Dynamic methylome modification is associated with mutational signatures in aging and the etiology of disease

Gayatri Kumar[1&2], Kashyap Krishnasamy[1&2], Naseer Pasha[1&2], Naveenkumar Nagarajan[1&2], Bhavika Mam[1&2], Mahima Kishinani[1], Vedanth Vohra[1], Renuka Jain[1&2], Villoo Morawala Patell[1,2,3]*

[1]Avesthagen Limited, Bangalore, India

[2]The Avestagenome Project® International Pvt Ltd, Bangalore, Karnataka, India-560005

[3]AGENOME LLC, USA

[*]Corresponding Author:

Dr.Villoo Morawala Patell,

Avesthagen Limited

Yolee Grande, 2nd Floor,

14, Pottery Road,

Bangalore, 560005, Karnataka, India

Email: villoo@avesthagen.com

Abstract

Reversible epigenetic changes within the loci of genes that regulate critical cell processes have recently emerged as important biomarkers of disease pathology. It is then natural to consider the consequences for population health risk of such epigenetic changes during the aging process. Specifically, the interplay between dynamic methylation changes that accompany aging and mutations that accrue in an individual's genome over time need further investigation. The current study investigated the role of dynamic methylation acting together with gene variants in an individual over time to gain insight into the evolving epigenome–genome interplay that affects biochemical pathways controlling physiological processes during aging.

We completed a whole-genome methylation and variant analysis in a non-smoking Zoroastrian-Parsi individual, collecting two samples, 12 years apart (at 53 and 65 years respectively) (ZPMetG-Hv2a-1A (old, t_0), ZPMetG-Hv2a-1B (recent, t_0+12)) and analysing them using a GridION Nanopore sequencer at 13X genome coverage overall. We further identified the single nucleotide variants (SNVs) and indels in known CpG islands by employing the Genome Analysis Tool Kit (GATK) and MuTect2 variant-caller pipeline with the GRCh37 (patch 13) human genome as reference.

We found 5258 disease-relevant genes that had been differentially methylated in this individual over 12 years. Employing the GATK pipeline, we found 24,948 genes, corresponding to 4,58,148 variants, specific to ZPMetG-Hv2a-1B, indicating the presence of variants that had accrued over time. A fraction of the gene variants (242/24948) occurred within the CpG regions that were differentially methylated, with 67/247 exactly coincident with a CpG site. Our analysis yielded a critical cluster of 10 genes that were each significantly methylated and had variants at the CpG site or the ±4 bp CpG region window. Kyoto Encyclopedia of Genes and Genomes (KEGG) enrichment network analysis as well as Reactome and STRING analysis of gene-specific variants indicated an impact on biological processes regulating the immune system, disease networks implicated in cancer and neurodegenerative diseases, and transcriptional control of processes regulating cellular senescence and longevity. Additional analysis of mutational signatures

51 indicated a majority of C>T transitions followed by T>C transitions in the more recent sample,
52 ZPMetG-Hv2a-1B.

53 Our current study provides additional insight into the aging methylome over time and the
54 interplay between different methylation and gene variants in the etiology of disease.

55

56 Introduction

57 Aging is a complex and time-dependent deterioration of physiological processes. Increased
58 human life expectancy has resulted in higher morbidity rates, as advanced age is the
59 predominant risk factor for several diseases, including cancer, dementia, diabetes, and
60 cardiovascular disease (CVD)[1]. In addition to molecular and cellular factors, such as cellular
61 senescence and telomere attrition, epigenetic changes that govern physiological processes
62 comprise a significant component of the aging process[2]. Methylation signatures specifically
63 influence gene expression without altering the DNA sequence, connecting intrinsic and extrinsic
64 signals. The most common methylation modifications involve the transfer of a methyl (CH_3)
65 group from *S*-adenosyl methionine (SAM) to the fifth position of cytosine nucleotides, forming
66 5-methylcytosine (5mC)[3]. Methylation is a dynamic event, in which reversal occurs at certain
67 sites, while progression with age causes methylation at many CpG sites in intergenic regions,
68 such as transcription start sites (TSSs[4]). The interplay between epigenetic events is manifested
69 in the regulation of major biological processes[5], such as development, differentiation, genomic
70 imprinting, and X chromosome inactivation (XCI).

71 DNA methylation, both heritable and dynamic, exhibits a strong correlation with age and age-
72 related outcomes. Additionally, the epigenome responds to a broad range of environmental
73 factors and is sensitive to environmental influences[6], including exercise, stress[7], diet[8], and sleep
74 patterns. Epigenetic profiling of identical twins[9] indicates that methylation events are
75 independent of genome homology, with global DNA methylation patterns differing between
76 older and younger twins, consistent with an age-dependent progression[10] of epigenetic
77 changes. Global methylation changes over an 11-year span in participants of an Icelandic

78 cohort, as well as age- and tissue-related alterations in certain CpG islands from an array of
79 1413 arbitrarily chosen CpG sites near gene promoters, further corroborate the evidence for
80 dynamic methylation patterns over time[11]. Promoter hypermethylation has been shown to
81 increase mutation rates, suggesting the influence of epigenetics on genetics. These studies
82 support the role of epigenome alterations in aging-associated genetic changes. It has been
83 reported that genes related to various tumors can be silenced by heritable epigenetic events
84 involving chromatin remodeling or DNA methylation (DNAm). In addition to dynamic
85 methylation changes over time, recent studies have also suggested that epigenetic markers,
86 and their maintenance, are controlled by genes and are tightly linked with DNA variants[12] that
87 accrue in an individual genome over time. In fact, the methylation of cytosine to 5mC in the
88 coding regions of genes increases the probability of mutations and consequent spontaneous
89 hydrolytic deamination resulting in C>T transitions.

90 The epigenetic milieu therefore provides a homeostatic mechanism at the molecular level that
91 allows phenotypic malleability in response to the changing internal and external environments.
92 The flexibility and dynamism of the methylome, as monitored through epigenetic profiling,
93 suggests the importance and influence of lifestyle changes in the control of gene regulation.
94 This also makes methylome studies an attractive means to identify epigenetic hotspots and
95 evaluate the impact of lifestyle changes on aging as well as provide potential targets for
96 therapeutic intervention[13,14]. Besides methylation changes, mutational signatures, including
97 somatic variations, can change the preponderance of methylation of CpG islands. CpG–SNP
98 interactions in the promoter regions have been found to be associated with various disorders,
99 including type 2 diabetes[15], breast cancer[16], coronary heart disease[17], and psychosis[13].

100 Therefore, comparison of comprehensive methylation patterns in healthy individuals at
101 multiple time points can determine whether such changes are early indicators of a late-onset
102 chronic disease. Identifying such indicators earlier can improve disease prognosis, with the
103 potential for reversibility. Our current epigenetic study is unique in analysing genome–
104 epigenome interactions from an individual of the dwindling endogamous Zoroastrian Parsi
105 community, which has a higher median life span and, as a result, a higher incidence of aging-
106 associated conditions, including neurodegenerative conditions, such as Parkinson's disease;

107 cancers of the breast, colon, gastrointestinal tract, and prostrate; auto-immune disorders; and
108 rare diseases.

109 To identify age- and disease-associated epigenetic changes in the Parsi population, we
110 performed an intra-individual methylome profile of a Zoroastrian-Parsi individual with
111 Nanopore-based sequencing technology using samples collected 12 years apart. The two
112 samples, ZPMetG-HV2a-1A (old) and ZPMetG-HV2a-1B (recent) (12 years apart) were
113 sequenced using Oxford Nanopore Technology (ONT). Comparing the samples, we found that
114 aging increased global methylation frequencies for the equivalent CpG sites, with an overall
115 increase of hypermethylated regions across genic regions compared with non-genic regions.
116 Some genes (108) were significantly hypermethylated, while 304 genes were significantly
117 hypomethylated. We also identified 10 unique CpG–SNP interactions in the form of variants
118 occurring at a CpG site and across the CpG region that may affect the biological processes
119 encoded in the genic regions. The enrichment in certain genes implicated pathways involved in
120 the immune response, cancer, neurogenerative diseases, and physiological processes regulating
121 cellular proliferation, senescence, aging, and longevity.

122

123 Materials and Methods

124 DNA Extraction

125 High-molecular-weight (HMW) genomic DNA was isolated from the buffy coat of EDTA-treated
126 whole blood using the Qiagen Whole Blood Genomic DNA extraction kit (cat. #69504). The
127 Extracted DNA Qubit™ dsDNA BR assay kit (cat. #Q32850) was used in preparation for the Qubit
128 2.0® Fluorometer (Life Technologies™).

129

130 DNA QC

131 HMW DNA of optimal quality is a prerequisite for Nanopore library preparation. The quality and
132 quantity of the gDNA was estimated using a Nanodrop spectrophotometer and a Qubit

133 fluorometer using the Qubit™ dsDNA BR assay kit (#Q32850) from Life Technologies,
134 respectively.

135 **DNA purification**

136 The DNA was subjected to column purification using the Zymoclean Large Fragment DNA
137 Recovery kit (Zymo Research, USA), followed by fragmentation using g-TUBE device (Covaris,
138 Inc.). The purified DNA samples were used for library preparation.

139

140 **Library preparation**

141 A total of 2 µg from each sample was used for Nanopore library preparation using the
142 Nanopore Ligation Sequencing kit (cat. #SQK-LSK109, Oxford Nanopore Technology, Oxford,
143 UK). Briefly, 2 µg of gDNA from each sample was end-repaired using an NEBnext Ultra II End
144 Repair kit, (New England Biolabs, MA, USA) and purified using 1x AmPure beads (Beckman
145 Coulter, USA). Adapter ligation (AMX) was performed at RT (20 ^{o}C) for 20 min using NEB Quick
146 T4 DNA ligase (New England Biolabs, MA, USA). The reaction mixture was purified using 0.6X
147 AmPure beads (Beckman Coulter, USA), and the sequencing library was eluted in 15 µl of
148 elution buffer provided in the ligation sequencing kit.

149

150 **Sequencing and sequence processing**

151 Sequencing was performed on a GridION X5 sequencer (Oxford Nanopore Technologies) using a SpotON
152 flow cell R9.4 (cat. #FLO-MIN106) as per the manufacturer's recommendation. Nanopore raw reads
153 ("fast5" format) were base called ("fastq5" format) using Guppy v2.3.4 software.

154 Genomic DNA samples were quantified using a Qubit fluorometer. For each sample, 100 ng of DNA was
155 fragmented to an average size of 350 bp by ultrasonication (Covaris ME220 ultrasonicator). DNA
156 sequencing libraries were prepared using dual-index adapters with the TruSeq Nano DNA Library Prep
157 kit (Illumina) as per the manufacturer's protocol. The amplified libraries were checked on a Tape Station
158 (Agilent Technologies) and quantified by real-time PCR using the KAPA Library Quantification kit (Roche)
159 with the QuantStudio-7flex Real-Time PCR system (Thermo Fisher Scientific). Equimolar pools of
160 sequencing libraries were sequenced using S4 flow cells in a Novaseq 6000 sequencer (Illumina) to
161 generate 2 x 150-bp sequencing reads for 30x genome coverage per sample.

162

163

164 **Extracting methylation base-called reads and QC**

165 Sequence information was encoded in signal-level data measured with a Nanopore sequencer.
166 The raw signal data in Fast5 format was then demultiplexed using the Guppy base-caller, and
167 the adapters were removed using Porechop to obtain the base-called reads. The quality of the
168 reads (in Fastq format) following adapter removal was evaluated using Fastqc (Supplementary
169 Figure 1). The adapter-trimmed reads in Fastq format were then mapped to the human
170 reference genome (GRCh37, patch 13, using the Minimap2 alignment program, version 2.17).
171 The aligned reads were then used to identify the methylated sites using Nanopolish software
172 (version 0.13.2). This tool provides a list of reads with the log-likelihood ratios (wherein a
173 positive value is evidence that the cytosine is methylated to 5-mC, and a negative value
174 indicates that an unreacted cytosine is present). From this file, we obtained the consolidated
175 methylation frequency file for both the sequenced samples. Average read lengths of 5.77 kb
176 and 12.15 kb were obtained for the ZPMetG-HV2a-1A (t0, old) and ZPMetG-HV2a-1B (t0+12,
177 recent) respectively.

178

179 **Differential methylation analysis**

180 The reads with methylated CpGs from the methylation frequency files of the two samples (ZPMetG-
181 HV2a-1A and ZPMetG-HV2a-1B) were compared using bedtools (Version.v2.29.2). Reads with
182 corresponding entries available in the two samples were taken further for differential methylation
183 analysis.

184 To the above, methylated reads common to the samples, ZPMetG-HV2a-1A and ZPMetG-HV2a-1B
185 annotations were provided from the GRCh37 GTF file.

186 Annotations were provided in the following manner:

187 - Promoter regions were defined as extending 2 kb upstream of genic regions
188 - TSS for genes on the "+" strand, column 2 and column 3 for genes on the "– "strand
189 - Genic and nongenic regions were extracted based on the tags provided in the third
190 column of the GTF file

191

192

193

194 **Hypomethylated and hypermethylated regions**

195 The porechopped Fastq reads for the two samples were mapped back to the reference genome
196 GRCh37 (patch 13) using the Minimap2 program, and the aligned outputs in the binary
197 alignment map (BAM) were then represented in BED format, which was carried out to identify
198 the regions sequenced in both samples. The unique regions in each sample that were
199 methylated corresponded to hypomethylated or hypermethylated regions. A region was tagged
200 as only "hypermethylated" if there were reads available for the given region in the sample
201 ZPMetG-HV2a-1A but not reported in the Nanopolish-generated methylation frequency file. It
202 was tagged as only "hypomethylated" if it was reported in the Nanopolish-generated
203 methylation frequency file for ZPMetG-HV2a-1A but no such corresponding record in ZPMetG-
204 HV2a-1B for a region sequenced with reads was available. The gene annotations for these
205 regions were obtained in the same manner as discussed in the previous section.

206

207 **Read mapping and variant calling for Illumina sequencing (ZPMetG-HV2a-1A and ZPMetG-**
208 **HV2a-1B)**

209 Single-nucleotide variants (SNVs) and indels were called using four different pipelines through a
210 combination of two read mappers and two variant callers. The GRCh37 (patch 13) human genome was
211 used as the reference genome to map the paired-end reads. The two read mappers used were BWA-
212 MEM (version 0.7.17) and Bowtie2 (version 2.4.1). The variant-calling pipeline was implemented using
213 GATK (version 7.3.1).

214 The GATK pipeline included additional read- and variant-processing steps, such as duplicate removal
215 using Picard tools, base quality score recalibration, indel realignment, and genotyping and variant
216 quality score recalibration using GATK, all performed according to GATK best-practice
217 recommendations[18]. The MuTect2 tool[19] was employed for identifying putative somatic variants,
218 following which Snpeff (build 2017-11-24) was used to annotate the variants and make functional
219 predictions. As described later in the Results section, variants identified using the BWA + GATK pipeline
220 were used for all downstream analyses. Variants in the intersection of all four pipelines (two read

221 mappers and two variant callers), in which the intersection is defined as variant calls for which the
222 chromosome, position, reference, and alternate fields in the VCF files were identical, were confidently
223 identified.

224

225 **Pathway mapping of genes to assess physiological implications**

226 A compendium of genes obtained from gene set enrichment analysis (GSEA) were extracted
227 that corresponded to tumor suppressors and oncogenes. Housekeeping genes were collated
228 from (source). Genes associated with aging-regulating pathways were obtained from KEGG and
229 other published work(s). KEGG pathway genes associated with different neurological and
230 physiological disorders were extracted. For this set of genes, epigenetic changes between the
231 two samples were examined. These epigenetic changes were measured in terms of differences
232 in methylation frequencies for ZPMetG-HV2a-1A and ZPMetG-HV2a-1B. The frequency of methylation
233 is given by the ratio of the number of methylated reads in a gene region ($\mathbf{m_r}$) and the total number of
234 reads detected in the gene region ($\mathbf{T_r}$).

235 $$\mathbf{f_{meth} = m_r / T_r}$$

236

237 The difference in methylation frequency ($\mathbf{\Delta f_{meth}}$) between the ZPMetG-HV2a-1B (t_0+12) and
238 ZPMetG-HV2a-1B (t_0 years) was obtained by subtracting the methylation frequency for t_0 from
239 that of t_0+12.

240 $$\mathbf{\Delta f_{meth} = (\Delta f_{meth})_{t0+12} - (\Delta f_{meth})_{t0}}$$

241

242 The degree of methylation was primarily classified as hypomethylated, hemi-hypomethylated,
243 hemi-hypermethylated, and hypermethylated based on the int $\mathbf{\Delta f_{meth}}$ intervals. Here, the term
244 "background" was also defined as corresponding to low-confidence scores of CpG methylation,
245 namely, $-0.2 \leq \mathbf{\Delta f_{meth}} < +0.2$. Hypomethylation was defined as a gradient of hypomethylation
246 ($\mathbf{\Delta f_{meth}} < 0.6$) and hemi-hypomethylation ($-0.6 \leq \mathbf{\Delta f_{meth}} < -0.2$). Hypermethylation was defined
247 as a gradient of hypermethylation ($\mathbf{\Delta f_{meth}}$ > than 0.6) and hemi-hypermethylation ($0.2 \leq \mathbf{\Delta f_{meth}} <$
248 0.6). The weighted difference in methylation frequency ($\mathbf{\Delta f_{meth_weighted}}$) for each read within a
249 gene region was obtained by factoring in the number of CpG islands ($\mathbf{cpg_{num}}$) detected for that

250 gene region. Thus, the weighted methylated frequency was obtained as a product of the
251 difference in methylation frequency and the number of CpG reads in the gene region.

252 $$\mathbf{\Delta f_{meth_weighted} = \Delta f_{meth} * cpg_{num}}$$

253

254 The normalized weighted methylation frequency score ($\mathbf{F_{meth_norm}}$) per degree of methylation
255 was obtained by normalizing the summed $\mathbf{\Delta f_{meth_weighted}}$ score with the total number of reads
256 ($\mathbf{T_r}$).

257

258 $$\mathbf{F_{meth_norm} = \Sigma\,(\Delta f_{meth_weighted})_{degree} / Tr}$$

259 The normalized weighted difference in methylation frequency score was assigned to a specific
260 gene and used in downstream analysis.

261

262 **Mutational signature analysis**

263 The non-negative matrix factorization (NMF) method described by Alexandrov et.al[20]., was used
264 to detect mutational signatures for disease associated KEGG gene sets in ZPMet-Hv2a-1A and
265 ZPMet-Hv2a-1A WGS. To determine the contribution of each signature in a sample, we used
266 flexible functions in the MutationalPatterns[21] routines, available in the R package, to assign the
267 mutations in each sample to the signatures identified by NMF.

268

269 ## Results

270

271 **Analysis of individual methylome modifications over time**

272 To account for differences in coverage depth, we extracted the nucleotide ranges for which
273 base-called entries were present in the methylation frequency files for both samples. This
274 enabled identifying any distinctive increase or decrease in methylation between the two
275 temporally separated samples. These comparable nucleotide ranges extracted from the
276 methylation base-called regions accounted for 18,758,642 regions, which covered 98% and 84%
277 of reported regions in fastq files of ZPMetG-HV2a-1A (old) and ZPMetG-HV2a-1B (recent)
278 respectively. The overall analysis pipeline is detailed in Figure 1A. We examined the overall

279 distribution in methylation frequency between ZPMetG-HV2a-1A (old) and ZPMetG-HV2a-1B
280 (recent) We applied a nonparametric statistical test (Mann–Whitney U test) and observed a
281 statistically significant difference between the two sample distributions (Figure 1B). The
282 chromosome-wise distribution of methylated frequencies in CpG dinucleotides, which was also
283 examined to discern changes in methylation between the samples, showed similar profiles for
284 genic and intergenic regions. Notable exceptions were observed for chromosomes 21, 22, and X
285 for genic regions (Supplementary Figure 2). The median of the distribution is given for each
286 chromosome within the entire frequency distribution (Figure 1B).
287

288 **Functional motif-based characterization of CpG's common to ZPMetG-Hv2a-1 and ZPMetG-**
289 **Hv2a-1**

290 The methylation base-called regions compared between ZPMetG-HV2a-1A (old) and ZPMetG-
291 HV2a-1B (recent) were mapped to GRCH37 (human reference genome from NCBI). Based on
292 the tag "gene", we extracted putative genic regions, while other regions (excluding even CDS,
293 mRNA, and tRNA) were extracted to examine the non-genic regions (Figure 2A). A majority of
294 the CpGs were found in the genic and gene promoter regions (Figure 2B). The distribution
295 across chromosomes for genic and non-genic regions in the two samples indicated that a
296 majority of the CpGs occurred within genic regions and were rarest within TSS regions. The
297 distribution of CpGs was similar across all chromosomes for the genic regions compared with
298 the distribution for TSS, non-genic, and genic promoter regions, which displayed a staggered
299 distribution across chromosomes (Figure 2C).

300

301 **Prioritization of gene families**

302 Characterization of the methylated regions indicated that 26,019 genes are found in the regions
303 methylated in ZPMetG-HV2a-1B (recent) compared to ZPMetG-HV2a-1A (old) and selective
304 filtering and prioritization of the 26,019 genes (normalized score) based on a curated dataset of
305 3050 gene families from the GSEA database[22], 168 from the Epigenetic Modifier Gene
306 database[23], 2833 housekeeping genes from HRT Atlas v1.0[24], and 3527 pathway- and disease-
307 specific genes from the KEGG database[25] yielded 5214 unique genes. Followers of the

308 Zoroastrian faith consider smoking a taboo, resulting in generations of the community that
309 have refrained from smoking. To address the epigenetic changes associated with smoking-
310 related genes and their relevance to the non-smoking Parsi community, we also curated a list of
311 44 documented genes that were differentially methylated in smokers. Taken together, our total
312 gene list was comprised of 5258 genes (Supplementary Table.2).

313 **Hierarchical clustering of differentially methylated regions in ZPMetG-Hv2a-1B compared to**
314 **ZPMetG-Hv2a-1A**

315 We performed hierarchical clustering of the differentially methylated genes by first obtaining
316 the frequency of CpG methylation across the entire gene length. Based on the weighted score
317 for CpG methylation (Materials and Methods), the CpG frequencies across genes were classified
318 as hypo-, hemi-hypo-, hemi-hyper-, and hypermethylated regions. The weighted score common
319 across all genes was defined as the background, which served as an internal control.

320 The hierarchical clustering of row-wise Z score of common differentially methylated gene
321 regions (Supplementary Table 3) in ZPMetG-Hv2a-1B identified a cluster of 103 genes that was
322 significantly hypermethylated and 307 genes that were hypomethylated (Figure 3A,
323 Supplementary Table 7). GSEA-based enrichment indicated an enrichment in genes coding for
324 transcription factors and protein kinases in the hyper- and hypomethylated clusters. Among the
325 hyper-methylated genes, homeodomain proteins were highly represented, while among the
326 hypomethylated genes, cell differentiation markers, tumor suppressors, and translocated
327 cancer genes predominated by comparison (Figure 3B, Supplementary Figure 3A)

328 PANTHER-based gene annotation classified these genes based on their molecular and protein
329 functions (Supplementary Figure 3B). Hypermethylated and hypomethylated clusters varied in
330 their functional classification with the presence of genes that function to increase transcription
331 regulator activity (Supplementary Figure 3B).

332

333 **Characterization of variants unique to ZPMetG-Hv2a-1B compared with ZPMetG-Hv2a-1A**

334 While dynamic methylation events constitute a major mechanism of epigenetic modification,
335 variants in the form of SNPs, in addition to dynamic methylation events, are known to be

336 critical to epigenome function through the modification of genomic information. To this end,
337 we employed the GATK variant characterization pipeline to identify variants specific to ZPMetG-
338 Hv2a-1B viz., which are variants that have accumulated in the individual because of aging. We
339 identified 4,58,148 variants corresponding to 24,948 unique genes.

340 We next sought to identify genes common to both dynamic methylated regions (CpGs) and
341 variants specific toZPMetG-Hv2a-1B. We found a direct correlation between gene length, CpG
342 count, and variant count. This association is specifically significant for the correlation between
343 variants and CpG (Pearson correlation coefficient, r=0.86, Figure 4A, Supplementary Figure 4A,
344 B). Analysis of the 5258 ZPMetG-Hv2a-1B-specific differentially methylated genes post
345 disease/GSEA prioritization filter for somatic variants yielded 4409 differentially methylated
346 genes (Supplementary Table 5) that are specific to ZPMetG-HV2a-1B (recent) and that harbour
347 somatic variants (Supplementary Figure 4C).

348 The location of variants with respect to their occurrence in CpG regions is crucial to the
349 resulting effect on epigenomic imprinting in the form of methylation changes. Variants that
350 occur exactly at the CpG site may modulate the resulting methylation, thereby effecting
351 methylation changes. In this context, we classified ZPMetG-HV2a-1B variants based on their
352 locus of incidence within the CpG region. We classified variant occurrence as those that occur
353 exactly within the CpG region and variants that occur within a distance window of 1–4 bp on
354 either side of the CpG region (±4 bp). Our analysis indicated 242 variants classified as "modifier"
355 variants occurring within the CpG region of 177 differentially methylated genes specific to
356 ZPMetG-Hv2a-1B (Figure 4B, Supplementary Table 6). A fraction of variants (67/242 or 27.68%)
357 occurred exactly at the CpG site and 175/242 variants (72%) occurred in the CpG region at a
358 locus window of ±4 bp (Fig. 4B). A vast majority of classified variants occurred on Chr. 1, with
359 most variants occurring within a ±4-bp window across all chromosomes except for Chr. 4, in
360 which more variants are observed exactly at the CpG site compared with the window size
361 (Supplementary Figure 5A). The number of variants within genes varied from 1–7 variants
362 (Supplementary Figure 5B). Characterization of the pathway association and interaction
363 networks differed between the genes harbouring variants at the CpG site compared with genes
364 with variants in the CpG region (Supplementary Figures 6 and 7).

365

366

367

368 **Interplay of the genome and epigenome in the context of aging and disease etiology**

369 The presence of genome variants can affect epigenetic modification. We therefore sought to
370 understand the biological processes of the genes in our study that carried both methylation
371 changes and harbored variants in the CpG regions. To this end, we compared 412 genes with
372 significant hypermethylation (108 genes) and hypomethylation (304 genes) changes
373 (Supplementary Table 7) for gene-specific variants within the CpG regions. Our analysis yielded
374 a critical cluster of 10 genes that were significantly methylated (hyper- or hypo-) and have
375 variants at the CpG site or within the ±4 bp CpG-region window (Figure 5A, Supplementary
376 Table 8). Only two genes (*CASP8* and *PCGF3*) were significantly hypomethylated and carried
377 variants at the CpG site, whereas both significantly hyper/hypo-methylated genes carried
378 variants in the CpG window region (3 hypermethylated genes and 6 hypomethylated genes).
379 The vast majority of genes were critical housekeeping genes, followed by transcription factors,
380 protein kinases, and genes implicated in neurodegenerative diseases (Figure 5B).

381 To identify the relevance of these 10 genes in pathways related to disease and homeostasis, we
382 clustered the 10 genes with 1000 genes that harboured variants and CpGs (Supplementary
383 Table 9, based on the normalized cumulative CpG.Variant score). Using Network analyst[26]-
384 based clustering and a KEGG-based disease enrichment module showed that the majority of the
385 clustered genes were implicated in cell proliferative pathways regulating cancer; cancer
386 subtypes; neurodegenerative diseases, such as Parkinson's disease, Alzheimer's disease,
387 Huntington's disease; and pathways implicated in cellular senescence and longevity-regulating
388 gene clusters (Figure 5C, Supplementary Table 10). Reactome analysis of the biological diversity
389 of gene function showed the association of the clustered genes with pathways involved in the
390 immune system, DNA repair, DNA transcriptional activity, the cell cycle, and disease-related
391 pathways (Figure.5D).

392

393

394

395 **Mutational signatures associated with aging in the ZPMetG-Hv2a-1B across all genes and**
396 **CpG-specific regions of functionally relevant genes**

397 We next proceeded to categorise mutational signatures that are characteristic combinations of
398 mutation types arising from specific mutagenesis processes that involve alterations related to
399 DNA replication, repair, and environmental exposure. Our analysis of showed the presence of
400 110,616 mutations, with a majority associated with C>T transitions, followed by T>C transitions.
401 The same analysis of 4409 prioritized genes that showed differential methylation and variants
402 specific to indicate a similar signature for C>T transitions, followed by an increase in C>A and
403 T>A mutations, while T>C mutations were reduced compared with the cumulative variants in
404 ZPMetG-Hv2a-1B (Figure 6A). Further categorization of the approach using gene sets that
405 consist of epigenetic modifiers, disease-specific lists, and smoking-related genes (sensitive to
406 environmental exposure) showed a high prevalence of C>T transitions, especially at the CpG
407 sites for gene sets regulating Alzheimer's and Parkinson's disease pathways, while smoking-
408 associated genes had a decreased incidence of C>T transitions at CpG sites but an increase in
409 overall C>T transitions (Figure 6B). Analysis of the complete 96 mutational signatures indicated
410 the prevalence of signatures for Parkinson's, Alzheimer's, and Huntington's diseases, while
411 mutational signatures for smoking-associated genes were not highly represented (Figure 6C).

412

413 **Discussion**

414 Aging is an important contributor to chronic diseases. Epigenetic modifications in the form of
415 methylation of CpG dinucleotides play a crucial role in regulating physiological processes, which
416 vary as age progresses. Differential methylation at CpG sites between younger and older
417 subjects has been shown to be associated with genes, including MTOR,
418 ULK1, ADCY6, IGF1R, CREB5, and RELA, that are linked to metabolic traits[27], and CREB5, RELA,

419 and ULK1 have been statistically associated with age. We found 5258 genes to be significantly
420 altered in terms of methylation differences over time in an individual assessed after 12 years.
421 DNAm levels of several CpG sites located within genes involved in longevity-regulating
422 pathways can be examined for DNAm in metabolic alterations associated with age progression.

423 Along with epigenetic modifications, variants at the CpG sites have been shown to adversely
424 affect gene function and activity, especially for disease-associated traits[28]. Many studies have
425 reported that CpG–SNP interactions are associated with different diseases, such as type 2
426 diabetes, breast cancer, coronary heart disease, and psychosis, that show a clear interaction
427 between genetic (SNPs) and epigenetic (DNA methylation) regulation. The introduction or
428 removal of CpG dinucleotides (possible sites of DNA methylation associated with the
429 environment) has been suggested as a potential mechanism through which SNPs influence gene
430 transcription and expression via epigenetics. Our analysis showed the presence of 242 variants,
431 classified as "modifier" variants, occurring within the CpG region of 177 differentially
432 methylated genes specific to ZPMetG-Hv2a-1B. These are genes involved in longevity-regulating
433 pathways, cellular proliferation, and transcriptional regulation.

434 Recent reports using tissue-specific methylation data describe a strong association between
435 C>T mutations and methylation at CpG dinucleotides in many cancer types, driving patterns of
436 mutation formation throughout the genome. In our analysis, the variants occurring at the CpG
437 sites were C>T transitions that occur within two genes, *CASP8* and *PCGF3*. Incidentally, in our
438 analysis *PCGF3* displays transitions not only at the CpG site but also within the CpG region.
439 *Pcgf3/5* mainly functions as a transcriptional activator, driving expression of many genes
440 involved in mesoderm differentiation, and *Pcgf3/5* is essential for regulating global levels of the
441 histone modifier H2AK119ub1 in embryonic stem cells[29]. *CASP8* methylation has been shown to
442 be relevant in cancers and may function as a tumor-suppressor gene in neuroendocrine lung
443 tumors[30]. Our study points to a crucial interplay between CpG methylation levels and variant
444 incidence. It is therefore critical to extend the study to other individuals to understand the
445 diversity of mutational signatures that accrue in CpG regions of functionally important genes to
446 further dissect the role of epigenome–genome interactions in homeostasis, disease, and aging.

447 Mutational signatures associated with aging and cancer subtypes have been studied and
448 indicate a strong correlation towards cytosine deamination at CpG sites, which results in an
449 increase in the frequency of C>T transitions with age of diagnosis, as age allows more time for
450 deamination events, leading to an accumulation of their effects[31]. In line with these
451 observations, we found that C>T transitions constitute most of the mutational changes in
452 ZPMetG-Hv2a-1B, indicating their accumulation over age. Specifically, we found an increase in
453 frequency of the same transitions at CpG sites in genes that participate in Alzheimer's and
454 Parkinson's disease pathways, but their relative contribution is decreased in smoking-
455 associated genes. Further analysis of the same signatures in a larger cohort corresponding to
456 disease and smoking status will be critical in identifying the combinatorial effect of mutational
457 signatures associated with epigenome modifications in the context of aging.

458 By means of personal methylome analysis, our study furthers understanding of the interactions
459 between the epigenome and genome in the form of CpG–gene variant interactions. This
460 approach promises to be helpful in identifying significant genomic loci that reflect the effects of
461 methylation modifications in combination with genetic variations that accrue in an individual,
462 thereby aiding in identifying associations with longevity and associated traits, such as the timing
463 of disease onset.

464
465

466 **References**

467 1. Morgan, A. E., Davies, T. J. & McAuley, M. T. The role of DNA methylation in ageing and
468 cancer. in *Proceedings of the Nutrition Society* vol. 77 412–422 (Cambridge University
469 Press, 2018).

470 2. Aunan, J. R., Watson, M. M., Hagland, H. R. & Søreide, K. Molecular and biological
471 hallmarks of ageing. *British Journal of Surgery* (2016) doi:10.1002/bjs.10053.

472 3. Moore, L. D., Le, T. & Fan, G. DNA methylation and its basic function.
473 *Neuropsychopharmacology* (2013) doi:10.1038/npp.2012.112.

474 4. Marttila, S. *et al.* Ageing-associated changes in the human DNA methylome: Genomic
475 locations and effects on gene expression. *BMC Genomics* (2015) doi:10.1186/s12864-
476 015-1381-z.

477 5. Chaligné, R. & Heard, E. X-chromosome inactivation in development and cancer. *FEBS*
478 *Letters* (2014) doi:10.1016/j.febslet.2014.06.023.

479 6. Vaiserman, A. Epidemiologic evidence for association between adverse environmental
480 exposures in early life and epigenetic variation: a potential link to disease susceptibility?
481 *Clinical Epigenetics* (2015) doi:10.1186/s13148-015-0130-0.

482 7. Burns, S. B., Szyszkowicz, J. K., Luheshi, G. N., Lutz, P. E. & Turecki, G. Plasticity of the
483 epigenome during early-life stress. *Seminars in Cell and Developmental Biology* (2018)
484 doi:10.1016/j.semcdb.2017.09.033.

485 8. Zhang, Y. & Kutateladze, T. G. Diet and the epigenome. *Nature Communications* (2018)
486 doi:10.1038/s41467-018-05778-1.

487 9. Fraga, M. F. *et al.* Epigenetic differences arise during the lifetime of monozygotic twins.
488 *Proc. Natl. Acad. Sci. U. S. A.* (2005) doi:10.1073/pnas.0500398102.

489 10. Talens, R. P. *et al.* Epigenetic variation during the adult lifespan: Cross-sectional and
490 longitudinal data on monozygotic twin pairs. *Aging Cell* (2012) doi:10.1111/j.1474-
491 9726.2012.00835.x.

492 11. Bjornsson, H. T. *et al.* Intra-individual change over time in DNA methylation with familial
493 clustering. *JAMA - J. Am. Med. Assoc.* (2008) doi:10.1001/jama.299.24.2877.

494 12. Bonder, M. J. *et al.* Disease variants alter transcription factor levels and methylation of
495 their binding sites. *Nat. Genet.* (2017) doi:10.1038/ng.3721.

496 13. Van Den Oord, E. J. C. G. *et al.* A whole methylome CpG-SNP association study of
497 psychosis in blood and brain tissue. *Schizophr. Bull.* (2016) doi:10.1093/schbul/sbv182.

498 14. Zhao, S. G. *et al.* The DNA methylation landscape of advanced prostate cancer. *Nat.*
499 *Genet.* (2020) doi:10.1038/s41588-020-0648-8.

500 15. Dayeh, T. A. *et al.* Identification of CpG-SNPs associated with type 2 diabetes and

501 differential DNA methylation in human pancreatic islets. *Diabetologia* (2013)
502 doi:10.1007/s00125-012-2815-7.

503 16. Harlid, S. *et al.* A candidate CpG SNP approach identifies a breast cancer associated ESR1-
504 SNP. *Int. J. Cancer* (2011) doi:10.1002/ijc.25786.

505 17. Chen, X. *et al.* Association of six CpG-SNPs in the inflammation-related genes with
506 coronary heart disease. *Hum. Genomics* (2016) doi:10.1186/s40246-016-0067-1.

507 18. Van der Auwera, G. A. *et al.* GATK Best Practices. *Curr. Protoc. Bioinformatics* (2002).

508 19. Tools, V. D. MuTect2. *GATK Man.* (2017).

509 20. Alexandrov, L. B., Nik-Zainal, S., Wedge, D. C., Campbell, P. J. & Stratton, M. R.
510 Deciphering Signatures of Mutational Processes Operative in Human Cancer. *Cell Rep.*
511 (2013) doi:10.1016/j.celrep.2012.12.008.

512 21. Blokzijl, F., Janssen, R., van Boxtel, R. & Cuppen, E. MutationalPatterns: Comprehensive
513 genome-wide analysis of mutational processes. *Genome Med.* (2018)
514 doi:10.1186/s13073-018-0539-0.

515 22. Subramanian, A. *et al.* Gene set enrichment analysis: A knowledge-based approach for
516 interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U. S. A.* (2005)
517 doi:10.1073/pnas.0506580102.

518 23. Nanda, J. S., Kumar, R. & Raghava, G. P. S. dbEM: A database of epigenetic modifiers
519 curated from cancerous and normal genomes. *Sci. Rep.* (2016) doi:10.1038/srep19340.

520 24. Hounkpe, B. W., Chenou, F., de Lima, F. & De Paula, E. V. HRT Atlas v1.0 database:
521 redefining human and mouse housekeeping genes and candidate reference transcripts
522 by mining massive RNA-seq datasets. *Nucleic Acids Res.* (2020) doi:10.1093/nar/gkaa609.

523 25. Tanabe, M. & Kanehisa, M. Using the KEGG database resource. *Curr. Protoc. Bioinforma.*
524 (2012) doi:10.1002/0471250953.bi0112s38.

525 26. Xia, J., Gill, E. E. & Hancock, R. E. W. NetworkAnalyst for statistical, visual and network-
526 based meta-analysis of gene expression data. *Nat. Protoc.* (2015)
527 doi:10.1038/nprot.2015.052.

528 27. Salas-Pérez, F. *et al.* DNA methylation in genes of longevity-regulating pathways:
529 Association with obesity and metabolic complications. *Aging (Albany. NY).* (2019)
530 doi:10.18632/aging.101882.

531 28. Zhou, D. *et al.* Polymorphisms involving gain or loss of CpG sites are significantly enriched
532 in trait-associated SNPs. *Oncotarget* (2015) doi:10.18632/oncotarget.5650.

533 29. Zhao, W. *et al.* Polycomb group RING finger proteins 3/5 activate transcription via an
534 interaction with the pluripotency factor Tex10 in embryonic stem cells. *J. Biol. Chem.*
535 **292**, 21527–21537 (2017).

536 30. Shivapurkar, N. *et al.* Differential inactivation of caspase-8 in lung cancers. *Cancer Biol.*

Ther. (2002) doi:10.4161/cbt.1.1.45.

31. Alexandrov, L. B. *et al.* Clock-like mutational processes in human somatic cells. *Nat. Genet.* (2015) doi:10.1038/ng.3441.

Figures

Figure 1

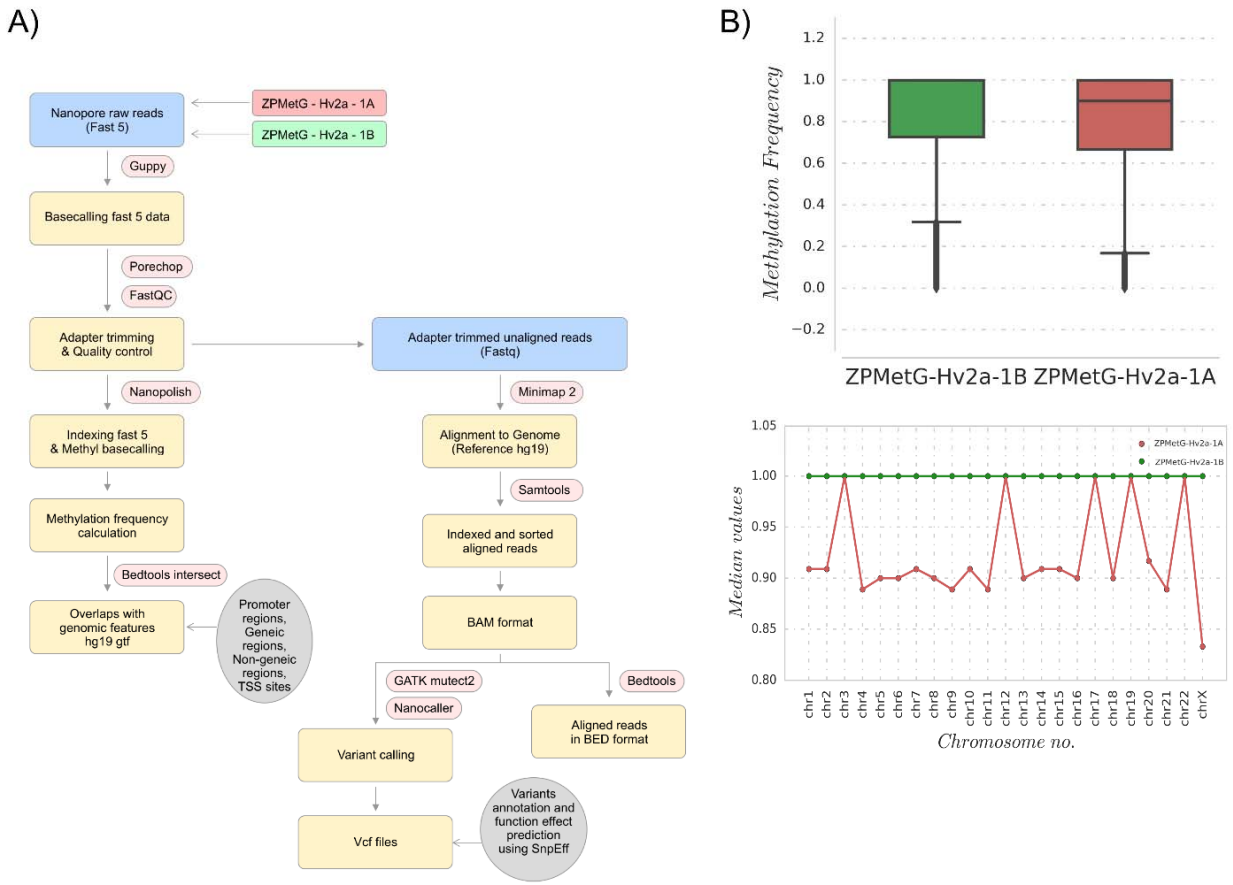
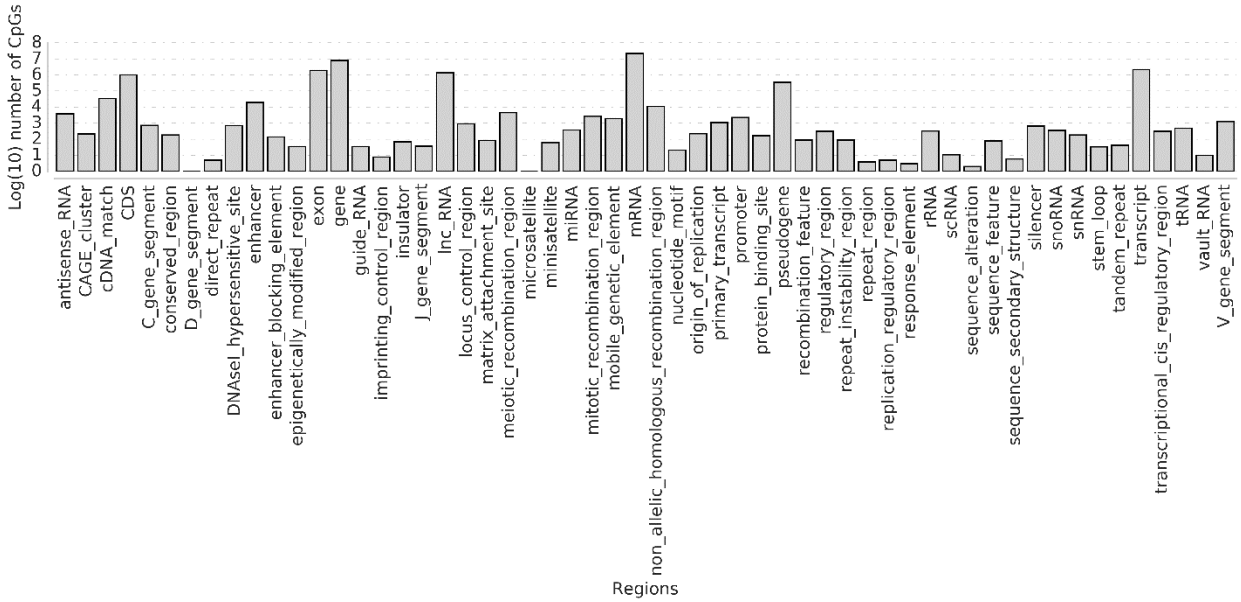


Figure 1. (A) Workflow for the analysis of individual methylome and mutational (SNP) variations for two samples from the same individual collected 12 years apart. **(B, upper)** Methylation frequency distributions for comparing base-called regions, shown for ZPMetG-Hv2a-1B (recent, green) and ZPMetG-Hv2a-1A (old, red). On the right is shown a histogram demonstrating the distribution to be non-Gaussian. **(B, lower)** Median values of the chromosome-wise methylation frequency distribution for ZPMetG-HV2a-1A (old, in red) and ZPMetG-HV2a-1B (recent, in green).

576 **Figure 2**

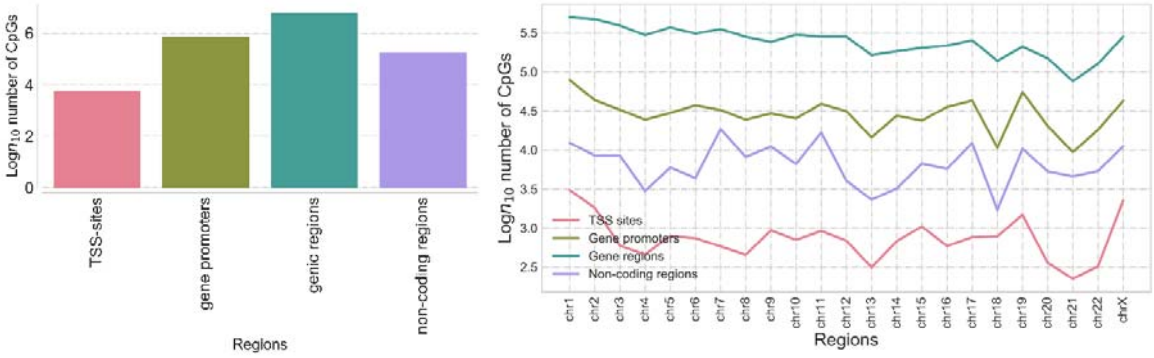
577

578 A)



579

580 B) C)



581

582

583 **Figure 2. (A)** The distribution and classification of methylation imprints (CpG) based on genome
584 architecture and classifiers**. (B)** The classification of common methylation imprints (CpG) based on
585 genomic characterization. **(C)** Chromosome-wise distribution of CpGs common to samples ZPMetG-
586 Hv2a-1 and ZPMetG-Hv2a-2 classified based on their distribution across genomic motifs.

587

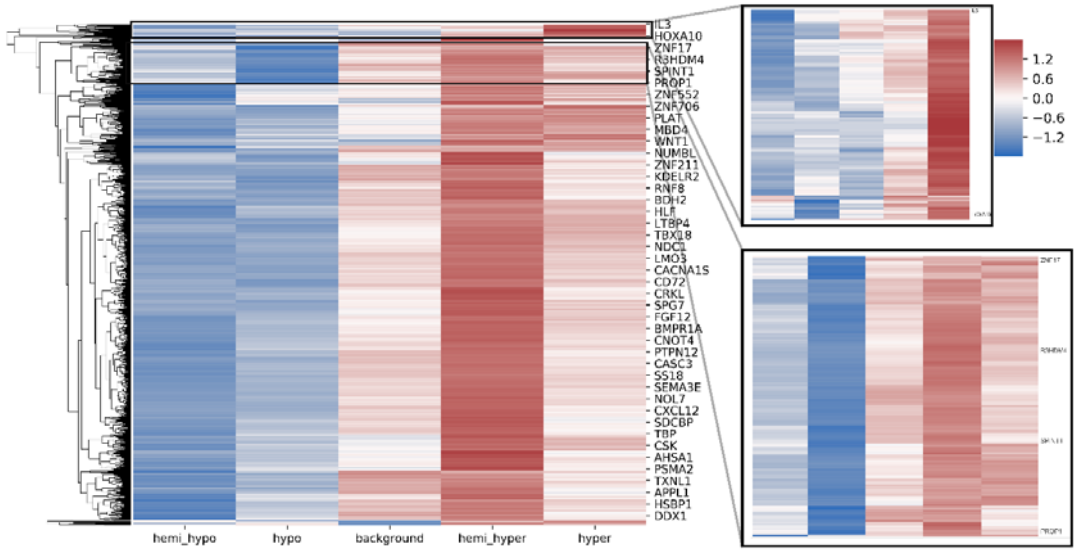
588

589

590 **Figure 3**

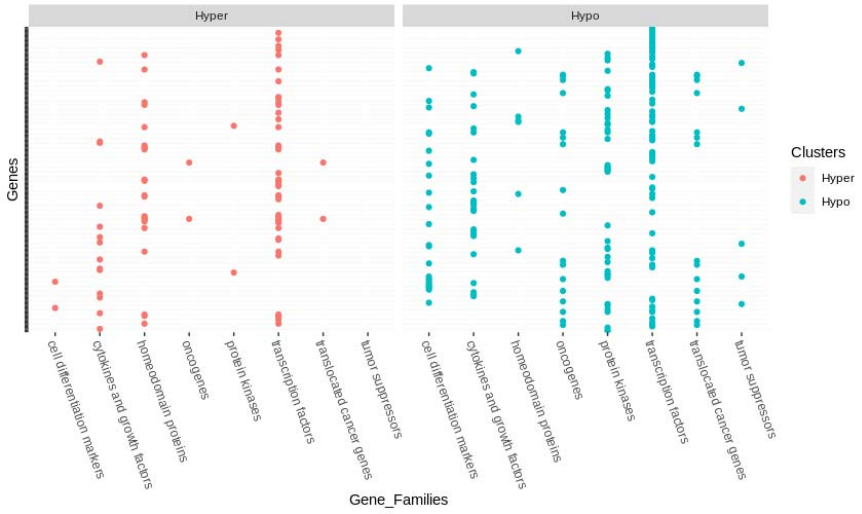
591

592 A)



593

594 B)



595

596

597 **Figure 3. (A)** Hierarchical clustering of row-wise Z scores of common differentially methylated gene
598 regions in sample s2. Clusters of hypermethylated (inlay figure; right, top) and hypomethylated (inlay
599 figure; right, bottom) genes. **(B)** GSEA-based classification of hyper- and hypomethylated clusters.

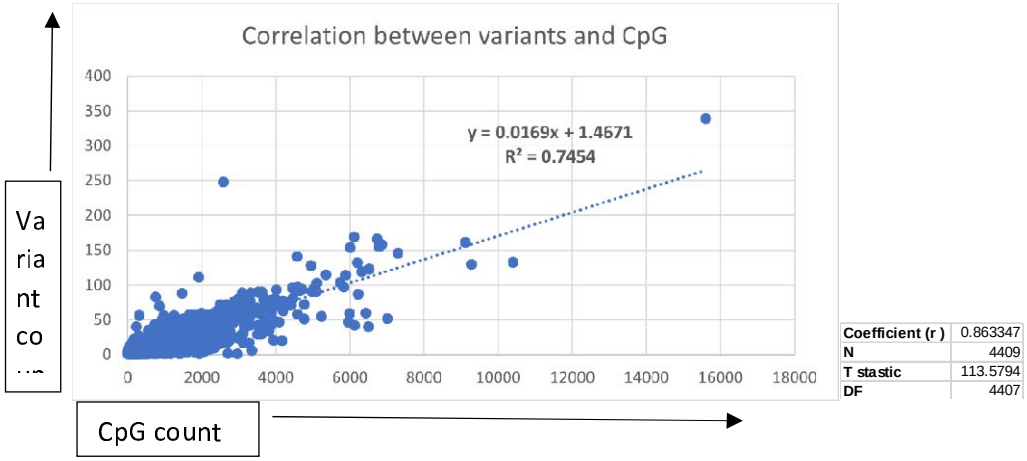
600

601

602 **Figure 4**

603

604 A)

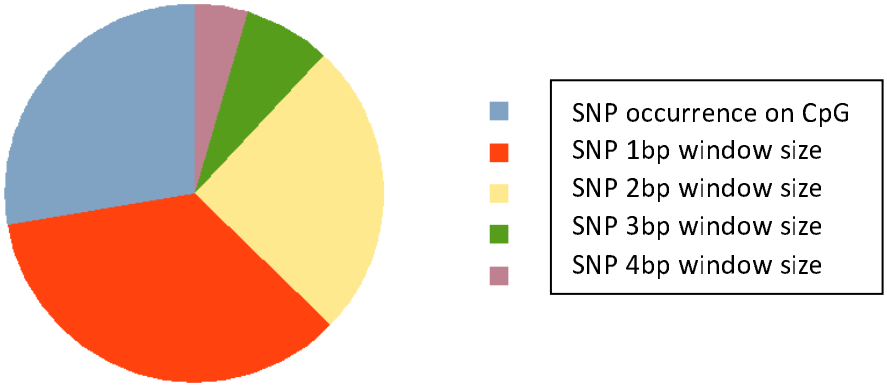


605

606

607

608 B)



609

610

611

612 **Figure 4. A)** Correlation between CpG and variant counts in ZPMetG-Hv2a-1B. The Pearson correlation
613 coefficient, r, is displayed at the side of the plot. **B)** Distribution of ZPMetG-HV2a-1B (recent) specific
614 variants across differentially methylated regions specific to ZPMetG-HV2a-1B (recent).

615

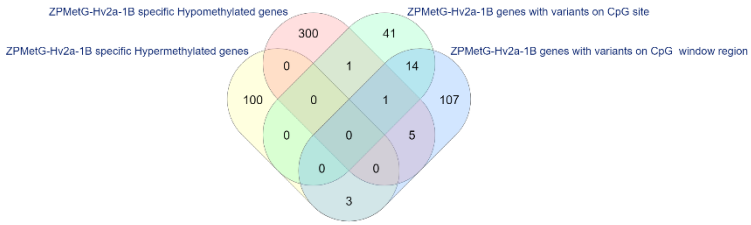
616

617

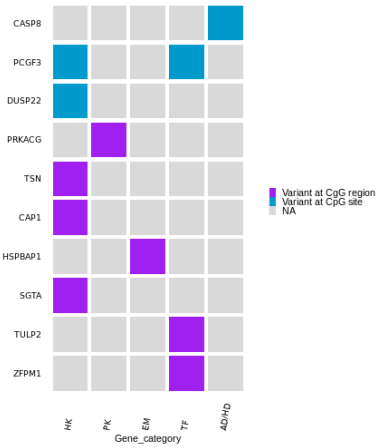
618

Figure 5

A)



B)



C)

D)

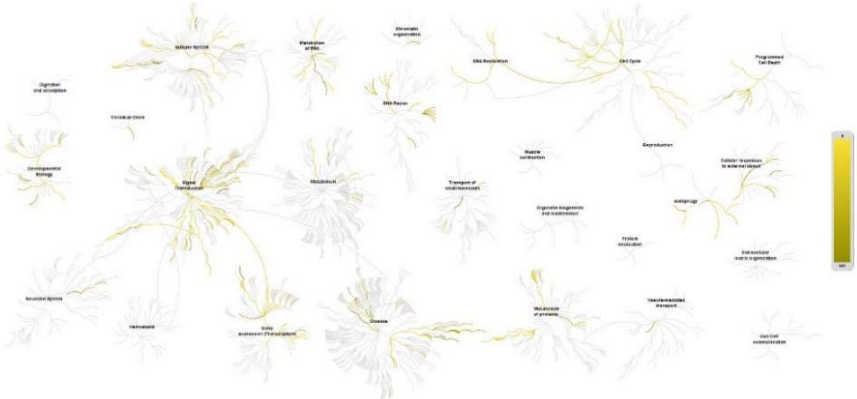
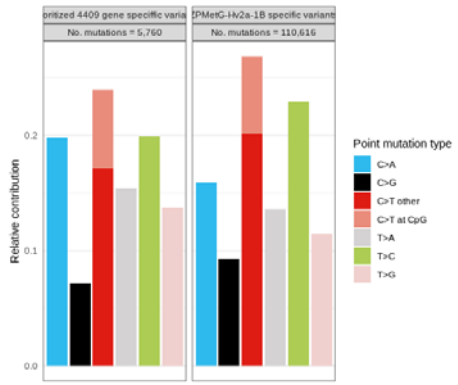


Figure 5. A) Distribution of CpG variants (exact, inexact) across significantly hyper- and hypomethylated genes in ZPMetG-Hv2a-1B **B)** Categorization of 10 critical genes based on biological activity. **C)** Network analysis using a KEGG-based disease-enrichment module for 10 critical and 1000 high-priority variants. **D)** Reactome analysis of biologically relevant pathways of 10 critical and 1000 high-priority variants.

632 **Figure 6**

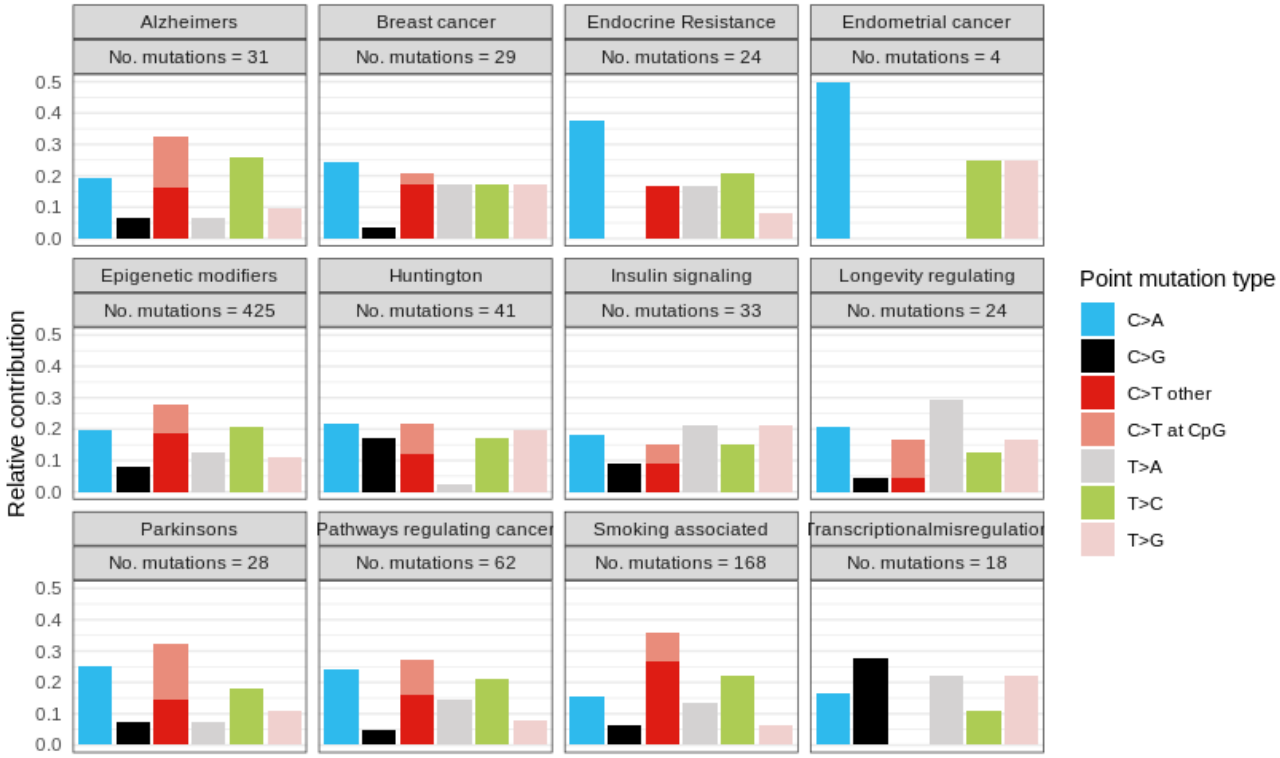
633

634 A)



635

636 B)



637

638

639

640

C)

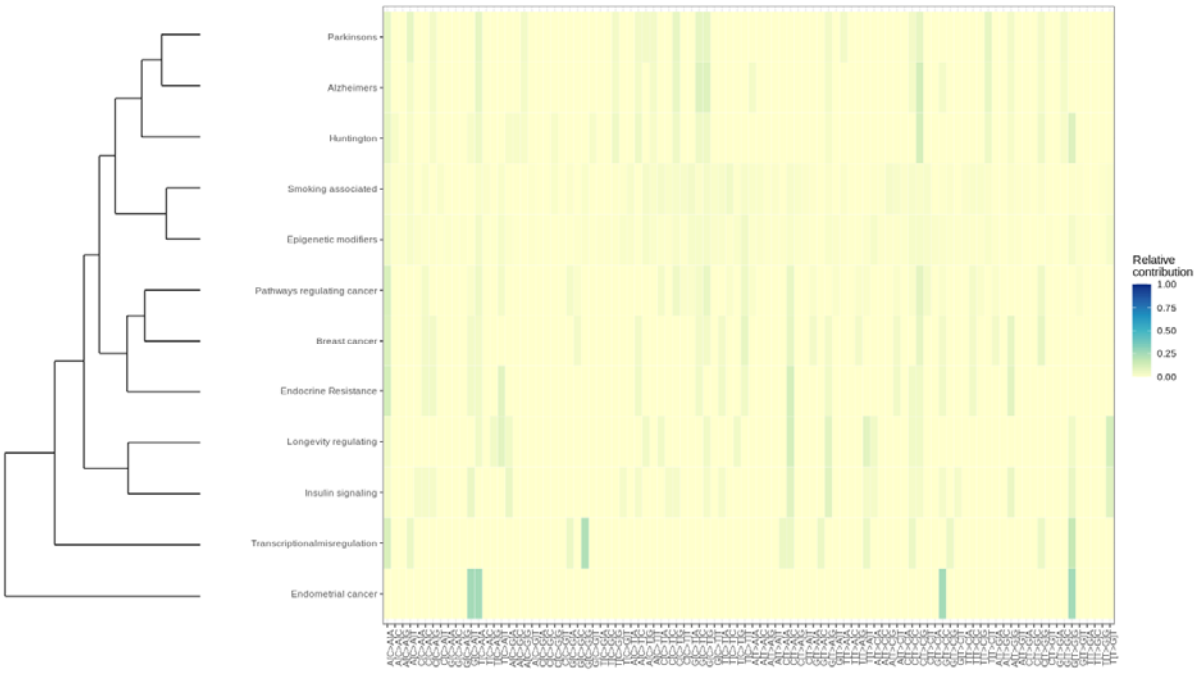


Figure 6. A) Distribution of mutational signatures across the genome and specific to CpG regions in 4409 prioritized genes and cumulative variants across ZPMetG-Hv2a-1B. **B)** Mutational signatures indicating transitions and transversions specific to ZPMetG-Hv2a-1B across different disease categories. **C)** Analysis of the relative contribution of the complete repertoire of mutational signatures (96) specific to ZPMetG-Hv2a-1B.