GTAGMe-seq: joint profiling of genetic variants, DNA methylation, GpC methyltransferase footprints, and 3D genome in the same DNA molecules

Hailu Fu[1], Haizi Zheng[1], Louis J. Muglia[1*], Li Wang[1#], Yaping Liu[1-5 #]

Affiliations:

[1] Division of Human Genetics, Cincinnati Children's Hospital Medical Center, Cincinnati, OH 45229, USA

[2] Division of Biomedical Informatics, Cincinnati Children's Hospital Medical Center, Cincinnati, OH 45229, USA

[3] Department of Pediatrics, University of Cincinnati College of Medicine, Cincinnati, OH 45229, USA

[4]Department of Electrical Engineering and Computing Sciences, University of Cincinnati College of Engineering and Applied Science, Cincinnati, OH 45229, USA

[5] University of Cincinnati Cancer Center, Cincinnati, OH 45219, USA

Corresponding Author's Email: L.W. wanglilove2@gmail.com, Y.L. lyping1986@gmail.com

* Present address: Burroughs Wellcome Fund, Research Triangle Park, NC 27614, USA

Abstract:

Cis-regulatory elements (CREs) coordinate to regulate the expression of their targeted genes, up to megabases away. However, the joint measurement of CREs' activities and their interactions over large genomic distances is limited by the current sequencing approaches. We report a method, GTAGMe-seq, which simultaneously captures the single nucleotide polymorphisms, DNA methylation, GpC methyltransferase footprints, and chromosome conformation changes from the same DNA molecules, together with transcriptome in the same assay. GTAGMe-seq shows high concordance with the state-of-art mono-omics assays across different molecular measurements. It reveals the coordinated GpC methyltransferase footprints at distal genomic segments in spatial proximity and different types of long-range allele-specific methyltransferase footprints.

Keywords: Cis-regulatory elements, joint profiling, DNA methylation, GpC methyltransferase footprints, chromosome conformation changes, 3D genome, transcriptome, coordinated GpC methyltransferase footprints, long-range allelic-specific CRE activity

48 **Main**

49 Cis-regulatory elements (CREs) play instrumental roles in the regulation of mammalian gene
50 expression[1]. Sets of transcriptional factors (TFs) bring together multiple CREs, which can be up
51 to megabases away, to initiate and maintain the expression of their targeted genes[2]. GpC
52 methyltransferase is recently utilized to footprint the TF bindings at CREs and reveals the
53 coordinated activities of CREs locally[3]. Since DNA is spatially organized into three-dimensional
54 (3D) space in nuclei, distal CREs can be brought into proximity through chromatin folding[4].
55 Therefore, it is possible that spatially proximal regions may also exhibit a coordinated GpC
56 methyltransferase footprint. However, conventional GpC methyltransferase based approaches,
57 such as NOMe-seq/dSMF[5,6], have limited power in detecting coordinated CREs activities across
58 large genomic distances due to the short reads profiled in next-generation sequencing (NGS).
59 Recent SMAC-seq combined methyltransferase with Nanopore sequencing to detect the
60 coordinated activities up to many kilobases[7], which, however, did not provide the spatial proximity
61 information to distinguish the cis- co-binding events by the same set of TFs with that due to the
62 trans-effects.

63 Moreover, recent studies suggested that allelic differences in genetic content can affect the local
64 CREs' activity by using heterozygous SNPs and their linked CpG methylation or GpC
65 methyltransferase footprint from the same reads[8–10]. In principle, the genetic differences between
66 two alleles may affect the coordination activities of both CREs that are distant in linear 2D
67 consideration of chromosomes but proximate in spatially in 3D. However, our current
68 understanding of the long-range allele-specific CRE activities is still limited by both computational
69 approaches and experimental assays. First, most GpC methyltransferase based assays heavily
70 depend on the sodium bisulfite treatment, which will lead to difficulties in distinguishing C>T SNPs
71 (65% SNPs in dbSNP) and C>T substitutions caused by bisulfite conversion. Our previously
72 developed Bis-SNP makes it possible to characterize the SNPs at limited known SNP loci with

73 high accuracy in single bisulfite sequencing[11]. However, the genome-wide genotyping accuracy
74 is still not comparable to the whole genome sequencing (WGS). Second, both the short fragment
75 sizes in NGS libraries and the lack of spatial proximity information in long-read sequencing will
76 prevent us from characterizing the spatial interactions between genetic contents and distal CRE
77 activities at two different alleles.

78 The chromosome conformation capture (3C) derived technologies, such as *in situ* Hi-C, have
79 been widely used to study the 3D genome organization[12]. In Hi-C experiments, spatial proximity
80 information is captured through restriction enzyme digestion and ligation of proximal genomic
81 segments. After the ligation, nucleosomes and TFs are still crosslinked with DNA inside the nuclei,
82 which can be footprinted by the exogenous GpC methyltransferase and detectable by the follow-
83 up bisulfite sequencing, similar to the covalent genetic variants and endogenous CpG
84 methylation[13,14]. Therefore, based on our previously developed NOMe-seq and Methyl-HiC[5,14], we
85 developed GTAGMe-seq (**G**enome, **T**ranscriptome, GpC **A**ccessibility, 3D **G**enome, and
86 **Me**thylome sequencing), to jointly profile multi-omics from the same DNA molecules, together
87 with the transcriptome in the same assay (Fig. 1a).

88 Briefly, after the crosslinking, we isolated nuclei and treated them with a methylation insensitive
89 restriction enzyme (e.g., *DpnII*) to digest DNA. We used proximity ligation to capture the long-
90 range interactions in the 3D genome topology. We further utilized GpC methyltransferase
91 (*M.CviPI*) to footprint the chromatin accessibility. After reversing crosslinks, we extracted RNA
92 molecules for total RNA-seq. We isolated the DNA molecules with the follow-up bisulfite
93 conversion and paired-end sequencing to obtain covalent endogenous cytosine methylation and
94 exogenous GpC methyltransferase footprint information. Finally, we improved our previously
95 developed computational method, Bis-SNP, to accurately identify the genome-wide single
96 nucleotide polymorphisms (SNPs) from the bisulfite converted reads (Details in Methods and
97 Supplementary Methods).

98 To demonstrate the performance of GTAGMe-seq, we applied it to the human lung fibroblast cell
99 line (IMR-90) and B-lymphoblastoid cell line (GM12878) (Supplementary Table 2-3), which have
100 both been comprehensively profiled publically[15]. First, we compared the measurement of the 3D
101 genome with *in situ* Hi-C results in the same cell lines. The contact matrix is highly similar to that
102 from *in situ* Hi-C (Fig. 1b, Supplementary Fig. 1d), with the indistinguishable distribution of contact
103 probabilities (Supplementary Fig. 1a). Across different resolutions, we also identified comparable
104 compartment score, insulation score, and topologically associated domains (TADs) from the two
105 datasets (Supplementary Fig. 1b-c,e). Next, we compared the measurement of DNA methylation
106 with WGBS from the same cell lines. To avoid the ambiguity of endogenous CpG methylation and
107 exogenous GpC methyltransferase footprint at GpCpG (GCG) sites, we excluded GCG sites for
108 the follow-up measurement of CpG methylation and GpC methyltransferase footprint as we did in
109 the NOMe-seq paper[5]. GTAGMe-seq captured 31-32 million HCGs (H=A, C, or T) in the human
110 genomes and the methylation level of these HCGs was highly consistent with that from WGBS
111 (Fig. 1c, Supplementary Fig. 1f-g). Similar to Methyl-HiC, GTAGMe-seq captured ~25% fewer
112 HCGs compared to WGBS. Third, we tested the measurement of GpC methyltransferase footprint
113 (GCH accessibility) at the known TFs binding sites. Notably, nucleosomes and TFs are both
114 crosslinked with DNA. Part of the open chromatin regions will still show the low GCH accessibility
115 level when TFs are bound, which is different from the conventional NOMe-seq without the
116 fixations. We observed a concordant lower level of GpC methyltransferase footprints at PolII
117 binding sites with a higher TF occupancy level measured by PolII ChIP-seq in the same cell line
118 (Fig. 1d). We also observed the expected GCH footprint patterns and HCG methylation level at
119 TF binding sites that usually bind distantly to the transcriptional starting sites (TSS) (Fig. 1e,
120 Supplementary Fig. 2). Moreover, we compared the measurement of the transcriptome with total
121 RNA-seq from the same cell lines. Messenger RNA (mRNA) is highly fragmented after the
122 formaldehyde fixation step (crosslink). With the computational correction of the possible bias at
123 fragmentation and G+C%, the transcriptome is generally concordant with that from total mRNA-

124 seq, in terms of absolute quantification (Fig. 1f, Supplementary Fig. 1h-i) and differentially
125 expressed genes (Supplementary Fig. 3). Finally, we compared the measurement of SNPs
126 (GM12878) with deep WGS from the 1000 genome project on the same individual (NA12878,
127 a.k.a. HG001). We evaluated the performance by the ground truth genotype results from the same
128 individual at the Genome in a Bottle Consortium (GIAB), which are well characterized by the
129 integration of multiple technologies. Results with improved Bis-SNP at GTAGMe-seq showed
130 similar high performance as that from WGS (Fig.1g, Supplementary Fig. 1j) and very high
131 accuracy at sites with genotype callable (Supplementary Fig. 1k). These results, taken together,
132 demonstrated that GTAGMe-seq can simultaneously and accurately profile 3D genome, DNA
133 methylation, GpC methyltransferase footprint, transcriptome, and SNPs in the biological samples.

134 We next asked if the genomic regions linearly separated but positioned proximity in 3D space
135 would have coordinated footprint status. We analyzed the GCH (H=A, C, or T) methyltransferase
136 footprint on GTAGMe-seq reads separated in chromatin loop anchors (Fig. 2a). Indeed, the GCH
137 methyltransferase footprint in GTAGMe-seq read pairs mapping to separated loop anchors
138 showed a similarly high correlation as that mapping to the local regions (<1kb, Fig. 2b). To ensure
139 the concordance is not due to the similar level of GCH methyltransferase footprint at the loop
140 anchors, we shuffled the links of read pairs at the loop anchors while maintaining the average
141 GCH methyltransferase footprint at loop anchors. The shuffled read pairs indeed showed
142 significantly much less correlation (Fig. 2a-b, Fisher's (1925) z test, $p<2.2e^{-16}$).

143 To further demonstrate the utility of GTAGMe-seq, we utilized it to analyze the long-range allele-
144 specific GpC methyltransferase footprint. We separated GTAGMe-seq reads into two groups by
145 their parent-of-origin at the heterozygous SNPs (SNP anchors, Fig. 3a). Since the sequencing
146 reads come from the same single DNA molecules, the other ends of the read pairs were also
147 separated into two alleles, even though they are not overlapped with the SNPs and mapped over
148 large genomic distances (non-SNP anchors, Fig. 3a). Thus, we are able to analyze the long-range

149 allele-specific GCH methyltransferase footprint. We identified three groups of allele-specific
150 footprint loci based on their differences of GCH accessibility between the two alleles at both
151 anchors (Group 1, n=2,149, FDR<0.05 at both anchors), or only at SNP anchors (Group 2,
152 n=3,285, FDR<0.05), or only at non-SNP anchors (Group 3, n=2,963, FDR<0.05). Further, to
153 understand the function of these long-rang allele-specific footprints, we checked the enrichment
154 of the anchors over different known TF binding sites. Group 1 showed the highest enrichment of
155 CTCF binding sites at both anchors (Fig. 3b, Supplementary Fig. 4), suggesting the disruption of
156 the CTCF binding by genetic variants at the SNP anchor may also disrupt the loop structure and
157 thus the binding activities at the non-SNP anchor in spatial proximity. Group 2 exhibited the
158 highest enrichment of PolII binding sites at the SNP anchors, but enrichment of different TFs at
159 the non-SNP anchors, which usually bind to the distal enhancer regions. Group 3 showed the
160 highest depletion of TF binding sites that are usually bound to the promoter regions.

161 Genome-wide association studies (GWAS) have associated common complex diseases with
162 hundreds of thousands of genetic variants. Due to the linkage disequilibrium (LD), many SNPs
163 near the tag SNPs in the GWAS studies are also identified to be significantly associated with the
164 traits. The local CRE activities, especially the imbalanced CREs activities linked with the switching
165 of risk and non-risk alleles nearby, are often utilized to prioritize the potential causal SNPs within
166 the LD blocks. The Group 3 long-range allele-specific footprint may suggest that the switching of
167 risk and non-risk allele does not need to affect the local chromatin accessibility but instead the
168 activities of distant CREs in spatial proximity. Indeed, we observed the overlap of SNP anchors
169 at Group 3 in GM12878 (B-lymphoblastoid cell) with three significant GWAS loci (rs1885276 with
170 $4.157e^{-09}$, rs6431262 with $2.023e^{-12}$, and rs3805683 with $2.267e^{-23}$) that are associated with
171 inflammatory bowel diseases (IBD)[16]. IBD is found to be driven by the malfunctions of the B cells,
172 but the annotation of chromatin states near these three regions are all heterochromatin regions
173 in GM12878 cells. Thus, instead of methylation or accessibility QTLs mapping that required

174 multiple samples, GTAGMe-seq here provides an alternative approach to link SNPs with the
175 variations of distal CREs in spatial proximity at a single assay.

176 Overall, we have developed a method to simultaneously study the genetic variants, DNA
177 methylation, GpC methyltransferase footprint, and 3D genome in the same DNA molecules
178 together with the transcriptome in the same assay. Combining multiple assays allowed us to
179 uncover the coordinated GpC methyltransferase footprint and long-range allele-specific footprint
180 at genomic regions that are far away in linear DNA sequences but spatially close in the nucleus.
181 While the library complexity in the current GTAGMe-seq protocol is still limited, this limitation may
182 be overcome in the future by the alternative bisulfite-free library construction strategies[17,18]. Future
183 extension of GTAGMe-seq to the single-cell level will allow us to further characterize the gene-
184 regulatory mechanism from genetic variants to molecule phenotypes at the heterogeneous
185 tissues.

186

187

188

189

190

191

192

193

194

195

196

197

198 **Methods**

199 **Cell culture**

200 The human lung fibroblast cell line IMR-90 was purchased from ATCC and cultured in EMEM
201 (ATCC) with 10% FBS (Gibco) and 1% penicillin/streptomycin. After reaching 70~80 confluence,
202 cells were detached with 0.5% Trypsin and harvested by centrifuge at 300g for 5min. Human B-
203 lymphoblastoid cell line GM12878 was obtained from CORIELL INSTITUTE and cultured with
204 RPMI-1640 supplemented with 15% FBS and 1% penicillin/streptomycin. GM12878 cells were
205 harvested at a density around 0.7M to 0.8M/mL by centrifuge at 300g for 5min.

206

207 **GTAGMe-Seq**

208 Cells were crosslinked with 1% formaldehyde for 10 min at room temperature, then quenched by
209 adding 0.2M Glycine. After washing with ice-cold PBS, cells were suspended in the nuclei
210 isolation buffer with RNase inhibitors and incubated on ice for 1hr. Subsequent to nuclei
211 membrane permineralization, chromatin was digested using 100U *DpnII* (NEB, R0543L) overnight
212 at 37 ℃ with RNase inhibitors. The next day, DpnII was inactivated, and nuclei suspension was
213 cooled to room temperature and subjected to biotin fill-in. Proximal segments were *in situ* ligated
214 using T4 DNA ligase (NEB, M0202) supplemented with RNase inhibitors. Then GpC
215 methyltransferase footprint was performed by incubating the nuclei with *M.CviPI* (NEB, M0227L),
216 and PBS was added to stop the reaction. After that, one-third of the nuclei were separated for
217 RNA isolation using MagMAX™ FFPE DNA/RNA Ultra Kit (Thermo A31881). The total RNA
218 library was prepared using SMARTer® Stranded Total RNA-Seq Kit v2 (TAKARA 634413). The
219 rest of the nuclei were reverse crosslinked, DNA was precipitated and sonicated to average 400bp
220 followed by 1xAMPure beads cleanup. Biotin pulldown was performed using streptavidin beads,
221 DNA on beads was end-repaired, d-A tailed, and ligated with truncated cytosine-methylated X-
222 gene Universal Stubby Adapter (IDT). 0.1%~0.5% unmethylated lambda DNA was spiked-in.

223 Bisulfite conversion was conducted using the EZ DNA Methylation-Gold Kit (D5006). Bisulfite
224 converted DNA library was amplified using truncated TruSeq™–Compatible Indexing Primer and
225 KAPA HiFi Uracil+ ReadyMix (Roche, KK2801), supplemented with $MgCl_2$. Library quality was
226 determined by qPCR and BioAnalyzer 2000 (Agilent Technologies). Pooling of multiplexed
227 sequencing samples, clustering, and sequencing was carried out as recommended by the
228 manufacturer on Illumina HiSeq XTen or NovaSeq S6000 with the 150 paired-end. RNA-seq
229 libraries or Phi-X (20%) were spiked in to overcome the imbalance of GC ratio of the bisulfite-
230 converted DNA libraries.

231

232 **GTAGMe-seq (DNA) data analysis**

233 Raw reads were first trimmed by using Trim Galore! (v0.6.6, with Cutadapt v2.10)[19] with
234 parameters "--paired_end --clip_R1 5 --clip_R2 5 --three_prime_clip_R1 5 --three_prime_clip_R2
235 5" to remove the adapters and low-quality reads. The clipped length was determined based on
236 the composition of base pairs along the sequencing cycle by FastQC (v0.11.9). The reference
237 genome (b37, human_g1k_v37.fa) was *in silico* converted to make C/T reference (all Cs were
238 converted to Ts) and G/A reference (all Gs were converted to As). Paired-end reads were mapped
239 by our previously developed bhmem (v0.37)[14] to each converted reference. Only uniquely
240 mapped and mapping quality passed reads (mapQ>30) on both ends were joined. Reads marked
241 as PCR duplicated reads were filtered. The reads with incompletely converted cytosine (WCH,
242 W=A or T, H=A, C, or T) were filtered out as previously described[11]. Joint read pairs with more
243 than 20kb insertion size were considered as long-range interactions and subjected to the following
244 interaction analysis. Bam files were further converted to .hic files for the following 3D genome
245 analysis similar to Hi-C data (details in the analysis of *in situ* Hi-C data). HCG methylation and
246 GCH accessibility level were calculated by bissnp_easy_usage.pl with Bis-SNP (v0.90) in NOMe-
247 seq mode. Only bases with a quality score of more than 5 were included in the downstream

248 methylation analysis. More details were implemented in Bhmem.java with parameters "-
249 outputMateDiffChr -buffer 100000".
250
251 **GTAGMe-seq (RNA) data analysis**
252 Raw reads were first trimmed as paired-end reads using Trim Galore! (v0.6.6, with Cutadapt
253 v2.10) with parameters "--paired_end --clip_R1 15 --clip_R2 15 --three_prime_clip_R1 5 --
254 three_prime_clip_R2 5" to remove the adapters and low-quality reads. The clipped length was
255 determined based on the composition of base pairs along the sequencing cycle by FastQC
256 (v0.11.9). Both alignment-based (STAR, v2.7.6a[20] ,and RNA-SeQC, v2.3.5[21]) or alignment-free
257 (Salmon, v1.5.1[22]) approaches were tried. Salmon was finally used for the bias correction,
258 visualization, and data analysis due to the slightly better concordance with the RNA-seq in
259 ENCODE. The reference transcriptome was obtained from Gencode (v33lift37) and indexed by
260 Salmon with k=31. The parameter used for Salmon quantification was: "quant --seqBias --gcBias
261 --posBias -l A --validateMappings". The differential expression analysis was performed on gene-
262 level by the edgeR (3.32.1)[23] package in R (4.0.5). Only genes with FDR<0.05 and log2 fold
263 change > 1 were considered as the differential expressed genes between IMR-90 and GM12878.
264 The same RNA-seq analysis pipeline was applied to the total RNA-seq from ENCODE.
265
266 **Analysis of WGBS data**
267 DNA methylation level in WGBS was directly downloaded from ENCODE (see details in
268 Supplementary Table 1) and extracted from the sites that only overlapped with HCGs in the
269 reference genome.
270
271 **Analysis of *in situ* Hi-C data**
272 Arrowhead in Juicer tools (v1.22.01)[24] is used to call TAD domains, with parameters "-r 10000 -k

273 KR -m 2000". The result is a BEDPE file containing all identified domains. FAN-C (v0.9.21)[25] is
274 used to calculate insulation scores, with subcommand insulation and parameters" @10kb@KR -
275 w 500000". To calculate A/B compartment scores, we first use the eigenvector command in Juicer
276 tools (v1.22.01) to obtain the first eigenvector of the Hi-C observed/expected correlation matrix,
277 with parameters KR and BP 500000. Then we calculate the average GC contents for each 500kb
278 bin. With the GC contents and the eigenvector aligned, we flip the sign of the eigenvector if
279 needed so that high GC content regions are associated with positive compartment scores
280 (compartment A), and low GC content regions are associated with negative compartment scores
281 (compartment B).
282

283 **Accurate identification of SNPs from GTAGMe-seq**

284 BisulfiteGenotyper in Bis-SNP (v0.90)[11] was utilized to identify SNPs with a genotype quality score
285 of more than 20 at GTAGMe-seq data (raw genotype). Further, VCFpostprocess was utilized to
286 filter out poor quality SNPs (at the regions with more than sequencing depth 250X, with strand
287 bias more than -0.02, with sequencing depth more than 40X and fraction of reads with mapping
288 quality 0 more than 0.1, genotype quality divided by sequencing depth less than 1, with 2 SNPs
289 nearby within the +/-10bp window). Finally, the filtered SNPs were utilized for imputation and
290 phasing at by Minimac4 (v1.0.2) and Eagle (v2.4.1) together with 1000G phase3 v5 panel (EUR
291 population). Since NA12878 was already profiled in 1000G phase3 project, to avoid the potential
292 bias, NA12878 was excluded from the panel for the imputation and phasing. Only biallelic SNPs
293 loci were utilized for the imputation and phasing. Other parameters during the SNP imputation
294 and phasing followed the similar steps as that recommended by Michigan Imputation Server[26].
295 After the imputation, only biallelic SNPs with minor allele frequency more than 1% and R^2 more
296 than 0.3 were kept for the following analysis. WGS genotyped results (NA12878) were
297 downloaded from the 1000G project (phase3)[27]. Gold standard genotype results (NA12878, a.k.a

298 HG001) were downloaded from the Genome in a Bottle (GIAB)[28] website. Bcftools (v1.10.2)[29]
299 was utilized to summarize the Ti/Tv ratio. Concordance in GATK (v4.1.9.0) was utilized to
300 evaluate the concordance among gold standard (GIAB), WGS (1000G), and GTAGMe-seq.
301

302 **Correlation analysis of GpC methyltransferase footprint in GTAGMe-seq**

303 Only read pairs with GCH methylation levels at both ends were kept for the analysis. For the
304 regions of interest, such as HiCCUPS loops anchor regions, Pearson Correlation Coefficient
305 (PCC) was calculated directly from the GCH methylation level of each read of the read pair, not
306 by the average GCH methylation level at each end of anchor regions. Only read pairs spanning
307 at least 20kb genomic distance were considered for the analysis. As a control, all the read pairs
308 within the same genomic regions, no matter whether they have long-range interaction or not, were
309 randomly shuffled 100 times and then subjected to PCC calculation. Fisher's (1925) z,
310 implemented in "cocor.indep.groups" at R package "cocor"21, was used to assess the significance
311 between observed PCC from interacted read pairs and PCC of the random read pairs. The scripts
312 for this analysis are provided in MethyCorAcrossHiccups.java with parameters "-
313 methyPatternSearch GCH -methyPatternCytPos 2 -bsConvPatternSearch WCH -
314 bsConvPatternCytPos 2 -useBadMate -useGeneralBedPe".
315

316 **Analysis of long-range allele-specific GpC methyltransferase footprint in GTAGMe-seq**

317 Only read pairs with one end overlapped with heterozygous SNPs, and GCH methylation levels
318 at both ends were kept for the analysis. At each heterozygous SNP position with enough reads
319 covered at both reference alleles and alternative alleles, the p-value was calculated by Fisher
320 Exact test and followed with FDR correction (Benjamini-Hochberg). Only loci with FDR<0.05 were
321 considered as significant allele-specific loci. For SNP anchor-Only or non-SNP anchor-Only
322 groups, the methylation differences between alleles should have FDR>0.95 at the insignificant

anchor. The scripts for this analysis is provided in LongRangeAsm.java with parameters "-minDist 1000 -coverageRef 2 -coverageAlt 2 -methyPatternSearch GCH -methyPatternCytPos 2 -bsConvPatternSearch WCH -bsConvPatternCytPos 2 -useBadMate -minMapQ 30".

Data availability

The GTAGMe-seq datasets generated in this study have been deposited in the Gene Expression Omnibus (GEO) under the accession number GSE189158. Previously published data used for this study are listed in Supplementary Table 1.

Code availability

All the source code is publicly available at GitHub:

https://bitbucket.org/dnaase/bisulfitehic/src/master/

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

All authors read and approved the manuscript for review.

Competing interests

Not applicable.

Funding

Y.L. is supported by the startup grant and Center for Pediatric Genomics award from Cincinnati Children's Hospital Medical Center.

Authors' contributions

Y.L. and L.W. conceived the study. H.F. and L.W. performed the experiment with input from Y.L. and L.J.M.. Y.L. and H.Z. performed the computational analysis. Y.L., H.F., L.W., and L.J.M. wrote the manuscript together. All authors read and approved the final manuscript.

Acknowledgments

The author thanks the input from Drs. Guoqiang Li and Bing Ren from Ludwig Institute for Cancer Research and University of California at San Diego, Drs. Matthew T. Weirauch and Raphael Kopan from Cincinnati Children's Hospital Medical Center. This work was supported by the computational resources from the Biomedical Informatics (BMI) high-performance computing cluster in CCHMC. This work also used the Extreme Science and Engineering Discovery Environment (XSEDE), which is supported by the National Science Foundation grant number ACI-1548562. This work used the XSEDE at the Pittsburgh Supercomputing Center (PSC) through allocation MCB190124P and MCB190006P.

370 **References**

371 1. Wittkopp, P. J. & Kalay, G. Cis-regulatory elements: molecular mechanisms and evolutionary
372 processes underlying divergence. *Nat. Rev. Genet.* **13**, 59–69 (2011).
373 2. Wasserman, W. W. & Sandelin, A. Applied bioinformatics for the identification of regulatory
374 elements. *Nat. Rev. Genet.* **5**, 276–287 (2004).
375 3. Sönmezer, C. *et al.* Molecular Co-occupancy Identifies Transcription Factor Binding
376 Cooperativity In Vivo. *Mol. Cell* **81**, 255–267.e6 (2021).
377 4. Yu, M. & Ren, B. The Three-Dimensional Organization of Mammalian Genomes. *Annu. Rev.*
378 *Cell Dev. Biol.* **33**, 265–289 (2017).
379 5. Kelly, T. K. *et al.* Genome-wide mapping of nucleosome positioning and DNA methylation
380 within individual DNA molecules. *Genome Res.* **22**, 2497–2506 (2012).
381 6. Krebs, A. R. *et al.* Genome-wide Single-Molecule Footprinting Reveals High RNA
382 Polymerase II Turnover at Paused Promoters. *Mol. Cell* **67**, 411–422.e4 (2017).
383 7. Shipony, Z. *et al.* Long-range single-molecule mapping of chromatin accessibility in
384 eukaryotes. *Nat. Methods* **17**, 319–327 (2020).
385 8. Xie, W. *et al.* Base-resolution analyses of sequence and parent-of-origin dependent DNA
386 methylation in the mouse genome. *Cell* **148**, 816–831 (2012).
387 9. Shoemaker, R., Deng, J., Wang, W. & Zhang, K. Allele-specific methylation is prevalent and
388 is contributed by CpG-SNPs in the human genome. *Genome Res.* **20**, 883–889 (2010).
389 10. Lee, I. *et al.* Simultaneous profiling of chromatin accessibility and methylation on human cell
390 lines with nanopore sequencing. *Nat. Methods* **17**, 1191–1199 (2020).
391 11. Liu, Y., Siegmund, K. D., Laird, P. W. & Berman, B. P. Bis-SNP: combined DNA methylation
392 and SNP calling for Bisulfite-seq data. *Genome Biol.* **13**, R61 (2012).
393 12. Rao, S. S. P. *et al.* A 3D map of the human genome at kilobase resolution reveals principles
394 of chromatin looping. *Cell* **159**, 1665–1680 (2014).

395 13. Selvaraj, S., R Dixon, J., Bansal, V. & Ren, B. Whole-genome haplotype reconstruction using
396 proximity-ligation and shotgun sequencing. *Nat. Biotechnol.* **31**, 1111–1118 (2013).
397 14. Li, G. *et al.* Joint profiling of DNA methylation and chromatin architecture in single cells. *Nat.*
398 *Methods* **16**, 991–993 (2019).
399 15. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human
400 genome. *Nature* **489**, 57–74 (2012).
401 16. de Lange, K. M. *et al.* Genome-wide association study implicates immune activation of
402 multiple integrin genes in inflammatory bowel disease. *Nat. Genet.* **49**, 256–261 (2017).
403 17. Liu, Y. *et al.* Bisulfite-free direct detection of 5-methylcytosine and 5-hydroxymethylcytosine
404 at base resolution. *Nat. Biotechnol.* **37**, 424–429 (2019).
405 18. Vaisvila, R. *et al.* Enzymatic methyl sequencing detects DNA methylation at single-base
406 resolution from picograms of DNA. *Genome Res.* (2021) doi:10.1101/gr.266551.120.
407 19. Krueger, F., James, F., Ewels, P., Afyounian, E. & Schuster-Boeckler, B.
408 *FelixKrueger/TrimGalore: v0.6.7 - DOI via Zenodo*. (Zenodo, 2021).
409 doi:10.5281/ZENODO.5127898.
410 20. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
411 21. DeLuca, D. S. *et al.* RNA-SeQC: RNA-seq metrics for quality control and process
412 optimization. *Bioinformatics* **28**, 1530–1532 (2012).
413 22. Patro, R., Duggal, G., Love, M. I., Irizarry, R. A. & Kingsford, C. Salmon provides fast and
414 bias-aware quantification of transcript expression. *Nat. Methods* **14**, 417–419 (2017).
415 23. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for
416 differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139–140
417 (2010).
418 24. Durand, N. C. *et al.* Juicer Provides a One-Click System for Analyzing Loop-Resolution Hi-C
419 Experiments. *Cell Syst* **3**, 95–98 (2016).
420 25. Kruse, K., Hug, C. B. & Vaquerizas, J. M. FAN-C: a feature-rich framework for the analysis

421 and visualisation of chromosome conformation capture data. *Genome Biol.* **21**, 303 (2020).
422 26. Das, S. *et al.* Next-generation genotype imputation service and methods. *Nat. Genet.* **48**,
423 1284–1287 (2016).
424 27. 1000 Genomes Project Consortium *et al.* A global reference for human genetic variation.
425 *Nature* **526**, 68–74 (2015).
426 28. Zook, J. M. *et al.* Extensive sequencing of seven human genomes to characterize benchmark
427 reference materials. *Scientific data* vol. 3 160025 (2016).
428 29. Li, H. A statistical framework for SNP calling, mutation discovery, association mapping and
429 population genetical parameter estimation from sequencing data. *Bioinformatics* **27**, 2987–
430 2993 (2011).

431
432
433
434
435
436
437
438
439
440
441
442
443
444
445

Figures

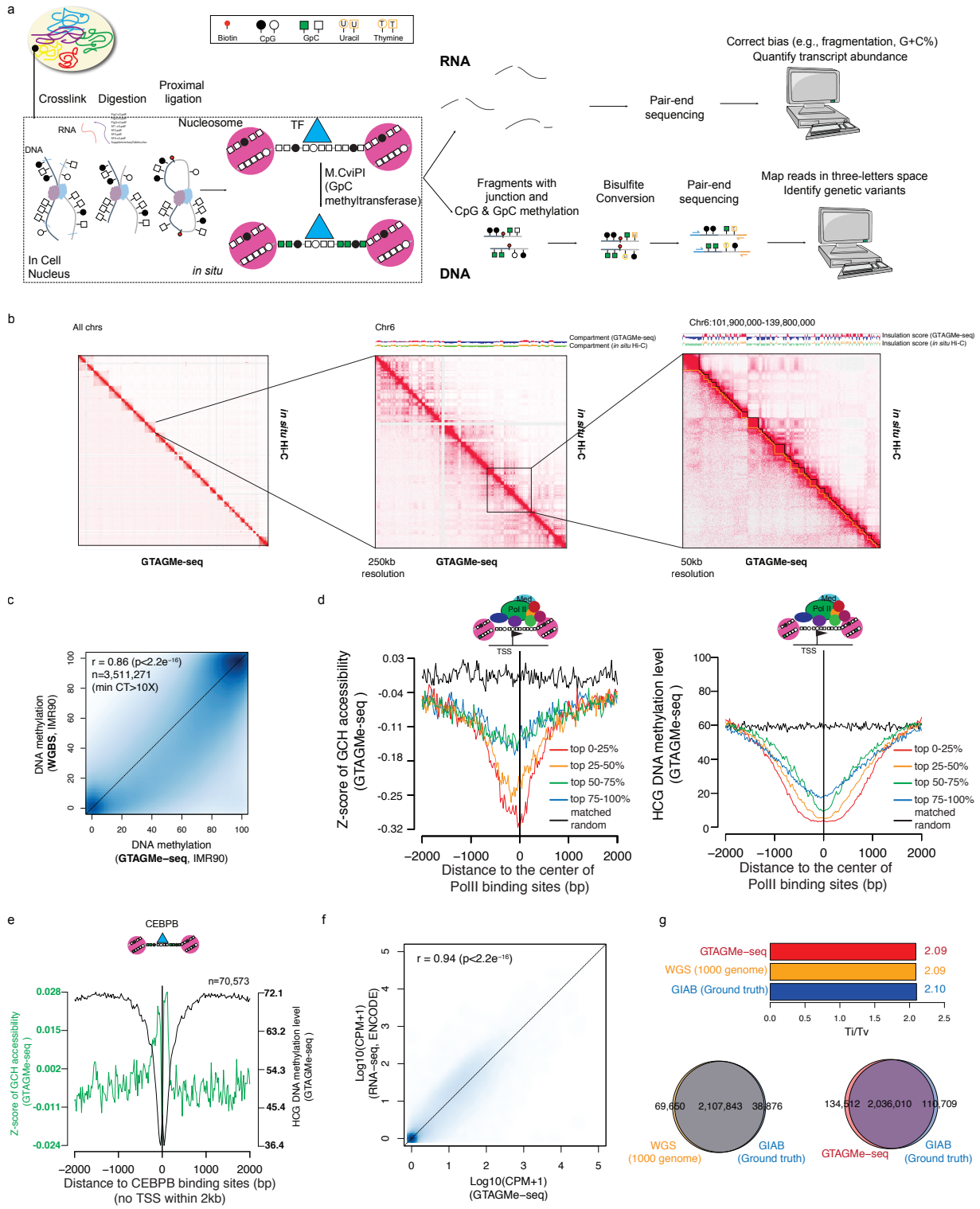

Fig.1. GTAGMe-seq generates high-quality multi-omics data. (a). The method scheme. (b).

449 The 3D genome concordance with Hi-C in IMR-90. From left to right: 1Mb resolution across all
450 chromosomes, 250kb resolution for chromosome 6 (include a comparison of compartment score
451 with Hi-C), 50kb resolution for chromosome 6 (include a comparison of insulation score with Hi-
452 C). (c). The methylation concordance with WGBS at HCG (H=A, C, or T) sites in IMR-90. (d).
453 Average GCH (H=A, C, or T) methyltransferase footprint level (left) and HCG methylation level
454 (right) near PolII binding sites. PolII binding sites are divided into quantiles based on the signal
455 strengths of PolII ChIP-seq in IMR-90. Matched random intervals are randomly chosen from the
456 same chromosome and interval lengths as the intervals in top 0-25% quantile. (e). Average GCH
457 methyltransferase footprint level and HCG methylation level around the distal CEBPB binding
458 sites in IMR-90. (f). Scatterplot of the concordance at the gene level with total RNA-seq
459 (ENCODE) in IMR-90. (g). The SNPs concordance with the ground truth SNPs data from Genome
460 in a Bottle (GIAB). The results are compared with the SNPs called from WGS (1000 genome
461 project, phase3 in NA12878). Ti/Tv ratio is also compared.
462
463
464
465
466
467
468
469
470
471
472

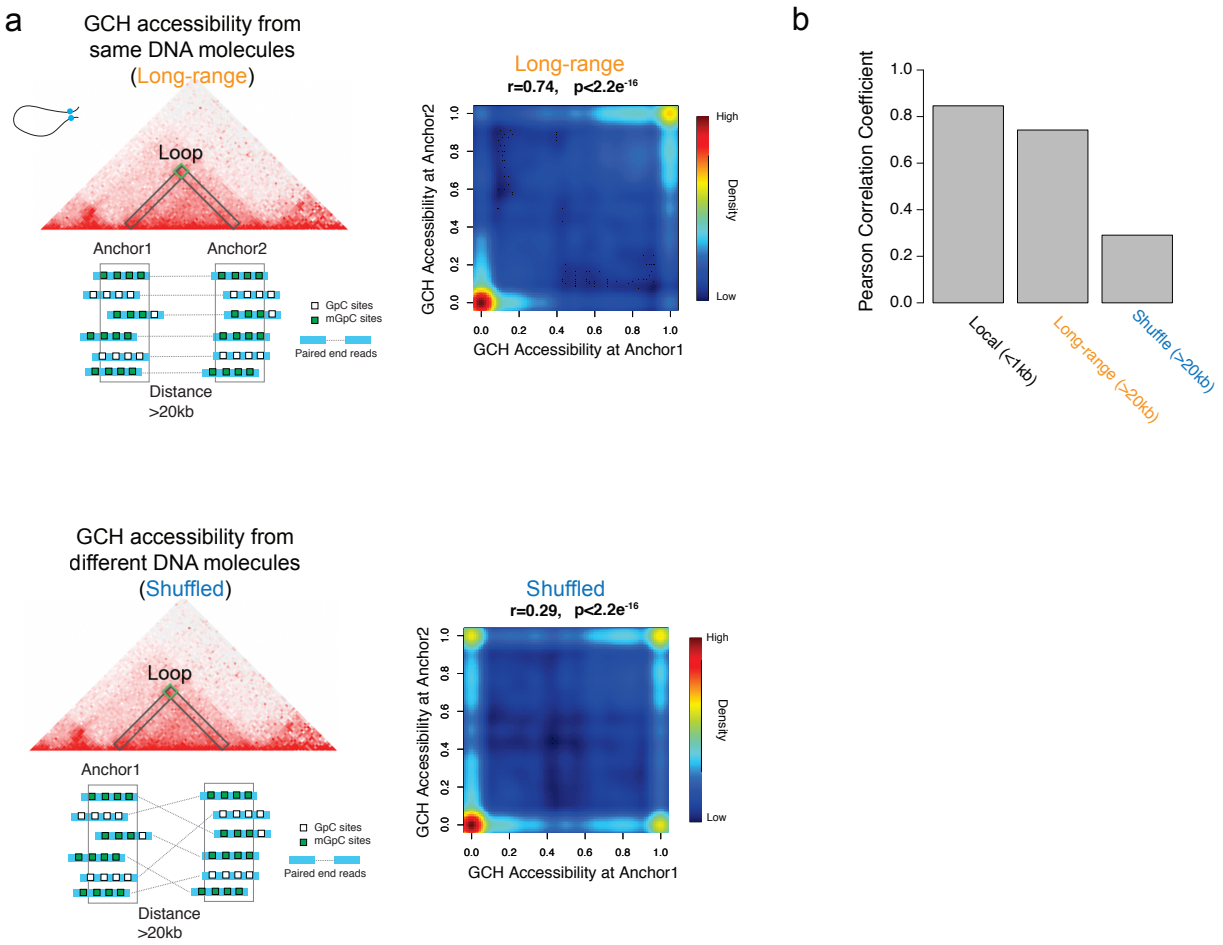

Fig.2. GTAGMe-seq reveals the coordinated GCH methyltransferase footprint at distal regions in spatial proximity. (a). The coordinated GCH methyltransferase footprint is much stronger at the read pairs from the same DNA molecules (top panel) than that from different DNA molecules but in the same regions (bottom panel). Each dot at the heatmap represents one read pair within the HiCCUPS anchor regions. (b). Average GCH methyltransferase concordance level across all the HiCCUPS anchor regions (>20kb, orange color), their matched local regions (utilize the read pairs from the same anchor and within 1kb distance, black color), or matched shuffled read pairs from the same HiCCUPS anchor regions (>20kb, blue color).

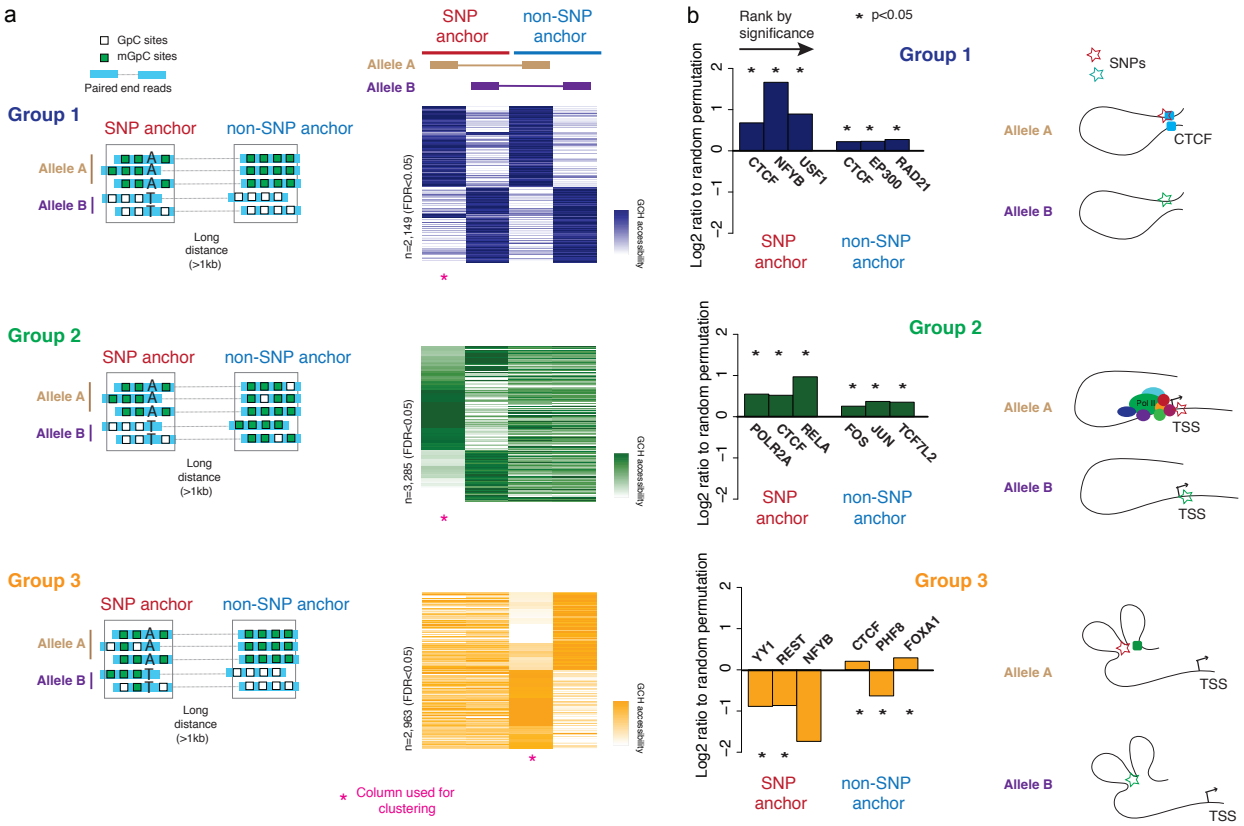

Fig.3. GTAGMe-seq reveals the long-range allele-specific GCH methyltransferase footprint. (a). Scheme and the heatmap representations of the three groups' long-range GCH methyltransferase footprint. The distance between SNP anchors and non-SNP anchors should be more than 1kb. Group 1 (purple) is the group of loci with significant differences of GCH methyltransferase footprint between Allele A and Allele B at both SNP anchors (FDR<0.05) and non-SNP anchors (FDR<0.05). Group 2 (green) is the group of loci with significant differences of GCH methyltransferase footprint between Allele A and Allele B only at SNP anchors (FDR<0.05) but not at non-SNP anchors (FDR>0.95). Group 3 (orange) is the group of loci with significant differences of GCH methyltransferase footprint between Allele A and Allele B only at non-SNP anchors (FDR<0.05) but not at SNP anchors (FDR>0.95). The rows of the heatmap are clustered by columns marked with the magenta star (*). (b). The log2 enrichment of TF binding sites at SNP anchors and non-SNP anchors. The log2 ratio enrichment level is calculated by the ratio with the

498 matched random intervals in the same chromosomes. The random permutation is repeated ten

499 times. Only the top 3 enriched TF binding sites are shown.

500