Inferring CpG methylation signatures accumulated along human history from genetic variation catalogs

Yichen Si[1] and Sebastian Zöllner[1,2]

[1]Department of Biostatistics, University of Michigan
[2]Department of Psychiatry, University of Michigan

Abstract

Understanding the DNA methylation patterns in the human genome is a key step to decipher gene regulatory mechanisms and model mutation rate heterogeneity in the human genome. While methylation rates can be measured e.g. with bisulfite sequencing, such measures do not capture historical patterns. Here we present a new method, Methylation Hidden Markov Model (MHMM), to estimate the accumulated germline methylation signature in human population history leveraging two properties: (1) Mutation rates of cytosine to thymine transitions at methylated CG dinucleotides are orders of magnitude higher than that in the rest of the genome. (2) Methylation levels are locally correlated, so the allele frequencies of neighboring CpGs can be used jointly to estimate methylation status. We applied MHMM to allele frequencies from the TOPMed and the gnomAD genetic variation catalogs. Our estimates are consistent with whole genome bisulfite sequencing (WGBS) measured human germ cell methylation levels at 90% of CpG sites, but we also identified $\sim 442,000$ historically methylated CpG sites that could not be captured due to sample genetic variation, and inferred methylation status for $\sim 721,000$ CpG sites that were missing from WGBS. Hypo-methylated regions identified by combining our results with experimental measures are 1.7 times more likely to recover known active genomic regions than those identified by WGBS alone. Our estimated historical methylation status can be leveraged to enhance bioinformatic analysis of germline methylation such as annotating regulatory and inactivated genomic regions and provide insights in sequence evolution including predicting mutation constraint.

Introduction

DNA methylation can directly modify protein binding sites or change chromatin 3D organization to regulate gene expression[1], and the majority of DNA methylation in mammalian cells is contributed by CpG methylation[2]. DNA methylation is also crucial for understanding mutation processes. In the human germline, the cytosine to thymine (C/T) mutation rate at methylated CpG sites is ten fold greater on average than that of unmethylated CpG sites[3, 4], leading to the observation that 99% of methylated CpG sites are mutated in at least one of 390k individuals [5]in genomAD[6], UK Biobank[7], and DiscovEHR[8].

Whole genome bisulfite sequencing (WGBS) has been the gold standard for measuring CpG methylation level since its introduction in 1992[9–11], and > 100 human tissues and cell lines have been profiled[12]. However, each of these datasets provides measurements for one sample of cells and cannot be extrapolated to population level methylation or long-term patterns of methylation. Such inference is limited by germline mutation bias[13], where a typically methylated C has mutated to a T. Individuals carrying a C/T mutation would be faithfully measured as unmethylated by bisulfite sequencing obfuscating the historical methylation at this locus or the methylation status of other individuals with a C allele. As methylated CpGs have high mutation rates, many such obfuscating mutations have reached high allele frequencies in the population. Across 45 million autosomal CpG sites across the genome, random individuals carry an expected 805,979 C/T mutations and homozygous T alleles at 135,222 sites, based on allele frequencies from Bravo[14]. This germline mutation bias reduces estimates of the mean methylation level, especially in small samples. Moreover, understanding methylation from an evolutionary perspective requires historical methylation information, which is never directly measurable. The closest approximations are methylation levels observed in sperms, oocytes, and germline cells at early developmental stages[4, 15]. Among them, methylation level in sperm has the strongest correlation with germline mutation rate and SNP density in current population samples[4] and is thus the best predictor for historical methylation. However, although researchers have reported high methylation rate and distinct methylation patterns in sperm cells even in the first published bisulfite sequencing study[9], existing sample sizes are too small (3 testis sperm samples available on Roadmap Epigenomics database[16] and 2 on ENCODE[17], all from different studies) to allow conclusive statements about human germ cell methylation[15, 18]. Because DNA methylation is dynamic during development and differs across tissues[15, 18, 19], combining information from WGBS datasets across tissues or cell lines will not mitigate the difficulties in studying germline methylation.

Given these challenges, computational approaches that aim to infer or impute DNA methylation have the potential to provide important information. Zhang et al.[20] use a variety of genome annotation, especially histone marks and regulatory elements, to train a statistical random forest predictor for methylation level in whole blood, for which epigenetic experiments have the largest sample sizes over multiple modalities. DeepCpG[21] combines DNA sequence context and incomplete methylation measures to impute missing methylation status using neural network. It has been applied only to infer single cell methylation states. Both are supervised

prediction methods that borrow information from a neighborhood including observed methylation levels at other CpG sites to infer missing methylation status at a focal position.

We propose a new method to infer germline methylation status using observed allele frequencies in publicly available variant catalogs at single base resolution called Methylation Hidden Markov Model (MHMM). Allele frequency is informative about germline methylation status at CpG site as methylated sites have very high mutation rates. As a result of their ~ 10 fold increase of mutation rate, some CpG sites that are consistently methylated in the population have mutated multiple times (recurrently) in the sample history, so that hyper-methylated regions are depleted of monomorphic sites and low-frequency variants. Along the DNA sequence, both methylated and unmethylated sites tend to form tight regions with high or low methylation rate so that information can usefully be shared locally. For instance, CpG islands, empirically defined as CpG dense segments, are highly enriched in genic region especially protein binding sites. In contrast, methylation of multiple consecutive CpGs in a promoter silences the corresponding genes[22, 23]. Combining this information, MHMM models methylation levels as hidden states and sample allele frequencies of C/T variants at individual CpG sites as observed states.

We apply MHMM to whole genome variants on 132,345 individuals from the TOPMed study[14] to estimate the historical methylation rate along the genome. Although our model uses information orthogonal to experimental measures or sample specific methylation status, our results are consistent with germline WGBS datasets at 90% of CpG sites. Estimated historical methylation level also recovers active genomic regions accurately, and we find that combining our estimates with experimental measures increases the enrichment of known active genomic regions. Since our results can also be interpreted as accumulated mutation burden at near base pair resolution, contrasting the observed and expected allele frequencies based on our model suggests CpG sites that are likely to be under purifying selection. Our software and inferred methylation levels are available at `https://github.com/Yichen-Si/cpghmm`.

Method

Hidden Markov Model To estimate the unobserved methylation status from polymorphism data, we build a continuous time, discrete state HMM. We model discretized methylation levels as hidden states, discretized allele frequencies as observations, and construct transition probability as a function of the physical distance between adjacent CpG sites. The stochastic process consists of all individual cytosines in CpG sites on both strands of a chromosome according to human reference genome GRCh38, and the emission models local sample frequency spectrum (SFS) as a result of the local methylation level.

We discretize methylation level into $K(\geq 2)$ states from hypo- to hyper-methylated, and the T-allele frequencies into M categories from monomorphically C to primarily T. Let Z_i and X_i indicate the hidden state (methylation level) and observation (allele frequencies) at location i respectively, so $Z_i \in [K] = \{1, \ldots, K\}$, $X_i \in [M]$. We explicitly include non-mutated (allele count AC= 0), singleton (AC= 1), doubletons (AC= 2), up to allele count 5 as separate categories, then choose log-linearly spaced break points to group higher allele frequencies. In practice, the two tails of the sample frequency spectrum (SFS) are most informative for our inference. We choose $K > 2$ because the historical methylation level averaged over time and population is likely to be fractional rather than binary as in a homogeneous cell sample. Choosing a state space with finer resolution also increases the robustness to mutation rate variation among methylated CpGs that depends on sequence context and other unmodeled factors[24].

We assume that the emission probability of a single observation conditional on its state follows a multinomial distribution, $P(X_i = m|Z_i = k) := E_{k,m}$, where $E_{k,\cdot}$ corresponding to a binned SFS specific for state k. We do not assume any structure of E and its estimates are dataset specific because methylation-level-stratified SFS depends on population demographics and sample size.

We parameterize transitions among states in two layers. At each position (the C in a CpG), we first generate its probability of moving out of the current state by the next position based on an exponential distribution of the distance until a change of state, then sample the next state if a transition occurs. We consider two adjacent CpG sites i and $i+1$ separated by d_i base pairs (bp). The probability for position $i+1$ to stay in the same state as position i conditional on $z_i = k$ is

$$P(z_{i+1} = z_i | z_i = k) = 1 - e^{-\theta_k \cdot d_i}.$$

Conditional on $z_{i+1} \neq z_i$ the probability to move to state $k' \neq k$ is $A_{k,k'}$, where A is a stochastic matrix satisfying $\sum_{k' \neq k} A_{k,k'} = 1, A_{k,k} = 0$ for all $k, k' \in [K]$. Taken together, the transition from k to $k' \neq k$ over d_i bp has probability

$$P(z_{i+1} = k' | z_i = k) = e^{-\theta_k \cdot d_i} A_{k,k'}.$$

We assume the K transition rates θ_k are independent, capturing the different length distribution of hyper- and hypo-methylation regions; we do not constrain conditional probabilities A to any structure, allowing the relation among states to be learnt entirely from the data.

Parameter estimation Unknown parameters in this model include A, E, and $\{\theta_k\}$, with $K^2 - 2K + KM$ degrees of freedom. Matrices A and E can be estimated with the Baum–Welch algorithm[25], but the maximum likelihood estimates for transition rate θ_k's do not have analytical forms. We use an approximate EM approach and update θ_k's iteratively by numerical optimization. We choose the Nelder–Mead method, a gradient-free, multi-dimensional optimization algorithm to search for a local optimum in the K-dimensional parameter space of θ_k's conditional on A and E in the M-step of each iteration. We estimate the parameters by running the above EM algorithm on each arm of the autosomes separately (excluding centromeres), then take the genome wide average of learned parameters weighted by the number of CpG sites in each chromosome arm.

Modeling observed allele frequencies as the sole result of underlying methylation level is subject to confounding from natural selection. Monomorphic CpG sites in a large sample could have low methylation levels and low mutation rates, or could be under purifying selection that removes C/T mutations[26]. To reduce the effect of selection on the inference of individual CpG sites, we compute a leave-one-out (LOO) likelihood of the hidden states at each CpG conditional on all but the focal site's allele frequency. In this way, we avoid assuming neutrality at every single site and instead assume it is rare to have a dense cluster of CpG sites all under selection that dominate a neighborhood. This assumption is mild because even in highly selected coding regions, fitness effect of CpG C/T mutations is limited by amino acid changes possible in the context. Computing LOO likelihoods only adds a fraction of overhead using similar computational tricks as in[27, 28].

Data application We train the above model and estimate marginal and LOO likelihoods for each CpG cytosine to be at each of the discretized methylation levels using autosomal allele frequencies from Bravo (132,345 genomes) and gnomAD (71,702 genomes) separately. We excluded $\sim$ 10M CpGs located in regions with low mappability based on the reference from the 1000 Genome Project[29] or low variant quality scores reported by the variant catalogs, resulting in $\sim$ 45M autosomal CpG sites.

We compare our results with methylation levels of 43 different tissues and cell lines from 81 publicly available WGBS datasets from the ENCODE portal[17] (https://www.encodeproject.org/, individual dataset identifiers are listed in supplementary). To best approximate the historical methylation we aim to estimate, we also calculate the weighted average methylation level based on two WGBS datasets, one from a testis sperm cell line and the other from an ovary tissue sample (data accessed at NCBI GEO database, accession GSE16256). As the ratio between male and female contribution to germline mutations is about 4:1[30], we weight the two datasets accordingly by 0.8 and 0.2. We acquire functional annotations of C/T mutations at CpG sites by running Ensembl VEP[31].

Combining experimental data and inference results to identify Hypo-methylated regions To asses the enrichment of hypo-methylated regions in regulatory regions, we first find genomic segments where all CpGs are hypo-methylated. We then connect adjacent segments heuristically if they are separated by less than 50 bp and less than 2 hyper-methylated CpGs.

Finally, we remove segments with CpG density lower than 5 CpGs per 100 bp, and define the remaining segments as hypo-methylated regions. Using this definition, we identify one set of hypo-methylated regions using MHMM methylation status and one set of hypo-methylated regions using germline WGBS dataset. For WGBS datasets, we assume CpGs with missing methylation status to be methylated. To get the consensus of experimental data and inference results, we take the intersection of overlapping pairs of hypo-methylated regions identified by MHMM and WGBS so that the resulting segments are located in hypo-methylated regions regardless of which method the methylation status are based on.

Results

Introduction We apply our Methylation Hidden Markov Model (MHMM) to infer historical germline methylation level using the TOPMed variant catalog[14] (freeze 8, 132,345 genomes) and replicate our observations in gnomAD[6] (v3.0, 71,702 genomes). We compute the probability distribution over discretized methylation levels at each of the 45 million autosomal CpG loci conditional on all observed CpG allele frequencies (AF) within 20Mb or up to the end of the chromosome arm.

We validate our estimates by showing that they are highly consistent with methylation status measured in testis sperm cells and investigate the differences between the two approaches. We show that differences between MHMM and WGBS indicate both limitations of our method and blind spots in WGBS. We then demonstrate potential applications of the method, focusing on 1) hypo-methylated regions inferred by our method, and show they are enriched for active/regulative genomic regions; 2) CpG sites located in inferred hyper-methylated regions but are monomorphic in the sample, and show they are enriched for sites where C/T mutations would cause severe functional consequences.

Whole genome methylation landscape measured by MHMM and WGBS

Summary of TOPMed MHMM estimates Applied to the TOPMed variant catalog, MHMM assigns most CpG sites to one of the two categories: 77.7% of CpG sites have > 0.9 probabilities to be in historically hyper-methylation regions, 12.0% of CpG sites have > 0.9 probability to be in historically hypo-methylation regions while the remaining 10.3% CpG sites have intermediate methylation level or cannot be confidently assigned to either category (Figuer A.3). The estimated local methylation rate is associated with local CpG densities, with hypo-methylated regions enriched in dense CpG neighborhoods (Figure A.2, A.3). The difference between maximum likelihood (MLE, using all observed CpG AF) and leave-one-out likelihood (LOO, excluding the focal site's own AF from its estimator) is generally small, above 0.1 (probability unit) at only 6.4% of sites. This difference is on average higher in regions with lower CpG density, ranging from 0.025 to 0.018 across density deciles.

Reproduce with gnomAD allele frequencies We applied the same model fitting process to gnomAD[6] AF based on 71,702 individuals primarily of European and African ancestries. Results from gnomAD are consistent with those from TOPMed in general as expected since AF do not differ qualitatively except for rare variants. There are 3.51% CpGs estimated as hyper-methylated using one dataset while estimated as hypo-methylated using the other; 94.1% of this discrepancy is contributed by rare variants that are discovered in one dataset but not in the other. The two datasets have different rare variant calling and quality control procedures, which affect the lower end of the SFS where most signal for our model comes from. Henceforth, we present results from TOPMed which has more homogeneous genotype qualities and a larger sample size; results for gnomAD are similar.

Summary of WGBS observations To evaluate the inferred methylation levels, we com-

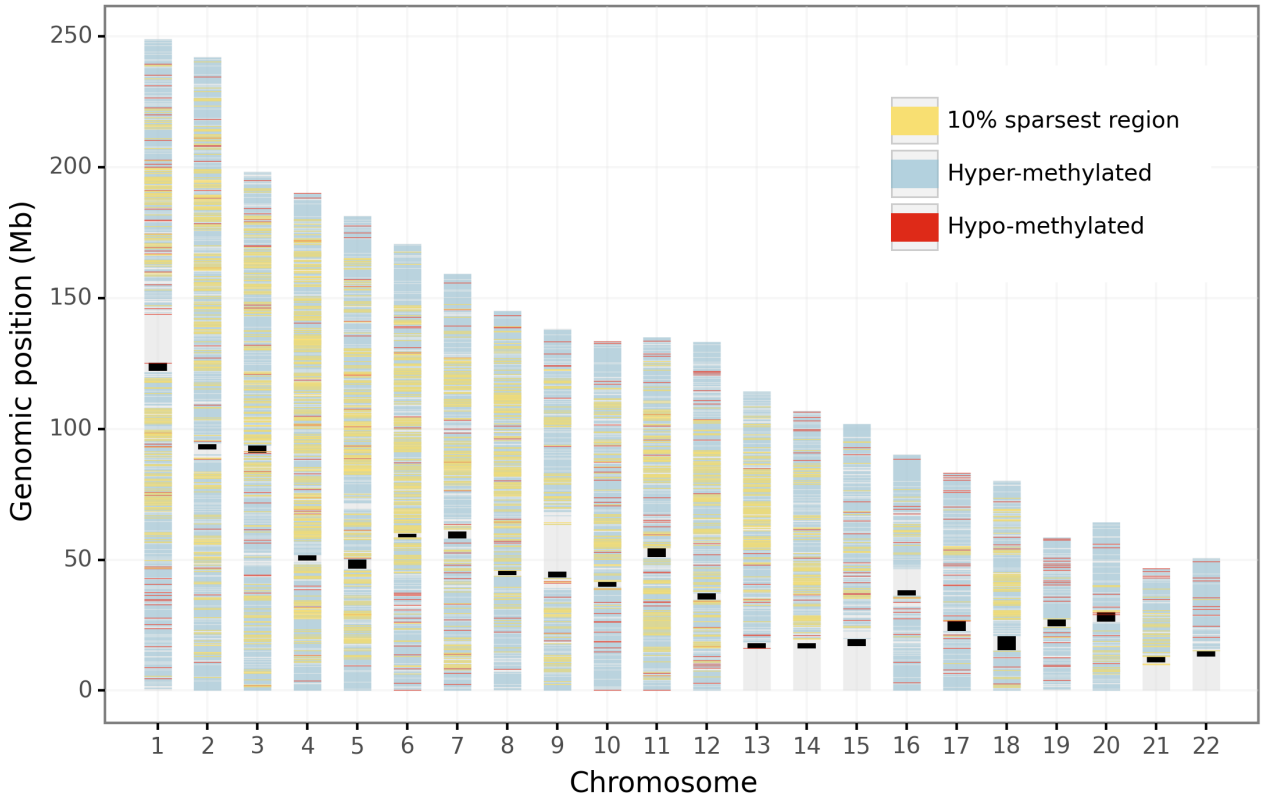

Figure 1: Whole genome landscape of germline methylation level inferred by MHMM. Hyper-methylated regions and hypo-methylated regions are colored by blue and red. Regions with low mappability and genotype quality are left uncovered. CpGs with the 10% lowest local CpG density where our estimates are less realizable are masked by yellow. Black boxes mark centromeres.

pare them to methylation levels measured by whole genome bisulfite sequencing (WGBS). To focus on the best available proxy for historical germline methylation levels, we use a weighted average (see Methods) of methylation level in a sperm cell line and an ovary tissue sample unless stated otherwise. In the averaged data, 76.3% of CpGs are methylated, 12.6% of CpGs are non-methylated, and the remaining 11% of CpGs' methylation levels are either missing or appear to be intermediate. The averaged methylation level is very similar to sperm cell line while the ovary tissue contributes more intermediate observations (18.1%), consistent with the fact that the ovary tissue is consist of multiple cell types.

We also notice an uneven distribution of WGBS missing values across the genome. While on average 1.6% of CpG sites have missing methylation status, CpG sites with higher local GC content or higher AF are enriched for missing values. Among 5.86M CpGs located in 65,551 autosome CpG islands[32] 7.7% are missing WGBS observations (4.8x enrichment), and among 176,257 CpGs with AF >0.5 3.2% are missing WGBS observations (2.0x enrichment).

Allele frequencies in hyper-methylated versus hypo-methylated regions The signal driving MHMM inference is the distorted site frequency spectrum (SFS) caused by the high mutation burden of CpGs consistently methylated in the germline. Figure 2 shows the SFS of CpGs stratified by if they are (hyper-)methylated according to both or only one of MHMM and WGBS, compared with SFS among non-methylated CpGs and non-CpG sites. CpGs iden-

tified as methylated by both methods have the most pronounced signature of high mutation rate and prevalent recurrent mutations, followed closely by hyper-methylated CpGs identified only by MHMM. Methylated CpGs identified only by WGBS have noticeably fewer common variants than the above two categories, though still substantially more than non-methylated CpGs. Only 5.9% of hyper-methylated CpGs remain monomorphic among 132k individuals (TOPMed freeze 8 from Bravo[14]), distinct from the hypo-methylated group where 61.3% remain monomorphic.

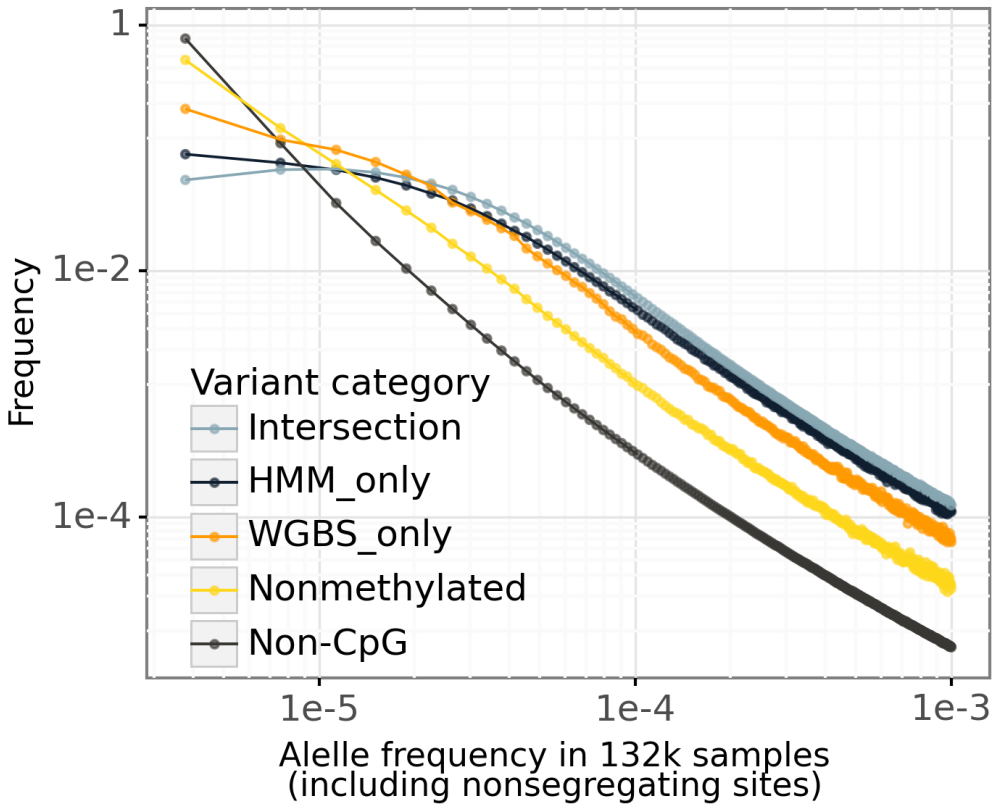

Figure 2: Sample frequency spectrum based on Bravo variant lists on a log-log scale. The SFS includes all autosome sites, including monomorphic sites (marked by the leftmost point in each line). The gray line summarizes all non-CpG sites (based on the reference genome), and the others stratify CpGs by their observed/inferred methylation status.

Comparison between MHMM and WGBS

Comparing inferred historical methylation with WGBS measured methylation, we see that among sites indicated by WGBS as methylated, 91% are inferred to be historically hyper-methylated and $\leq$0.7% are inferred to be hypo-methylated; among sites indicated by WGBS as non-methylated, 82% are inferred to be historically hypo-methylated and $\leq$8% are inferred to be hyper-methylated. Among the remaining 10% of sites measured as having intermediate methylation level (due to heterozygous donors and/or differential methylation between sperm and ovary), 89% are inferred to be intermediates or hyper-methylated (Figure 3). Overall MHMM predictions and experimental observations are consistent to germline cell methylation levels at 90% of the sites. We extend this comparison to other 80 samples from diverse tissues

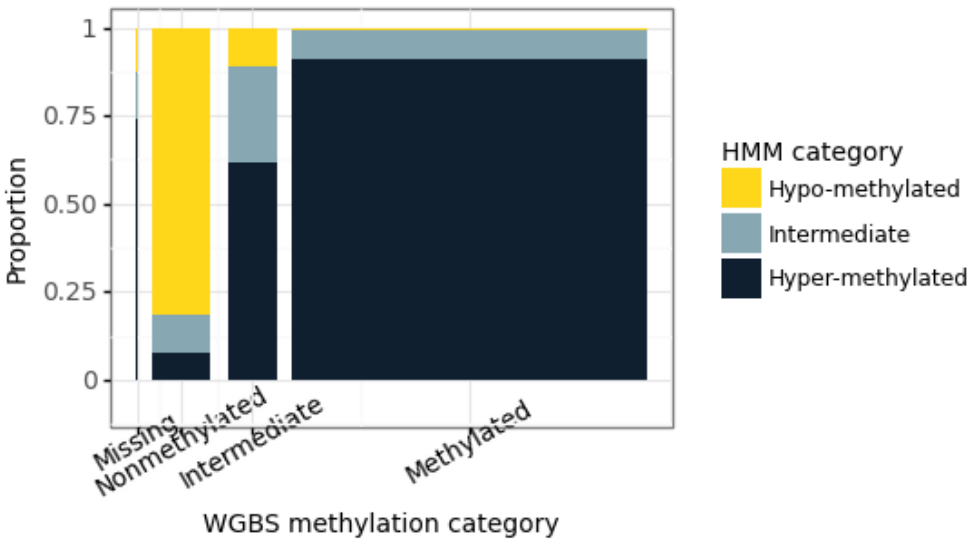

Figure 3: Compare HMM estimated methylation level with WGBS measures in germlines. Width of each bar is proportional to the number of CpGs in the corresponding WGBS methylation category, i.e. the denominator of the proportion in y-axis. There are 0.64% sites missing from the WGBS dataset, 74.1% and 12.5% of them are inferred to be hyper- and hypo-methylated respectively, close to global average.

and cell lines (Table A.2) and observe that the level of experimentally measured methylation varies between 40% and 85%. Consistency of experimental measures with MHMM is driven by this overall methylation rate (Figure A.5, A.6).

To better understand the 10% discordance between MHMM and WGBS, we identify sequence properties associated with such differences. We observe that discordance depends on local CpG density. In the 10% sparsest regions (less than 10 CpG per 1kb), 74% of measured non-methylated sites are inferred to be intermediate or hyper-methylated. In contrast, in the 10% densest regions ($\geq$60 CpG per 1kb), 6% of measured non-methylated sites are inferred to be intermediate or hyper-methylated. This difference in concordance is likely driven by the fact that the MHMM integrates information across neighboring CpGs but the correlation of methylation levels decays rapidly with distance[20, 21]. As GC content is higher in coding sequence and near transcription start sites compared to intronic and intergenic regions[23], our estimates and WGBS agree 95.3% of the times in known active regions (Figure A.7).

Germline mutation bias The high T-allele frequency in many methylated sites creates a challenge for using WGBS to infer historically methylated CpGs. Because experimental techniques correctly read a mutated CpG where the sample donor carries a T allele as non-methylated, inference of historical or population level methylation based on a small sample of current day individuals is biased by their germline mutations[13]. Here we assess this bias in a WGBS dataset of a testis sperm cell line to disentangle this mutation effect from differential methylation between tissues.

We categorize CpG sites by their T-allele frequencies from Bravo and count how many sites are identified by WGBS as methylated in each AF window (Figure A.8). Monomorphic (AF = 0) sites and ultra-rare variants (AF $< 10^{-4}$) are more likely to be non-methylated than

intermediate frequency variants. This is consistent with the fact that non-methylated sites have lower mutation rates and so are less likely to be polymorphic. However, common variants (AF >1%) are often measured as partially methylated, and almost all high AF variants (AF >50%) are measured as non-methylated.

Further, while methylation status is similar among nearby CpGs, CpGs with high T-allele frequency measured as nonmethylated by WGBS are typically surrounded by methylated CpGs. For each CpG site measured as non-methylated, we assess the methylation status of 5 immediate CpG neighbors both upstream and downstream as measured by WGBS. Among sites with AF <0.01, only 8.0% have at least one of their 10 neighbors methylated. This is expected, as methylation is locally correlated. In contrast, among non-methylated focal sites with AF >0.9, ≥95.3% of sites have at least one of their 10 neighbors methylated. For 61.0% of such sites, all of their 10 neighbor CpGs are methylated (Figure A.9), suggesting that these sites are historically methylated but measured to be non-methylated only because the sample donor carries the T allele.

Germline bias in WGBS partially explains the discordance between our MHMM estimates and the observed methylation status. ≤1% of MHMM estimated hyper-methylated CpGs with AF ≤10% are measured as non-methylated by WGBS, while 9.1% of variants with AF 10-50% and 93.6% of variants with AF>90% are measured as non-methylated by WGBS. In addition, 37.2% of estimated hyper-methylated CpGs with AF 10-90% show intermediate methylation level by WGBS, compared to only 7.7% of estimated hyper-methylated CpGs with AF below 10%, consistent with the large proportion of heterozygous genotypes in the AF range (Figure 4). If we project the proportion of methylated CpGs among sites with intermediate AF to those with high AF, we expect ∼367,000 historically methylated CpG sites to be mutated in the donor and therefore their historical methylation status is not measurable.

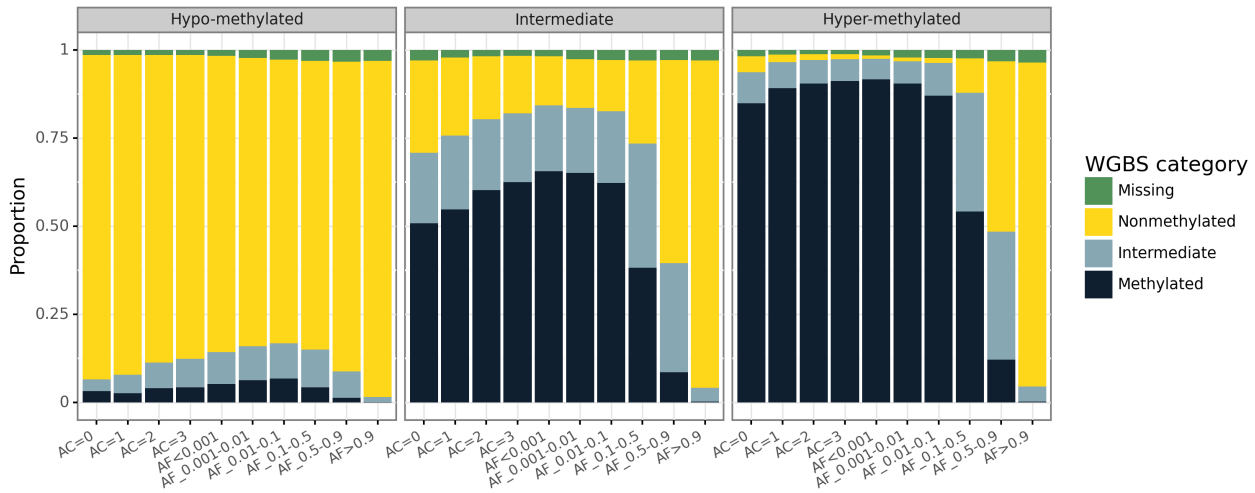

Figure 4: Composition of WGBS measured methylation status among CpGs in each MHMM inferred state (three panels) and allele frequency categories (x-axis). X-axis shows the allele count (AC) or allele frequency (AF) of the T allele at CpG sites (data from Bravo), from monomorphic (left most) to high frequency (right). The high discordance between inferred methylation status and measured methylation status for high T allele frequency illustrates the germline mutation bias of WGBS.

Combining inferred hypo-methylated regions and WGBS measures to identify active chromatin regions

Since hypo-methylated CpG islands are more likely to contain regulatory functions we assessed the functional annotation of hypo-methylated regions identified by combining MHMM and WGBS (see Method) of length >50 bp, 64.9% of which do not overlap with traditionally defined CpG islands. MHMM and WGBS identify similar number (5.4M v.s. 5.3M) of hypo-methylated CpGs overall, and 5.0M CpGs are assigned to hypo-methylated regions by either method. Among these 5.0M CpGs, 85% are located in hypo-methylated regions identified by both methods. We assessed the proportion of these regions that overlap with genomic features for active elements, including promoters, proximal enhancers, open chromatin, and transcription factor (TF) binding sites (Table1). Among hypo-methylated regions identified by MHMM, 43.9% overlapped with at least one of these features compared to 27.9% of hypo-methylated regions identified by WGBS. Hypo-methylated regions supported by both MHMM and WGBS are even more likely to overlap with known active elements than those supported by either of the methods alone. Overall, 46.7% of hypo-methylated regions identified by both methods overlap with at least one of those active elements.

Feature	**Pr(overlap with known features)**					
	MHMM ∩	**Marginal**		**CpG islands**		
	WGBS	**MHMM**	**WGBS**	**Hypo**	**Other**	**Hyper**
Promoter	0.2697	0.2497	0.1526	0.4221	0.0327	0.0035
Proximal enhancer	0.2794	0.2573	0.1592	0.3700	0.0372	0.0055
ATAC-seq peak (open chromatin)	0.4011	0.3810	0.2352	0.3488	0.0670	0.0217
TF ChIP-seq (TF binding sites)	0.1874	0.1875	0.1138	0.1618	0.0327	0.0136
Any of above	0.4669	0.4387	0.2785	0.6831	0.1300	0.0431

Table 1: MHMM and WGBS-identified hypo-methylated regions are enriched for known active elements. Column 2 is the probabilities of overlapping with known active elements among hypo-methylated regions identified by both methods; column 3-4 are the marginal probabilities when using each method alone; column 5-7 are the probabilities for CpG islands stratified by methylation status estimated by MHMM.

Applications

Inferred methylation level separates known CpG islands into two classes

CpG islands are defined as regions that have high local GC content and enriched CG dinucleotides compared to the expectation if nucleotides are distributed randomly[32]. Although 70% of annotated promoters in the human genome are located in such high CpG density regions[33], whether a CpG island can function as transcription regulator heavily depends on its methylation status[23]. We observe that known CpG islands fall into two distinct groups characterized by their historical methylation level, consistent with previous studies evaluating sequence conservation[34]. Among 25,743 autosome CpG islands, 77.1% are primarily (>90%) hypo-methylated while 13.4% are primarily (>90%) hyper-methylated. Among historically

hypo-methylated CpG islands 42.2% overlap with known promoters and 37.0% overlap with known proximal enhancers[35]. Among historically hyper-methylated CpG islands, only 0.3% overlap with known promoters and 0.6% overlap with known proximal enhancers (Table 1).

Unmutated variants in hyper-methylated regions

The segregation pattern of genetic variations in the current population sample is shaped not only by the mutation rate, but also by purifying selection. We classify CpG sites by the potential effect of C/T mutations using the Ensembl Variant Effect Predictor (VEP)[31] and stratify SFS among hyper-methylated CpGs by the functional consequences of the C/T mutation (Figure A.1). Among all hyper-methylated CpGs 94.5% are polymorphic, but putative loss of function (LoF) mutations are under strong purifying selection with 72.1% sites being polymorphic and only 7.0% reaching allele frequency $\geq$0.01%. This mutational constraint is the motivation for our leave-one-out estimator (see Method), especially in coding sequences (Figure A.4).

Since only 5.5% hyper-methylated CpGs remain monomorphic, observing a monomorphic site in a hyper-methylated neighborhood suggests the CpG may be maintained by purifying selection. Table A.1 shows the proportion of monomorphic site by mutation consequences and methylation status. Among CpG sites where a C/T mutation would be synonymous, 6.7% of hyper-methylated CpGs remain monomorphic, compared to 62.0% among hypo-methylated CpG sites, a $>$ 9-fold difference. Among CpG sites where a C/T mutation would potentially cause the loss of function of the protein, 28.4% of hyper-methylated CpGs remain monomorphic, compared to 77.5% among hypo-methylated CpG sites, $<$ 3-fold difference.

Discussion

We developed a Methylation Hidden Markov Model (MHMM) to infer historical germline methylation at CpG dinucleotides. Our method leverages the high mutation rates and the consequently distorted sample frequency spectrum (SFS) at methylated cytosines and the local correlation of methylation status. We applied MHMM to allele frequencies from the TOPMed and the gnomAD databases and estimated whole autosome methylation. Comparing our estimates with methylation levels in 6 germline tissue samples measured by whole genome bisulfite sequencing (WGBS) indicates 90% concordance in identifying hyper-methylated and hypo-methylated regions.

The differences between estimates by MHMM and experimental measures by WGBS are driven both by limitations of MHMM and blind spots in WGBS. As MHMM aggregates the signal from nearby CpGs, it has very limited information for estimating the methylation level of isolated CpGs. Indeed, MHMM differs from WBGS at about 74% hypo-methylated CpGs in the 10% lowest density regions but only at 5% hypo-methylated CpGs in the 10% highest density regions. On the other hand, WGBS's precision is limited in two scenarios: First, originally hyper-methylated CpGs that now have high derived allele frequencies in a population sample may not be identified by WGBS[13]. Experimental measures from a small number of samples reflect the personal methylation status of those samples and thus depends on these samples' genotypes while MHMM estimates reflect the historical population methylation level. As a result, MHMM infers locally homogeneous methylation status, while WGBS identifies individual mutated CpG sites correctly shown as unmethylated, interrupting otherwise methylated regions.

Second, sequencing based techniques often have low coverage in GC-rich regions and thus low quality methylation calls. This GC content bias is more significant with PCR amplification which is currently standard in WGBS[11]. In the testis sperm data (GSM1127119) for example, we have observed a 4.8 fold enrichment of missing data in known CpG islands compared to the global missing rate. The polymorphism data underlying our methylation inference is primarily from PCR-free sequencing protocols and are thus less affected by GC content bias.

Thus, MHMM creates more precise calls of population methylation status in GC-dense regions, and combining MHMM inference with the experimental data improves our ability to identify germline open chromatin and regulatory elements that are often hypo-methylated: MHMM identified hypo-methylated regions are 1.6 times more likely to contain known active genomic regions than WGBS-identified hypo-methylated regions, and hypo-methylated regions supported by both methods are more likely to contain known active genomic regions than those supported by either method alone. Since MHMM is orthogonal to WGBS experiments in the source of both information and bias, one can further explore alternative strategies to integrate the two methods based on local GC density and allele frequency.

Comparing our method with other computational methods that estimate methylation is difficult, as these methods typically use tissue or cell type specific genomic features. For example, Zhang *et al.* (2015) achieved 91.9% prediction accuracy of methylation level in blood cells, which have the largest sample size and most complete genomic annotations, by using various

functional features including each focal site's neighboring methylation status.

Based on the high precision of our method, we can also draw more general conclusions about CpG methylation in the human genome. First, our allele frequency based inference highlights the strong deviation of SFS among highly mutable methylated CpG from the expected SFS under the low, genome-wide averaged mutation rate. The high consistency of the current day experimental methylation level in random individuals with the accumulated signature estimated by MHMM also suggests that human germline methylation pattern is relatively stable in the past.

Second, we show that some CpG sites are maintained by purifying selection[5]. A CpG site where its own allele frequency substantially differs from that predicted by its neighborhood is 3.7 times more likely to be a potential loss of function mutation site. To account for purifying selection in our inference, we calculate two statistics, the marginal and leave-one-out likelihoods, for methylation status at single base pair resolution. We adopted a heuristic combination of the two statistics that effectively reduces confounding from purifying selection while incurring minimal loss of information at the majority of CpG sites that are near neutral.

Although we emphasize the population level interpretation of the estimated methylation signature, a future application is to compare our results with more individual measures from human and closely related species. For instance, our inferred historical germline methylation level could serve as a baseline to be compared with that measured in specific tissue types to identify patterns acquired during developmental processes; or with that measured in a small sample from a distinct population to investigate germline variation across ancestries; or with that measured in chimpanzees, gorillas, or orangutans to detect local fast evolution.

When applying our method to study germline methylation patterns in different human groups or non-human species, high quality sequencing data are critical. An intrinsic limitation of our method is its reliance on accurate rare variant genotyping to capture the number of monomorphic sites and ultra-rare alleles. This lower tail of the SFS is the most informative about germline mutation rate, though the exact relationship depends on the population history and sample size. In polymorphism data from tens of thousands of genomes, parallel mutations distort the whole SFS, whereas in a smaller sample or a sample from a population with smaller effective population size, the signal may only come from whether a site is polymorphic or not. To be robust to sample demographics and genotyping artifacts, we do not constrain the SFS by population genetics parameters, but instead let the model learn the difference of SFS across genomic regions from the data.

Overall, we demonstrate that we can leverage the accumulated mutation burden at CpG sites to infer germline methylation level averaged across the human population and over past generations, and thus reveal methylation patterns hidden from experimental measures. Our results provide new information for interpreting non-coding regions by identifying potential regulatory elements and CpGs under mutation constraint.

Availability

Our software and inferred methylation levels as a track for the UCSC Genome Browser are available at https://github.com/Yichen-Si/cpghmm.

Acknowledgements: This work was supported by R01 HG011031 and R01 HG005855 to SZ.

Author contribution statement: Y.S. and S.Z. developed the ideas and wrote the manuscript. Y.S. developed the method, implemented the software, and performed the analyses.

References

1. Mattei, A. L., Bailly, N. & Meissner, A. DNA methylation: a historical perspective. *Trends in Genetics* **38.** Publisher: Elsevier, 676–707. ISSN: 01689525. http://www.cell.com/article/S0168952522000713 (2022) (July 2022).

2. Patil, V., Ward, R. L. & Hesson, L. B. The evidence for functional non-CpG methylation in mammalian cells. *Epigenetics* **9.** Publisher: Taylor & Francis, 823–828. ISSN: 1559-2294. http://www.tandfonline.com/doi/abs/10.4161/epi.28741 (2022) (June 2014).

3. Cooper, D. N., Mort, M., Stenson, P. D., Ball, E. V. & Chuzhanova, N. A. Methylation-mediated deamination of 5-methylcytosine appears to give rise to mutations causing human inherited disease in CpNpG trinucleotides, as well as in CpG dinucleotides. *Human genomics* **4.** Publisher: Hum Genomics, 406–10. ISSN: 1479-7364. https://pubmed.ncbi.nlm.nih.gov/20846930/ (2022) (Aug. 2010).

4. Zhou, Y. *et al.* The Impact of DNA Methylation Dynamics on the Mutation Rate During Human Germline Development. *G3 Genes|Genomes|Genetics* **10,** 3337–3346. ISSN: 2160-1836. https://academic.oup.com/g3journal/article/10/9/3337/6060130 (2021) (Sept. 2020).

5. Agarwal, I. & Przeworski, M. Mutation saturation for fitness effects at human CPG sites. *eLife* **10.** ISSN: 2050084X. https://doi.org/10.1101/2021.06.02.446661 (2021).

6. Karczewski, K. J. *et al.* The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581.** Publisher: Nature Research, 434–443. ISSN: 14764687. https://doi.org/10.1038/s41586-020-2308-7 (2021) (May 2020).

7. Szustakowski, J. D. *et al.* Advancing human genetics research and drug discovery through exome sequencing of the UK Biobank. en. *Nature Genetics* **53.** Publisher: Nature Publishing Group, 942–948. ISSN: 1546-1718. https://www.nature.com/articles/s41588-021-00885-0 (2023) (July 2021).

8. Dewey, F. E. *et al.* Distribution and clinical impact of functional variants in 50,726 whole-exome sequences from the DiscovEHR study. *Science* **354.** Publisher: American Association for the Advancement of Science, aaf6814. https://www.science.org/doi/full/10.1126/science.aaf6814 (2023) (Dec. 2016).

9. Frommer, M. *et al.* A genomic sequencing protocol that yields a positive display of 5-methylcytosine residues in individual DNA strands. *Proceedings of the National Academy of Sciences* **89.** Publisher: National Academy of Sciences, 1827–1831. ISSN: 0027-8424. http://www.pnas.org/cgi/doi/10.1073/pnas.89.5.1827 (2021) (Mar. 1992).

10. Harris, R. A. *et al.* Comparison of sequencing-based methods to profile DNA methylation and identification of monoallelic epigenetic modifications. *Nature Biotechnology* **28.** Publisher: Nature Publishing Group, 1097–1105. ISSN: 1087-0156. http://www.nature.com/articles/nbt.1682 (2021) (Oct. 2010).

11. Olova, N. *et al.* Comparison of whole-genome bisulfite sequencing library preparation strategies identifies sources of biases affecting DNA methylation data. *Genome Biology* **19.** Publisher: BioMed Central Ltd., 33. ISSN: 1474-760X. http://www.ncbi.nlm.nih.gov/pubmed/29544553 (2021) (Dec. 2018).

12. Davis, C. A. *et al.* The Encyclopedia of DNA elements (ENCODE): data portal update. *Nucleic acids research* **46.** Publisher: Oxford University Press, D794–D801. ISSN: 1362-4962. http://www.ncbi.nlm.nih.gov/pubmed/29126249 (2022) (Jan. 2018).

13. Wulfridge, P., Langmead, B., Feinberg, A. P. & Hansen, K. D. Analyzing whole genome bisulfite sequencing data from highly divergent genotypes. *Nucleic Acids Research* **47.** Publisher: Oxford Academic, e117–e117. ISSN: 0305-1048. https://academic.oup.com/nar/article/47/19/e117/5545001 (2021) (Nov. 2019).

14. Zachary, A *et al.* Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program (2019).

15. Gkountela, S. *et al.* DNA Demethylation Dynamics in the Human Prenatal Germline. *Cell* **161.** Publisher: Cell Press, 1425–1436. ISSN: 00928674. https://linkinghub.elsevier.com/retrieve/pii/S0092867415005607 (2021) (June 2015).

16. Bernstein, B. E. *et al.* The NIH Roadmap Epigenomics Mapping Consortium. en. *Nature Biotechnology* **28.** Number: 10 Publisher: Nature Publishing Group, 1045–1048. ISSN: 1546-1696. https://www.nature.com/articles/nbt1010-1045 (2023) (Oct. 2010).

17. Luo, Y. *et al.* New developments on the Encyclopedia of DNA Elements (ENCODE) data portal. eng. *Nucleic Acids Research* **48,** D882–D889. ISSN: 1362-4962 (Jan. 2020).

18. Li, C. *et al.* DNA methylation reprogramming of functional elements during mammalian embryonic development. *Cell Discovery* **4.** Publisher: Nature Publishing Group, 41. ISSN: 2056-5968. http://www.nature.com/articles/s41421-018-0039-9 (2021) (Dec. 2018).

19. Lokk, K. *et al.* DNA methylome profiling of human tissues identifies global and tissue-specific methylation patterns. *Genome biology* **15.** Publisher: BioMed Central, r54. ISSN: 1474-760X. http://www.ncbi.nlm.nih.gov/pubmed/24690455 (2021) (Apr. 2014).

20. Zhang, W., Spector, T. D., Deloukas, P., Bell, J. T. & Engelhardt, B. E. Predicting genome-wide DNA methylation using methylation marks, genomic position, and DNA regulatory elements. *Genome Biology* **16.** arXiv: 1308.2134, 14. ISSN: 1474760X. https://genomebiology.biomedcentral.com/articles/10.1186/s13059-015-0581-9 (Dec. 2015).

21. Christof Angermueller, Lee, H. J., Reik, W. & Stegle, O. DeepCpG: Accurate prediction of single-cell DNA methylation states using deep learning. *Genome Biology* **18.** Publisher: BioMed Central Ltd., 1–13. ISSN: 1474760X. https://genomebiology.biomedcentral.com/articles/10.1186/s13059-017-1189-z (2021) (Apr. 2017).

22. Razin, A. CpG methylation, chromatin structure and gene silencing—a three-way connection. *The EMBO Journal* **17.** Publisher: John Wiley & Sons, Ltd, 4905–4908. ISSN: 14602075. http://emboj.embopress.org/cgi/doi/10.1093/emboj/17.17.4905 (2021) (Sept. 1998).

23. Deaton, A. M. & Bird, A. CpG islands and the regulation of transcription. *Genes & Development* **25.** Publisher: Cold Spring Harbor Laboratory Press, 1010–1022. ISSN: 0890-9369. http://genesdev.cshlp.org/lookup/doi/10.1101/gad.2037511 (2021) (May 2011).

24. Carlson, J. *et al.* Extremely rare variants reveal patterns of germline mutation rate heterogeneity in humans. *bioRxiv.* ISBN: 4146701805936, 108290. ISSN: 2041-1723. https://www.nature.com/articles/s41467-018-05936-5 (2017).

25. Baum, L. E. & Petrie, T. Statistical Inference for Probabilistic Functions of Finite State Markov Chains. *The Annals of Mathematical Statistics* **37.** Publisher: Institute of Mathematical Statistics, 1554–1563. ISSN: 0003-4851. https://www.jstor.org/stable/2238772 (2023) (1966).

26. Boukas, L., Bjornsson, H. T. & Hansen, K. D. Promoter CpG Density Predicts Downstream Gene Loss-of-Function Intolerance. *American Journal of Human Genetics* **107.** Publisher: Cell Press, 487–498. ISSN: 15376605. (2021) (Sept. 2020).

27. Li, Y. & Abecasis, G. R. MaCH: Using Sequence and Genotype Data to Estimate Haplotypes and Unobserved Genotypes. *Genetic Epidemiology* **119.** arXiv: NIHMS150003 ISBN: 7136777690, 5124–5136. ISSN: 0895-3309 (2010).

28. Kang, H. M., Zaitlen, N. A. & Eskin, E. EMINIM: An Adaptive and Memory-Efficient Algorithm for Genotype Imputation. *Journal of Computational Biology* **17,** 547–560. ISSN: 1066-5277. http://genetics.cs.ucla.edu/eminim. (Mar. 2010).

29. Auton, A. *et al.* A global reference for human genetic variation. *Nature* **526,** 68–74. ISSN: 14764687 (2015).

30. Ségurel, L., Wyman, M. J. & Przeworski, M. Determinants of Mutation Rate Variation in the Human Germline. en. *Annual Review of Genomics and Human Genetics* **15,** 47–70. ISSN: 1527-8204, 1545-293X. https://www.annualreviews.org/doi/10.1146/annurev-genom-031714-125740 (2022) (Aug. 2014).

31. McLaren, W. *et al.* The Ensembl Variant Effect Predictor. *Genome biology* **17.** Publisher: BioMed Central Ltd., 122. ISSN: 1474-760X. http://www.ncbi.nlm.nih.gov/pubmed/27268795 (2021) (June 2016).

32. Nassar, L. R. *et al.* The UCSC Genome Browser database: 2023 update. eng. *Nucleic Acids Research,* gkac1072. ISSN: 1362-4962 (Nov. 2022).

33. Saxonov, S., Berg, P. & Brutlag, D. L. A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters. *Proceedings of the National Academy of Sciences* **103.** Publisher: National Academy of Sciences, 1412–1417. ISSN: 0027-8424. http://www.pnas.org/cgi/doi/10.1073/pnas.0510310103 (2021) (Jan. 2006).

34. Cohen, N. M., Kenigsberg, E. & Tanay, A. Primate CpG Islands Are Maintained by Heterogeneous Evolutionary Regimes Involving Minimal Selection. en. *Cell* **145,** 773–786. ISSN: 0092-8674. https://www.sciencedirect.com/science/article/pii/S009286741100482X (2023) (May 2011).

35. Kristjánsdóttir, K. *et al.* Population-scale study of eRNA transcription reveals bipartite functional enhancer architecture. *bioRxiv,* 426908. https://doi.org/10.1101/426908 (2018).

36. Barrett, T. *et al.* NCBI GEO: archive for functional genomics data sets—update. *Nucleic Acids Research* **41,** D991–D995. ISSN: 0305-1048. https://doi.org/10.1093/nar/gks1193 (2023) (Jan. 2013).