

Cost-effective solutions for high-throughput enzymatic DNA methylation sequencing

Amy Longtin^{1,2}, Marina M. Watowich^{1,2}, Baptiste Sadoughi^{3,4}, Rachel M. Petersen^{1,2}, Sarah F. Brosnan⁵, Kenneth Buetow^{3,4,6}, Qiuyin Cai⁷, Michael D. Gurven⁸, Heather M. Highland⁹, Yi-Ting Huang¹⁰, Hillard Kaplan¹¹, Thomas S. Kraft¹², Yvonne A. L. Lim¹³, Jirong Long⁷, Amanda D. Melin^{14,15}, Jamie Roberson¹⁰, Kee-Seong Ng¹⁶, Jonathan Stieglitz¹⁷, Benjamin C. Trumble^{3,4,6}, Vivek V. Venkataraman¹⁴, Ian J. Wallace¹⁸, Jie Wu⁷, Noah Snyder-Mackler^{3,4}, Angela Jones¹⁰, Alexander G. Bick^{2,10,19}, and Amanda J. Lea^{1,2,10,*}

¹Department of Biological Sciences, Vanderbilt University, Nashville, TN, USA

²Evolutionary Studies Initiative, Vanderbilt University, Nashville, TN, USA

³Center for Evolution and Medicine, Arizona State University, Tempe, AZ, USA

⁴School of Human Evolution and Social Change, Arizona State University, Tempe, AZ, USA

⁵Departments of Psychology & Philosophy, Neuroscience Institute, Center for Behavioral Neuroscience, and the Language Research Center, Georgia State University, Atlanta, GA, USA

⁶Institute of Human Origins, Arizona State University, Tempe, AZ, USA

⁷Division of Epidemiology, Department of Medicine, Vanderbilt University Medical Center, Nashville, TN, USA

⁸Department of Anthropology, University of California, Santa Barbara, CA, USA

⁹Department of Epidemiology, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA

¹⁰Vanderbilt Genetics Institute, Vanderbilt University School of Medicine, Nashville, TN, USA

¹¹Institute for Economics and Society, Chapman University, Orange, CA, USA

¹²Department of Anthropology, University of Utah, Salt Lake City, UT, USA

¹³Department of Parasitology, Faculty of Medicine, Universiti Malaya, Kuala Lumpur, Malaysia

¹⁴Department of Anthropology & Archaeology, University of Calgary, 2500 University Dr NW, Calgary, AB, Canada.

¹⁵Department of Medical Genetics, 3330 Hospital Drive NW, HMRB 202, Calgary, AB T2N 4N1, Canada.

¹⁶Department of Medicine, Faculty of Medicine, Universiti Malaya, Kuala Lumpur, Malaysia

¹⁷Toulouse School of Economics, Toulouse, France

¹⁸Department of Anthropology, University of New Mexico, Albuquerque, New Mexico, USA

¹⁹Division of Genetic Medicine, Department of Medicine, Vanderbilt University Medical Center, Nashville, TN, USA

*Correspondence to: amanda.j.lea@vanderbilt.edu

ABSTRACT

Characterizing DNA methylation patterns is important for addressing key questions in evolutionary biology, geroscience, and medical genomics. While costs are decreasing, whole-genome DNA methylation profiling remains prohibitively expensive for most population-scale studies, creating a need for cost-effective, reduced representation approaches (i.e., assays that rely on microarrays, enzyme digests, or sequence capture to target a subset of the genome). Most common whole genome and reduced representation techniques rely on bisulfite conversion, which can damage DNA resulting in DNA loss and sequencing biases. Enzymatic methyl sequencing (EM-seq) was recently proposed to overcome these issues, but thorough benchmarking of EM-seq combined with cost-effective, reduced representation strategies has not yet been performed. To do so, we optimized Targeted Methylation Sequencing protocol (TMS)—which profiles ~4 million CpG sites—for miniaturization, flexibility, and multispecies use at a cost of ~\$80. First, we tested modifications to increase throughput and reduce cost, including increasing multiplexing, decreasing DNA input, and using enzymatic rather than mechanical fragmentation to prepare DNA. Second, we compared our optimized TMS protocol to commonly used techniques, specifically the Infinium MethylationEPIC BeadChip (n=55 paired samples) and whole genome bisulfite sequencing (n=6 paired samples). In both cases, we found strong agreement between technologies ($R^2 = 0.97$ and 0.99, respectively). Third, we tested the optimized TMS protocol in three non-human primate species (rhesus macaques, geladas, and capuchins). We captured a high percentage (mean=77.1%) of targeted CpG sites and produced methylation level estimates that agreed with those generated from reduced representation bisulfite sequencing ($R^2 = 0.98$). Finally, we applied our protocol to profile age-associated DNA methylation variation in two subsistence-level populations—the Tsimane of lowland Bolivia and the Orang Asli of Peninsular Malaysia—and found age-methylation patterns that were strikingly similar to those reported in high income cohorts, despite known differences in age-health relationships between lifestyle contexts. Altogether, our optimized TMS protocol will enable cost-effective, population-scale studies of genome-wide DNA methylation levels across human and non-human primate species.

INTRODUCTION

Characterizing and understanding epigenomic variation is important for evolutionary and developmental biology, geroscience, and biomedicine. DNA methylation—the covalent addition of methyl groups to cytosines—is a semi-malleable and environmentally-responsive epigenetic modification integral to gene regulation in many species, including our own (1). Because DNA methylation moderates gene expression throughout the life course, it is critical for processes such as development (2–4), cell programming (5), tissue specificity (6), aging (7–11), and disease progression (12–14). For example, changes in DNA methylation are considered a “hallmark” of the aging process, with most studies reporting age-associated gains in methylation in hypomethylated regions (e.g., promoters and transcribed regions) and age-associated losses in methylation in hypermethylated regions (e.g., heterochromatic regions, Polycomb-repressed regions) (15–17). These age-related patterns are so consistent that DNA methylation patterns have been used to construct molecular clocks that reliably predict chronological age (18,19). Further, because DNA methylation is known to respond to environmental inputs, it has been implicated as a mechanism through which diverse experiences can “get under the skin” to impact long-term physiology and health (e.g., famine (20–24), psychosocial stress (25–29), or infection (30–33)).

To profile genome-wide DNA methylation at scale, most studies rely on reduced representation methods: human studies have largely favored microarrays, while non-human studies have favored reduced representation bisulfite sequencing (RRBS) due to the lack of species-specific microarrays (34). Both methods quantify DNA methylation at a subset (1-5%) of CpGs in the genome, and thus provide a cost-effective strategy relative to genome-wide assays (e.g., whole genome bisulfite sequencing (WGBS)). For example, the Infinium MethylationEPIC v2.0 BeadChip, or EPIC array, covers ~930K CpG sites including functional elements identified by the ENCODE project (35), DNase hypersensitive sites, and putatively important sites for human disease and development (36,37). In contrast, RRBS fragments DNA using the *Msp*1 enzyme that cuts DNA at CCGG motifs, which following size selection, enriches for 1-5% of the genome with high CpG content such as CpG islands and gene bodies (34,38). Importantly, both microarrays and RRBS rely on sodium bisulfite, which converts unmethylated cytosines to thymine while leaving methylated cytosines protected from conversion. This chemical reaction requires high pHs and temperatures, which can cause unwanted DNA fragmentation and damage,

especially to unmethylated cytosines (39). Ultimately, such damage can create difficulties during library preparation as well as biases in the downstream data (39–41).

Enzymatic methyl sequencing (EM-seq) offers a useful alternative to bisulfite sequencing with several key benefits: EM-seq relies on enzymatic rather than chemical conversion of unmethylated cytosines to thymine, resulting in substantially less DNA damage (40). As a result, whole genome EM-seq has been shown to recover more CpGs sites, have lower duplication rates, have better between-replicate correlations, and require less DNA input than WGBS (40). However, existing EM-seq benchmarked protocols rely on whole genome rather than reduced representation strategies, hindering their adoption especially for population-scale studies. To address this gap, Twist Biosciences recently created a hybrid capture panel that targets ~4 million CpG sites in the human genome and is compatible with EM-seq. The first generation Twist methylation capture probe set uses ~550k probes to target functionally relevant CpG sites (e.g., those in enhancers, gene bodies, and near transcription start sites) and to cover 95% of CpG sites included on the widely used EPIC array (42–45). Importantly, off the shelf, this protocol is similar or lower in cost to existing reduced representation approaches. For example, it provides coverage of approximately four times as many CpG sites relative to the EPIC array at one fourth the cost—a ~16-fold gain in the data-to-price ratio.

Here, we aimed to develop and benchmark an optimized and further cost-reduced version of the targeted methylation sequencing (TMS) approach suitable for population-scale studies, including both human and non-human primate (NHP) studies (Figure 1A). To do so, we built upon the off the shelf TMS protocol (Figure 1B), which recommends 8 plexing of samples per capture reaction and 200 ng of DNA input, and tested four multiplexing strategies (12, 24, 48, and 96 plex, using 200 ng of sample input; Figure 1C). We also tested five DNA input amounts (25, 50, 100, 200, and 400ng, using the 12-plex strategy) and other minor protocol modifications such as varying the annealing temperature during hybrid capture and varying the method used for DNA fragmentation (Figure 1C). Following optimization, we assessed: 1) the robustness of our protocol through a direct comparison with the EPIC array and WGBS; 2) the extension of optimized TMS for use in NHP species; and 3) the utility of our protocol to uncover biological effects of interest (see Table 1 for sample sizes and sample information; Figure 1C). Specifically, we tested for age effects on DNA methylation in two non-industrial populations that exhibit minimal evidence for age-related increases in cardiometabolic diseases—the Tsimane of Bolivia

and the Orang Asli of Peninsular Malaysia (46–48). We were interested in the degree to which these groups exhibit epigenomic patterns similar to what has been reported for Western, post-industrial settings, or alternatively, if they exhibit novel patterns that may be associated with their favorable aging trajectories. Overall, we found that we were able to miniaturize and optimize the TMS protocol to ~\$80 per sample, while maintaining data quality and comparability to existing methods, extending to NHP species, and applying the protocol to uncover age effects in two under-represented human populations.

Table 1: Study populations and sample information for each experiment (names of experiments are as described in Results). NHP = non human primate, F = female, M = male, VUMC = Vanderbilt University Medical Center. See also Table S1 for sample metadata and read depth.

Population	Tissue type	Sample size	Sample sex	Mean sample age (range)	Experiment
Tsimane (human)	Whole blood	n=192	103 F; 89 M	49.6 years old (18.0–83.6)	1, 2, 3, 6, & 8
VUMC (human)	Whole blood	n=55	Not available	Not available	5
Orang Asli (human)	White blood cells	n=88	46 F; 42 M	35.3 years old (18–78)	4 & 9
Rhesus macaque (NHP, <i>Macaca mulatta</i>)	Heart (16), kidney (16), adrenal (16), spleen (16), lung (16), liver (16)	n=96	39 F; 57 M	10.57 years old (3.18–19.93)	7 & 8
Gelada (NHP, <i>Theropithecus gelada</i>)	Whole blood	n=68	21 F; 47 M	Unknown; all animals >5 years old	7
Capuchin (NHP, <i>Sapajus apella</i>)	Whole blood	n=28	19 F; 9 M	19.4 years old (9.0–41.0)	7
Total		n=527			

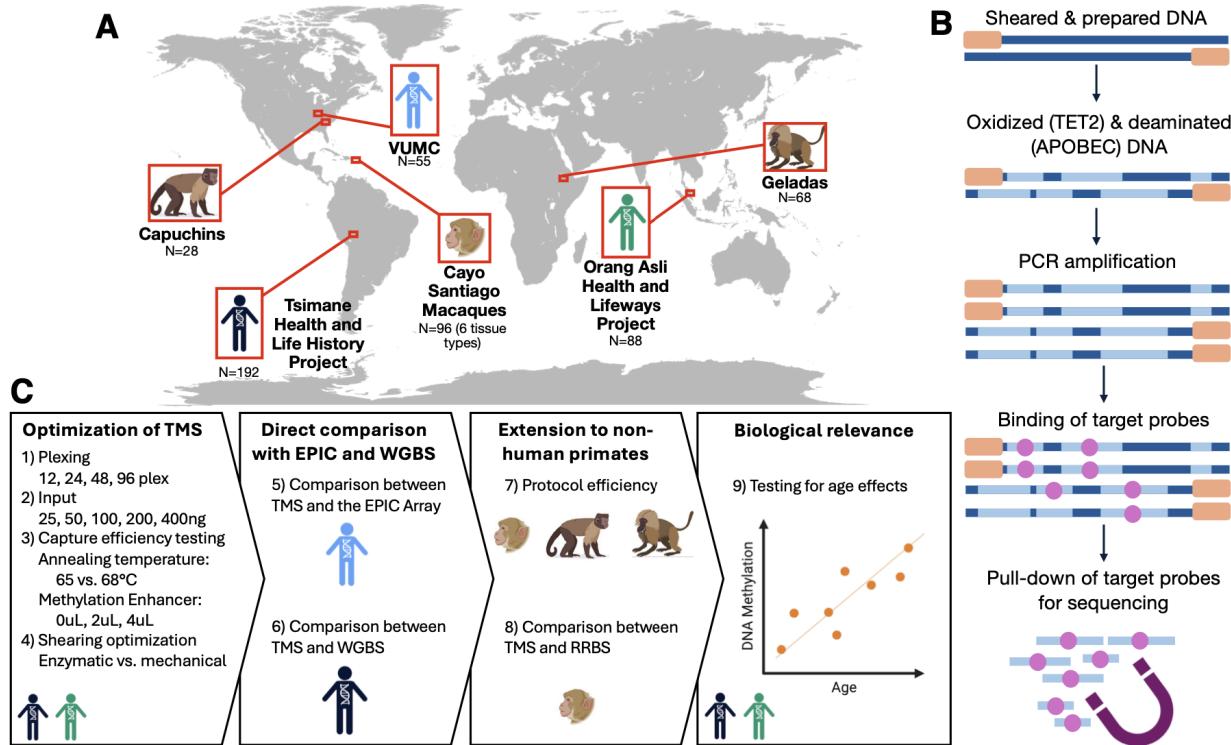


Figure 1. Experimental design and study populations.

(A) To optimize the TMS protocol, we used samples from three human and three NHP populations: the Tsimane of Bolivia, a Vanderbilt University Medical Center cohort, the Orang Asli of Malaysia, rhesus macaques from Cayo Santiago in Puerto Rico, capuchin monkeys from captive sites throughout the United States, and gelada monkeys from Ethiopia.

(B) The TMS protocol begins with DNA shearing and adapter ligation. Next, two enzymes, TET2 and APOBEC, are used to oxidize and deaminate the DNA. TET2 recognizes methyl groups attached to cytosines and converts them to Ca/g. APOBEC follows TET2 and converts the unmethylated cytosines to uracils. Following PCR amplification (which converts uracils to thymines), hybrid capture is used to enrich for targeted regions of the genome. Samples are then submitted for high throughput sequencing.

(C) Overview of experiments and analyses. The samples used for each set of experiments are noted by a population-specific icon.

RESULTS

Data quality is robust to a range of multiplexing strategies, input amounts, and protocol modifications

Experiments 1 & 2: Varying multiplexing strategies and input amounts

Using DNA from Tsimane individuals, we tested four multiplexing strategies (12, 24, 48, and 96 plex, using 200ng of sample input) and five DNA input amounts (25, 50, 100, 200, and 400ng, using the 12-plex strategy). Raw quality control metrics such as percent CHH methylation (a proxy for the rate at which unmethylated cytosines are converted to thymine) and mapping efficiency (percent of reads uniquely mapped to the genome) were high for all samples. Mapping efficiency was consistent across plexing strategies (average mapping efficiency: 12-plex = 71.9%, 24-plex = 72.9%, 48-plex = 72.5%, and 96-plex = 73.5%; ANOVA: F-value =

0.843, p-value = 0.472; Figure 2A) but affected by input amount, with higher DNA input having greater higher mapping efficiency (ANOVA: F-value = 13.57, p-value < 0.001, Figure 2B, Table S2). CHH methylation was consistently well below 1%, indicative of high conversion rate across all plexing and input strategies (range=0.1-0.27%; Figure S1, Table S3 & S4) (49).

After filtering for CpG sites with >5x coverage that were within the Twist probe set (+/- 200bp) and that were covered in the majority of samples in a given experiment, we retained an average of 4,197,008 CpG sites (s.d.=546,767) across plexing experiments and 4,051,941 CpG sites (s.d. = 93,106) across input experiments (Tables S5 & S6). On average, this represented 96.42% and 92.19% coverage of the TMS probe set across the plexing and input experiments, respectively (Table S7 & S8). In addition to consistently recovering the expected set of CpGs, we also observed extremely repeatable methylation levels across the plexing and input experiments (all $R^2 > 0.99$; Table S9 & S10). The CpGs covered by our experiments were distributed across diverse genomic annotations, and the median DNA methylation levels within a given annotation displayed expected patterns (Figure 2C & D) (50). For example, we observed high levels of methylation in quiescent and heterochromatin regions and low levels of methylation in promoters and transcribed regions.

Experiments 3 & 4: Optimizing capture efficiency and DNA shearing strategies

In experiments 1 and 2, we used the recommended 65°C annealing temperature during the hybrid capture step—where prepared DNA is bound to the capture probe set to select CpG sites of interest—and the recommended 2uL of methylation enhancer, which increases the efficiency of this reaction. Here, we found that ~3/4 of all of our mapped reads were “on-target”, meaning that they overlapped with the designed probe set and represented successful hybrid capture (Tables S7 & S8). This suggests that ~¼ of reads are “off target” and randomly distributed across the genome rather than within our regions of interest (Tables S9 & S10). We therefore performed a third experiment to test two protocol modifications that might decrease the off-target proportion: we increased the annealing temperature (testing 65°C or 68°C) and we varied the amount of methylation enhancer (testing 0uL, 2uL, or 4uL).

In experiment 3, we found that increasing the annealing temperature from 65°C to 68°C resulted in a lower proportion of off-target reads (ANOVA: F-value = 84.2, p-value < 0.0001; Figure 2E, Figure S2, Table S11). Across samples annealed at 65°C, an average of 78.5% of

reads were on-target, while this number rose to 84.2% at 68°C. However, this increase in capture efficiency came at a cost to the breadth of CpG sites covered: across samples annealed at 65°C, we observed coverage of on average 92.0% of the probe set, while this number fell to 72.2% for samples annealed at 68°C (Figure 2F; Tables S12 & S13). This suggests that higher annealing temperatures lead to greater but more specific binding during the hybrid capture step, and thus the increased capture efficiency comes at the expense of recovering all the expected CpG sites. In general, we did not find meaningful differences across methylation enhancer amounts and we therefore excluded this reagent from downstream experiments (Figure 2E & F). Given the loss of certain genomic regions at 68°C, downstream experiments focused on a 65°C annealing temperature.

We next performed a fourth experiment focused on protocol optimization, in which we varied the strategies used to fragment genomic DNA prior to EM-seq library preparation: specifically, we tested mechanical shearing via Covaris sonication against enzymatic shearing with the NEBNext UltraShear reagent. Mechanical shearing is the current standard approach but is expensive, requires special equipment, and is labor intensive. Conversely, enzymatic fragmentation is cheaper, does not require special equipment, and is more compatible with automation. For experiments 3 and 4, we used the 96-plex strategy and 200 ng of sample input, since experiments 1 and 2 suggested that data quality does not suffer from higher plexing strategies.

Enzymatic shearing resulted in a similar number of covered sites as was previously observed with mechanical shearing (n=4,591,123 and 4,523,981 filtered CpG sites for the 10 and 20 minute protocols, respectively). Average site-specific methylation levels were also highly concordant between approaches (mechanical versus 10 min enzymatic: $R^2 = 0.9875$; mechanical versus 20 min enzymatic: $R^2=0.9876$; 10 min versus 20 min enzymatic: $R^2=0.9944$; Figures S3 & S4). This was also true when we focused on a subset of paired samples processed using both methods (n=3; mechanical versus 10 min enzymatic: average $R^2 = 0.971$; mechanical versus 20 min enzymatic: average $R^2=0.971$; 10 min versus 20 min enzymatic: average $R^2=0.987$, Table S14). From these experiments, we concluded that enzymatic fragmentation can be substituted into the protocol with no loss to data quality.

We also used these data, which represent our “best” protocol (96-plex, 200ng input, 65°C annealing, no methylation enhancer, enzymatic shearing), to understand a critical aspect of

experimental design—how many reads one would need to generate to achieve a given average coverage per CpG site (Figure S5). In general, we observe a 1:2 relationship between the number of mapped, paired end reads (in millions) and average coverage, such that ~25M mapped paired end reads translates to ~50x average coverage per CpG site.

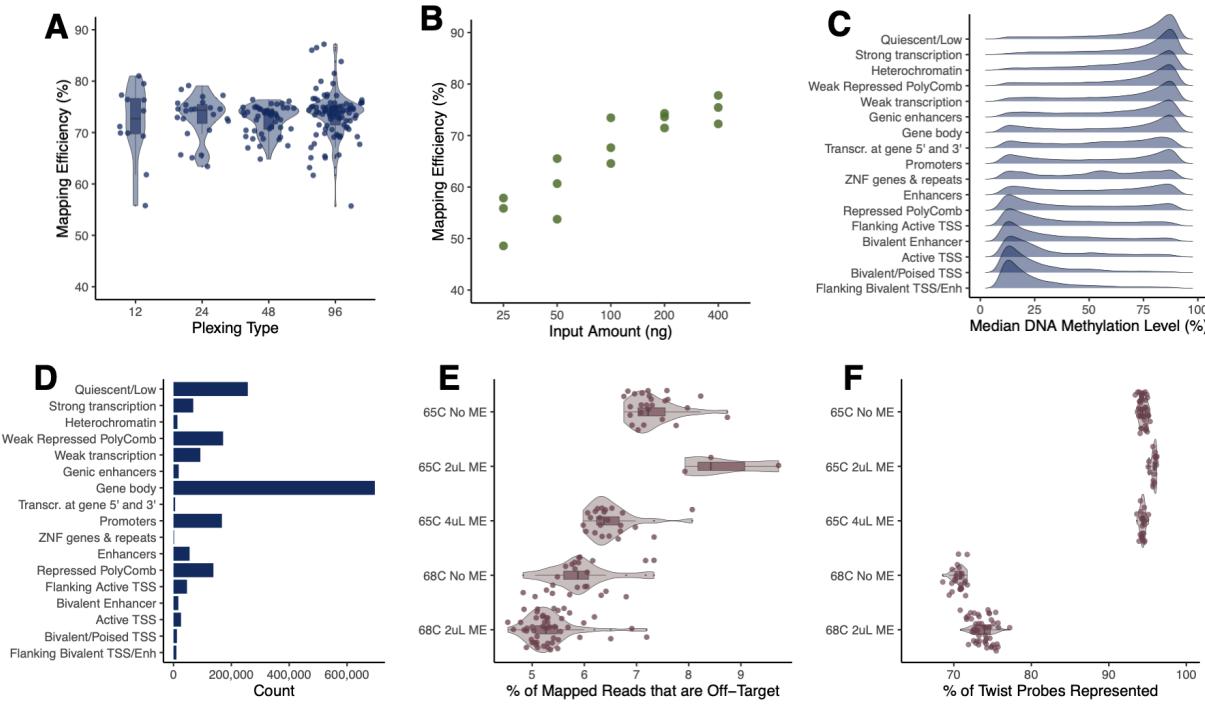


Figure 2. Optimized TMS produces high-quality DNA methylation data across a range of plexing strategies, input amounts, and protocol modifications.

(A) High (>70%) mean mapping efficiency across plexing strategies. Each point represents a sample within a plexing strategy and the y-axis represents the percent of reads uniquely mapped per sample.

(B) Mapping efficiency increases as input amount increases. Each point represents a 12-plex pool made with varying DNA input amounts per sample, the y-axis represents the percent of reads uniquely mapped per sample.

(C) Median DNA methylation levels for reads located within different chromHMM genomic annotations from NIH Roadmap Epigenomics (using data from the 96-plex, 200 ng input from experiment 1).

(D) The number of CpG sites falling within different chromHMM genomic annotations (using data from the 96-plex, 200 ng input from experiment 1).

(E) Percent of reads that are not within Twist probes (off-target reads) following protocol modifications to annealing temperature and methylation enhancer (ME) volume. For each set of protocol conditions, the x-axis represents the percent of mapped reads that do not overlap with the Twist probe set.

(F) Percent of Twist probes that are represented following protocol modifications to adjust the annealing temperature and ME volume. For each set of protocol conditions, the x-axis represents the percentage of Twist probes that were represented by at least 1 read.

Epigenomic profiles measured with TMS recapitulate those measured with the EPIC array and WGBS

Experiment 5: Comparison of TMS to the EPIC array

To ensure that TMS could perform comparably to the most popular current reduced-representation technology (the EPIC array), we generated paired data for 55 samples

using both platforms (and following the 96-plexing, 200 ng input TMS protocol from experiment 1). After filtering, we analyzed 682,295 CpG sites common to both technologies, and found high concordance between per-site DNA methylation levels averaged across all individuals in the dataset ($R^2 = 0.97$; Figure 3A). This remained true when we subsetted to 235,234 variably methylated sites (i.e., sites with methylation levels $>10\%$ and $<90\%$; mean $R^2 = 0.83$; Figure 3B). Because methylation patterns are relatively consistent within the human genome, we also confirmed that these correlations were higher for EPIC-TMS data generated from the same sample compared to EPIC-TMS comparisons made between random pairs of samples (mean R^2 for non-variable sites: 0.92, mean R^2 for variable sites only: 0.75; Figure S6).

Of note, these analyses reconfirmed a known bias in the EPIC array data (51), which does not allow for methylation levels of 100% (such that the correlation between average TMS- and EPIC-measured DNA methylation levels is slightly off the $x=y$ line; Figure 3A). This is because EPIC-derived DNA methylation levels are represented as beta values, calculated as the ratio of the intensity of the methylated bead type to the total locus intensity plus an offset value. Due to the addition of the offset value, beta values of 1 are mathematically impossible.

Experiment 6: Comparison of TMS to WGBS

For further validation, we also generated WGBS data for 6 samples included in experiment 3 (96-plexing, 200 ng input, 65°C annealing temperature, no ME, mechanical fragmentation). After filtering and merging with the TMS data, we retained 3,078,771 CpG sites covered by both the TMS and WGBS approaches. For these sites, the average methylation levels observed across technologies was highly correlated ($R^2: 0.9871$; Figure 3C). We also found that the genome-wide distribution of DNA methylation levels derived from WGBS was more similar to TMS than to the EPIC array, specifically in that it included many sites with average methylation levels of 100% or close to 100%, as expected (Figure 3D, Figure S7).

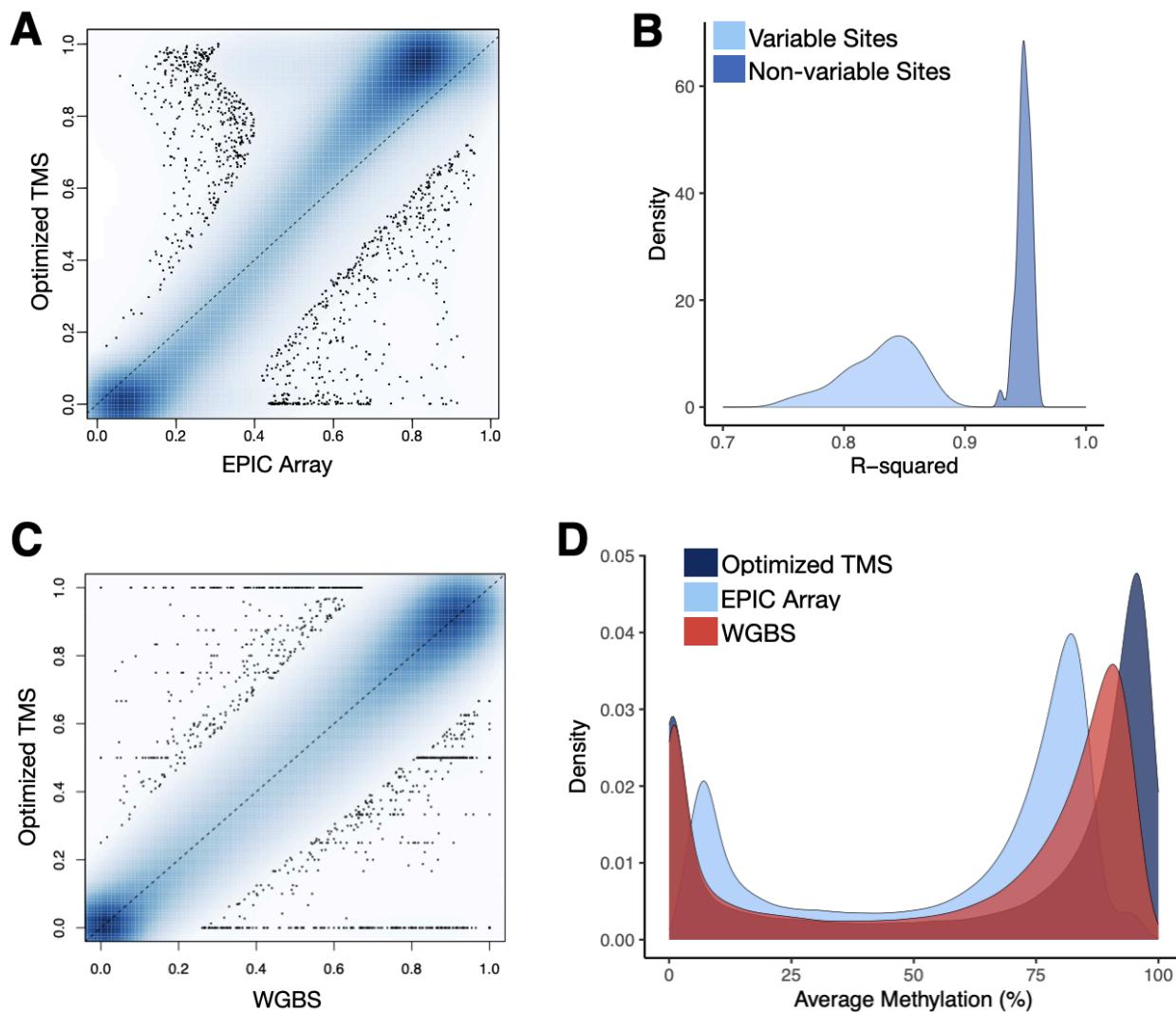


Figure 3. Optimized TMS recapitulates DNA methylation levels measured with the EPIC array and WGBS.

(A) Correlation in DNA methylation levels for EPIC array versus TMS ($R^2=0.97$). Site-level methylation averaged across 6 samples measured using the EPIC array (x-axis) and 96-plex, 200 ng input TMS (y-axis). Each point represents a site measured across both technologies and R^2 value was generated using linear modeling. Sites filtered to $>5X$ coverage in $>75\%$ of samples within each technology.

(B) Histogram of R^2 values calculated for each individual sample (i.e., comparing DNA methylation levels measured on both technologies for a given sample). R^2 values are provided when all CpG sites common to both technologies are included, as well as when only variably methylated CpG sites are included.

(C) Correlation in DNA methylation levels for WGBS versus TMS ($R^2=0.9871$). Site-level methylation averaged across 6 samples measured using WGBS (x-axis) and 96-plex, 200 ng input TMS (y-axis). Each point represents a site measured across both technologies and R^2 value was generated using linear modeling. Sites filtered to $>5X$ coverage in $>75\%$ of samples within each technology.

(D) Density plot of the average DNA methylation levels detected for matched sites between the three technologies (713,282 sites). Notably, the EPIC array is biased against DNA methylation levels of 100%, as previously observed by Shu et al. (2021) (51) and explained by the equation used to calculate beta values.

TMS can be effectively applied to non-human primate species

Experiment 7: Applying TMS to capuchin, rhesus macaque, and gelada samples

To enable epigenomic analyses in our close primate relatives, we also tested whether TMS (96-plex, 200ng input protocol from experiment 1) could be effectively applied to three NHP species: capuchins (*Sapajus apella*; n=28 samples from blood), rhesus macaques (*Macaca mulatta*; n=96 samples from 6 tissues (see Figure S8 and Table S15), and geladas (*Theropithecus gelada*; n=68 samples from blood). While the probe set is designed from the human genome, NHP species share high levels of sequence homology with humans, especially in coding regions and regions near genes (52), leading us to hypothesize that a majority of CpG sites would be recovered. We mapped the Twist probe sequences to each of the NHP genomes to confirm this intuition, and from this analysis expected to capture 3.0-4.8 million CpG sites across the three species (Figure 4B). Importantly, for the rhesus macaque samples, we also generated paired RRBS data and compared our TMS results to a technology that does not rely on hybrid capture.

When examining initial quality control metrics, we found that all three NHP species had high mapping efficiencies (average = 81.96% for capuchins, 82.62% for geladas, and 81.35% for macaques; Figure 4A). Further, the average CHH methylation levels were all extremely low (<1%), again suggesting high conversion rates (Figure S9). Following filtering, we recovered $\sim\frac{1}{2}$ to $\frac{3}{4}$ of expected CpG sites in the NHP datasets (3,343,133 in capuchin, 5,387,280 in gelada, and 5,486,073 in rhesus macaque). The number of sites recovered scales generally with divergence time (capuchins share a common ancestor with humans 35-45 million years ago, geladas and rhesus macaques share a common ancestor with humans 23-28 million years ago) (53). In all species, we were able to reliably measure more sites than would be typical of RRBS (see below), and we note that some of the between-species variation in performance could be explained by heterogenous read depth (Table S16) as well as reference assembly quality.

When examining average DNA methylation levels across species, we found that, as expected, all exhibited bimodal genome-wide profiles similar to humans (Figure 4F). Further, because the rhesus macaque samples were derived from 6 different tissue types (Figure S8, Table S15), we also confirmed that samples displayed expected tissue-specific epigenetic patterns. Specifically, we demonstrated that a Principal Components Analysis (PCA) was able to reliably separate samples by tissue type (Figure 4C), as has been observed in previous studies using both bisulfite sequencing and the EPIC array (54-56).

Experiment 8: Comparison of TMS to RRBS

Studies of NHP species have historically relied on RRBS because of the species-specificity of microarray technologies and the cost barrier of WGBS (57–59). To test how our optimized TMS protocol compares to RRBS, we generated paired data for all 96 rhesus macaque samples. After filtering both datasets to 721,766 common CpG sites, we found a high concordance of the average DNA methylation levels estimated by both technologies ($R^2=0.97$; Figure 4D, Figure S10). This remained true when we subsetted specifically to 92,692 variably methylated CpG sites (i.e., sites with average DNA methylation levels >0.1 and <0.9 ; $R^2=0.5945$; Figure 4E).

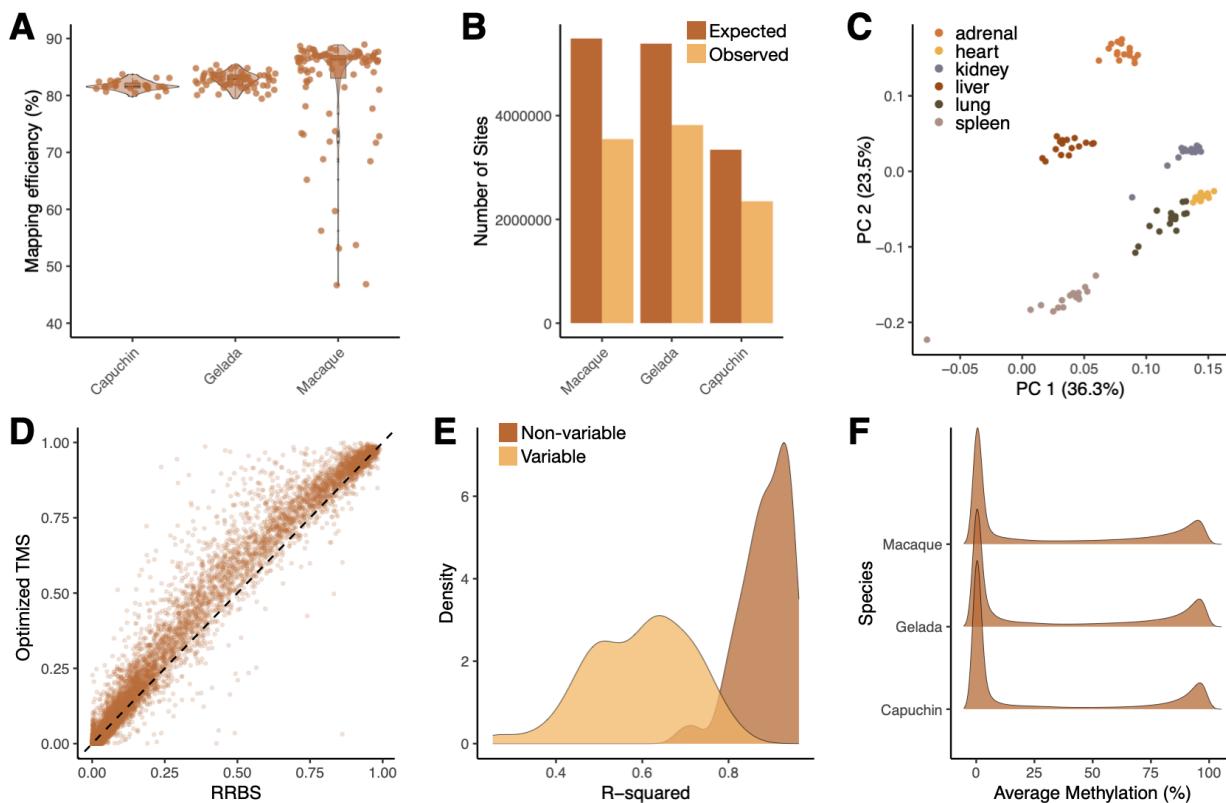


Figure 4. Optimized TMS performs well in non-human primate species and when compared to RRBS.

(A) Optimized TMS in NHPs results in high mapping efficiencies despite the use of human-specific probes. Here, each of the species are mapped to their own genome. TMS data from experiment 1 (96 plex, 200ng input) are included for comparison. We hypothesize that low mapping efficiency in certain rhesus macaque samples is due to variations in types of tissues processed.

(B) Number of expected and observed CpG sites covered in each NHP genome. Expected sites were derived from mapping the Twist probes to each NHP genome, while observed sites represent those detected with an average coverage $>5X$ in $>75\%$ of samples.

(C) Principal components analysis of TMS-derived DNA methylation levels from rhesus macaque samples spanning six distinct tissues.

(D) Similar levels of DNA methylation are detected on a site-by-site basis using RRBS (x-axis) and optimized TMS (y-axis) ($R^2=0.97$).

(E) Density plot of LM R^2 when comparing data generated via optimized TMS and RRBS for the same rhesus macaque sample. R^2 values are provided when all CpG sites common to both technologies are included, as well as when only variably methylated CpG sites are included.

(F) Density curves of the average methylation calculated at a given site across samples for each of the NHP species. Curves show expected bimodal distribution in which much of the CpG sites are either hypomethylated or hypermethylated.

TMS reveals consistent age-associated DNA methylation patterns in diverse populations

We next sought to test whether TMS could reliably detect biological signatures of aging and—because we generated data from several human populations—provide novel insights into how diverse environments and ecologies affect human molecular aging. Much of the human data we generated in optimizing TMS came from two Indigenous, non-industrial, subsistence-level populations—the Tsimane of Bolivia, who practice small-scale horticulture and foraging, and the Orang Asli of Peninsular Malaysia, who are comprised of multiple ethnolinguistic subgroups practicing mixed subsistence strategies (60–63). Both groups have participated in long-term, integrated studies of anthropology and health (47,48), which have revealed minimal evidence for age-associated increases in non-communicable diseases such as cardiovascular disease, obesity, and hypertension that are common in high income, post-industrial contexts (64,65).

Consequently, we were interested in understanding the degree to which age-associated DNA methylation patterns in these groups broadly recapitulate what has been reported in cohorts from high-income countries. More generally, very little population-scale work has characterized epigenomic aging in populations outside of high-income countries (but see (29,66,67)).

When we tested for age effects on DNA methylation levels in each population (focusing on variably methylated CpG sites and using beta-binomial models (68)), we found that 22% of tested CpGs sites ($n=229,727$, $FDR<5\%$) were significantly associated with age in the Orang Asli ($n=88$, age range = 18–78). Of the significant age-associated sites, 40% gained methylation with age and 60% lost methylation with age, recapitulating global patterns observed in previous studies (15,69). In the Tsimane ($n=94$, age range = 18–75), 0.21% of tested CpG sites ($n=1,979$) were significantly associated with age, with the lower number of age-associated sites potentially driven by the more restricted age distribution in this sample. Specifically, the Tsimane dataset included 83% of samples >40 years old compared to the Orang Asli dataset, which included 31% of samples >40 years. In the Tsimane, we again observed a global bias toward hypomethylation with age (62% of significant age-associated sites).

Next, we compared the age effect estimated for each CpG site in the Tsimane and Orang Asli with age effects estimated from a modern, post-industrial Swedish cohort previously published in (70) ($n=421$, age range = 14–94, data generated on the Human Methylation450 BeadChip). Across the 64,084 variably methylated CpG sites measured in all datasets, estimates

of the age effect were similar (Pearson correlation: Tsimane v. Orang Asli: $R = 0.24$, $p < 10^{-16}$; Orang Asli v. Swedish: $R = 0.59$, $p < 10^{-16}$; Tsimane v. Swedish: $R = 0.35$, $p = 0$). Among CpG sites that were significantly associated with age (FDR < 10%), we observed even higher correlations in age effects (Tsimane v. Orang Asli: $n=259$, $R = 0.72$, $p = 1.07 \times 10^{-42}$; Orang Asli v. Swedish: $n=20642$, $R = 0.76$, $p = 0$; Tsimane v. Swedish: $n=424$, $R = 0.66$, $p = 3.30 \times 10^{-54}$; Figure S11). Interestingly, the overlap of significant age-associated CpGs sites was more consistent between the two subsistence-level populations than between either subsistence-level population and the Swedish cohort (Fisher's exact test: Tsimane v. Orang Asli: $OR = 2.22$, $CI = 1.8 - 2.7$, $p = 7.81 \times 10^{-18}$; Orang Asli v. Swedish: $OR = 2.18$, $CI = 2.0 - 2.3$, $p = 2.71 \times 10^{-205}$; Tsimane v. Swedish: $OR = 1.42$, $CI = 1.06 - 1.94$, $p = 0.02$).

As methylation changes during aging are known to differ by genomic context, we investigated the extent to which age-associated CpG sites were more likely to show gains (hypermethylation) versus losses (hypomethylation) with age within different functional annotations (Figure 5) (59,70). Despite their diverse lifestyles, epidemiological patterns, and genetic backgrounds, we found strikingly similar age patterns between the three populations. For example, all three populations exhibited age-related hypomethylation in quiescent regions and regions weakly repressed by Polycomb complexes, as well as age-related hypermethylation in promoters, bivalent TSS regions, and enhancers (Figure 5; Table S17). However, there were also some functional annotations that showed patterns specific to subsistence-level lifestyles (e.g., enrichment of hypermethylation with age in gene bodies) or to certain populations (e.g., depletion of hypermethylation with age in heterochromatin in Tsimane), suggesting that while some regions of the genome may exhibit “universal” epigenomic changes with age, others may be modified by environmental or lifestyle factors.

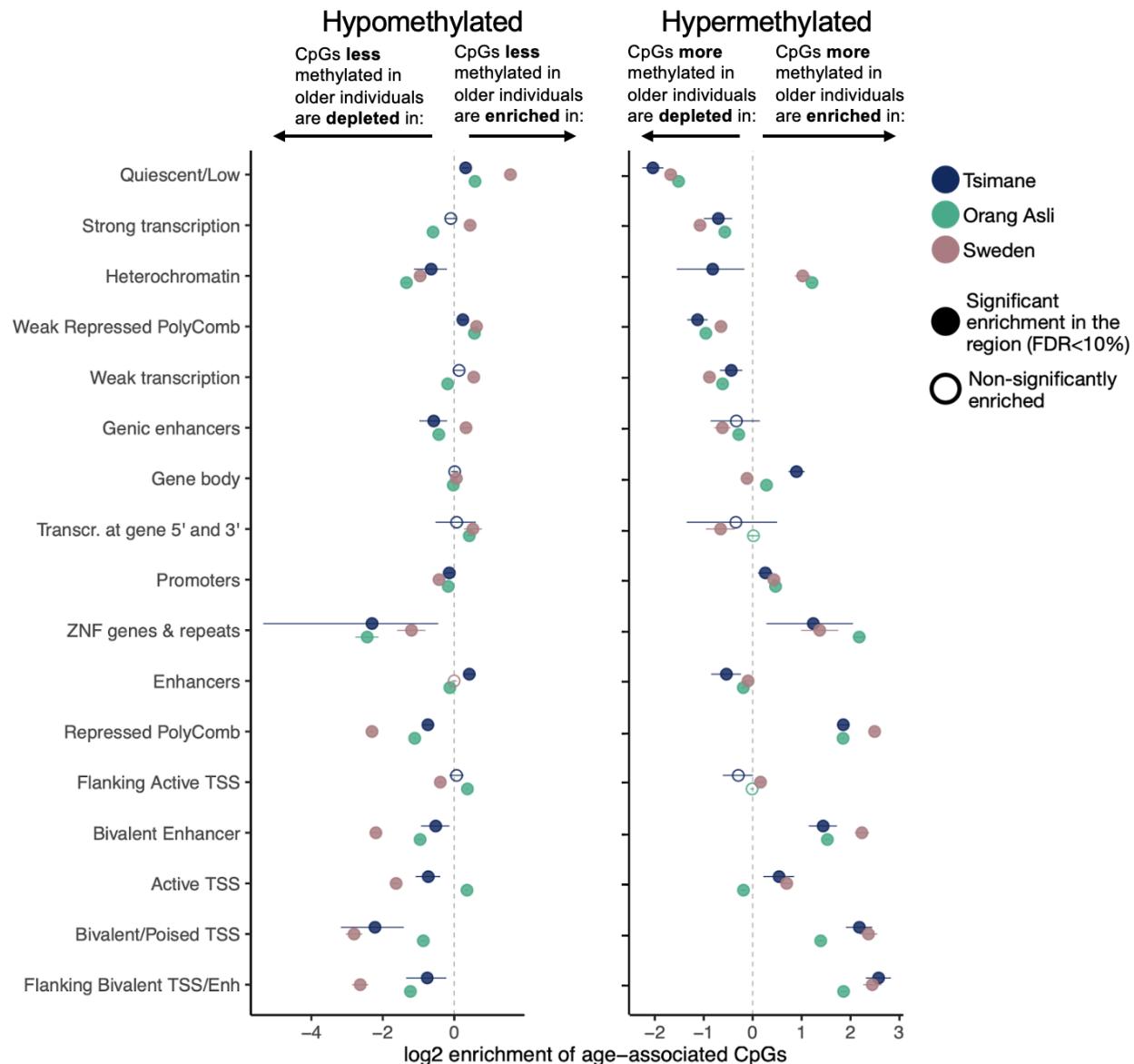


Figure 5: TMS reveals that age-associated DNA methylation patterns are broadly consistent across small scale, subsistence-level populations and post-industrial populations. After quantifying the CpG sites that were differentially methylated with age, we tested the extent to which CpG sites that were hypomethylated with age (i.e., sites with lower methylation levels in older individuals) and hypermethylated with age (i.e., sites with higher methylation levels in older individuals) were more or less likely to fall in different genomic regions. The y-axis is ordered by median methylation levels (as in Figure 2C-D), such that regions higher on the y-axis values have higher average methylation levels regardless of age. We performed enrichment tests for the three populations in which we quantified age effects, the Tsimane (experiment 1 - 96-plex condition, 200ng DNA input, mechanical fragmentation method), the Orang Asli (experiment 4 - 96-plex condition, 200ng DNA input, enzymatic fragmentation method), and a cohort from a high-income country (Sweden) using summary statistics from Johansson et al. (2013).

DISCUSSION

We developed and thoroughly benchmarked a multiplexed, cost-effective version of the TMS protocol and applied it to diverse human populations as well as multiple NHP species. We recommend an optimal protocol for future work (96-plex, 200ng input, 65°C annealing, no

methylation enhancer, enzymatic shearing), but found that data quality remained high acrossplexing strategies, input amounts, and protocol modifications. Importantly, the 96-plex version of the TMS protocol—including sequencing to achieve ~50x coverage per CpG site on the Illumina NovaSeq X—can currently be performed for ~\$80/sample (with roughly half being reagents and labor, and the other half being sequencing on the NovaSeq X platform). Relative to the commonly used EPIC array for human studies, which costs ~\$400 per sample and profiles ~ $\frac{1}{4}$ the number of sites covered by TMS, this represents massive savings enabling larger-scale, population-based studies.

We found high concordance between TMS-derived DNA methylation levels and those derived from other commonly used methods—namely the EPIC array, WGBS, and RRBS. WGBS is the gold standard for comprehensive DNA methylation measurement, but is prohibitively expensive for most studies. RRBS has filled in as a more cost-effective alternative, but due to the stochastic nature of the *Msp*1 digest followed by size selection, not all CpG sites are reliably covered across individuals and missing data can impede downstream analyses (Figure S12). Both WGBS and RRBS also require somewhat specialized laboratory and bioinformatics expertise to execute. As a result of these challenges, microarray-based methods (such as the EPIC array) have become the most commonly used approach in human genomics. Consequently, many popular bioinformatics pipelines and specialized algorithms for DNA methylation data (e.g., epigenetic clocks or cell type deconvolution (72,73)) are currently keyed to microarrays. While DNA methylation levels derived from TMS are strongly correlated with the EPIC array, it is important to keep in mind that: 1) a small subset of sites are not covered by both technologies, and 2) because microarrays output beta values (equivalent to methylated signal/(total signal + an offset)), the relationship between TMS- and EPIC-derived values can not be exactly 1:1 (as noted previously (51)). It is of course possible to recalibrate or rederive popular algorithms such as epigenetic clocks, but we caution that care will be needed when applying microarray-based algorithms to TMS data (as would be true for WGBS or RRBS as well).

The study of DNA methylation in NHP species is deeply important to our understanding of gene regulatory evolution (74–76), comparative aging (58,59,77,78), and environmental impacts on phenotype (59,79). For example, both captive and field-based NHP studies have strongly contributed to our understanding of how social and ecological inputs “get under the

skin” to influence fitness-related traits through changes in DNA methylation (80,81). Because microarrays are generally species-specific (but see (82,83)), NHP studies have struggled to identify easy to use, reduced representation approaches for large-scale studies, with most previous work relying on RRBS (59,78,79,81). Although RRBS is easily adapted for non-human species, TMS can work with smaller amounts of input DNA than bisulfite-based protocols (40), which can be critical for studies of wild or endangered species. While TMS uses capture probes designed from the human genome, NHPs share high levels of sequence similarity, which we show is sufficient to reliably capture 2-3 million CpG. Even though not all ~4 million CpG sites are captured, this still represents a more consistent and cost-effective approach relative to the alternatives. Notably, we find that TMS is effective in both catarrhine (monkeys of Africa and Asia) and platyrhine (monkeys of Central and South America) species, suggesting it may be effective in other members of these clades.

We applied our optimized TMS protocol to profile DNA methylation levels in two subsistence-level, small-scale, Indigenous populations—the Tsimane of Bolivia and the Orang Asli of Peninsular Malaysia. Both populations have partnered with anthropologists and biologists in long-term studies of behavior, health, and genomics (via The Tsimane Health and Life History Project and The Orang Asli Health and Lifeways Project, respectively (47,48)). These studies have revealed minimal evidence for age-associated increases in cardiometabolic disease in subsistence-level contexts (see also (64,84,85)). Conditions such as cardiovascular disease, type 2 diabetes, and hypertension are widely regarded as being inevitable “diseases of aging” in Western societies (86–88). Further, analyses across a gradient of market-integration and acculturation with the Orang Asli have suggested that lifestyle can directly modify age-related health patterns (89). We therefore wondered to what degree the well-established “hallmarks” of epigenomic aging, derived almost entirely from studies of high-income cohorts, were reflected in these populations (16,90). Our results reveal more similarities than differences in how the epigenomes of the Tsimane, Orang Asli, and high-income individuals change with age. This may reflect recent realizations that much of the epigenome is not functionally important in certain developmental stages and/or cell types, and is thus not actively maintained with age. Instead, DNA methylation patterns throughout much of the genome are thought to decay via a stochastic but predictable process, whereby consistently hypermethylated regions lose methyl marks with age and consistently hypomethylated regions gain methyl marks with age (71,91). However, we

find that more functionally important regions of the genome such as enhancers and actively transcribed regions display heterogeneity both in the direction of age effects and how this manifests across populations, leaving clear scope for future studies of environmental and lifestyle effects on epigenomic aging patterns (as well as extensions such as studies of epigenomic age acceleration (66,92,93)). Together, our optimized TMS protocol has the potential to add value and enable larger-scale studies in the many fields that query DNA methylation patterns, such as genetic medicine, developmental biology, evolutionary biology, anthropology, public health, geroscience, and more.

METHODS

Study populations, sample collection, and DNA extraction

Tsimane of Bolivia. The Tsimane are an Indigenous horticulturalists population spread across >90 villages in the Bolivian lowlands and totaling approximately 17,000 people (47). We extracted DNA from 192 venous whole blood (WB) samples collected between the years of 2010–2021 as part of the Tsimane Health and Life History Project (THLHP) and stored long-term at -80C. The THLHP has continuously collected demographic, behavioral, environmental, and health data along with the provision of medical services for over two decades (94) (University of New Mexico IRB: #07-157; University of California, Santa Barbara IRB: #3-21-0652). The sample set for this project included 103 females and 89 males, with a mean age of 54.3 years old (range 18.0–83.6 years old) (see Table 1). Genomic DNA was extracted using the Zymo *Quick-DNA* 96 kit (Zymo Research #D3012) following the manufacturer's instructions.

Orang Asli of Peninsular Malaysia. The Orang Asli consist of ~19 ethnolinguistic groups and a total population of ~210,000 people (48). They traditionally subsist on a mixture of hunting, gathering, fishing, small-scale farming, and trade of forest products (60,61). We extracted DNA from 88 white blood cell (WBC) samples that were collected in 2023 as part of the Orang Asli Health and Lifeways Project (Vanderbilt University IRB #212175). These samples were derived from venous blood draws followed with washing with QIAGEN PureGene red blood cell lysis. Samples were stored in liquid nitrogen upon collection, and at -80C for longer term storage. The Orang Asli sample included 46 females and 42 males, with a mean age of 35.3 years old (range 18–78 years old; Table 1). Genomic DNA was extracted using the Zymo

Quick-DNA/RNA MagBead kit (Zymo Research #R2131) following the manufacturer's instructions.

Vanderbilt University Medical Center cohort. We were granted access to de-identified EPIC array data (Infinium MethylationEPIC v2.0 Kit) and TMS data from 55 paired human whole blood samples. These samples were sourced from a healthy cohort recruited through the Vanderbilt University Medical Center (VUMC) in Nashville, TN USA. Due to IRB restrictions, demographic data or other metadata were not available for these samples.

Rhesus macaques. We obtained extracted DNA from rhesus macaque tissue samples (n=96) collected by the Cayo Biobank Research Unit in partnership with the University of Puerto Rico's Caribbean Primate Research Center (CPRC) (95–99). Beginning in 2016, samples were collected from individuals living on the island of Cayo Santiago, an NIH-managed free-range colony of provisioned rhesus macaques. Specifically, as part of an ongoing population management plan designed by CPRC, select individuals were culled and tissues from all major organ systems were systematically harvested, stored in a fixative buffer, and frozen at -80C (IACUC #A400117). This data set consists of samples from six different tissue types: adrenal, heart, kidney, lung, liver, and spleen, with 16 samples from each type and samples coming from 23 unique individuals (Table S3). This dataset includes samples from 11 females and 12 males, ages 3.2 to 19.9 years old (mean 10.6 years old), collected from 2016–2019 (Table 1; Table S3). Genomic DNA was extracted using the Zymo *Quick-DNA/RNA MagBead* kit (Zymo Research #R2131) following the provided manufacturer's protocols.

Geladas. We extracted DNA from whole blood from 68 geladas; 21 were female and 47 were male and all were considered adult (i.e., over 5 years old, the typical age of reproductive maturation in this species (100)) (Table 1). Gelada samples were collected as part of the Simien Mountains Gelada Research Project (SMGRP) which, since 2017, has carried out annual capture-and-release campaigns to collect morphometric data and whole blood samples from wild Ethiopian geladas. Samples were stored in liquid nitrogen upon collection, and at -80C for longer term storage. This research was conducted with approval from the Ethiopian Wildlife and Conservation Agency (EWCA), and the Institutional Animal Care and Use Committees (IACUCs) at the University of Washington (protocol 4416-01) and Arizona State University (20-1754 R) (see Chiou et al. for more details on sample collection (101)). Genomic DNA was

extracted using the Qiagen DNeasy Blood & Tissue kits (Qiagen #69581) following the provided protocols.

Capuchins. Blood was collected from individuals in the captive capuchin monkey colony at Georgia State University in January 2023. Of the 28 capuchins, 19 were female and 9 were male with an average age of 19.4 years old (range 9–41 years old; Table 1). A trained veterinarian anesthetized the monkeys using 13 mg/kg Ketamine, delivered intramuscularly. Whole blood samples were collected in 6mL EDTA tubes, stored at 4°C, and shipped to Arizona State University where they were flash frozen into 0.5mL aliquots and stored at -80°C until used for analysis. DNA was extracted using the Qiagen DNeasy Blood & Tissue kits (Qiagen #69581) following the manufacturer's protocols. Blood collection was approved by the GSU IACUC (protocol A20018).

Overview of TMS library preparation

We used the Qubit dsDNA assay to determine the quantity of all extracted DNA. The samples were normalized to the desired input amount, and libraries were prepared using the NEBNext® Enzymatic Methyl-seq kit (P/N: E7120L) following a modified version of the manufacturer's protocol that included 9 cycles of PCR for the final library amplification followed by a 0.65X bead cleanup. 71.4 ng of each resulting library was pooled and captured using the Human Methylome panel from Twist Biosciences following the manufacturer's instructions (P/N: 105521). The final post-capture PCR reaction was split into 2 reactions per pool and cleaned with a 1X bead cleanup and then combined. Pool quality was assessed post-hybridization using the Agilent Bioanalyzer and quantified using a qPCR-based method with the KAPA Library Quantification Kit (P/N: KK4873) and the QuantStudio 12K instrument.

Prepared library pools were sequenced on the NovaSeq 6000 at the Vanderbilt Technologies for Advanced Genomics (VANTAGE) Core. We used 150 bp paired-end sequencing and targeted 30-50M paired-end reads per sample. Real Time Analysis Software (RTA) and NovaSeq Control Software (NCS) (1.8.0; Illumina) were used for base calling. MultiQC (v1.7; Illumina) was used for data quality assessments. For each sample, we applied the Illumina DRAGEN Methylation Pipeline v4.1.23 using the custom bed file from Twist Biosciences. The deliverables from DRAGEN consist of FASTQs, bams, cytosine reports (which

include counts of methylated and unmethylated reads per CpG site), and methyl and mapping metric reports.

TMS library preparation for experiments 1 & 2: Varying multiplexing strategies and input amounts

To determine whether TMS can be effectively multiplexed beyond the recommended 8-plex, we used 96 Tsimane samples to test four different multiplexing strategies during capture: 12-, 24-, 48-, and 96-plex. To test whether TMS is robust to DNA input amounts, we tested five input amounts: specifically, 25, 50, 100, 200, and 400 ng of sample were used as input into the EM-seq library prep. Here, we kept the plexing strategy constant (12-plex) and used three Tsimane samples, each represented three times within each pool and included three replicates of a control sample (HG01879 from the 1000 Genomes Project) (102).

TMS library preparation for experiments 3 & 4: Optimizing capture efficiency and DNA shearing strategies

To optimize the capture efficiency of Twist target sites, we tested the use of two different annealing temperatures—65° and 68° C—along with the use of a methylation enhancer (ME)—produced by Twist Biosciences (Catalog #103557) consisting of Tris EDTA buffer to block the binding of off-target probes thereby improving on-target capture efficiency. The specific combinations we explored were: testing a 65°C annealing temperature with 0uL (n=192), 2uL (n=96), and 4uL (n=96) of ME and testing a 68°C annealing temperature with 0uL (n=96) and 2uL (n=192) of ME. These experiments were conducted with 96-plexed Tsimane samples (n=192), and using 200 ng of sample input.

Next, we tested the use of an enzymatic fragmentation method to replace the Covaris (LE220) mechanical shearing method to fragment the DNA. Mechanical shearing is known to decrease library quality through damage to DNA; however, previous use of enzymatic shearing methods have been shown to remove methyl groups from methylated CpG sites, thus biasing detected methylation levels. Here, we aim to test whether enzymatic fragmentation, which is not currently recommended by the TMS protocol, impacts site-level DNAm estimates through a comparison with samples prepared using mechanical shearing. We performed the optimized TMS with enzymatic shearing using 4uL of NEBNext UltraShear (NEB #M7634S/L) for 10 or

20 minutes. This experiment was conducted using 96-plexed samples from the Orang Asli (n=88) and using 200 ng of sample input.

TMS and RRBS library preparation for experiments 7 and 8

To evaluate the efficacy of optimized TMS on three NHP species—macaques, geladas, and capuchins—we applied the 96-plex protocol design from experiment 1 with 200 ng input. To compare rhesus macaque TMS to RRBS, we generated libraries using 150 ng of DNA input in combination with 1ng of lambda phage DNA and 1uL of Msp1—a digestive enzyme which cuts CCGG nucleotide motifs. Next, using NEBNext Ultra II reagents, we performed end repair and adapter ligation to the DNA fragments produced by Msp1 digestion. We then performed bisulfite conversion on the fragments using the EZ-96 DNA Methylation Lightning MagPrep kit (Zymo Research #D5046) following the manufacturer directions. The samples were then PCR amplified for 16 cycles with unique dual indexed sequencing primers. We selected for 180-2000 bp fragments and sequenced on an Illumina NovaSeq S2 flow cell with 2x51bp sequencing (78,103).

Low-level processing of TMS data

For experiments 1, 2, 7, and 8, we used a custom bioinformatics pipeline to process all FASTQ files into counts of methylated versus unmethylated cytosines at each CpG site. For experiments 3, 4, 5, and 6, we used Illumina’s Dynamic Read Analysis for GENomics (DRAGEN) pipeline (104) to process all FASTQ files into counts of methylated versus unmethylated cytosines at each CpG site. Importantly, both our custom pipeline and DRAGEN follow the same general steps and rely on the Bismark suite (105), making them highly comparable. We also processed 7 samples from experiment 4 using both methods to empirically confirm that our custom pipeline and the Illumina DRAGEN pipeline produced near identical results (Figure S13).

For our custom pipeline, we first trimmed adapters using Trimmomatic (version 0.39) (106) and TrimGalore (version 0.6.6) (107) for human and NHP samples, respectively. Following trimming, we used Bismark (version 0.24.0) (105) to map reads to each species’ respective genomes (hg38 for human, mmul10 for rhesus macaque, cimit for capuchin, and tgel1 for gelada). We retained only uniquely mapped reads and used the methylation extractor function

within Bismark to extract counts of methylated versus unmethylated cytosines at each cytosine. These files were further filtered for CpG contexts only.

For all samples, run through either the custom or DRAGEN pipeline, we extracted two measures of data quality that are automatically calculated by Bismark: the percent of reads that mapped uniquely to the reference genome and the average methylation percentage for cytosines in a CHH context. The latter value serves as a commonly used estimate of the efficiency with which a given protocol converts unmethylated cytosines to thymine, because cytosines located outside of CpG contexts are extremely unlikely to be methylated in the mammalian genome (108,109). For experiments 1 and 2, we tested whether multiplexing strategy and input amount impacted mapping efficiency and percent CHH methylation using a one-way ANOVA test, followed by a pairwise t-test in the case of significance, with the ‘aov’ and ‘pairwise.t.test’ functions in the ‘stats’ R package (110).

For each study, we used the BSseq R package (111) to compile count matrices (derived from our custom pipeline or DRAGEN) across samples and to perform region, coverage, and missingness filtering. For experiments 1, 3, 4, 5, and 6 we used built-in functions in BSseq to filter for sites within the probes regions (+/- 200 bp) and for sites covered at >5X in >75% of samples. We made slight modifications to this filtering pipeline for other experiments. For experiment 2, where n=3 for each input amount, we relaxed our missingness filter to sites with at least one read observed in at least $\frac{2}{3}$ samples. For experiment 7, which focused on NHP genomes for which the probe set coordinates (which are provided in hg38) are irrelevant, we did not perform region filtering. The number of sites analyzed for each experiment (reported in the main text and in Figure S2) therefore varies slightly depending on sample size, sequencing coverage, and other factors that impact which CpG sites passed our filters.

To confirm the fidelity of optimized TMS, we also checked whether CpGs captured by the protocol were distributed as expected throughout different genomic regions (e.g., promoters, enhancers) and that the average methylation levels in different genomic regions were as expected (e.g., quiescent regions being lowly methylated, repressed polycomb being highly methylated). To do so, we annotated each CpG site by whether it fell into a gene body, promoter, or non-genic region, and by chromatin state. We used hg38 gene body coordinates from Ensembl’s ‘biomaRt’ package in R, and we defined promoter regions as the 2000 bp region upstream of TSSs. We annotated CpGs as falling in chromatin states as defined by hg38 ChromHMM annotations from

NIH's Roadmap Epigenomics Project (50). We then counted the number of CpG sites that fell in each region (Figure 2C) and evaluated the median methylation across samples (Figure 2D).

Quantifying capture efficiency

A subset of our experiments sought to understand and optimize two measures of efficiency of the hybrid capture step: 1) how many of the expected CpG sites from the probe set passed our filtering parameters and were thus analyzable and 2) how many of the reads we generated for a given sample were on-target and putatively captured by the probe set, rather than representing off-target randomly sequenced DNA fragments that do not contribute to analyzable data as they are often sparsely shared between samples. For #1, we used the bedtools (version 2.28.0) (112) intersect command to determine the proportion of CpG sites that are within +/- 200bp with at least 1 probe (using a bed file available on the Twist Biosciences website (<https://www.twistbioscience.com/resources/data-files/twist-human-methylome-panel-target-bed-file>)). For #2, we used the bedtools function bamtobed to convert the mapped reads for each sample into a bed file; because we used a paired end sequencing strategy, each bed coordinate included a fragment start position from R1 and a fragment end position from R2. We then used the bedtools intersect command to determine the proportion of mapped read pairs that are within 200bp of at least 1 Twist probe.

Comparing DNA methylation measurements between TMS, the EPIC array, and WGBS

We used our filtered BSSeq object from experiment 5 to compare to data from the EPIC array generated for 55 paired human samples (average number of CpG sites measured with EPIC = 936,280; average call rate = 0.999). We downloaded the EPIC CpG coordinates from the Illumina website and merged with the TMS CpG locations, resulting in a shared dataset of 682,295 CpG sites passing filters and common to both technologies. We then performed two analyses to understand consistency. First, we calculated the average per-site methylation level across all samples included in the TMS or EPIC array datasets, respectively. We then ran a linear model testing the relationship between the two sets of average methylation levels using the 'lm' function in the 'stats' package in R. Second, we used the 'lm' function to estimate the R^2 value comparing per-site methylation levels for estimates derived from each technology for a given individual (i.e., not averaged across the dataset). This resulted in a distribution of 55 R^2 values.

Because all humans share canonical methylation patterns, we also compared this distribution to a distribution of 55 R^2 values derived from the same analysis after sample identity was permuted. We used the ‘t.test’ function in the ‘stats’ package in R to ask whether these distributions were significantly different.

We used a very similar strategy to compare ~30x WGBS data generated for six paired Tsimane samples with TMS data generated from experiment 1 (96-plex, 200 ng input). First, we performed low level processing of the WGBS data using Illumina’s DRAGEN pipeline and merged this with our filtered TMS data, resulting in 3,078,771 CpG sites common to both datasets. We calculated the average methylation level across samples reported for each site and technology and ran a linear model using the ‘lm’ function in the ‘stats’ package in R to calculate the R^2 value. We did not compare individual-based R^2 values to permuted values for this experiment, given the small number of individuals.

Understanding TMS performance in NHP species and comparing DNA methylation measurements between TMS and RRBS

To estimate the number of CpG sites that we expected to recover when applying the human probe set to each NHP species, we converted the probe bed file to a FASTA file using the bedtools command ‘getfasta’ (112) and the hg38 reference genome. We then used Bismark to map the FASTA file to each non-human primate’s respective genome. From the mapped bam file, we used the ‘bamToBed’ function in bedtools to extract coordinates for the mapped probes and to add a +/- 200 bp offset. Finally, we applied the ‘getfasta’ function in bedtools to extract the sequence for the mapped regions (plus the 200bp buffer) from the non-human primate genome and to count the number of CpG sites in this region set.

Similar to the comparisons between TMS and the EPIC array, we used paired RRBS data for the 96 rhesus macaque samples to directly compare methylation data generated using TMS versus RRBS. To do so, we processed the RRBS data using the same custom pipeline and filtering parameters described for TMS data, with the only modification being that we used the ‘—rrbs’ parameter in Trim Galore to remove unmethylated cytosines artificially introduced during library preparation from the 3’ end of fragments. We merged the filtered TMS and RRBS datasets, resulting in 721,766 CpG sites common to both technologies. As described for the TMS-EPIC array comparison, we then 1) calculated the average per-site methylation level across

all samples included in each dataset and compared these vectors using linear models and 2) estimated the R^2 value for methylation level estimates derived from each technology for a given individual, and used a t-test to compare this distribution to a distribution for the same analysis where sample identity was permuted (Figure S14).

Testing for age-associated changes in DNA methylation in the Tsimane and Orang Asli

We tested the extent to which DNA methylation varied with age using data from two experiments: 1) 96-plex, 200 ng input data from experiment 1 for the Tsimane of Bolivia and 2) 96-plex, 200 ng input, enzymatically sheared data from experiment 4 for the Orang Asli of Peninsular Malaysia. Subsistence-level groups generally show little age-related decline in cardiometabolic health (113), but it remains unknown if their relative lack of age-related chronic disease extends to molecular phenotypes because these populations are chronically understudied and underrepresented, especially in genomics. To evaluate age-effects on DNA methylation levels, we focused on sites that passed the filtering parameters described above and that were variably methylated (median methylation $>10\%$ and $<90\%$). This resulted in 936,547 and 1,001,011 testable sites in the Tsimane and Orang Asli, respectively. We used the ‘betabin’ function from the R ‘aod’ package (114) to test for age effects on the proportion of methylated / total read counts at each CpG site in each population, controlling for self-reported sex, batch, and cell type composition (proportion of neutrophils and lymphocytes, including B, CD4-T, and CD8-T cells). We extracted the p-value for the age effect for each tested site, and corrected for multiple hypothesis testing using the Benjamini-Hochberg false discovery rate approach implemented in the ‘p.adjust’ function in R.

Cell type composition was estimated via deconvolution in the R package ‘EpiDISH’ (115), using the ‘centDHSbloodDMC.m’ reference and the ‘RPC’ method. We focused on the cell populations referenced above because they matched empirically derived estimates available for the same samples for the Orang Asli ($R^2 > 0.5$, $p < 0.05$). Specifically, information about the relative proportion of lymphocytes, neutrophils, eosinophils, monocytes, and basophils was obtained for each Orang Asli sample from a 5-part white blood cell differential via the HemoCue WBC DIFF system and compared to deconvolution estimates.

We compared the age-related effects we found in Tsimane and Orang Asli to those reported by Johansson et al. (2013) (70), which quantified DNA methylation throughout aging in

a Swedish cohort using the 450K array. As Johansson et al. provided age estimates by Illumina ID and chromosome/position (i.e., site), we used UCSC’s Genome Browser LiftOver to convert genomic coordinates to hg38. We then were able to combine this dataset with our TMS data based on genomic coordinate position. This allowed us to test for correlations of age effect sizes amongst CpGs with data in all three datasets (Johansson, TMS from Tsimane, and TMS from Orang Asli). To understand how age effects varied across regions of the genome with distinct functions, we annotated each CpG site by the chromatin states in which it fell using ChromHMM annotations from NIH’s Roadmap Epigenomics Project (50). We then asked—for each population separately—whether CpGs significantly hypermethylated with age (FDR < 10%) were enriched in each chromatin state compared to all other sites and age-associated hypermethylated sites that did not fall into that region using the ‘fisher.test’ function in R. We performed this test for all regions and did the same for hypomethylated sites. All statistical analyses were performed using R version 4.2.2.

DATA AND CODE AVAILABILITY

All NHP data generated as part of this study has been deposited in NCBI’s Sequence Read Archive under accession number PRJNA1156067.

The human genomic data generated as part of this study comes from Indigenous participants from the Tsimane Health and Life History Project (THLHP) and the Orang Asli Health and Lifeways Project (OA HeLP). Both THLHP and OA HeLP’s highest priority is the minimization of risk to study participants. Both projects adhere to the “CARE Principles for Indigenous Data Governance” (Collective Benefit, Authority to Control, Responsibility, and Ethics) and are also committed to the “FAIR Guiding Principles for scientific data management and stewardship” (Findable, Accessible, Interoperable, Reusable). To adhere to these principles while minimizing risks, genomic data from both projects are available via restricted access. For OA HeLP, these requests can be made via dbGap (accession number TBD). For THLHP, these requests can be made via email following the instructions provided here:

<https://tsimane.anth.ucsb.edu/data.html>. In both cases, requests for de-identified genomic data will take the form of an application that details the exact uses of the data and the research questions to be addressed, procedures that will be employed for data security and privacy, potential benefits to the study communities, and procedures for assessing and minimizing

stigmatizing interpretations of the research results. Both projects are committed to open science and the leadership is available to assist interested investigators in preparing data access requests.

All scripts used to perform the analyses described here are available at [“https://github.com/alongtin15/TMS-Cost-effective-solutions-for-high-throughput-enzymatic-DNA-methylation-sequencing.”](https://github.com/alongtin15/TMS-Cost-effective-solutions-for-high-throughput-enzymatic-DNA-methylation-sequencing)

ACKNOWLEDGMENTS

First, we thank the Tsimane and Orang Asli study participants and communities for their involvement and support. We also thank all members of the research teams associated with the Tsimane Health and Life Histories Project and the Orang Asli Health and Lifeways Project. Second, we thank the members of the Lea Lab at Vanderbilt University, the SMack Lab at Arizona State University, and the Vanderbilt Technologies for Advanced Genomics (VANTAGE) team for their feedback, expertise, and support in completing this work. Third, we are grateful to the research infrastructure provided by Vanderbilt University’s Advanced Computing Center for Research and Education and Arizona State University’s Sol Computing clusters. Fourth, we thank Dr. Rex Howard and Dr. Michael Hart, the GSU veterinarians, for their assistance in collecting the capuchin blood samples. Finally, we acknowledge funding from the National Institute of General Medical Sciences (R35-GM147267), National Institute on Ageing (R01AG054442 and R61AG078529), the National Science Foundation (BCS-2142090), the Canadian Institute for Advanced Research (Azrieli Global Scholars Program), the French National Research Agency under the Investments for the Future (Investissements d’Avenir) programme (ANR-17-EURE-0010), the Kinship Foundation, (Searle Scholars Program), and the Pew Charitable Trusts (Pew Biomedical Scholars Program).

SUPPLEMENTAL INFORMATION

Supplemental information provided in a separate document.

REFERENCES

1. Reik W. Stability and flexibility of epigenetic gene regulation in mammalian development. *Nature*. 2007 May 24;447(7143):425–32.

2. Duhl DM, Vrieling H, Miller KA, Wolff GL, Barsh GS. Neomorphic agouti mutations in obese yellow mice. *Nat Genet*. 1994 Sep;8(1):59–65.
3. Morgan HD, Sutherland HG, Martin DI, Whitelaw E. Epigenetic inheritance at the agouti locus in the mouse. *Nat Genet*. 1999 Nov;23(3):314–8.
4. Dolinoy DC, Weidman JR, Waterland RA, Jirtle RL. Maternal genistein alters coat color and protects Avy mouse offspring from obesity by modifying the fetal epigenome. *Environ Health Perspect*. 2006 Apr;114(4):567–72.
5. Mohn F, Schübeler D. Genetics and epigenetics: stability and plasticity during cellular differentiation. *Trends Genet*. 2009 Mar;25(3):129–36.
6. Lokk K, Modhukur V, Rajashekhar B, Märkens K, Mägi R, Kolde R, et al. DNA methylome profiling of human tissues identifies global and tissue-specific methylation patterns. *Genome Biol*. 2014 Apr 1;15(4):r54.
7. Horvath S. DNA methylation age of human tissues and cell types. *Genome Biol*. 2013;14(10):R115.
8. Horvath S, Raj K. DNA methylation-based biomarkers and the epigenetic clock theory of ageing. *Nat Rev Genet*. 2018 Jun;19(6):371–84.
9. Jylhävä J, Pedersen NL, Hägg S. Biological Age Predictors. *EBioMedicine*. 2017 Jul;21:29–36.
10. Horvath S, Gurven M, Levine ME, Trumble BC, Kaplan H, Allayee H, et al. An epigenetic clock analysis of race/ethnicity, sex, and coronary heart disease. *Genome Biol*. 2016 Aug 11;17(1):171.
11. Bocklandt S, Lin W, Sehl ME, Sánchez FJ, Sinsheimer JS, Horvath S, et al. Epigenetic predictor of age. *PLoS One*. 2011 Jun 22;6(6):e14821.
12. Hao G, Youssef NA, Davis CL, Su S. The role of DNA methylation in the association between childhood adversity and cardiometabolic disease. *Int J Cardiol*. 2018 Mar 15;255:168–74.
13. Robertson KD. DNA methylation and human disease. *Nat Rev Genet*. 2005 Aug;6(8):597–610.
14. Costantino S, Mohammed SA, Ambrosini S, Paneni F. Epigenetic processing in cardiometabolic disease. *Atherosclerosis*. 2019 Feb;281:150–8.
15. Pal S, Tyler JK. Epigenetics and aging. *Sci Adv*. 2016 Jul;2(7):e1600584.
16. López-Otín C, Blasco MA, Partridge L, Serrano M, Kroemer G. The hallmarks of aging. *Cell*. 2013 Jun 6;153(6):1194–217.

17. Ciccarone F, Tagliatesta S, Caiafa P, Zampieri M. DNA methylation dynamics in aging: how far are we from understanding the mechanisms? *Mech Ageing Dev*. 2018 Sep;174:3–17.
18. Seale K, Horvath S, Teschendorff A, Eynon N, Voisin S. Making sense of the ageing methylome. *Nat Rev Genet*. 2022 Oct;23(10):585–605.
19. Simpson DJ, Chandra T. Epigenetic age prediction. *Aging Cell*. 2021 Sep;20(9):e13452.
20. Painter RC, de Rooij SR, Bossuyt PM, Simmers TA, Osmond C, Barker DJ, et al. Early onset of coronary artery disease after prenatal exposure to the Dutch famine. *Am J Clin Nutr*. 2006 Aug;84(2):322–7.
21. Heijmans BT, Tobi EW, Stein AD, Putter H, Blauw GJ, Susser ES, et al. Persistent epigenetic differences associated with prenatal exposure to famine in humans. *Proc Natl Acad Sci U S A*. 2008 Nov 4;105(44):17046–9.
22. Tobi EW, Lumey LH, Talens RP, Kremer D, Putter H, Stein AD, et al. DNA methylation differences after exposure to prenatal famine are common and timing- and sex-specific. *Hum Mol Genet*. 2009 Nov 1;18(21):4046–53.
23. Tobi EW, Goeman JJ, Monajemi R, Gu H, Putter H, Zhang Y, et al. DNA methylation signatures link prenatal famine exposure to growth and metabolism. *Nat Commun*. 2014 Nov 26;5:5592.
24. Wang Z, Song J, Li Y, Dong B, Zou Z, Ma J. Early-Life Exposure to the Chinese Famine Is Associated with Higher Methylation Level in the INSR Gene in Later Adulthood. *Sci Rep*. 2019 Mar 4;9(1):3354.
25. Murgatroyd C, Patchev AV, Wu Y, Micale V, Bockmühl Y, Fischer D, et al. Dynamic DNA methylation programs persistent adverse effects of early-life stress. *Nat Neurosci*. 2009 Dec;12(12):1559–66.
26. Laubach ZM, Greenberg JR, Turner JW, Montgomery TM, Pioon MO, Sawdy MA, et al. Early-life social experience affects offspring DNA methylation and later life stress phenotype. *Nat Commun*. 2021 Jul 20;12(1):4398.
27. Unternaehrer E, Luers P, Mill J, Dempster E, Meyer AH, Staehli S, et al. Dynamic changes in DNA methylation of stress-associated genes (OXTR, BDNF) after acute psychosocial stress. *Transl Psychiatry*. 2012 Aug 14;2(8):e150.
28. Cunliffe VT. The epigenetic impacts of social stress: how does social adversity become biologically embedded? *Epigenomics*. 2016 Dec;8(12):1653–69.
29. McDade TW, Ryan CP, Jones MJ, Hoke MK, Borja J, Miller GE, et al. Genome-wide analysis of DNA methylation in relation to socioeconomic status during development and early adulthood. *Am J Phys Anthropol*. 2019 May;169(1):3–11.

30. Pacis A, Tailleux L, Morin AM, Lambourne J, MacIsaac JL, Yotova V, et al. Bacterial infection remodels the DNA methylation landscape of human dendritic cells. *Genome Res.* 2015 Dec;25(12):1801–11.
31. Zhang X, Justice AC, Hu Y, Wang Z, Zhao H, Wang G, et al. Epigenome-wide differential DNA methylation between HIV-infected and uninfected individuals. *Epigenetics.* 2016 Oct 2;11(10):750–60.
32. Sagonas K, Meyer BS, Kaufmann J, Lenz TL, Häslер R, Eizaguirre C. Experimental Parasite Infection Causes Genome-Wide Changes in DNA Methylation. *Mol Biol Evol.* 2020 Aug 1;37(8):2287–99.
33. McDade TW, Ryan C, Jones MJ, MacIsaac JL, Morin AM, Meyer JM, et al. Social and physical environments early in development predict DNA methylation of inflammatory genes in young adulthood. *Proc Natl Acad Sci U S A.* 2017 Jul 18;114(29):7611–6.
34. Gu H, Smith ZD, Bock C, Boyle P, Gnirke A, Meissner A. Preparation of reduced representation bisulfite sequencing libraries for genome-scale DNA methylation profiling. *Nat Protoc.* 2011 Apr;6(4):468–81.
35. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature.* 2012 Sep 6;489(7414):57–74.
36. Zhang Z, Zeng C, Zhang W. Characterization of the Illumina EPIC Array for Optimal Applications in Epigenetic Research Targeting Diverse Human Populations. *Epigenetics Commun [Internet].* 2022 Dec 1;2. Available from: <http://dx.doi.org/10.1186/s43682-022-00015-9>
37. Pidsley R, Zotenko E, Peters TJ, Lawrence MG, Risbridger GP, Molloy P, et al. Critical evaluation of the Illumina MethylationEPIC BeadChip microarray for whole-genome DNA methylation profiling. *Genome Biol.* 2016 Oct 7;17(1):208.
38. Gu H, Bock C, Mikkelsen TS, Jäger N, Smith ZD, Tomazou E, et al. Genome-scale DNA methylation mapping of clinical samples at single-nucleotide resolution. *Nat Methods.* 2010 Feb;7(2):133–6.
39. Tanaka K, Okamoto A. Degradation of DNA by bisulfite treatment. *Bioorg Med Chem Lett.* 2007 Apr 1;17(7):1912–5.
40. Vaisvila R, Ponnaluri VKC, Sun Z, Langhorst BW, Saleh L, Guan S, et al. Enzymatic methyl sequencing detects DNA methylation at single-base resolution from picograms of DNA. *Genome Res.* 2021 Jul;31(7):1280–9.
41. Feng S, Zhong Z, Wang M, Jacobsen SE. Efficient and accurate determination of genome-wide DNA methylation patterns in *Arabidopsis thaliana* with enzymatic methyl sequencing. *Epigenetics Chromatin.* 2020 Oct 7;13(1):42.
42. Twist Human Methylome Panel [Internet]. [cited 2024 May 7]. Available from: <http://dx.doi.org/10.1186/s43682-022-00015-9>

https://www.twistbioscience.com/sites/default/files/resources/2022-06/DOC-001317_ProductSheet_HumanMethylPanel-Rev1-singles.pdf

43. Twist Target Capture EM-seq [Internet]. [cited 2024 May 7]. Available from: https://www.twistbioscience.com/sites/default/files/resources/2022-06/DOC-001222_Protoctol_TargetedMethylationSequencing-REV4-singles.pdf
44. NEBNext EM-seq Library Prep Protocol [Internet]. [cited 2024 May 7]. Available from: https://www.twistbioscience.com/sites/default/files/resources/2023-06/DOC-001224-Protocol_NEBNextEnzMethylLibraryPrep-REV5-singles.pdf
45. Twist Bioscience's Methylation Detection System [Internet]. 2023 [cited 2024 Aug 9]. Available from: <https://www.youtube.com/watch?v=GhgugFoIDJo>
46. Vasunilashorn S, Crimmins EM, Kim JK, Winking J, Gurven M, Kaplan H, et al. Blood lipids, infection, and inflammatory markers in the Tsimane of Bolivia. *Am J Hum Biol.* 2010 Nov-Dec;22(6):731–40.
47. Gurven M, Stieglitz J, Trumble B, Blackwell AD, Beheim B, Davis H, et al. The Tsimane Health and Life History Project: Integrating anthropology and biomedicine. *Evol Anthropol.* 2017 Apr;26(2):54–73.
48. Wallace IJ, Lea AJ, Lim YAL, Chow SKW, Sayed IBM, Ngui R, et al. Orang Asli Health and Lifeways Project (OA HeLP): a cross-sectional cohort study protocol. *BMJ Open.* 2022 Sep 20;12(9):e058660.
49. Genereux DP, Johnson WC, Burden AF, Stöger R, Laird CD. Errors in the bisulfite conversion of DNA: modulating inappropriate- and failed-conversion frequencies. *Nucleic Acids Res.* 2008 Dec;36(22):e150.
50. Roadmap Epigenomics Consortium, Kundaje A, Meuleman W, Ernst J, Bilenky M, Yen A, et al. Integrative analysis of 111 reference human epigenomes. *Nature.* 2015 Feb 19;518(7539):317–30.
51. Shu C, Zhang X, Aouizerat BE, Xu K. Comparison of methylation capture sequencing and Infinium MethylationEPIC array in peripheral blood mononuclear cells. *Epigenetics Chromatin.* 2020 Nov 23;13(1):51.
52. Warren WC, Harris RA, Haukness M, Fiddes IT, Murali SC, Fernandes J, et al. Sequence diversity analyses of an improved rhesus macaque genome enhance its biomedical utility. *Science [Internet].* 2020 Dec 18;370(6523). Available from: <http://dx.doi.org/10.1126/science.abc6617>
53. Raaum RL. Molecular evidence on primate origins and evolution. In: *Handbook of Paleoanthropology.* Berlin, Heidelberg: Springer Berlin Heidelberg; 2015. p. 1083–135.
54. Blake LE, Roux J, Hernando-Herraez I, Banovich NE, Perez RG, Hsiao CJ, et al. A comparison of gene expression and DNA methylation patterns across tissues and species.

Genome Res. 2020 Feb;30(2):250–62.

55. Slieker RC, Roost MS, van Iperen L, Suchiman HED, Tobi EW, Carlotti F, et al. DNA Methylation Landscapes of Human Fetal Development. *PLoS Genet.* 2015 Oct;11(10):e1005583.
56. Ziller MJ, Gu H, Müller F, Donaghey J, Tsai LTY, Kohlbacher O, et al. Charting a dynamic DNA methylation landscape of the human genome. *Nature.* 2013 Aug 22;500(7463):477–81.
57. Lea AJ, Vilgalys TP, Durst PAP, Tung J. Maximizing ecological and evolutionary insight in bisulfite sequencing data sets. *Nat Ecol Evol.* 2017 Aug;1(8):1074–83.
58. Anderson JA, Johnston RA, Lea AJ, Campos FA, Voyles TN, Akinyi MY, et al. High social status males experience accelerated epigenetic aging in wild baboons. *eLife* [Internet]. 2021 Apr 6;10. Available from: <http://dx.doi.org/10.7554/eLife.66128>
59. Watowich MM, Costa CE, Chiou KL, Goldman EA, Petersen RM, Patterson S, et al. Immune gene regulation is associated with age and environmental adversity in a nonhuman primate. *Mol Ecol.* 2024 Jul 20;e17445.
60. Carey I. *Orang Asli: The Aboriginal Tribes of Peninsular Malaysia.* Oxford University Press; 1976. 376 p.
61. Endicott K. Introduction. In: *Malaysia's Original People.* NUS Press Pte Ltd; 2018. p. 1–38.
62. Ringhofer L. Exploring an indigenous world in the Bolivian Amazon: The case of the 'Tsimane'. In: *Fishing, Foraging and Farming in the Bolivian Amazon.* Dordrecht: Springer Netherlands; 2010. p. 45–82.
63. Pérez-Llorente I, Paneque-Gálvez J, Luz AC, Macía MJ, Guèze M, Domínguez-Gómez JA, et al. Changing indigenous cultures, economies and landscapes: The case of the 'Tsimane', Bolivian Amazon. *Landsc Urban Plan.* 2013 Dec;120:147–57.
64. Gurven M, Blackwell AD, Rodríguez DE, Stieglitz J, Kaplan H. Does blood pressure inevitably rise with age?: longitudinal evidence among forager-horticulturalists. *Hypertension.* 2012 Jul;60(1):25–33.
65. Pontzer H, Wood BM, Raichlen DA. Hunter-gatherers as models in public health. *Obes Rev.* 2018 Dec;19 Suppl 1:24–35.
66. McDade TW, Ryan CP, Adair LS, Lee NR, Carba DB, MacIsaac JL, et al. Association between infectious exposures in infancy and epigenetic age acceleration in young adulthood in metropolitan Cebu, Philippines. *Am J Hum Biol.* 2023 Nov;35(11):e23948.
67. Ryan CP, Rege RJ, Lee NR, Carba DB, Kobor MS, MacIsaac JL, et al. Maternal epigenetic clocks measured during pregnancy do not predict gestational age at delivery or offspring birth outcomes: a replication study in metropolitan Cebu, Philippines. *Clin Epigenetics.*

2022 Jun 22;14(1):78.

68. Comprehensive R Archive Network (CRAN) [Internet]. [cited 2024 Aug 7]. aod: Analysis of Overdispersed Data. Available from: <https://CRAN.R-project.org/package=aod>
69. Jones MJ, Goodman SJ, Kobor MS. DNA methylation and healthy human aging. *Aging Cell*. 2015 Dec;14(6):924–32.
70. Johansson A, Enroth S, Gyllensten U. Continuous Aging of the Human DNA Methylome Throughout the Human Lifespan. *PLoS One*. 2013 Jun 27;8(6):e67378.
71. Tong H, Dwaraka VB, Chen Q, Luo Q, Lasky-Su JA, Smith R, et al. Quantifying the stochastic component of epigenetic aging. *Nat Aging*. 2024 Jun;4(6):886–901.
72. Zoller J, Horvath S. MammalMethylClock R package: software for DNA methylation-based epigenetic clocks in mammals. *Bioinformatics* [Internet]. 2024 May 2;40(5). Available from: <http://dx.doi.org/10.1093/bioinformatics/btae280>
73. Pelegí-Sisó D, de Prado P, Ronkainen J, Bustamante M, González JR. methylclock: a Bioconductor package to estimate DNA methylation age. *Bioinformatics*. 2021 Jul 19;37(12):1759–60.
74. Vilgalys TP, Rogers J, Jolly CJ, Mukherjee S, Tung J. Evolution of DNA Methylation in Papio Baboons. *Mol Biol Evol*. 2019 Mar 1;36(3):527–40.
75. Anderson JA, Vilgalys TP, Tung J. Broadening primate genomics: new insights into the ecology and evolution of primate gene regulation. *Curr Opin Genet Dev*. 2020 Jun;62:16–22.
76. Housman G, Gilad Y. Prime time for primate functional genomics. *Curr Opin Genet Dev*. 2020 Jun;62:1–7.
77. Maegawa S, Lu Y, Tahara T, Lee JT, Madzo J, Liang S, et al. Caloric restriction delays age-related methylation drift. *Nat Commun*. 2017 Sep 14;8(1):539.
78. Chiou KL, Montague MJ, Goldman EA, Watowich MM, Sams SN, Song J, et al. Rhesus macaques as a tractable physiological model of human ageing. *Philos Trans R Soc Lond B Biol Sci*. 2020 Nov 9;375(1811):20190612.
79. Lea AJ, Altmann J, Alberts SC, Tung J. Resource base influences genome-wide DNA methylation levels in wild baboons (*Papio cynocephalus*). *Mol Ecol*. 2016 Apr;25(8):1681–96.
80. Provençal N, Suderman MJ, Guillemin C, Massart R, Ruggiero A, Wang D, et al. The signature of maternal rearing in the methylome in rhesus macaque prefrontal cortex and T cells. *J Neurosci*. 2012 Oct 31;32(44):15626–42.
81. Anderson JA, Lin D, Lea AJ, Johnston RA, Voyles T, Akinyi MY, et al. DNA methylation

signatures of early-life adversity are exposure-dependent in wild baboons. *Proc Natl Acad Sci U S A.* 2024 Mar 12;121(11):e2309469121.

82. Arneson A, Haghani A, Thompson MJ, Pellegrini M, Kwon SB, Vu H, et al. A mammalian methylation array for profiling methylation levels at conserved sequences. *Nat Commun.* 2022 Feb 10;13(1):783.
83. Maciejewski E, Horvath S, Ernst J. Cross-species and tissue imputation of species-level DNA methylation samples across mammalian species. *bioRxiv* [Internet]. 2023 Nov 27; Available from: <http://dx.doi.org/10.1101/2023.11.26.568769>
84. Campbell B, Gray PB, Leslie P. Age-related changes in body composition among Turkana males of Kenya. *Am J Hum Biol.* 2005 Sep-Oct;17(5):601–10.
85. Nirmala A. Age variation in blood pressure: effect of sex and urbanization in a genetically homogeneous caste population of Andhra Pradesh. *Am J Hum Biol.* 2001 Nov-Dec;13(6):744–52.
86. Palmer AK, Jensen MD. Metabolic changes in aging humans: current evidence and therapeutic strategies. *J Clin Invest* [Internet]. 2022 Aug 15;132(16). Available from: <http://dx.doi.org/10.1172/JCI158451>
87. Guo J, Huang X, Dou L, Yan M, Shen T, Tang W, et al. Aging and aging-related diseases: from molecular mechanisms to interventions and treatments. *Signal Transduct Target Ther.* 2022 Dec 16;7(1):391.
88. Jaul E, Barron J. Age-Related Diseases and Clinical and Public Health Implications for the 85 Years Old and Over Population. *Front Public Health.* 2017 Dec 11;5:335.
89. Watowich MM, Arner AM, Wang S, John E, Kahumbu JC, Kinyua P, et al. The built environment is more predictive of cardiometabolic health than other aspects of lifestyle in two rapidly transitioning Indigenous populations [Internet]. *medRxiv.* 2024 [cited 2024 Aug 29]. p. 2024.08.26.24312234. Available from: <https://www.medrxiv.org/content/10.1101/2024.08.26.24312234v1.abstract>
90. López-Otín C, Blasco MA, Partridge L, Serrano M, Kroemer G. Hallmarks of aging: An expanding universe. *Cell.* 2023 Jan 19;186(2):243–78.
91. Meyer DH, Schumacher B. Aging clocks based on accumulating stochastic variation. *Nat Aging.* 2024 Jun;4(6):871–85.
92. Qiao X, Straight B, Ngo D, Hilton CE, Owuor Olungah C, Naugle A, et al. Severe drought exposure in utero associates to children's epigenetic age acceleration in a global climate change hot spot. *Nat Commun.* 2024 May 16;15(1):4140.
93. Cheng M, Conley D, Kuipers T, Li C, Ryan CP, Taeubert MJ, et al. Accelerated biological aging six decades after prenatal famine exposure. *Proc Natl Acad Sci U S A.* 2024 Jun 11;121(24):e2319179121.

94. Lea AJ, Garcia A, Arevalo J, Ayroles JF, Buetow K, Cole SW, et al. Natural selection of immune and metabolic genes associated with health in two lowland Bolivian populations. *Proc Natl Acad Sci U S A*. 2023 Jan 3;120(1):e2207544120.
95. Munds RA, Cooper EB, Janiak MC, Lam LG, DeCasien AR, Bauman Surratt S, et al. Variation and heritability of retinal cone ratios in a free-ranging population of rhesus macaques. *Evolution*. 2022 Aug;76(8):1776–89.
96. Testard C, Brent LJN, Andersson J, Chiou KL, Negron-Del Valle JE, DeCasien AR, et al. Social connections predict brain structure in a multidimensional free-ranging primate society. *Sci Adv*. 2022 Apr 15;8(15):eabl5794.
97. Turcotte CM, Mann EHJ, Stock MK, Villamil CI, Montague MJ, Dickinson E, et al. The ontogeny of sexual dimorphism in free-ranging rhesus macaques. *Am J Biol Anthropol*. 2022 Feb;177(2):314–27.
98. Watowich MM, Chiou KL, Montague MJ, Cayo Biobank Research Unit, Simons ND, Horvath JE, et al. Natural disaster and immunological aging in a nonhuman primate. *Proc Natl Acad Sci U S A* [Internet]. 2022 Feb 22;119(8). Available from: <http://dx.doi.org/10.1073/pnas.2121663119>
99. Pavez-Fox MA, Negron-Del Valle JE, Thompson IJ, Walker CS, Bauman SE, Gonzalez O, et al. Sociality predicts individual variation in the immunity of free-ranging rhesus macaques. *Physiol Behav*. 2021 Nov 1;241:113560.
100. Roberts EK, Lu A, Bergman TJ, Beehner JC. Female reproductive parameters in wild geladas (*theropithecus gelada*). *Int J Primatol*. 2017 Feb;38(1):1–20.
101. Chiou KL, Janiak MC, Schneider-Crease IA, Sen S, Ayele F, Chuma IS, et al. Genomic signatures of high-altitude adaptation and chromosomal polymorphism in geladas. *Nat Ecol Evol*. 2022 May;6(5):630–43.
102. 1000 Genomes Project Consortium, Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, et al. A global reference for human genetic variation. *Nature*. 2015 Oct 1;526(7571):68–74.
103. Snyder-Mackler N. Paperpile. [cited 2024 Jun 14]. Reduced Representation Bisulfite Sequencing (RRBS) with NEB Reagents. Available from: <https://paperpile.com/app/p/f98e0f74-f414-0f69-bf7a-abe4e85222ec>
104. Paperpile [Internet]. [cited 2024 Jun 14]. Dynamic Read Analysis for GENomics (DRAGEN). Available from: <https://paperpile.com/app/p/8d00ea63-cf76-0737-8118-4a62f6a02e17>
105. Krueger F, Andrews SR. Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. *Bioinformatics*. 2011 Jun 1;27(11):1571–2.
106. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence

data. *Bioinformatics*. 2014 Aug 1;30(15):2114–20.

107. Krueger F, James F, Ewels P, Afyounian E, Weinstein M, Schuster-Boeckler B, et al. *FelixKrueger/TrimGalore*: v0.6.10 - add default decompression path [Internet]. Zenodo; 2023. Available from: <https://zenodo.org/record/7598955>

108. Feng S, Cokus SJ, Zhang X, Chen PY, Bostick M, Goll MG, et al. Conservation and divergence of methylation patterning in plants and animals. *Proc Natl Acad Sci U S A*. 2010 May 11;107(19):8689–94.

109. Kint S, De Spiegelaere W, De Kesel J, Vandekerckhove L, Van Criekinge W. Evaluation of bisulfite kits for DNA methylation profiling in terms of DNA fragmentation and DNA recovery using digital PCR. *PLoS One*. 2018 Jun 14;13(6):e0199091.

110. Paperpile [Internet]. [cited 2024 Jun 14]. R: A Language and Environment for Statistical Computing. Available from: <https://paperpile.com/app/p/f9161119-c341-02c8-b6bc-dcc0a9113158>

111. Park Y, Wu H. Differential methylation analysis for BS-seq data under general experimental design. *Bioinformatics*. 2016 May 15;32(10):1446–53.

112. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*. 2010 Mar 15;26(6):841–2.

113. Kaplan H, Thompson RC, Trumble BC, Wann LS, Allam AH, Beheim B, et al. Coronary atherosclerosis in indigenous South American Tsimane: a cross-sectional cohort study. *Lancet*. 2017 Apr 29;389(10080):1730–9.

114. Comprehensive R Archive Network (CRAN) [Internet]. [cited 2024 Aug 7]. aod: Analysis of Overdispersed Data. Available from: <https://CRAN.R-project.org/package=aod>

115. Zheng SC, Breeze CE, Beck S, Teschendorff AE. Identification of differentially methylated cell types in epigenome-wide association studies. *Nat Methods*. 2018 Dec;15(12):1059–66.