Dynamic Methylome Modification associated with mutational signatures in ageing and etiology of disease

Gayathri 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], 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

Epigenetic markers and reversible change in the loci of genes regulating critical cell processes, have recently emerged as important biomarkers in the study of disease pathology. The epigenetic changes that accompany ageing in the context of population health risk needs to be explored. Additionally, the interplay between dynamic methylation changes that accompany ageing in relation to mutations that accrue in an individual's genome over time needs further investigation. Our current study captures the role for variants acting in concurrence with dynamic methylation in an individual analysed over time, in essence reflecting the genome-epigenome interplay, affecting biochemical pathways controlling physiological processes. In our current study, we completed the whole genome methylation and variant analysis in one Zoroastrian-Parsi non-smoking individual, collected at an interval of 12 years apart (at 53 and 65 years respectively) (ZPMetG-Hv2a-1A (old, t_0), ZPMetG-Hv2a-1B (recent, t_0+12)) using Grid-ion Nanopore sequencer at 13X genome coverage overall. We further identified the Single Nucleotide Variants (SNVs) and indels in known CpG islands by employing GATK and MuTect2 variant caller pipeline with GRCh37 (patch 13) human genome as the reference.

We found 5258 disease relevant genes differentially methylated across 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, indicative of variants that accrued over time. 242/24948 gene variants occurred within the CpG regions that were differentially methylated with 67/247 exactly occurring on the CpG site. Our analysis yielded a critical cluster of 10 genes which are significantly methylated and have variants at the CpG site or the ±4bp CpG region window. KEGG enrichment network analysis, Reactome and STRING analysis of mutational signatures of gene specific variants indicated an impact in biological process regulating immune system, disease networks implicated in cancer and neurodegenerative diseases and transcriptional control of processes regulating cellular senescence and longevity.

Our current study provides an understanding of the ageing methylome over time through the interplay between differentially methylated genes and variants in the etiology of disease.

52 ## Introduction

53 Aging is a complex and time-dependent deterioration of physiological processes. Increased
54 human life expectancy have also resulted in higher morbidity rates since advanced age is a
55 predominant risk factor for several diseases including cancer, dementia, diabetes, and
56 cardiovascular disease (CVD)[1]. In addition to molecular and cellular factors like cellular
57 senescence, telomere attrition, epigenetic changes that govern such processes comprise a
58 significant component of the aging process[2]. Methylation signatures more specifically influence
59 gene expression without altering the DNA sequence bridging the intrinsic and extrinsic signals.
60 The most common methylation modifications involves the transfer of a methyl (CH_3) group
61 from *S*-adenosyl methionine (SAM) to the fifth position of cytosine nucleotides, forming 5-
62 methylcytosine (5mC)[3]. Methylation is a dynamic event wherein instances of reversal in
63 methylation occur at certain sites, while progression with age causes methylation at many CpG
64 sites in inter-genic regions such as TSS[4] (Transcription start sites), etc. The interplay of epigenetic
65 events manifests in the regulation of major biological processes[5], such as development,
66 differentiation, genomic imprinting and X chromosome inactivation (XCI).
67
68 DNA methylation both heritable and dynamic exhibits strong correlation with age and age-
69 related outcomes. Additionally, the epigenome responds to a broad range of environmental
70 factors and is sensitive to environmental influences[6] including exercise, stress[7], diet[8] and shift
71 work. Epigenetic profiling of identical twins[9] indicate methylation events to be independent of
72 genome homology with global DNA methylation patterns differing in the older compared to
73 younger twins, consistent with an age-dependent progression[10] of epigenetic change. Global
74 methylation changes over an 11-year span in participants of an Icelandic cohort, and age- and
75 tissue-related alterations in some CpG islands from an array of 1413 arbitrarily chosen CpG sites
76 near gene promoters, further corroborate the evidence for dynamic methylation patterns over
77 time[11]. Promoter hyper-methylation has been shown to increase mutation rates suggesting
78 influence of epigenetics over genetics. These studies argue in support for the role of epigenome
79 changes in aging associated changes. It has been noted that genes related to various tumors can
80 be silenced by heritable epigenetic events involving chromatin remodeling or DNAm (DNA

81 methylation). In addition to dynamic methylation changes over time, recent studies have also
82 suggested that epigenetic markers, or their maintenance, are controlled by genes and are and
83 tightly linked with DNA variants[12] that accrue in an individual genome over time. In fact,
84 methylation of cytosine to 5mC in the coding regions of genes increases the proneness to
85 mutations, consequently spontaneous hydrolytic deamination resulting in C to T transitions.

86 The epigenetic milieu therefore provides a homeostatic mechanism at the molecular level which
87 allows phenotypic malleability in response to the changing internal and external environments.
88 The flexibility and dynamic changes in the methylome highlight the importance and influence of
89 lifestyle changes in the control of gene regulation through epigenetic profiling. This also makes
90 methylome studies an attractive source to identify hotspots to evaluate the impact of lifestyle
91 changes on aging as well as provide potential targets for therapeutic intervention[13,14]. Besides
92 methylation changes, mutational signatures that include somatic variations can change the
93 preponderance of methylation of CpG islands. CpG-SNPs in the promoter regions have been
94 found to be associated with various disorders, including type 2 diabetes[15], breast cancer[16],
95 coronary heart disease[17], and psychosis[13].
96
97 Therefore, comparison of comprehensive methylation patterns in healthy individuals at multiple
98 time points can further our understanding of whether such changes can be early indicators of a
99 late onset chronic disease. Identifying such indicators earlier, can improve disease prognosis with
100 potential for reversibility. Our current epigenetic study is unique in analysing the genome-
101 epigenome interactions from an individual of the endogamous dwindling Zoroastrian Parsi
102 community, have higher median life span accompanied by a higher incidence of aging-associated
103 conditions, neurodegenerative conditions like Parkinsons disease, cancers of the breast, colon,
104 gastrointestinal, prostrate, auto-immune disorders, and rare diseases.
105
106 To identify the age and disease associated epigenetic changes in Parsi population, we performed
107 an intra-individual methylome profile of a Zoroastrian Parsi individual with Nanopore-based
108 sequencing technology separated by a time span of 12 years. The two samples from the same
109 non-smoking Zoroastrian Parsi individual, ZPMetG-HV2a-1A (old) and ZPMetG-HV2a-1B (recent)

110 (12 years apart) were sequenced through Oxford Nanopore Technology (ONT). Between the
111 samples, we found that aging increased global methylation frequencies for the equivalent CpG
112 sites, with an overall increase of hyper-methylated regions across genic regions as compared with
113 non-genic regions. 108 genes were significantly hypermethylated and 304 genes were
114 significantly hypomethylated. We also identified 10 unique CpG-SNP interactions in the form of
115 variants occurring at the CpG site and across the CpG region that may affect the biological process
116 of these occurrences in the genic regions. Enrichment of pathways associated with the genes
117 showed the implication of pathways involved in immune responses, pathways implicated in
118 cancer, neurogenerative diseases and physiological processes regulating cellular proliferation,
119 senescence regulating ageing and longevity.

120

121 Materials and Methods

122 **DNA Extraction:**

123 The high molecular weight genomic DNA was isolated from the buffy coat of EDTA-whole blood using
124 the Qiagen whole Blood Genomic DNA extraction kit (#69504). Extracted DNA Qubit™ dsDNA BR assay
125 kit (#Q32850) for Qubit 2.0® Fluorometer (Life Technologies™).

126

127 **DNA QC**

128 High molecular weight (HMW) DNA of optimal quality is a prerequisite for Nanopore library
129 preparation. Quality and quantity of the gDNA was estimated using Nanodrop
130 Spectrophotometer and Qubit fluorometer using Qubit™ dsDNA BR assay kit (#Q32850) from Life
131 Technologies respectively.

132

133 **DNA purification**

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

137 **Library preparation**

138 A total of 2 µg from each sample was used for nanopore library preparation using ligation sequencing kit
139 (SQK-LSK109). Briefly, 2 µg of gDNA from each sample was end-repaired (NEBnext ultra II end repair kit,
140 New England Biolabs, MA, USA) and purified using 1x AmPure beads (Beckmann Coulter, USA). Adapter
141 ligation (AMX) was performed at RT (20 ^{o}C) for 20 minutes using NEB Quick T4 DNA Ligase (New England
142 Biolabs, MA, USA). The reaction mixture was purified using 0.6X AmPure beads (Beckmann Coulter, USA)
143 and sequencing library was eluted in 15 µl of elution buffer provided in the ligation sequencing kit (SQK-
144 LSK109) from Oxford Nanopore Technology (ONT).

145

146 **Sequencing and sequence processing**

147 Sequencing was performed on GridION X5 (Oxford Nanopore Technologies, Oxford, UK) using
148 SpotON flow cell R9.4 (FLO-MIN106) as per manufacturers recommendation. Nanopore raw
149 reads (*'fast5'* format) were base called (*'fastq5'* format) using Guppy v2.3.4 software.
150 Genomic DNA samples were quantified using the Qubit fluorometer. For each sample, 100 ng of
151 DNA was fragmented to an average size of 350 bp by ultrasonication (Covaris ME220
152 ultrasonicator). DNA sequencing libraries were prepared using dual-index adapters with the
153 TruSeq Nano DNA Library Prep kit (Illumina) as per the manufacturer's protocol. The amplified
154 libraries were checked on a Tape Station (Agilent Technologies) and quantified by real-time PCR
155 using the KAPA Library Quantification kit (Roche) with the QuantStudio-7flex Real-Time PCR
156 system (Thermo). Equimolar pools of sequencing libraries were sequenced using S4 flow cells in
157 a Novaseq 6000 sequencer (Illumina) to generate 2 x 150-bp sequencing reads for 30x genome
158 coverage per sample.

159

160 **Extracting methylation basecalls reads and QC**

161 The sequence information is encoded in signal-level data measured by the Nanopore sequencer. The raw
162 signal data in Fast5 format is then demultiplexed using the Guppy basecaller and the adapters are
163 removed using Porechop to obtain the basecalled reads. The quality of the reads (in Fastq format)
164 post adapter removal is evaluated using Fastqc (Supplementary Figure.1). The adapter trimmed
165 reads in Fastq format are then mapped to the Human reference genome (GRCh37 (patch 13)
166 using minimap2 (Version 2.17). The aligned reads obtained are then used to identify the
167 methylated sites using Nanopolish software (Version 0.13.2). This tool provides a list of reads

168 with the log-likelihood ratios (wherein a positive value is evidence that the cytosine is methylated
169 to 5-mC and a negative value indicates a Cytosine is present). From this file, we obtain the
170 consolidated methylation frequency file for both the sequenced samples. Average read length of
171 5.77 kb and 12.15 kb was obtained from ZPMetG-HV2a-1A (t0, old) and ZPMetG-HV2a-1B (t0+12,
172 recent) respectively.

173

174 **Differential methylation analysis**

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

179 To the above, methylated reads common to the samples, ZPMetG-HV2a-1A and ZPMetG-HV2a-
180 1B annotations were provided from the GRCh37 GTF file.

181 Annotations were provided in the following manner –

182 - Promoter regions were identified as 2Kb upstream of genic regions.
183 - TSS (Transcription start sites) for genes on '+' strand, column 2 and column 3 for genes
184 on '- 'strand
185 - Genic and Non-genic regions were extracted based on the tags provided in the third
186 column of the GTF file.

187

188 **Hypomethylated and Hypermethylated regions**

189 The porechopped Fastq reads for the two samples were mapped back to the reference genome
190 GRCh37 (patch 13) using minimap2 and the aligned outputs in BAM (Binary Alignment Map) then
191 represented in bed format. This was carried out to identify the regions sequenced in both the
192 samples. The unique regions in each sample that were methylated would correspond to
193 hypomethylated or hypermethylated regions. A region would be tagged as only Hypermethylated
194 if there were reads available for the given region in the sample ZPMetG-HV2a-1A but not
195 reported in the Nanopolish generated methylation frequency file. It would be tagged as only

196 Hypomethylated if it is reported in the Nanopolish generated methylation frequency file of
197 ZPMetG-HV2a-1A but no such corresponding record in ZPMetG-HV2a-1B for a region sequenced
198 with reads available. The gene annotations for these regions were obtained in the same manner
199 as discussed in the previous section.
200
201 **Read mapping and Variant calling for illumina (ZPMetG-HV2a-1A and ZPMetG-HV2a-1B)**
202 Single Nucleotide Variants (SNVs) and indels were called using four different pipelines through a
203 combination of two read mappers and two variant callers. The GRCh37 (patch 13) human genome
204 was used as the reference genome to map the paired end reads. The two read mappers used
205 were BWA-MEM (Version 0.7.17) and Bowtie2 (version 2.4.1). The variant calling pipeline was
206 implemented using GATK (Version 7.3.1).
207 The GATK pipeline included additional read and variant processing steps such as duplicate
208 removal using Picard tools, base quality score recalibration, indel realignment, and genotyping
209 and variant quality score recalibration using GATK, all used according to GATK best practice
210 recommendations[18]. The MuTect2 tool[19] was employed for identifying the putative somatic
211 variants, following which Snpeff (build 2017-11-24) was used to annotate the variants along with
212 functional prediction. As described later in the Results section, variants identified using the BWA+
213 GATK pipeline were used for all downstream analysis. Variants in the intersection of all four
214 pipelines (two read mappers and two variant callers) were confidently identified, where the
215 intersection is defined as variant calls for which the chromosome, position, reference, and
216 alternate fields in the VCF files were identical.
217
218 **Pathway mapping of genes to assess physiological implications**
219 A compendium of genes from GSEA were extracted corresponding to tumor suppressors, oncogenes, etc.
220 Housekeeping genes were collated from (source). Genes associated with aging regulating pathways were
221 obtained from KEGG and other published work(s). KEGG pathway genes associated with different
222 neurological and physiological disorders were extracted. For this set of genes, the epigenetic changes
223 between the two samples were examined. The epigenetic changes were measured in terms of differences
224 in methylation frequencies for ZPMetG-HV2a-1A and ZPMetG-HV2a-1B**.** The frequency of methylation is

given by the ratio of the number of methylated reads in a gene region (**m_r**) and the total number of reads detected in the gene region (**T_r**).

$$f_{meth} = m_r / T_r$$

The difference in methylation frequency (**Δf_{meth}**) between the subject at ZPMetG-HV2a-1B (t_0+12) and ZPMetG-HV2a-1B (t_0 years) was obtained by subtracting the methylation frequency for t_0 from that of t_0+12.

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

The degree of methylation was primarily classified as 'hypomethylated', 'hemi-hypomethylated', 'hemi-hypermethylated' and 'hypermethylated' based on the int **Δf_{meth}** intervals. Here, a term 'background' was also defined corresponding to low-confidence scores of CpG methylation i.e. **Δf_{meth}**, greater than or equal to -0.2 but lesser than 0.2. Hypomethylation has been defined as a gradient of hypomethylation (**Δf_{meth}** < than 0.6) and hemi-hypomethylation (-0.6 <=**Δf_{meth}** <-0.2). Hypermethylation has been defined as a gradient of hypermethylation (**Δf_{meth}** > than 0.6) and hemi-hypermethylation (0.2 <=**Δf_{meth}**<0.6). The weighted difference in methylation frequency (**$\Delta f_{meth_weighted}$**) for each read within a gene region was obtained by factoring in the number of CpG islands (**cpg_{num}**) detected for that gene region. Thus, the weighted methylated frequency was obtained as a product of the difference in methylation frequency and the number of CpG reads in the gene region.

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

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

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

The normalized weighted difference in methylation frequency score was assigned to a specific gene and taken further for downstream analysis.

254 ## Results

255

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

257 To account for difference in coverage depth, we extracted the nucleotide ranges for which base
258 called entries are present in the methylation frequency files for both samples. This enables
259 identifying any distinctive increase or decrease in methylation between the two temporally
260 separated samples. These comparable nucleotide ranges extracted from the methylation base
261 called regions accounted for 18,758,642 regions which covered 98% and 84% of reported regions
262 in fastq files of ZPMetG-HV2a-1A (old) and ZPMetG-HV2a-1B (recent) respectively. The overall
263 analysis pipeline is elaborated in Figure.1A. We examined the overall distribution in the
264 methylation frequency between ZPMetG-HV2a-1A (old) and ZPMetG-HV2a-1B (recent) We
265 applied a non-parametric statistical test (Mann Whitney U) and observed a statistically significant
266 difference between the two sample distributions (Figure.1B). The chromosome wise distribution
267 of the methylated frequencies in CpG dinucleotides was also examined to discern changes in
268 methylation between the samples showed similar profiles for genic and intergenic regions.
269 Notable exceptions were observed for Chr.21, 22 and X for genic regions (Supplementary Figure
270 2). The median of the distribution is given in Fig.1B for each chromosome while the entire
271 frequency distributions (Figure.1B).

272

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

275

276 The methylation base called regions compared between ZPMetG-HV2a-1A (old) and ZPMetG-
277 HV2a-1B (recent) were mapped to GRCH37 (The human reference genome from NCBI). Based on
278 the tag 'gene', we extracted putative genic regions, while other regions (excluding even CDS
279 (coding sequence), mRNA, tRNA etc) were extracted to examine the non-genic regions (Figure.
280 2A). Majority of the CpGs were found in the genic region and the gene promoter regions (Figure.
281 2B). The distribution across chromosomes for genic and non-genic regions in the two samples
282 indicated majority of the CpGs on genic regions and the lowest for TSS regions. The distribution

283 of CpG's was similar across all chromosomes for the genic regions compared to the distribution
284 for TSS, non-genic and genic promoters that displayed a staggered distribution across
285 chromosomes (Figure. 2C)

286

287 **Prioritization of genes families:**

288

289 Characterization of the methylated regions indicated 26,019 genes that are found in the regions
290 methylated in ZPMetG-HV2a-1B (recent) compared to ZPMetG-HV2a-1A (old) and Selective
291 filtering and prioritization of the 26,019 genes (based normalized score) based on a curated
292 dataset of 3050 gene families from the Gene Set Enrichment Analysis GSEA database[20], 168
293 Epigenetic modifier gene database[21], 2833 Housekeeping genes from HRT Atlas v1.0[22], 3527
294 pathway and disease specific genes from KEGG database[23] yielded 5214 unique genes. It is
295 cultural practice that followers of the Zoroastrian faith consider smoking a taboo, resulting in
296 generations of the community that refrained from smoking. To address the epigenetic changes
297 associated with smoking related genes and its relevance in the non-smoking Parsi community,
298 we also curated a list of 44 documented genes that were differentially methylated in smokers.
299 Taken together, out total gene list comprised of 5258 genes (Supplementary Table.2).

300

301 **Hierarchical clustering of Differentially Methylated Regions in ZPMetG-Hv2a-1B compared to ZPMetG-**
302 **Hv2a-1**

303 We performed hierarchical clustering of the differentially methylated genes by first obtaining the
304 frequency of CpG methylation across the entire gene length. Based on the weighted score for
305 CpG methylation (materials and methods), the CpG frequencies across genes were classified as
306 Hypo, Hemi-Hypo, Hemi-Hyper and Hyper methylated regions. The weighted score common
307 across all genes was classified as background that served as an internal control.

308 The hierarchical clustering of row-wise Z score of common Differentially Methylated Gene
309 regions (Supplementary Table.3) in ZPMetG-Hv2a-1B identified a cluster of 103 genes that were
310 significantly hyper-methylated and 307 genes that were hypo-methylated (Figure. 3A,

311 Supplementary Table.7). GSEA based enrichment indicated an enrichment of genes coding for
312 transcription factors and protein kinases in the hyper and hypomethylated clusters. Among
313 Hypermethylated genes, Homeodomain proteins were highly represented, while among the
314 hypomethylated genes, cell differentiation markers, tumor suppressors and translocated cancer
315 genes dominated in comparison to each other (Figure.3B, Supplementary Figure.3A)

316 PANTHER based gene annotation classified these genes based on their molecular and protein
317 function (Supplementary Figure 3B). Hypermethylated and Hypomethylated clusters varied in
318 their functional classification with a presence of genes that function to increase transcription
319 regulator activity (Supplementary Figure 3B)

320

321 **Characterization of variants unique to ZPMetG-Hv2a-1B compared to ZPMetG-Hv2a-1**

322 While dynamic methylation events constitute a major mechanism of epigenetic modification,
323 variants in the form of SNPs in addition to dynamic methylation events are known to be critical
324 to epigenome function through the modification of genomic information. To this end, we
325 employed the GATK variant characterization pipeline to identify variants specific to ZPMetG-
326 Hv2a-1B viz., variants that have accumulated in the individual because of ageing. We identified
327 4,58,148 variants corresponding to 24,948 unique genes to ZPMetG-HV2a-1B (recent).

328 We next sought to identify genes common to both dynamic methylated regions (CpGs) and
329 variants specific to ZPMetG-Hv2a-1B. We found a direct correlation between gene length, CpG
330 count and variant count. This association is specifically significant to the correlation between
331 variants and CpG (Pearson correlation coefficient, r:0.86) (Figure.4A, Supplementary Figure 4A,
332 B). Analysis of the 5258 ZPMetG-Hv2a-1B specific differentially methylated genes post
333 disease/GSEA prioritisation filter for somatic variants yielded 4409 differentially methylated
334 genes (Supplementary Table.5) that are specific to ZPMetG-HV2a-1B (recent) that harbour
335 somatic variants (Supplementary Figure 4C).

336 The location of variants with respect to their occurrence on the CpG region is crucial to the
337 resultant effect on the epigenomic imprinting in the form of methylation changes. Variants that
338 occur exactly on the CpG site may modulate the resultant methylation, thereby effecting

339 methylation changes. In this context, we classified ZPMetG-HV2a-1B variants based on their loci
340 of incidence on the CpG region. We classified variant occurrence as those that occur exactly on
341 the CpG region and variants that occur at a distance window of 1-4bp either side of the CpG
342 region (±4bp). Our analysis indicated 242 variants classified as "Modifier variants" occurring
343 within the CpG region of 177 Differentially Methylated Genes specific to ZPMetG-Hv2a-1B
344 (Figure. 4B, Supplementary Table.6). 27.68% (67/242 variants) occur exactly at the CpG site and
345 72% (175/242 variants) occur in the CpG region at a loci window of ±4bp (Fig. 4B). Majority of
346 classified variants occur in Chr.1 with most variants occurring at ±4bp window size across all
347 chromosomes except for Chr.4 where more variants are observed exactly at the CpG site
348 compared to the window size (Supplementary Figure 5A). The number of variants within genes
349 varied between 1-7 variants (Supplementary Figure 5B). Characterization of the pathway
350 association and interaction networks differed between the genes harbouring variants at the CpG
351 site compared to genes with variants in the CpG region (Supplementary Figure.6,7)

352 **Interplay of genome and epigenome in the context of ageing and disease etiology**

353 The presence of variants in the genome can affect the epigenetic modification. We therefore
354 sought to understand the biological processes of the genes in our study that carried both
355 methylation changes and harboured variants in the CpG region. To this end, we compared 412
356 genes with significant hyper (108 genes), hypo methylation (304 genes) changes (Supplementary
357 Table.7) for gene specific variants within the CpG region. Our analysis yielded a critical cluster of
358 10 genes which are significantly methylated (hyper/hypo) and have variants at the CpG site or
359 the ±4bp CpG region window (Figure.5A, Supplementary Table 8). Only 2 genes (*CASP8, PCGF3*)
360 significantly hypomethylated carried variants at the CpG site, whereas both significantly
361 hyper/hypo methylated genes carried variants at the CpG window region (3 Hypermethylated
362 genes and 6 hypomethylated genes). A vast majority of genes were critical housekeeping genes,
363 followed by transcription factors, protein kinases and genes implicated in neurodegenerative
364 diseases (Figure. 5B).

365 To identify the relevance of these 10 genes in pathways related to disease and homeostasis, we
366 clustered the 10 genes with 1000 genes that harboured variants and CpG's (Supplementary

367 Table.9, based on normalized cumulative CpG.Variant score). Using Network analyst[24] based
368 clustering using KEGG based disease enrichment module showed that the majority of the
369 clustered genes are implicated in cell proliferative pathways regulating cancer, cancer subtypes,
370 neurodegenerative disease like Parkinsons disease, Alzheimers disease, Huntington disease and
371 pathways implicated in cellular senescence and longevity regulating gene clusters (Figure 5C,
372 Supplementary Table.10). Reactome analysis of the biological diversity of the gene function
373 showed their association to pathways involved in immune system, DNA repair, DNA
374 transcriptional activity, cell cycle and disease related pathways (Figure.5D).

375

376 **Mutational signatures associated with ageing (ZPMetG-Hv2a-1B) across all genes and CpG**
377 **specific region of functionally relevant genes**

378 We next proceeded to categorise mutational signatures that are characteristic combinations of
379 mutation types arising from specific mutagenesis process that involve alterations related to DNA
380 replication, repair and environmental exposure. Our analysis of ZPMetG-Hv2a-1B showed the
381 presence of 110,616 with a majority associated with C>T transitions followed by T>C transitions.
382 Analysis of the same in 4409 prioritized genes that show differential methylation and variants
383 specific to ZPMetG-Hv2a-1B indicated a similar signature for C>T transitions, followed by an
384 increase in C>A and T>A mutations, while T>C mutations were reduced compared to the
385 cumulative variants in ZPMetG-Hv2a-1B (Figure.6A). Further categorisation of the approach using
386 gene sets that consist of epigenetic modifiers, disease specific lists and smoking related
387 (environmental exposure) showed a high prevalence of C>T transitions especially at the CpG site
388 for gene sets regulating Alzheimers and Parkinsons pathways, while smoking associated genes
389 had a decreased incidence of C>T transitions at CpG site but an increase in overall C>T transitions
390 (Figure.6B)

391

392

393

394 **Discussion:**

395 Aging is one of the important contributors to chronic diseases. Epigenetic modifications in the
396 form of methylation of CpG di-nucleotides, play a crucial role in regulating physiological
397 processes which vary as age progresses. Differential methylation at CpG sites between younger
398 and older subjects have been shown associated with genes - MTOR, ULK1, ADCY6, IGF1R, CREB5
399 and RELA linked to metabolic traits[25], and CREB5, RELA, and ULK1 were statistically associated
400 with age. We found 5258 genes significantly altered in terms of methylation differences over
401 time in an individual assessed at a time of 12 years. DNAm levels of several CpG sites located at
402 genes involved in longevity-regulating pathways can be examined for DNAm in metabolic
403 alterations associated with age progression.

404 Along with epigenetic modifications, variants in the CpG sites have been shown to adversely
405 affect gene function and activity, especially in the context of disease associated traits[26]. Many
406 studies have reported that CpG-SNPs are associated with different diseases, such as type 2
407 diabetes, breast cancer, coronary heart disease, and psychosis which show a clear interaction
408 between genetic (SNPs) and epigenetic (DNA methylation) regulation. The introduction or
409 removal of CpG dinucleotides (possible sites of DNA methylation associated with the
410 environment) has been suggested as a potential mechanism through which SNPs can influence
411 gene transcription and expression via epigenetics. Our analysis showed the presence of 242
412 variants classified as "Modifier variants" occurring within the CpG region of 177 Differentially
413 Methylated Genes specific to ZPMetG-Hv2a-1B. These are genes involved in longevity regulating
414 pathways, cellular proliferation, and transcriptional regulation.

415 Recent reports using tissue-specific methylation data describe a strong association between C>T
416 mutations and methylation at CpG dinucleotides in many cancer types, driving patterns of
417 mutation formation throughout the genome. In our analysis, the variants occurring at the CpG
418 sites were C>T transitions, that constitute two genes *CASP8, PCGF3*. Incidentally *PCGF3* in our
419 analysis displays transitions not only at the CpG site but also transitions within the CpG region.
420 *Pcgf3/5* mainly function as transcriptional activators driving expression of many genes involved
421 in mesoderm differentiation and *Pcgf3/5* are essential for regulating global levels of the histone

422 modifier H2AK119ub1 in Embryonic Stem cells[27]. *CASP8* methylation has been shown to be
423 relevant in cancers and may function as a tumor suppressor gene in neuroendocrine lung
424 tumors[28]. Our study points to a crucial interplay between CpG methylation levels and variant
425 incidence. It is therefore critical to extend the study to other individuals to understand the
426 diversity of mutational signatures that accrue in CpG regions of functionally important genes to
427 further dissect the epigenome-genome interactions in the context of homeostasis, disease and
428 ageing.

429 Mutational signatures associated to ageing and cancer subtypes have been studied and indicate
430 a strong correlation towards cytosine deamination at CpG sites, that result in an increase of C>T
431 transitions with age of diagnosis, as age allows more time for deamination events leading to
432 accumulation of their effects[29]. In line with these observations, we find that C>T transitions
433 constitute most of the mutational changes in ZPMetG-Hv2a-1B, indicative of their accumulation
434 over age. In specific, we find an increase of the same transitions at CpG sites in genes that
435 participate in Alzheimers and Parkinsons disease pathways but their relative contribution is
436 decreased in smoking associated genes. Further analysis of the same signatures in different
437 cohorts corresponding to diseases and smoking status would be critical in identifying the
438 combinatorial effect of mutational signatures associated with epigenome modifications in the
439 context of ageing.

440 Our study therefore adds to an important understanding in personal methylome analysis in the
441 evidence for interactions between epigenome and genome in the form of CpG:variant
442 interactions. The approach would be significant in identifying significant genome loci that reflect
443 the effects of methylation modifications in combination with genetic variation that accrue in an
444 individual, therefore aid in identifying associations with longevity and associated traits like
445 disease onset.

446
447
448
449
450
451

Declarations:

Ethics approval and consent to participate

The samples of peripheral blood collected in this study involve human healthy donors and were obtained with their informed consent. were in accordance with the ethical standards of the institution (Avesthagen Limited, Bangalore, India) and in line with the 1964 Helsinki declaration and its later amendments. The Zoroastrian Parsi adult participant (>18 years) underwent height and weight measurements and answered an extensive questionnaire designed to capture their medical, dietary, and life history. The subject provided written informed consent for the collection of samples and subsequent analysis. This study was approved by the Avesthagen Ethics Committee constituted under the Department of Biotechnology, Government of India (BLAG-CSP-033).

Availability of data and materials

The authors declare that all data and materials are available upon request.

Competing interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Funding

The project was funded by the grant awarded to Dr.Villoo Morawala-Patell "Cancer risk in smoking subjects assessed by next-generation sequencing profile of circulating free DNA and RNA" (GG-0005) by the Foundation for a Smoke-Free World, New York, USA.

Authors contributions

VMP conceptualized, designed, guided the experiments and analysis; VMP founded The Avestagenome Project™ and provided access to the dataset for this study; NP, GK, NN, BM and KK analysed the sequences, performed bioinformatics analysis, interpreted the results and

plotted the graphs and figures; MK, VV provided support in the form of literature mining and database curation. VMP, NP, GK, NN and KK drafted the manuscript. All authors reviewed the manuscript.

Acknowledgements

We thank the Foundation for Smoke-Free World, who is advancing global progress in smoking cessation and harm reduction, for funding this project and Dr. Derek Yach for his support of this project. We thank Genotypic technologies and Dr.Raja Mugasimangalam, Bangalore for their excellent sequencing services and technical assistance. We thank the Zoroastrian-Parsi community of India for their enthusiastic cooperation. We thank Dr.Sami Gazder, Kouser Sonnekhan, and the The Avestagenome Project™ project team.

496 **References**

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

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

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

504 4. Marttila, S. *et al.* Ageing-associated changes in the human DNA methylome: Genomic
505 locations and effects on gene expression. *BMC Genomics* (2015) doi:10.1186/s12864-
506 015-1381-z.

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

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

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

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

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

519 10. Talens, R. P. *et al.* Epigenetic variation during the adult lifespan: Cross-sectional and
520 longitudinal data on monozygotic twin pairs. *Aging Cell* (2012) doi:10.1111/j.1474-
521 9726.2012.00835.x.

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

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

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

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

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

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

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

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

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

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

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

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

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

547 23. Tanabe, M. & Kanehisa, M. Using the KEGG database resource. *Curr. Protoc. Bioinforma.*
548 (2012) doi:10.1002/0471250953.bi0112s38.

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

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

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

557 27. Zhao, W. *et al.* Polycomb group RING finger proteins 3/5 activate transcription via an
558 interaction with the pluripotency factor Tex10 in embryonic stem cells. *J. Biol. Chem.*
559 **292**, 21527–21537 (2017).

560 28. Shivapurkar, N. *et al.* Differential inactivation of caspase-8 in lung cancers. *Cancer Biol.*
561 *Ther.* (2002) doi:10.4161/cbt.1.1.45.

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

564

565

Figures

Figure 1: Process workflow and Methylation frequency distribution in an individual separated by two time points

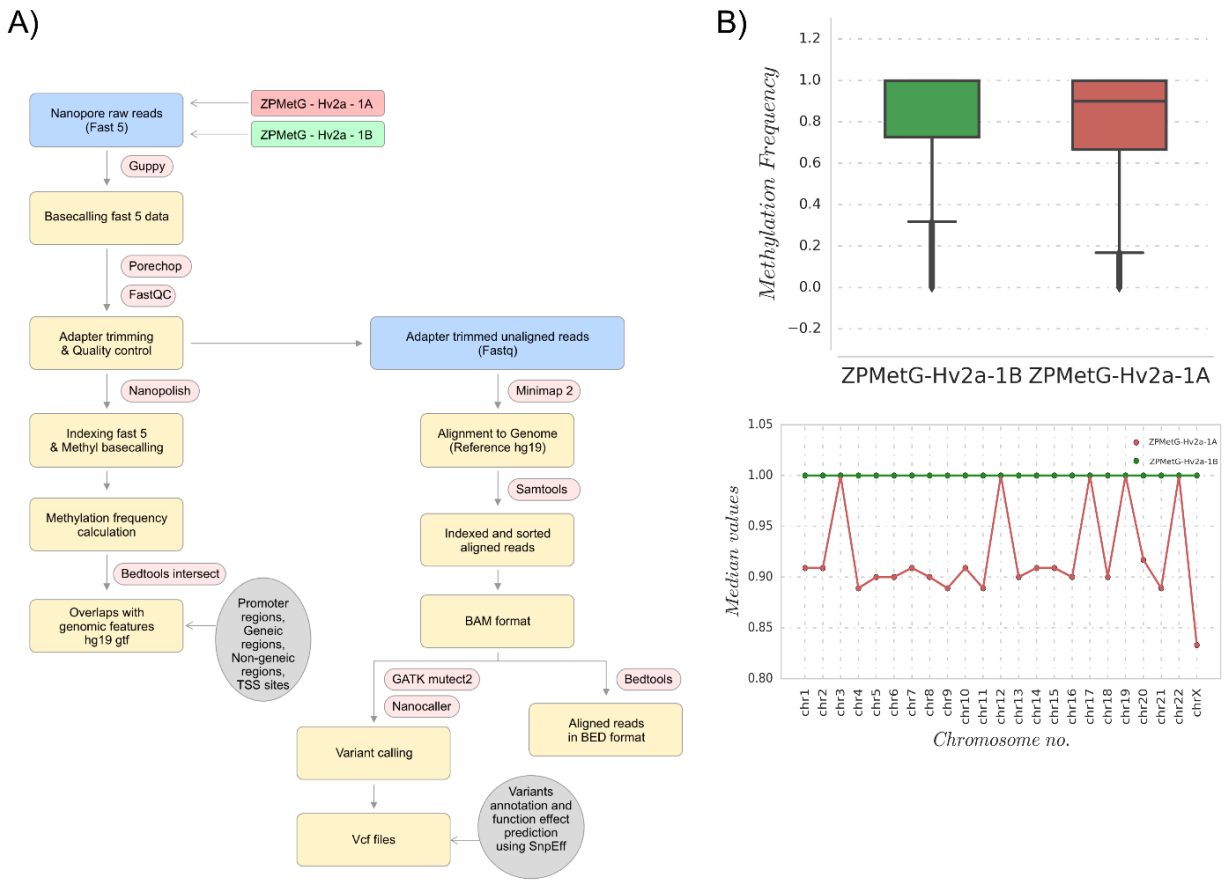
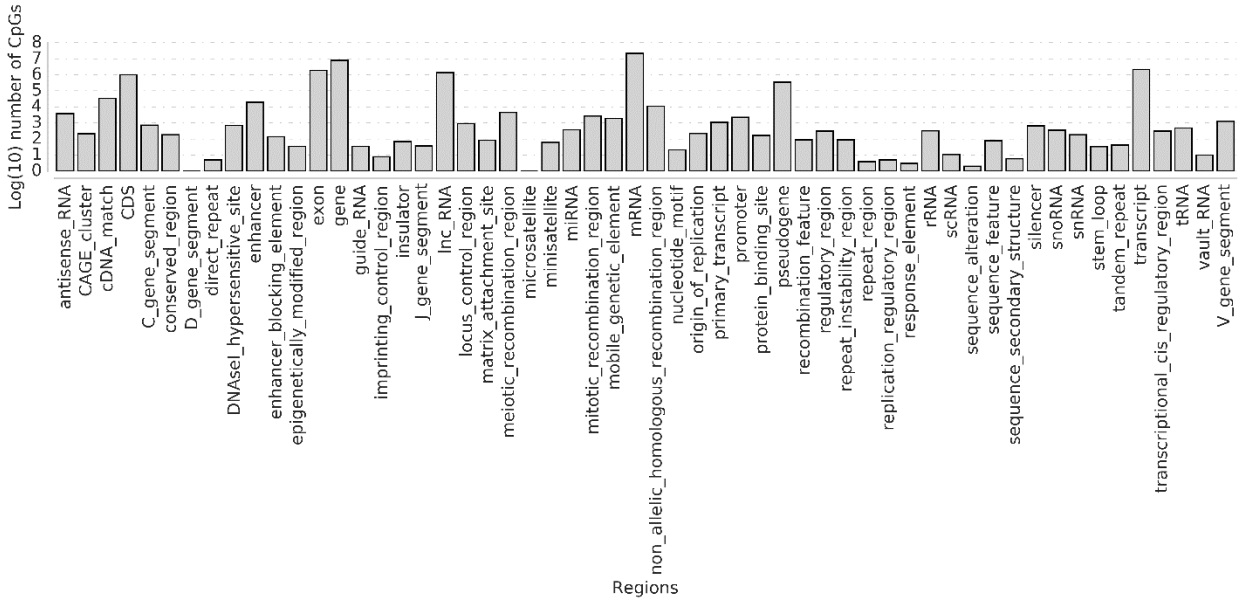


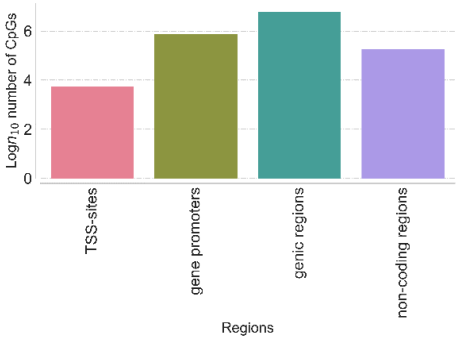
Figure 1: (A) Flowchart representing the workflow for analysis of individual methylome and mutational signature (SNP) variations across an individual separated by a duration of 12 years. **(B)** Methylation frequency distribution for compared base called regions shown for ZPMetG-Hv2a-1B (recent, in green) and ZPMetG-Hv2a-1A (old, in red). On the right is shown the histogram, identifying the nature of distribution to be non-Gaussian. Median values of the methylation frequency distribution chromosome wise for ZPMetG-HV2a-1A (old, in red) and ZPMetG-HV2a-1B (recent, in green)

Figure 2: Distribution and classification of common methylation imprints (CpG) based on genome architecture

A)



B)



C)

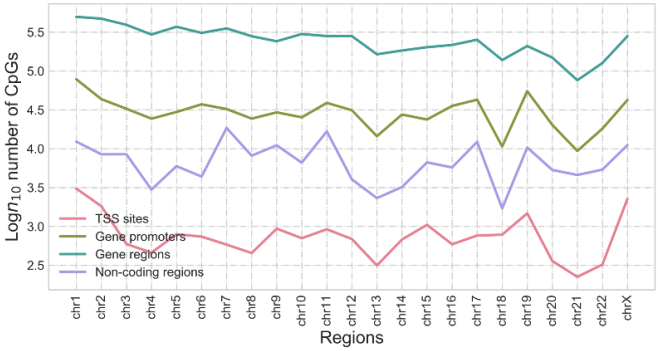
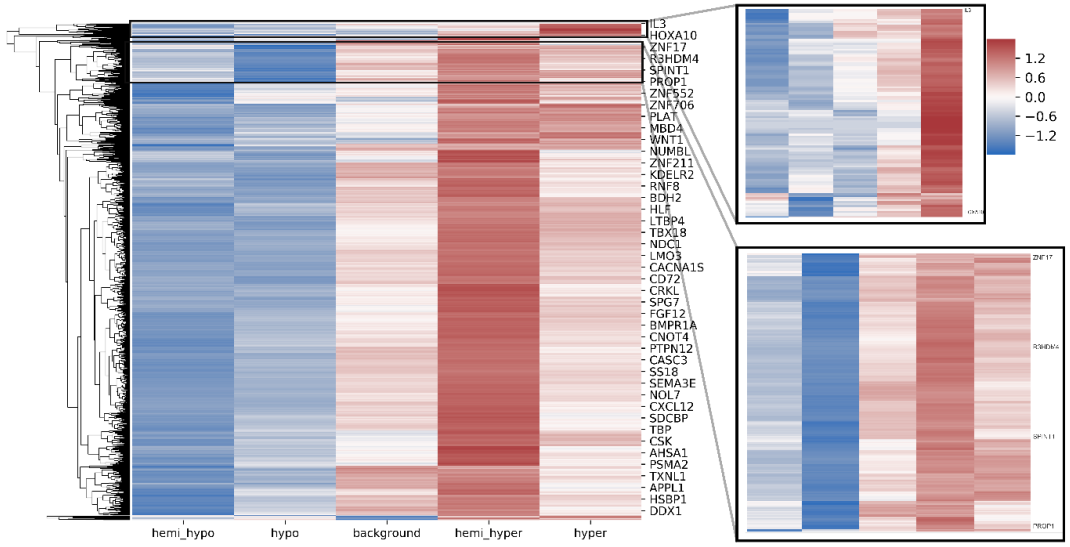


Figure 2: (A) Histogram representing the distribution and classification of methylation imprints (CpG) based on genome architecture and classifiers**. (B)** Histogram depicting classification of common methylation imprints (CpG) based on genomic characterization (inlay, left); Chromosome wise distribution of CpG numbers common to ZPMetG-Hv2a-1 and ZPMetG-Hv2a-2 classified based on distribution across genomic motifs (inlay, right)

Figure 3: Heatmap of Differentially Methylated Genic Regions observed in ZPMetG-HV2a-1B (recent) compared to ZPMetG-HV2a-1A (old)

A)



B)

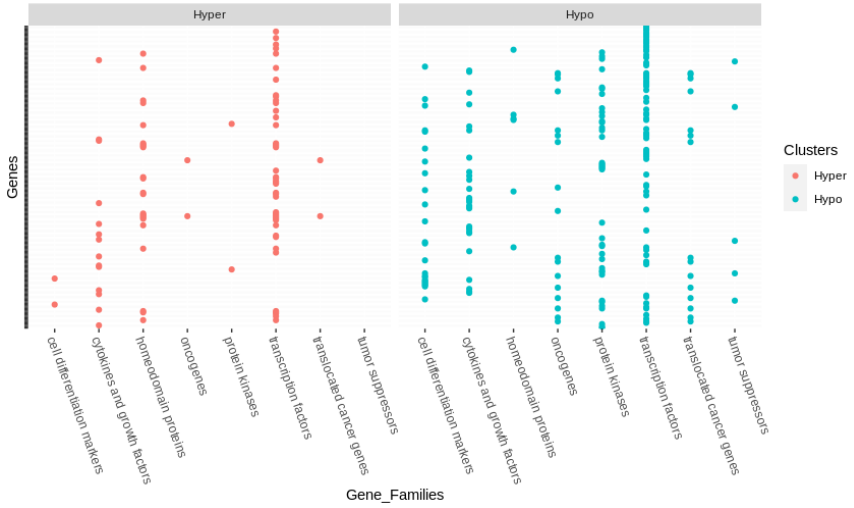
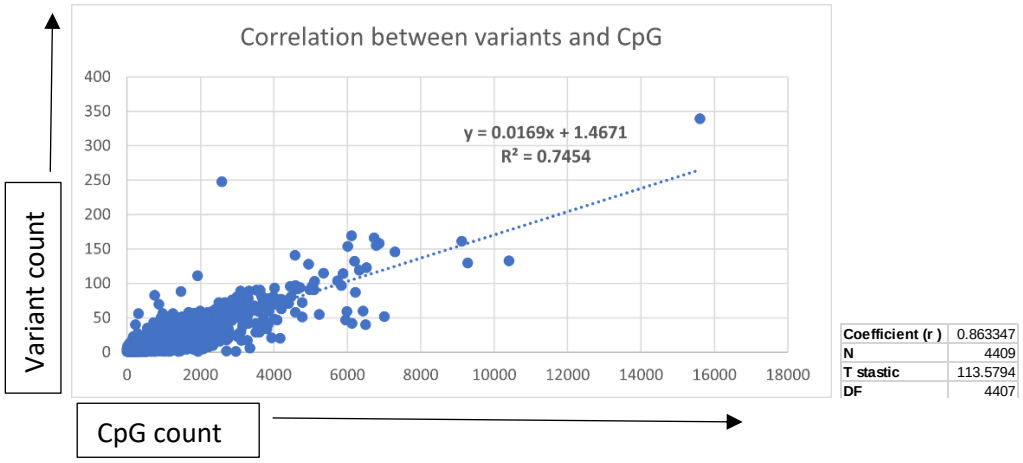


Figure 3: Heatmap displaying hierarchical clustering of row-wise Z score of common Differentially Methylated Gene regions in ZPMetG-Hv2a-1B. Clusters of Hypermethylated genes (inlay figure; right, top) and Hypomethylated (inlay figure; right, bottom). (B) GSEA based classification of hyper and hypomethylated clusters

Figure 4: Distribution of SNPs across CpG islands (Methylated regions) in ZPMetG-HV2a-1B (recent)

A)



B)

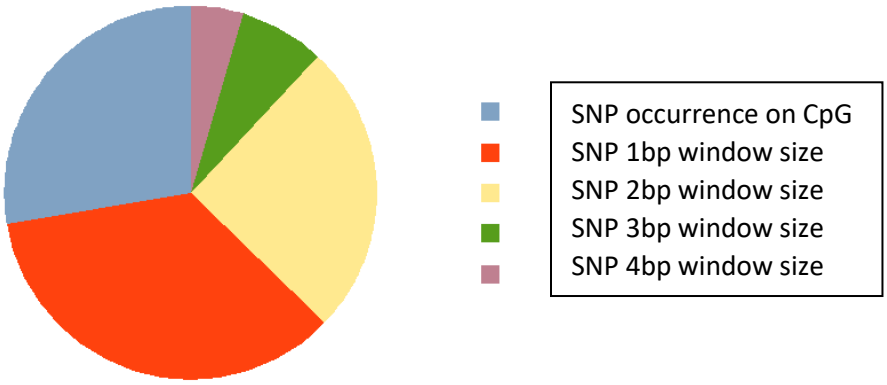
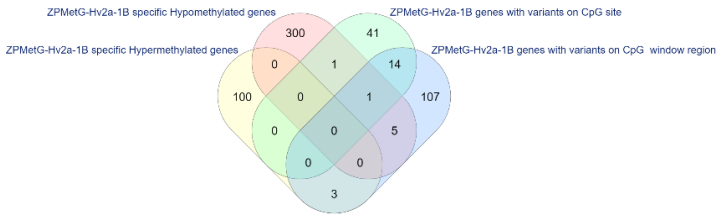


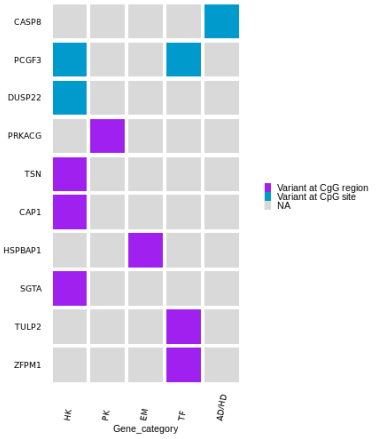
Figure 4: A) Correlation between CpG and variant counts occurring in ZPMetG-Hv2a-1B. Pearson correlation coefficient,r is displayed on the side B) Distribution of ZPMetG-HV2a-1B(recent) specific variants across Differentially Methylated Regions specific to ZPMetG-HV2a-1B (recent)

Figure 5: Biological effect and pathway association of ZPMetG-HV2a-1B (recent) variants occurring exactly in the CpG sites

A)



B)



C)

D)

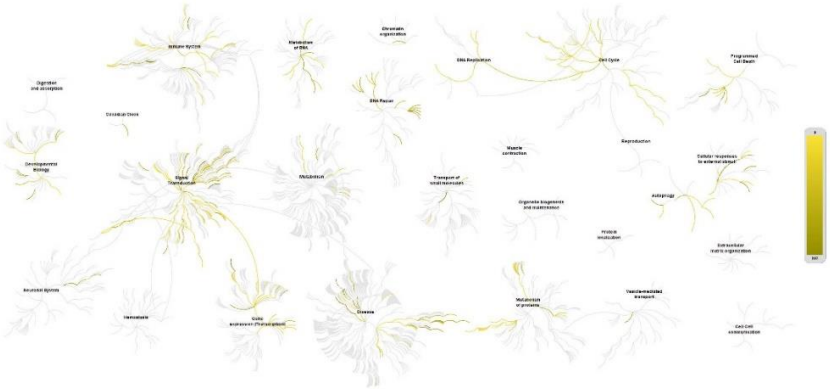
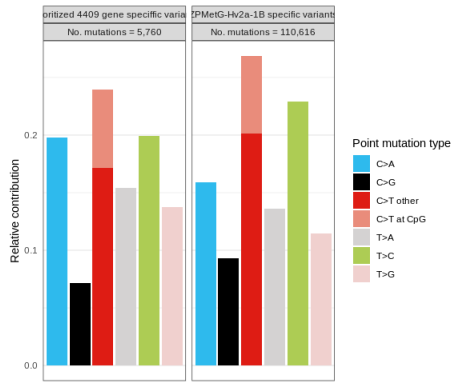


Figure 5: A) Distribution of CpG variants (exact, non-exact) across significant hyper and hypomethylated genes in ZPMetG-Hv2a-1B B) Categorization of 10 critical genes based on biological acitivity C) Network analyst using KEGG based disease enrichment module for 10 critical and 1000 high priority variants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

Figure 6: Assessment of mutational signatures and its distribution among CpG sites of ZPMetG-HV2a-1B (recent) in context of disease etiology

A)



B)

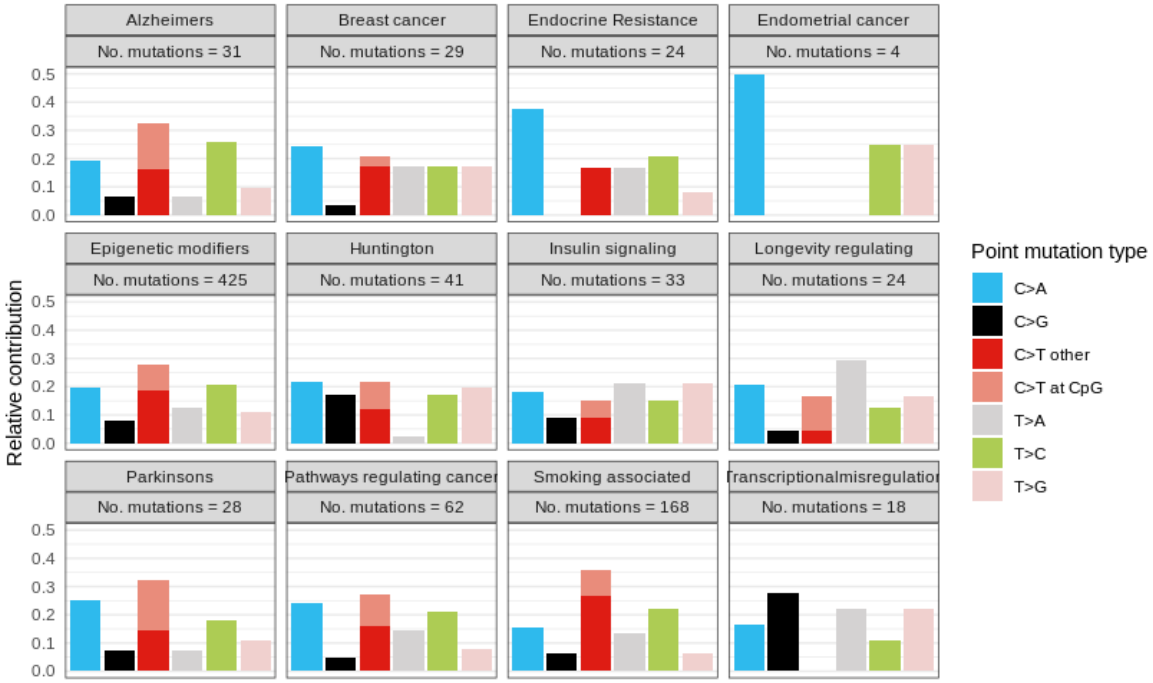


Figure 6: A) Distribution of mutational signatures across the genome and specific to CpG regions in 4409 prioritized genes and cumulative variants across ZPMetG-Hv2a-1B B) Mutational signatures indicative of transitions and transversions specific to ZPMetG-Hv2a-1B across different disease categories.