (Liao et al. 2023).

METHODS

Genome assembly and scaffolding

The genome assemblies used in this study were produced for the draft human pangenome, combining PacBio HiFi long-read and Illumina short-read sequencing for de novo assembly with parental short-read data used to for phasing of these assemblies. Phased contig assembly was performed using hifiasm (Cheng et al. 2021). Post-assembly, the RagTag toolset was employed for reference-guided scaffolding, organizing contigs into structured, phased chromosomal assemblies (Alonge et al. 2022). Validation was performed using BioNano optical mapping, in order to detect and amend potential assembly inaccuracies (Lam et al. 2012). For each of the 10 assemblies derived from the five diploid samples, we created a whole-genome pairewise alignment BAM file with hg38 using minimap2 (Li 2018).

Whole Genome Bisulfite Sequencing (WGBS)

The cytosine methylation was estimated by mapping to reference genome hg38 or the individual assemblies using the standard WGBS pipeline Bismark as described by Lee et al. (Krueger and Andrews 2011; Lee et al. 2020a).

PacBio HiFi Methylation

PacBio single-molecule high fidelity (HiFi) Circular Consensus Sequencing (CCS) reads with kinetics tags can be generated using pbccs (https://github.com/nlhepler/pbccs). We then predicted cytosine methylation probability per molecule in all CpG context by applying

589 primrose to PacBio HiFi reads (https://github.com/PacificBiosciences/jasmine). The CpG
590 methylation probability was saved as the ML and MM tags in the BAM file following
591 standard BAM format convention (http://samtools.github.io/hts-specs/SAMtags.pdf).
592 Unaligned HiFi BAM files were aligned to human reference genome hg38 using PacBio
593 pbmm2 (https://github.com/PacificBiosciences/pbmm2), and CpG methylation percentage
594 on the reference genome were calculated using PacBio pb-CpG-tools
595 (https://github.com/PacificBiosciences/pb-CpG-tools). We ran the methylation percentage
596 calculation using default model "PB-model".

597 Nanopore ONT methylation

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

607 Methylation comparison

608 With genome-wide methylation percentage of all three methods calculated, we made
609 pairwise correlation heatmap, PCA plot and calculated the mean average error, 20%

610 concordance, and Pearson correlation using R (Wickham 2009; Emerson et al. 2013). We
611 also applied DSS to define pairwise differential methylation regions (DMRs) (Wu et al.
612 2015; Feng et al. 2014).

613 Long reads phasing

614 We used HiPhase for HiFi reads phasing (Holt et al. 2024). To determine in each phase
615 block which haplotype is paternal, and which is maternal, we downloaded HiFi reads used
616 for constructing paternal and maternal assemblies from hifiasm output gfa files (HG00621,
617 HG00741, HG01952, HG01978, HG03516). We labeled the haplotype paternal if the
618 number of reads labeled with it from the same phase block used for paternal assembly is
619 >10 times higher than that used for maternal assembly, otherwise maternal. Once we
620 separated HiFi reads to paternal and maternal with whatshap (Patterson et al. 2015), we
621 aligned them to the individual paternal and maternal assemblies, accordingly using pbmm2
622 (https://github.com/PacificBiosciences/pbmm2). At last, we calculated the genome-wide
623 methylation percentage on individual assemblies using pb-CpG-tools
624 (https://github.com/PacificBiosciences/pb-CpG-tools).

625 To phase ONT reads according to the individual assemblies, we created a diploid VCF file
626 for each sample using dipcall (https://github.com/lh3/dipcall) (Li et al. 2018) and
627 haptagged all reads using haplotype information from the VCF file. We then separated
628 paternal and maternal reads using whatshap (Patterson et al. 2015) and aligned them to
629 paternal and maternal assembly using minimap2 (Li 2018, 2021), respectively. At last, we
630 converted the bam file to methylbed files with modbam2bed (https://github.com/epi2me-
631 labs/modbam2bed).

632 ## Individual genome visualization

633 For each pairwise genome alignment between hg38 and individual genome, we created a
634 genome-align track file using the pairwise genome alignment BAM file (Zhuo et al. 2023).
635 We built individual genome assembly of both maternal and paternal chromosomes of the
636 five samples (10 assemblies in total) and aligned each of them to the hg38 to create
637 genome-align tracks. We also phased HiFi reads using HiPhase using alignment to hg38 and
638 assigned haplotypes to paternal/maternal for each phasing block based on hifiasm
639 assembly construction graph (Holt et al. 2024; Cheng et al. 2021). We then aligned
640 maternal and paternal HiFi reads to the maternal/paternal assembly separately for each of
641 the five samples. Applying the genome-align track we anchored gene, repeat, and CpG
642 island annotations from hg38 onto the individual genomes. We colored each CpG site by
643 their methylation prediction (red: methylated; blue: unmethylated) per HiFi read using the
644 modbed tracks and displayed piled up methylation level using the methylC track. Both
645 modbed and methylC tracks were mapped to the individual genome. Combining genome-
646 align, methylC, and modbed track files, we created three datahubs on the WashU
647 Epigenome browser representing HiFi, ONT, and WGBS data
648 (https://epigenomegateway.wustl.edu/browser2022/?genome=hg38&hub=https://wangc
649 luster.wustl.edu/~xzhuo/hifi_methylation/hifi.all.json,
650 https://epigenomegateway.wustl.edu/browser2022/?genome=hg38&hub=https://wangcl
651 uster.wustl.edu/~xzhuo/hifi_methylation/ont.all.json,
652 https://wangcluster.wustl.edu/~xzhuo/hifi_methylation/wgbs_individual/wgbs.all.json)
653 (Li et al. 2022; Zhou et al. 2014; Li et al. 2023).

654 After aligning WGBS reads to the individual assembly as described above, we also ploted
655 the WGBS methylation of non-reference *Alu* insertion at chr11:59565869 using
656 Methylartist (Ewing et al. 2020; Cheetham et al. 2022).

657 Structural variation and TE annotation

658 We obtained the structural variations of each sample from HPRC consortium (Download)
659 (Liao et al. 2023). We merged all the called SVs from the 32 samples using bcftools merge
660 (Danecek et al. 2021). To further annotate transposable element insertions among all
661 structural variations, we ran MELT-LRA and PALMER2 on HiFi assemblies and xTEA on the
662 same PacBio HiFi library mapped to hg38 (Gardner et al. 2017; Zhou et al. 2024; Chu et al.
663 2021). For *Alu*, L1 and SVA insertions, we included MEIs called by at least two of three
664 methods. MELT-LRA did not call HERVK insertion, and we manually verified 25 out of 34
665 HERVK insertions identified by either xTEA or PALMER2. We summarized MEI caller
666 output in supplementary (Supplemental Table S1). We then intersected the coordinates of
667 TE insertions with SV and annotated which SVs are TE insertions (Supplemental Table S2).

668 We developed a script to extract non-reference insertions. Using a BED file of annotated
669 insertion sites, it can extract inserted and soft-clipped sequences from raw reads and
670 generate a new unmapped modBAM file of inserted sequences with base modification tags
671 (MM and ML) (bam.extractInsertion.py). We can then align the unmapped modBAM file to
672 their respected TE consensus (L1HS, *Alu*Y, SVA_F, LTR5_Hs and HERVK-int from dfam
673 database) to create CpG methylation pileup results (Wheeler et al. 2012; Hubley et al.
674 2015).

675 To calculate average methylation of certain regions, we extracted methylation prediction
676 value of each CpG site of each read from bam files with methylation tags with a python
677 script (python bam.Mmtag.regions.oo.py -b <bam> -r <insertion.bed> -o
678 <ins.methylC.reldist.txt> -f <bed>) available in Github repository
679 (https://github.com/xzhuo/modbamUtil)

680 To evaluate the relationship between CpG methylation and indels, we aggregated
681 methylation prediction of all 32 samples and calculated the average CpG methylation of
682 each CpG Island (CGI) (Quinlan and Hall 2010).

683 reference TE methylation

684 We extracted 52802 *AluY*, 2746 *AluYa5*, 36 *AluYa8*, 2111 *AluYb8*, and 301 *AluYb9* from
685 hg38 using repeatMasker annotation. similarly, we extracted 124 SVA_E, 331 SVA_F, 481
686 LTR5 as reference TEs. We also downloaded 146 active L1 on hg38 from L1 Base2
687 (Penzkofer et al. 2017). We then take their DNA sequences from the hg38 reference and
688 their corresponding methylation from piled up bedMethyl files and generated new
689 modBam files to bind methylation percentage to the sequence as modified base
690 (bam.creation.py). At last, we mapped the new BAM file to TE consensus using Minimap2 to
691 align CpG methylation percentage to the TE consensus (L1HS, *Alu*Y, SVA_F, LTR5_Hs and
692 HERVK-int) (Wheeler et al. 2012; Hubley et al. 2015).

693 Reference or non-reference insertions are relative to the reference genome chosen for
694 alignment, and non-reference insertions can be aligned to the genome if we chose the
695 individual assembly as the “reference”. Therefore, the sample process can be applied to
696 extract “non-ref” insertions on the hg38 after align long reads to the individual assembly.

697 ## Insertion and flanking methylation calculation

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

702 ## CpG island methylation and INDEL frequency calculation

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

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

712 We also repeated the same calculation using same methylation data and INDEL callset from
713 1000 genome project and HGSVC project (The 1000 Genomes Project Consortium 2015;
714 Ebert et al. 2021).

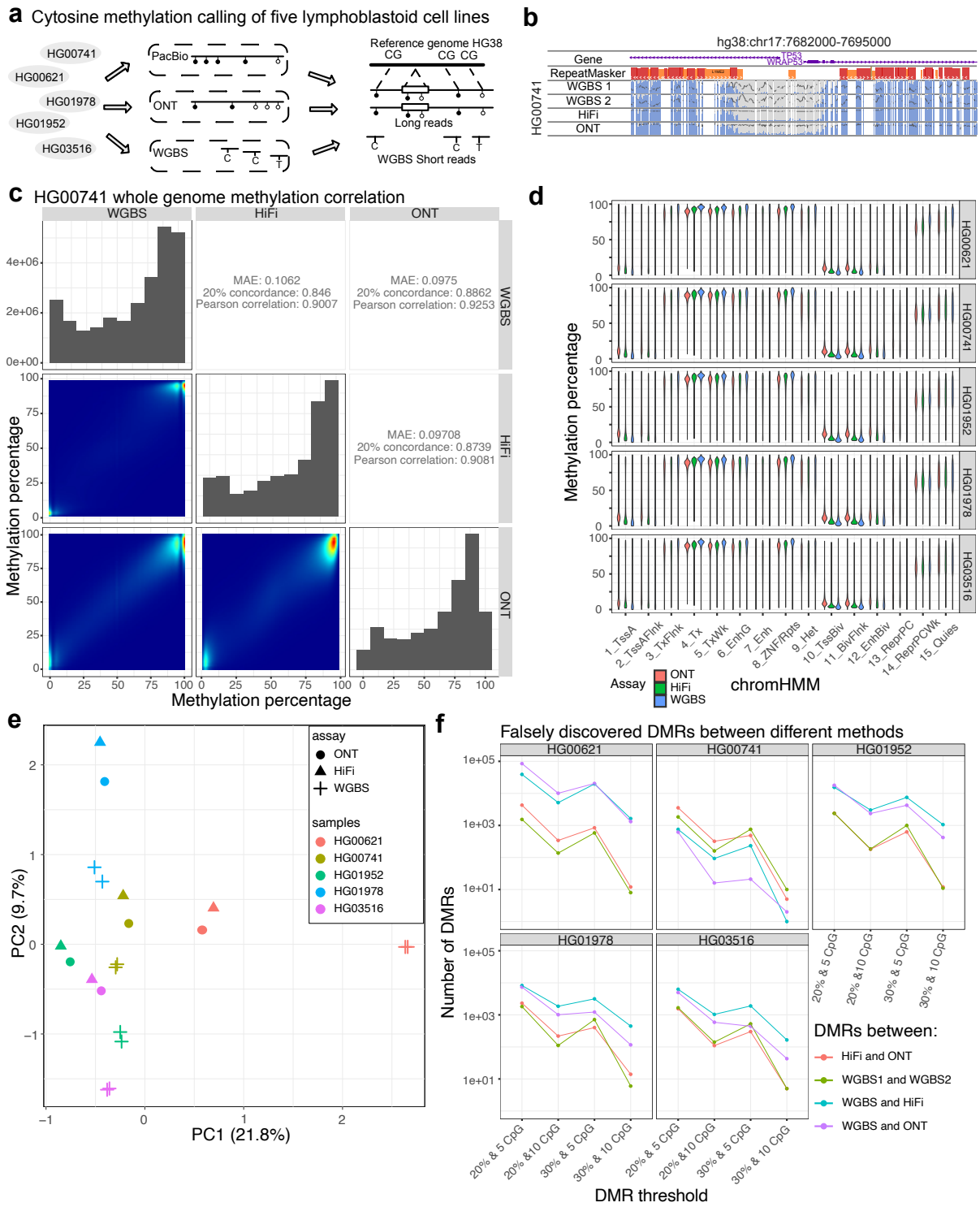
715 # DATA ACCESS

716 All 32 PacBio HiFi methylation data of HPRC year1 samples are available at https://s3-us-
717 west-2.amazonaws.com/human-pangenomics/index.html?prefix=submissions/560047a6-
718 6d16-4b0c-aac9-7d0c83e2188e--HIFI-METHYLATION-READS/ .

719 The Structural variation annotation is available at https://s3-us-west-
720 2.amazonaws.com/human-pangenomics/index.html?prefix=submissions/B581EBA7-
721 8BDE-4C7C-9DEA-78B99A051155--Yale_HPP_Year1_Variant_Calls/ .

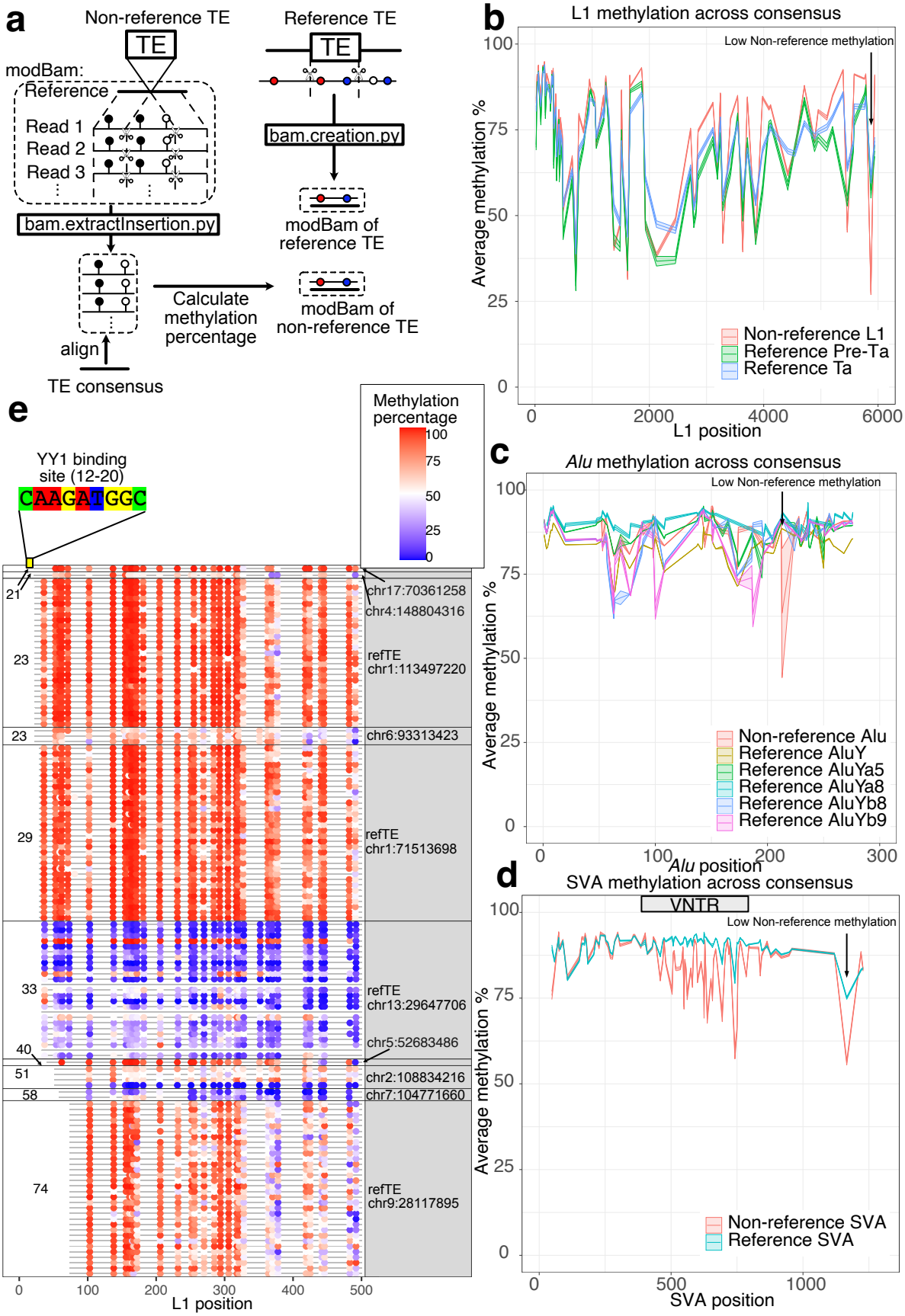
722

Figures


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
TP53
WRAP53
RepeatMasker
WGBS 1
WGBS 2
HiFi
ONT
HG00741
c HG00741 whole genome methylation correlation
WGBS
HiFi
ONT
MAE: 0.1062
20% concordance: 0.846
Pearson correlation: 0.9007
MAE: 0.0975
20% concordance: 0.8862
Pearson correlation: 0.9253
MAE: 0.09708
20% concordance: 0.8739
Pearson correlation: 0.9081
Methylation percentage
d
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
Assay
ONT
HiFi
WGBS
chromHMM
e
assay
ONT
HiFi
WGBS
samples
HG00621
HG00741
HG01952
HG01978
HG03516
PC2 (9.7%)
PC1 (21.8%)
f
Falsely discovered DMRs between different methods
Number of DMRs
DMR threshold
20% & 5 CpG
20% &10 CpG
30% & 5 CpG
30% & 10 CpG
DMRs between:
HiFi and ONT
WGBS1 and WGBS2
WGBS and HiFi
WGBS and ONT

724

725 **Figure 1. a**: Study design of methylation calling benchmarking. We performed CpG
726 methylation calling for five HPRC samples (HG00621, HG00741, HG01952, HG01978, and
727 HG03516), using PacBio, ONT, and WGBS. We mapped all reads to the human reference
728 genome hg38 to compare their differences. **b**: WashU Epigenome Browser screenshot of
729 the two WGBS replicates, HiFi, and ONT methylation of HG00741 on the 10 kb window
730 around the TP53 gene promoter. All methylC tracks Y-axis percentage was from 0% to
731 100%. **c**: Whole genome methylation percentage correlation between WGBS, HiFi, and ONT
732 methylation of HG00741. The whole genome methylation percentage distribution was
733 represented as a density plot. The correlation heatmap and metrics (Mean Average Error,
734 20% concordance, and Pearson correlation) were plotted on the left-bottom and right-top,
735 respectively. **d**: Methylation percentage distribution of the GM12878 15 chromHMM states
736 across all five samples. The 15 chromHMM states were listed on the X-axis and methylation
737 calling results from different methods were colored differently. **e**: PCA plot of the top two
738 principal components of the whole genome methylation from the five HPRC samples. For
739 each sample, we performed two replicates of WGBS in addition to HiFi and ONT
740 methylation calling. We used different shapes to represent different methods (circle,
741 triangle, and cross for ONT, HiFi, and WGBS) and different colors to represent different
742 samples (red, yellow, green, blue, and purple for HG00621, HG00741, HG01952, HG01978,
743 and HG03516, respectively). **f**: Number of falsely discovered differentially methylated
744 regions (DMRs) using different threshold cut-offs. Four different thresholds (20% average
745 methylation difference with at least 5 CpG sites, 20% average methylation difference with
746 at least 10 CpG sites, 30% average methylation difference with at least 5 CpG sites, and
747 30% average methylation difference with at least 10 CpG sites) were represented on the X-
748 axis. The number of DMRs was represented on the Y-axis. The number of DMRs between
749 different methods under different conditions was plotted using different colors.


a
Non-reference TE
Reference TE
TE
modBam:
Reference
Read 1
Read 2
Read 3
bam.extractInsertion.py
bam.creation.py
modBam of reference TE
Calculate methylation percentage
modBam of non-reference TE
align
TE consensus
b
L1 methylation across consensus
Low Non-reference methylation
Average methylation %
L1 position
Non-reference L1
Reference Pre-Ta
Reference Ta
c
Alu methylation across consensus
Low Non-reference methylation
Average methylation %
Alu position
Non-reference Alu
Reference AluY
Reference AluYa5
Reference AluYa8
Reference AluYb8
Reference AluYb9
d
SVA methylation across consensus
VNTR
Low Non-reference methylation
Average methylation %
SVA position
Non-reference SVA
Reference SVA
e
Methylation percentage
YY1 binding site (12-20)
CAAGATGGC
chr17:70361258
chr4:148804316
refTE chr1:113497220
chr6:93313423
refTE chr1:71513698
refTE chr13:29647706
chr5:52683486
chr2:108834216
chr7:104771660
refTE chr9:28117895
L1 position

750

751 **Figure 2**. **a**: Schematic representation of our pipeline to extract reference and non-
752 reference insertions along with their methylation annotation. We developed a script
753 (bam.extractInsertion.py) to extract truncated DNA sequences from reads while retaining
754 positional methylation tags (subseq modBam file). We then aligned this modBam to the
755 corresponding TE consensus and calculated the methylation profile of the TE insertion. We
756 also extracted sequences from the reference genome and the methylation percentage from
757 the bedMethyl file to create another modBam file of reference TEs and their methylation
758 (bam.creation.py). The methylation would be anchored to the TE consensus after
759 alignment. **b**: Methylation profile of L1 insertions. X-axis represented positions on the L1Hs
760 consensus, Y-axis represented methylation percentage with 95% standard error of the
761 mean illustrated as shaded areas. Reference L1pre-Ta, reference L1Ta, and non-reference
762 L1 insertions were labeled with different colors. Only full-length elements (truncation <
763 50bp on both ends) were included for profiling. **c**: Methylation profile of *Alu* insertions with
764 95% standard error of mean. X-axis represented positions on the *AluY* consensus, Y-axis
765 represented methylation percentage. *AluY*, *AluYa5*, *AluYa8*, *AluYb8*, *AluYb9* on the reference
766 and non-reference *Alu* insertions were labeled with different colors. **d**: Methylation profile
767 of SVA insertions with 95% standard error of mean. X-axis represented positions on the
768 SVA_F consensus, Y-axis represented methylation percentage. Reference SVA and non-
769 reference SVA insertions were labeled with different colors. The VNTR of SVA consensus
770 were labeled on the graph with a gray box. **e**: The first 500bp methylation of L1 insertions
771 whose first 20-50bp were truncated. Start position was labeled on the left, and the
772 insertion coordinates on hg38 were labeled on the right. The "refTE" indicates it was an L1
773 insertion on the hg38 reference. Each row represented one of the 32 samples. The
774 methylation percentage of each CpG site was illustrated by the blue-red (low to high) color
775 scale.

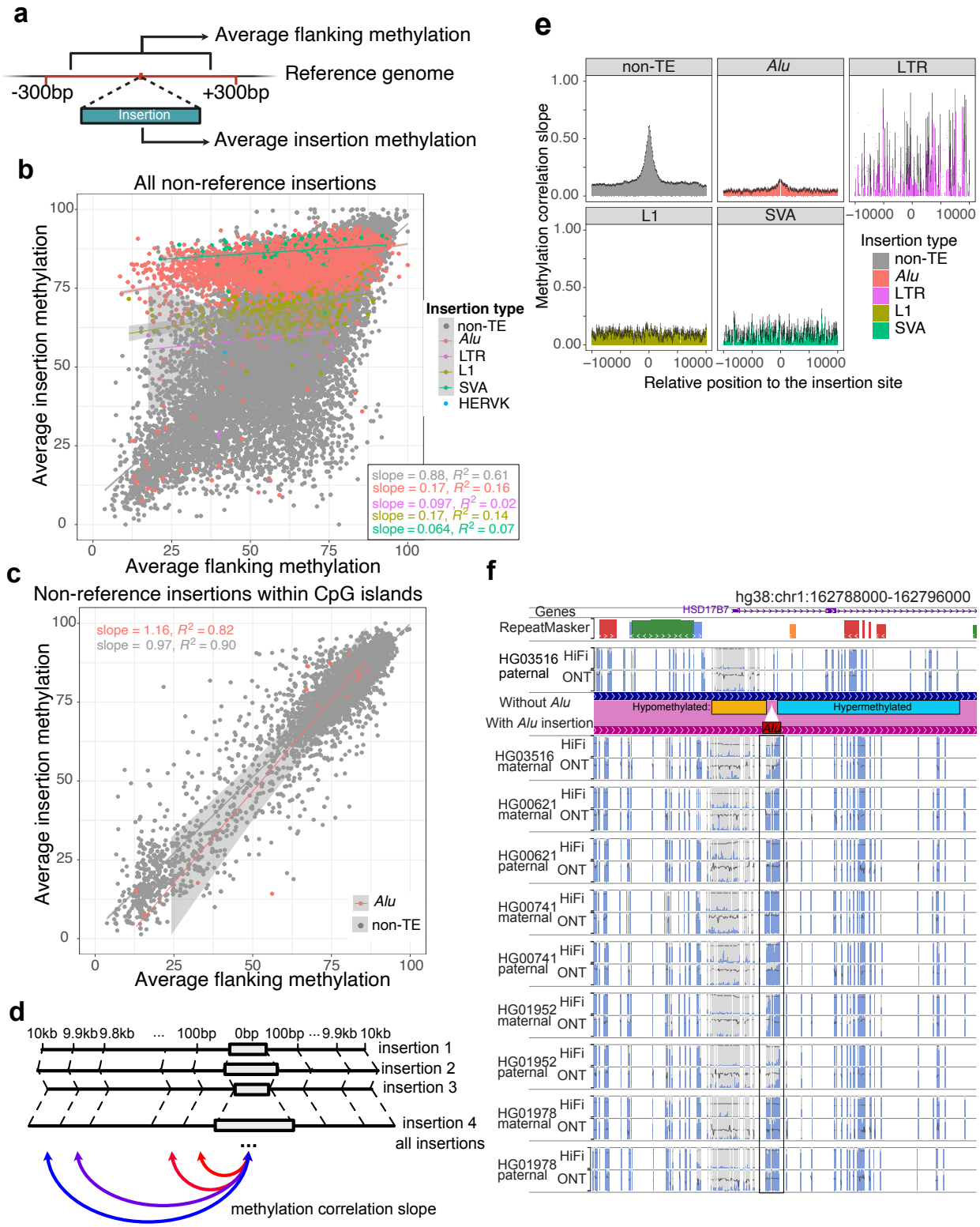
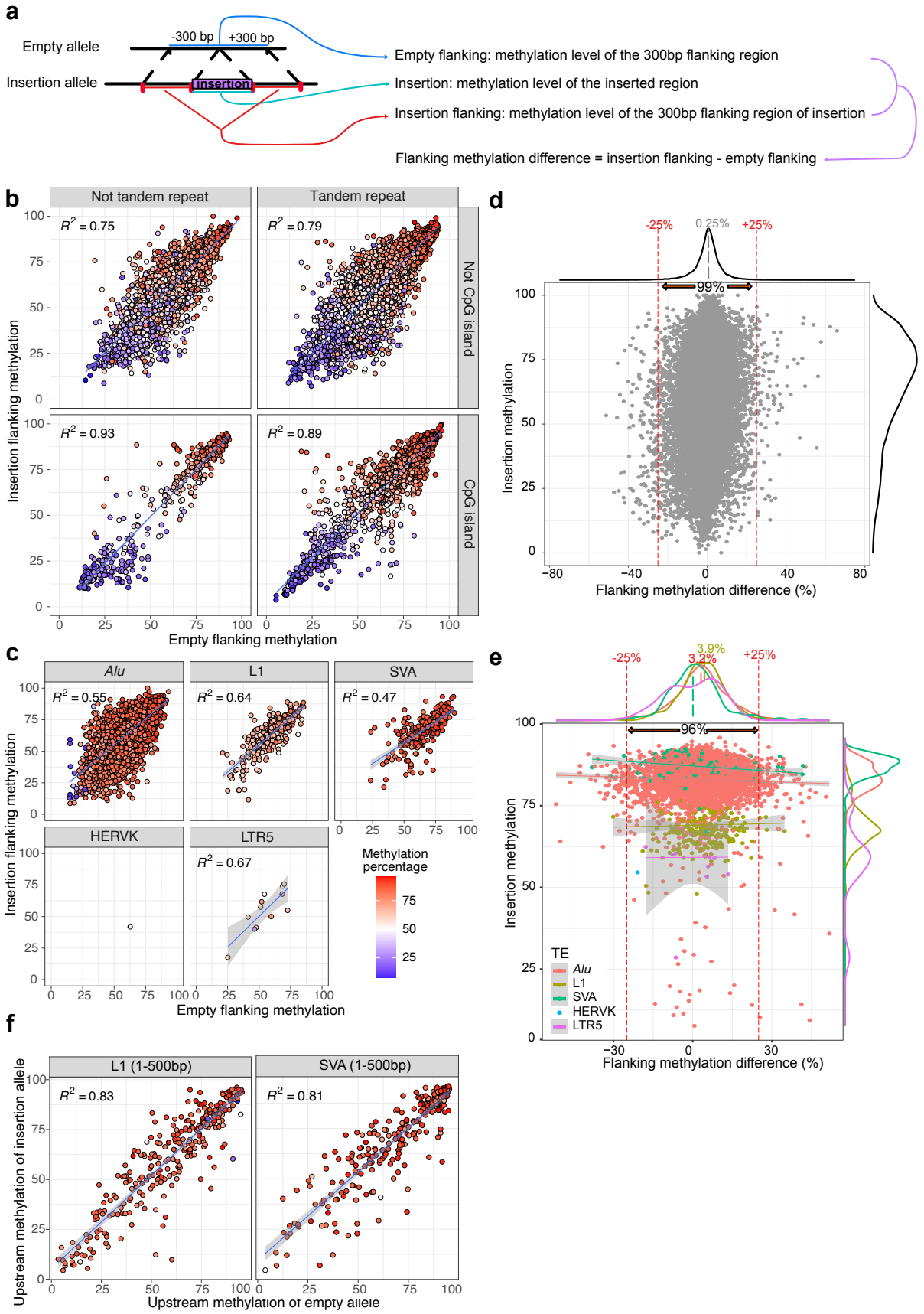


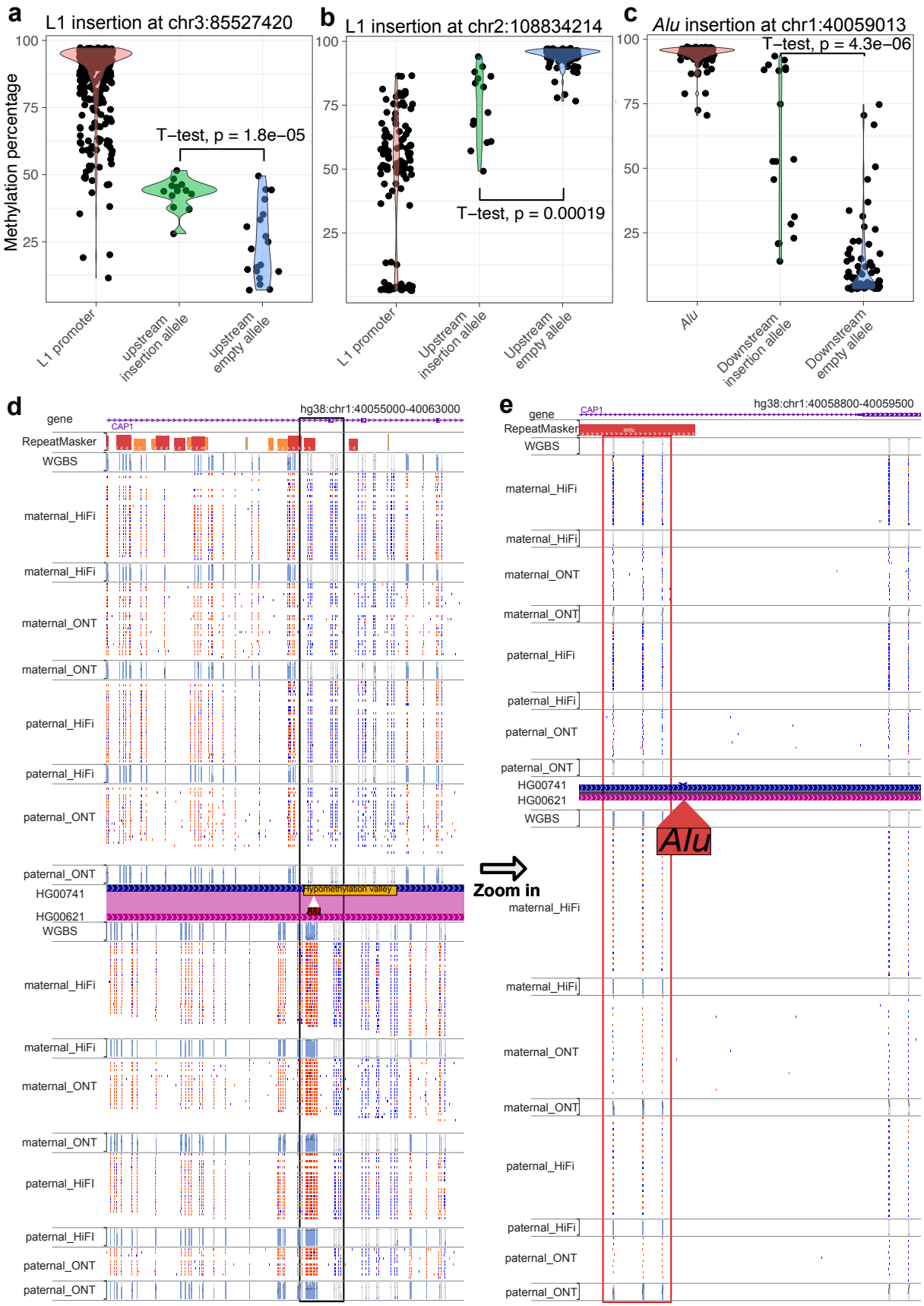
Figure 3. a: Schematic representation of average methylation calculation of all insertions and associated 300bp flanking regions. **b**: Methylation correlation between insertions and 300bp flanking regions. The average methylation of insertions was represented on the Y-

780 axis, while the methylation of flanking regions on the X-axis. *Alu*, LTR, L1, SVA, and non-TE
781 insertions were represented by red, green, blue, and purple dots, respectively, whereas
782 non-TE insertions were labeled grey, serving as a background. Linear regression lines with
783 standard errors were also plotted and labeled with same color scheme. **c**: Methylation
784 correlation between insertions within CpG islands and 300bp flanking regions. The average
785 methylation of insertions was represented on the Y-axis, while the methylation of flanking
786 regions on the X-axis. *Alu*, and non-TE insertions were represented in the same way as they
787 are in 3b. **d**: Schematic representation of insertion-flanking methylation regression as a
788 function of insertion-flanking distance. For each distance from 100bp to 10kb, we
789 performed linear regression between insertion and flanking average methylation and
790 estimated the linear regression slope and standard errors. **e**: Linear regression slope
791 between insertion and flanking methylation as a function of the insertion-flanking distance
792 from 10kb upstream to 10kb downstream. Standard errors were represented by error
793 bars. *Alu*, LTR, L1, SVA and non-TE insertions were plotted in separated plots with different
794 colors. **f**: The WashU Epigenome browser views of an *Alu* insertion between
795 hypomethylated and hypermethylated regions. The insertion was absent from HG03516
796 paternal haplotype but present in all the other nine haplotypes. A genome-align track was
797 used to illustrate the insertion site. The *Alu* insertion methylation was boxed, showing a
798 gradual change of methylation from left to right in some of them. Both HiFi and ONT
799 methylC tracks displayed methylation percentage from 0%-100%.


a
Empty allele
-300 bp
+300 bp
Insertion allele
Insertion
Empty flanking: methylation level of the 300bp flanking region
Insertion: methylation level of the inserted region
Insertion flanking: methylation level of the 300bp flanking region of insertion
Flanking methylation difference = insertion flanking - empty flanking
b
Not tandem repeat
Tandem repeat
Not CpG island
CpG island
$R^2 = 0.75$
$R^2 = 0.79$
$R^2 = 0.93$
$R^2 = 0.89$
Insertion flanking methylation
Empty flanking methylation
c
Alu
L1
SVA
HERVK
LTR5
$R^2 = 0.55$
$R^2 = 0.64$
$R^2 = 0.47$
$R^2 = 0.67$
Methylation percentage
Insertion flanking methylation
Empty flanking methylation
d
-25%
0.25%
+25%
99%
Insertion methylation
Flanking methylation difference (%)
e
-25%
3.9%
3.2%
+25%
96%
Insertion methylation
Flanking methylation difference (%)
TE
Alu
L1
SVA
HERVK
LTR5
f
L1 (1-500bp)
SVA (1-500bp)
$R^2 = 0.83$
$R^2 = 0.81$
Upstream methylation of insertion allele
Upstream methylation of empty allele

800

801 **Figure 4. a**: Schematic representation of a polymorphic insertion with aligned long reads.
802 The allele without the insertion site (top) and the allele with the insertion site (bottom)
803 were represented as reference allele and insertion allele. The average methylation level of
804 the insertion, insertion flanking regions, and empty flanking regions were calculated from
805 reads from all 32 samples. We then calculated flanking methylation difference by
806 subtracting the methylation level of empty flanking region from that of insertion flanking
807 region. **b**: Methylation level of empty flanking (X-axis), insertion flanking (Y-axis), and non-
808 TE inserted regions (blue-red color scale as shown in Fig. 4c). Insertions were separated
809 into four facets by whether they are tandem repeats and whether they are CpG islands.
810 Linear regression R^2 values were shown on each panel. **c**: Similar to Fig. 4b, but the
811 methylation level of different TE insertions was presented and separated by the TE
812 insertion types. R^2 were displayed in the top left corner on each panel. **d**: Dot plot
813 illustrated the relationship between methylation level of insertion (Y-axis) and methylation
814 difference between empty and insertion alleles of the flanking regions (X-axis). The average
815 methylation difference of all non-TE insertions was labeled at 0.25%. Red dotted lines
816 indicated the methylation difference threshold of +/-25%, and 99% of non-TE insertions
817 fall within this range. The insertion methylation density was displayed on the right side of
818 the figure. **e**: TE insertion flanking methylation differences plotted in the sample way as in
819 Fig. 4d. Different TE insertions were distinguished by different colors. The percentage of all
820 insertions within 25% differences and the average flanking methylation differences of
821 different TEs were labeled. The insertion methylation density was shown on the right of
822 the figure. The average methylation of each TE class was labeled with the same color
823 scheme. **f**: Methylation level of 300 bp upstream of the empty allele (X-axis), 300 bp
824 upstream of the insertion allele(Y-axis), and the first 500 bp inserted methylation (blue-red
825 color scale) of L1 and SVA insertions displayed similar to Figure 4b and c.


a
L1 insertion at chr3:85527420
Methylation percentage
T−test, p = 1.8e−05
L1 promoter
upstream insertion allele
upstream empty allele
b
L1 insertion at chr2:108834214
T−test, p = 0.00019
L1 promoter
Upstream insertion allele
Upstream empty allele
c
Alu insertion at chr1:40059013
T−test, p = 4.3e−06
Alu
Downstream insertion allele
Downstream empty allele
d
hg38:chr1:40055000-40063000
gene
CAP1
RepeatMasker
WGBS
maternal_HiFi
maternal_ONT
paternal_HiFi
paternal_ONT
HG00741
HG00621
hypomethylation valley
Alu
Zoom in
e
hg38:chr1:40058800-40059500
Alu

826

Figure 5. a: The violin plot illustrated the methylation levels of the L1 promoter at chr3:85527420. It compared the upstream CpG methylation of the insertion allele and the empty allele. Each dot represented the methylation percentage of a specific CpG site in a particular sample. A t-test was conducted to compare the matching upstream methylation levels, and the p-value is indicated on the figure. **b**: The violin plot of L1 promoter methylation at chr2:108834214 is similar to Fig.5a. The upstream CpG methylation of the insertion allele and the upstream methylation of the empty allele were shown. Each dot represented the methylation percentage of a specific CpG site in a particular sample. A t-test was performed and labeled accordingly. **c**: The violin plot of *Alu* methylation at chr1:40059013 is similar to Fig. 5a. The downstream CpG methylation of insertion allele and the downstream methylation of empty allele were shown. Each dot represented the methylation percentage of a specific CpG site in a particular sample. A t-test was performed and labeled accordingly. **d**: The WashU Epigenome Browser view of a hypermethylated *Alu* insertion within a hypomethylated valley that was present in HG00621 but absent in HG00741 characterized by WGBS, HiFi, and ONT methylation. Long-reads separated by their haploid type displayed the methylation percentage with the blue-red (low-high) two color scale, and all the methylC tracks displayed methylation percentages ranging from 0% to 100%. **e**: Zoom-in view of the methylation of the same *Alu* element from Fig. 5d. The boxed 3 CpG sites within 200bp of the insertion site underwent a change in methylation status, transitioning from hypomethylation to hypermethylation in association with the *Alu* insertion.

Reference

Alonge M, Lebeigle L, Kirsche M, Jenike K, Ou S, Aganezov S, Wang X, Lippman ZB, Schatz MC, Soyk S. 2022. Automated assembly scaffolding using RagTag elevates a new tomato system for high-throughput genome editing. *Genome Biology* **23**: 258.

Batzer MA, Deininger PL. 2002. Alu repeats and human genomic diversity. *Nature Reviews Genetics* **3**: 370–379.

Bird A. 2002. DNA methylation patterns and epigenetic memory. *Genes Dev* **16**: 6–21.

Bird AP. 1986. CpG-rich islands and the function of DNA methylation. *Nature* **321**: 209–213.

Bodea GO, Botto JM, Ferreiro ME, Sanchez-Luque FJ, de los Rios Barreda J, Rasmussen J, Rahman MA, Fenlon LR, Jansz N, Gubert C, et al. 2024. LINE-1 retrotransposons contribute to mouse PV interneuron development. *Nat Neurosci* 1–11.

Boissinot S, Chevret P, Furano AV. 2000. L1 (LINE-1) Retrotransposon Evolution and Amplification in Recent Human History. *Molecular Biology and Evolution* **17**: 915–928.

Cheetham SW, Kindlova M, Ewing AD. 2022. Methylartist: tools for visualizing modified bases from nanopore sequence data. *Bioinformatics* **38**: 3109–3112.

Cheng H, Concepcion GT, Feng X, Zhang H, Li H. 2021. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nat Methods* **18**: 170–175.

Choi JY, Purugganan MD. 2018. Evolutionary Epigenomics of Retrotransposon-Mediated Methylation Spreading in Rice. *Molecular Biology and Evolution* **35**: 365–382.

Chu C, Borges-Monroy R, Viswanadham VV, Lee S, Li H, Lee EA, Park PJ. 2021. Comprehensive identification of transposable element insertions using multiple sequencing technologies. *Nat Commun* **12**: 3836.

Cokus SJ, Feng S, Zhang X, Chen Z, Merriman B, Haudenschild CD, Pradhan S, Nelson SF, Pellegrini M, Jacobsen SE. 2008. Shotgun bisulphite sequencing of the Arabidopsis genome reveals DNA methylation patterning. *Nature* **452**: 215–219.

Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, Whitwham A, Keane T, McCarthy SA, Davies RM, et al. 2021. Twelve years of SAMtools and BCFtools. *GigaScience* **10**: giab008.

Ebert P, Audano PA, Zhu Q, Rodriguez-Martin B, Porubsky D, Bonder MJ, Sulovari A, Ebler J, Zhou W, Serra Mari R, et al. 2021. Haplotype-resolved diverse human genomes and integrated analysis of structural variation. *Science* **372**: eabf7117.

Emerson JW, Green WA, Schloerke B, Crowley J, Cook D, Hofmann H, Wickham H. 2013. The Generalized Pairs Plot. *Journal of Computational and Graphical Statistics* **22**: 79–91.

Ernst J, Kellis M. 2012. ChromHMM: automating chromatin-state discovery and characterization. *Nature methods* **9**: 215–216.

Ewing AD, Smits N, Sanchez-Luque FJ, Faivre J, Brennan PM, Richardson SR, Cheetham SW, Faulkner GJ. 2020. Nanopore Sequencing Enables Comprehensive Transposable Element Epigenomic Profiling. *Molecular Cell* **80**: 915-928.e5.

Fazzari MJ, Greally JM. 2004. Epigenomics: beyond CpG islands. *Nat Rev Genet* **5**: 446–455.

Feng H, Conneely KN, Wu H. 2014. A Bayesian hierarchical model to detect differentially methylated loci from single nucleotide resolution sequencing data. *Nucleic Acids Res* **42**: e69.

Flusberg BA, Webster DR, Lee JH, Travers KJ, Olivares EC, Clark TA, Korlach J, Turner SW. 2010. Direct detection of DNA methylation during single-molecule, real-time sequencing. *Nat Methods* **7**: 461–465.

897 Gardner EJ, Lam VK, Harris DN, Chuang NT, Scott EC, Pittard WS, Mills RE, Consortium T
898 1000 GP, Devine SE. 2017. The Mobile Element Locator Tool (MELT): population-
899 scale mobile element discovery and biology. *Genome Res* **27**: 1916–1929.

900 Gerdes P, Chan D, Lundberg M, Sanchez-Luque FJ, Bodea GO, Ewing AD, Faulkner GJ,
901 Richardson SR. 2023. Locus-resolution analysis of L1 regulation and
902 retrotransposition potential in mouse embryonic development. *Genome Res* **33**:
903 1465–1481.

904 Grandi FC, Rosser JM, Newkirk SJ, Yin J, Jiang X, Xing Z, Whitmore L, Bashir S, Ivics Z, Izsvák
905 Z, et al. 2015. Retrotransposition creates sloping shores: a graded influence of
906 hypomethylated CpG islands on flanking CpG sites. *Genome Research* **25**: 1135–
907 1146.

908 Han L, Su B, Li W-H, Zhao Z. 2008. CpG island density and its correlations with genomic
909 features in mammalian genomes. *Genome Biology* **9**: R79.

910 Hodges E, Smith AD, Kendall J, Xuan Z, Ravi K, Rooks M, Zhang MQ, Ye K, Bhattacharjee A,
911 Brizuela L, et al. 2009. High definition profiling of mammalian DNA methylation by
912 array capture and single molecule bisulfite sequencing. *Genome Res* **19**: 1593–
913 1605.

914 Holt JM, Saunders CT, Rowell WJ, Kronenberg Z, Wenger AM, Eberle M. 2024. HiPhase:
915 jointly phasing small, structural, and tandem repeat variants from HiFi sequencing.
916 *Bioinformatics* **40**: btae042.

917 Hubley R, Finn RD, Clements J, Eddy SR, Jones TA, Bao W, Smit AFA, Wheeler TJ. 2015. The
918 Dfam database of repetitive DNA families. *Nucleic Acids Res* **44**: D81–D89.

919 Jones PA. 2012. Functions of DNA methylation: islands, start sites, gene bodies and
920 beyond. *Nat Rev Genet* **13**: 484–492.

921 Krueger F, Andrews SR. 2011. Bismark: a flexible aligner and methylation caller for
922 Bisulfite-Seq applications. *Bioinformatics* **27**: 1571–1572.

923 Lam ET, Hastie A, Lin C, Ehrlich D, Das SK, Austin MD, Deshpande P, Cao H, Nagarajan N,
924 Xiao M, et al. 2012. Genome mapping on nanochannel arrays for structural variation
925 analysis and sequence assembly. *Nat Biotechnol* **30**: 771–776.

926 Lanciano S, Philippe C, Sarkar A, Pratella D, Domrane C, Doucet AJ, van Essen D, Saccani
927 S, Ferry L, Defossez P-A, et al. 2024. Locus-level L1 DNA methylation profiling
928 reveals the epigenetic and transcriptional interplay between L1s and their
929 integration sites. *Cell Genomics* **4**: 100498.

930 Laurent L, Wong E, Li G, Huynh T, Tsirigos A, Ong CT, Low HM, Sung KWK, Rigoutsos I,
931 Loring J, et al. 2010. Dynamic changes in the human methylome during
932 differentiation. *Genome Res* **20**: 320–331.

933 Lee HJ, Hou Y, Chen Y, Dailey ZZ, Riddihough A, Jang HS, Wang T, Johnson SL. 2020a.
934 Regenerating zebrafish fin epigenome is characterized by stable lineage-specific
935 DNA methylation and dynamic chromatin accessibility. *Genome Biology* **21**: 52.

936 Lee Y, Lee J, Kim J, Kim Y-J. 2020b. Insertion variants missing in the human reference
937 genome are widespread among human populations. *BMC Biology* **18**: 167.

938 Li D, Purushotham D, Harrison JK, Hsu S, Zhuo X, Fan C, Liu S, Xu V, Chen S, Xu J, et al.
939 2022. WashU Epigenome Browser update 2022. *Nucleic Acids Research* **50**: W774–
940 W781.

941 Li D, Zhuo X, Harrison JK, Liu S, Wang T. 2023. Modbed track: Visualization of modified
942 bases in single-molecule sequencing. *Cell Genomics* **3**: 100455.

943 Li H. 2018. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**:
944 3094–3100.

945 Li H. 2021. New strategies to improve minimap2 alignment accuracy ed. C. Alkan.
946 *Bioinformatics* **37**: 4572–4574.

947 Li H, Bloom JM, Farjoun Y, Fleharty M, Gauthier L, Neale B, MacArthur D. 2018. A synthetic-
948 diploid benchmark for accurate variant-calling evaluation. *Nat Methods* **15**: 595–
949 597.

950 Liao W-W, Asri M, Ebler J, Doerr D, Haukness M, Hickey G, Lu S, Lucas JK, Monlong J, Abel
951 HJ, et al. 2023. A draft human pangenome reference. *Nature* **617**: 312–324.

952 Lister R, Pelizzola M, Dowen RH, Hawkins RD, Hon G, Tonti-Filippini J, Nery JR, Lee L, Ye Z,
953 Ngo Q-M, et al. 2009. Human DNA methylomes at base resolution show widespread
954 epigenomic differences. *Nature* **462**: 315–322.

955 Liu Y, Rosikiewicz W, Pan Z, Jillette N, Wang P, Taghbalout A, Foox J, Mason C, Carroll M,
956 Cheng A, et al. 2021. DNA methylation-calling tools for Oxford Nanopore
957 sequencing: a survey and human epigenome-wide evaluation. *Genome Biology* **22**:
958 295.

959 Maunakea AK, Nagarajan RP, Bilenky M, Ballinger TJ, D'Souza C, Fouse SD, Johnson BE,
960 Hong C, Nielsen C, Zhao Y, et al. 2010. Conserved role of intragenic DNA
961 methylation in regulating alternative promoters. *Nature* **466**: 253–257.

962 Meissner A, Mikkelsen TS, Gu H, Wernig M, Hanna J, Sivachenko A, Zhang X, Bernstein BE,
963 Nusbaum C, Jaffe DB, et al. 2008. Genome-scale DNA methylation maps of
964 pluripotent and differentiated cells. *Nature* **454**: 766–770.

965 Moore LD, Le T, Fan G. 2013. DNA Methylation and Its Basic Function.
966 *Neuropsychopharmacol* **38**: 23–38.

967 Ni P, Nie F, Zhong Z, Xu J, Huang N, Zhang J, Zhao H, Zou Y, Huang Y, Li J, et al. 2023. DNA
968 5-methylcytosine detection and methylation phasing using PacBio circular
969 consensus sequencing. *Nature Communications* **14**: 4054.

970 Nunn A, Otto C, Fasold M, Stadler PF, Langenberger D. 2022. Manipulating base quality
971 scores enables variant calling from bisulfite sequencing alignments using
972 conventional bayesian approaches. *BMC Genomics* **23**: 477.

973 Patterson M, Marschall T, Pisanti N, van Iersel L, Stougie L, Klau GW, Schönhuth A. 2015.
974 WhatsHap: Weighted Haplotype Assembly for Future-Generation Sequencing
975 Reads. *Journal of Computational Biology* **22**: 498–509.

976 Penzkofer T, Jäger M, Figlerowicz M, Badge R, Mundlos S, Robinson PN, Zemojtel T. 2017.
977 L1Base 2: more retrotransposition-active LINE-1s, more mammalian genomes.
978 *Nucleic Acids Research* **45**: D68–D73.

979 Pollard MO, Gurdasani D, Mentzer AJ, Porter T, Sandhu MS. 2018. Long reads: their
980 purpose and place. *Human Molecular Genetics* **27**: R234–R241.

981 Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic
982 features. *Bioinformatics* **26**: 841–842.

983 Roadmap Epigenomics Consortium, Kundaje A, Meuleman W, Ernst J, Bilenky M, Yen A,
984 Heravi-Moussavi A, Kheradpour P, Zhang Z, Wang J, et al. 2015. Integrative analysis
985 of 111 reference human epigenomes. *Nature* **518**: 317–330.

986 Saha K, Nielsen GI, Nandani R, Zhang Y, Kong L, Ye P, An W. 2024. YY1 is a transcriptional
987 activator of the mouse LINE-1 Tf subfamily. *Nucleic Acids Research* gkae949.

988 Sanchez-Luque FJ, Kempen M-JHC, Gerdes P, Vargas-Landin DB, Richardson SR, Troskie
989 R-L, Jesuadian JS, Cheetham SW, Carreira PE, Salvador-Palomeque C, et al. 2019.
990 LINE-1 Evasion of Epigenetic Repression in Humans. *Molecular Cell* **75**: 590-
991 604.e12.

992 Schatz MC. 2017. Nanopore sequencing meets epigenetics. *Nat Methods* **14**: 347–348.

993 Sentmanat MF, Elgin SCR. 2012. Ectopic assembly of heterochromatin in Drosophila
994 melanogaster triggered by transposable elements. *Proceedings of the National*
995 *Academy of Sciences* **109**: 14104–14109.

996 Sigurpalsdottir BD, Stefansson OA, Holley G, Beyter D, Zink F, Hardarson MÞ, Sverrisson
997 SÞ, Kristinsdottir N, Magnusdottir DN, Magnusson OÞ, et al. 2024. A comparison of
998 methods for detecting DNA methylation from long-read sequencing of human
999 genomes. *Genome Biol* **25**: 69.

1000 Simpson JT, Workman RE, Zuzarte PC, David M, Dursi LJ, Timp W. 2017. Detecting DNA
1001 cytosine methylation using nanopore sequencing. *Nat Methods* **14**: 407–410.

1002 Smallwood SA, Tomizawa S, Krueger F, Ruf N, Carli N, Segonds-Pichon A, Sato S, Hata K,
1003 Andrews SR, Kelsey G. 2011. Dynamic CpG island methylation landscape in
1004 oocytes and preimplantation embryos. *Nat Genet* **43**: 811–814.

1005 Smith ZD, Hetzel S, Meissner A. 2024. DNA methylation in mammalian development and
1006 disease. *Nat Rev Genet* **26**: 7–30.

1007 Stergachis AB, Debo BM, Haugen E, Churchman LS, Stamatoyannopoulos JA. 2020. Single-
1008 molecule regulatory architectures captured by chromatin fiber sequencing. *Science*
1009 **368**: 1449–1454.

1010 Sved J, Bird A. 1990. The expected equilibrium of the CpG dinucleotide in vertebrate
1011 genomes under a mutation model. *Proceedings of the National Academy of*
1012 *Sciences of the United States of America* **87**: 4692–4696.

1013 The 1000 Genomes Project Consortium. 2015. A global reference for human genetic
1014 variation. *Nature* **526**: 68–74.

1015 The ENCODE Project Consortium. 2012. An integrated encyclopedia of DNA elements in
1016 the human genome. *Nature* **489**: 57–74.

1017 Tse OYO, Jiang P, Cheng SH, Peng W, Shang H, Wong J, Chan SL, Poon LCY, Leung TY,
1018 Chan KCA, et al. 2021. Genome-wide detection of cytosine methylation by single
1019 molecule real-time sequencing. *Proc Natl Acad Sci USA* **118**.
1020 https://www.pnas.org/content/118/5/e2019768118 (Accessed February 9, 2021).

1021 Turker MS. 2002. Gene silencing in mammalian cells and the spread of DNA methylation.
1022 *Oncogene* **21**: 5388–5393.

1023 Turker MS. 1999. The establishment and maintenance of DNA methylation patterns in
1024 mouse somatic cells. *Seminars in Cancer Biology* **9**: 329–337.

1025 Vaisvila R, Ponnaluri VKC, Sun Z, Langhorst BW, Saleh L, Guan S, Dai N, Campbell MA,
1026 Sexton BS, Marks K, et al. 2021. Enzymatic methyl sequencing detects DNA
1027 methylation at single-base resolution from picograms of DNA. *Genome Res* **31**:
1028 1280–1289.

1029 van Dijk EL, Jaszczyszyn Y, Naquin D, Thermes C. 2018. The Third Revolution in Sequencing
1030 Technology. *Trends in Genetics* **34**: 666–681.

1031 Weber M, Davies JJ, Wittig D, Oakeley EJ, Haase M, Lam WL, Schübeler D. 2005.
1032 Chromosome-wide and promoter-specific analyses identify sites of differential DNA
1033 methylation in normal and transformed human cells. *Nat Genet* **37**: 853–862.

1034 Wheeler TJ, Clements J, Eddy SR, Hubley R, Jones TA, Jurka J, Smit AFA, Finn RD. 2012.
1035 Dfam: a database of repetitive DNA based on profile hidden Markov models.
1036 *Nucleic Acids Research* **41**: D70–D82.

1037 Wickham H. 2009. *Ggplot2*. Springer, New York http://link.springer.com/10.1007/978-0-
1038 387-98141-3.

1039 Wu H, Xu T, Feng H, Chen L, Li B, Yao B, Qin Z, Jin P, Conneely KN. 2015. Detection of
1040 differentially methylated regions from whole-genome bisulfite sequencing data
1041 without replicates. *Nucleic Acids Res* **43**: e141.

1042 Wyler M, Stritt C, Walser J-C, Baroux C, Roulin AC. 2020. Impact of Transposable Elements
1043 on Methylation and Gene Expression across Natural Accessions of Brachypodium
1044 distachyon. *Genome Biology and Evolution* **12**: 1994–2001.

1045 Xu L, Seki M. 2020. Recent advances in the detection of base modifications using the
1046 Nanopore sequencer. *J Hum Genet* **65**: 25–33.

1047 Zhang Y, Shu J, Si J, Shen L, Estecio MRH, Issa J-PJ. 2012. Repetitive elements and enforced
1048 transcriptional repression co-operate to enhance DNA methylation spreading into a
1049 promoter CpG-island. *Nucleic Acids Research* **40**: 7257–7268.

1050 Zhou W, Liang G, Molloy PL, Jones PA. 2020. DNA methylation enables transposable
1051 element-driven genome expansion. *PNAS* **117**: 19359–19366.

1052 Zhou W, Mumm C, Gan Y, Switzenberg JA, Wang J, Oliveira PD, Kathuria K, Losh SJ,
1053 McDonald TL, Bessell B, et al. 2024. A personalized multi-platform assessment of
1054 somatic mosaicism in the human frontal cortex. 2024.12.18.629274.
1055 https://www.biorxiv.org/content/10.1101/2024.12.18.629274v1 (Accessed March
1056 26, 2025).

1057 Zhou X, Li D, Lowdon RF, Costello JF, Wang T. 2014. methylC Track: visual integration of
1058 single-base resolution DNA methylation data on the WashU EpiGenome Browser.
1059 *Bioinformatics* **30**: 2206–2207.

1060 Zhuo X, Hsu S, Purushotham D, Kuntala PK, Harrison JK, Du AY, Chen S, Li D, Wang T. 2023.
1061 Comparing genomic and epigenomic features across species using the WashU
1062 Comparative Epigenome Browser. *Genome Res* **33**: 824–835.

1063

1064