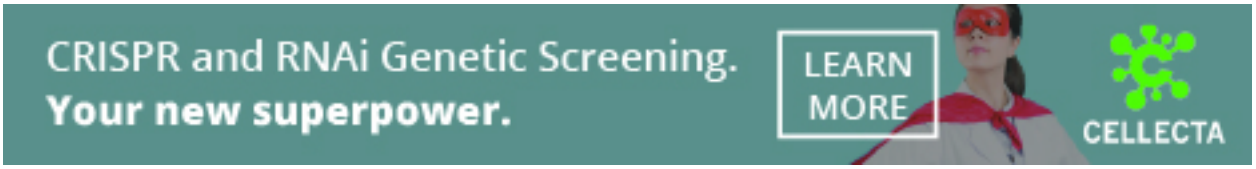
Genome skimming with nanopore sequencing precisely determines global and transposon DNA methylation in vertebrates

Christopher Faulk

Genome Res. 2023 33: 948-956 originally published online July 13, 2023
Access the most recent version at doi:10.1101/gr.277743.123



Method

Genome skimming with nanopore sequencing precisely determines global and transposon DNA methylation in vertebrates

Christopher Faulk

Department of Animal Science, College of Food, Agricultural and Natural Resource Sciences, University of Minnesota, Minneapolis, Minnesota 55455, USA

Genome skimming is defined as low-pass sequencing below 0.05× coverage and is typically used for mitochondrial genome recovery and species identification. Long-read nanopore sequencers enable simultaneous reading of both DNA sequence and methylation and can multiplex samples for low-cost genome skimming. Here I present nanopore sequencing as a highly precise platform for global DNA methylation and transposon assessment. At coverage of just 0.001×, or 30 Mb of reads, accuracy is sub-1%. Biological and technical replicates validate high precision. Skimming 40 vertebrate species reveals conserved patterns of global methylation consistent with whole-genome bisulfite sequencing and an average mapping rate >97%. Genome size directly correlates to global DNA methylation, explaining 39% of its variance. Accurate SINE and LINE transposon methylation in both the mouse and primates can be obtained with just 0.0001× coverage, or 3 Mb of reads. Sample multiplexing, field portability, and the low price of this instrument combine to make genome skimming for DNA methylation an accessible method for epigenetic assessment from ecology to epidemiology and for low-resource groups.

[Supplemental material is available for this article.]

Genome skimming refers to unbiased low-pass sequencing below 0.05× coverage and is used to reconstruct mitochondrial genomes as well as identify species and parasites (Twyford and Ness 2017; Papaiakovou et al. 2023; Zhang et al. 2023). Oxford Nanopore Technologies' nanopore sequencers have been used for genome skimming to produce mitochondrial genomes with success, and at least one study used genome skimming to report DNA methylation (Euskirchen et al. 2017; De Vivo et al. 2022). An important biomarker, DNA methylation varies with age, tissue, species, and environmental exposures, making it a useful measure from ecology to epidemiology (Laubach et al. 2018).

Global DNA methylation measures the ratio of 5′methylcytosines (5 mCs) versus total cytosines reported as a percentage. Current methods to measure global methylation have drawbacks in either cost or accuracy. Antibody-based methods do not have the resolution to distinguish small-magnitude changes often seen in biologically significant epidemiological exposures (Breton et al. 2017). Other methods such as reduced representation bisulfite sequencing (RRBS) and whole-genome bisulfite sequencing (WGBS), although accurate, are expensive, biased in target location, and rely upon bisulfite conversion, which causes downstream bioinformatic challenges.

Nanopore sequencers produce long reads from native genomic DNA, typically in the 10- to 50-kb range, and simultaneously report the presence of DNA methylation genome-wide (Faulk 2022). Long reads enable more contiguous genome assembly and easier reconstruction of repetitive regions (Nurk et al. 2022). Alignment is faster with fewer errors and has the ability to span large transposons and detect structural variants and single-nucleotide variants. Applied to genome skims, long reads enable more accurate recovery of mitochondrial genomes with lower sequencing depth (Cai et al. 2022). Sample multiplexing, field portability, and the low price of the instrument all reduce cost and enable sequencing by low-resource groups (Jahan et al. 2021).

Genome skimming is unbiased in most genomic regions; however, some regions are present in multiple copies, and this enrichment will be reflected in the skimmed read data set. Mitochondria can be present in thousands of copies per nuclear genome and are therefore a common analysis target because a skimming level of 0.1× (300 Mb) would result in hundreds of full-length copies of mitogenomes. Transposons are another overrepresented category in the genome. Repeats make up ∼50% of a mammalian genome on average, with some families present in more than 1 million copies per nuclear genome (Wanner and Faulk 2021). The difficulty in mapping transposons is why WGBS does not provide a true "whole-genome" measure, as its mapping rate is typically ∼50% versus the nanopore mapping rate of >97% we found here. Although few studies have focused on transposons in genome skimming data, the potential is high for transposon biology (Choudhury et al. 2017; Baeza et al. 2021; Sarkar et al. 2023). From an epigenetic perspective, transposons have long been used as proxies for global DNA methylation in epidemiology, so validation of their methylation to global DNA methylation holds promise for their continued use as biomarkers with a lower cost per sample (Faulk et al. 2013).

The aims of this study are (1) to report global DNA methylation levels using genome skimming in the nucleus, in mitochondria,

Corresponding author: cfaulk@umn.edu
Article published online before print. Article, supplemental material, and publication date are at https://www.genome.org/cgi/doi/10.1101/gr.277743.123.

33:948–956 Published by Cold Spring Harbor Laboratory Press; ISSN 1088-9051/23; www.genome.org

Table 1. DNA methylation at varying coverage levels, bootstrapped subsampling

Depth	11.2× (34 Gb)	10× (30 Gb)	1× (3 Gb)	0.1× (300 Mb)	0.01× (30 Mb)	0.03× (10 Mb)	0.001× (3 Mb)	0.0001× (300 kb)
Mean	82.87	82.86	82.85	82.84	82.77	82.56	82.58	81.91
Bootstraps	1	10	10	10	10	10	10	10
Minimum		82.85	82.74	82.47	81.80	81.88	78.90	62.82
Maximum		82.87	82.92	83.27	83.33	83.26	85.06	91.97
Range		0.0192	0.1718	0.7990	1.5301	1.3754	6.1525	29.1491
Standard deviation		0.0060	0.0600	0.2595	0.4735	0.4709	2.0102	9.3787
Standard error		0.0019	0.0190	0.0821	0.1497	0.1489	0.6357	2.9658

and, specifically, in transposons and (2) to confirm the precision of this method through replication and across species.

Results

Coverage depth estimation

Before designing a genome skimming experiment, it is necessary to determine the minimum level of sequencing depth to obtain sufficient precision for statistical power. A single chimpanzee genome was sequenced to 11.2× depth using a nanopore PromethION instrument and global DNA methylation assessed at 82.87%. Reads were subsampled at coverage levels representing the range of genome skimming expectations, from 0.1× (300 Mb) down to 0.0001× (300 kb), and bootstrapped 10 times to calculate error (Table 1). At 0.01× (30 Mb), coverage of a primate genome results in an error of <1% difference from the true value (Fig. 1).

Biological and technical replication

Reproducibility is crucial for assays, which may vary over a small range, and is measured by precision, that is, how close repeated measures are to each other. In the following experiment, mice were sequenced to skimming-level coverage between 0.0025× and 0.016× (median, 0.0058×) or by read coverage, from 8 Mb–133 Mb (median, 47 Mb) reads per sample (Table 2). Three tissues per mouse were sequenced with five mice serving as biological replicates. For controls, whole-genome amplified DNA is expected to have near zero DNA methylation and was used as the low methylated control, whereas CpG methylase–treated DNA is expected to have near 100% methylation and was used as the high methylated control. Multiple aliquots of the same two control reactions were individually barcoded along with the mouse tissues and run on a single MinION flow cell. These served as technical replicates as well as range controls.

DNA methylation was highly consistent within tissues and significantly higher in testes versus cortex ($P<0.05$) (Fig. 2). Quality control validated the experimental protocol (Fig. 3A–D). Methylation values were robust to differences in total bases and average length sequenced, with a fourfold difference in total reads from the lowest group to the highest and a twofold difference in read length between groups. Quality scores did not affect methylation accuracy either, with cortex having a significantly lower, but still high-quality, average q-score between tissues (29.1 vs. 30.2). The lowest q-score occurred in the low methylation control, likely owing to the poorer quality of the in vitro whole-genome amplified DNA. Both controls performed as expected. Technical replication proved very precise with a 0.15% and 0.11% standard deviation for the high and low controls, respectively. All reads reliably mapped to the mouse genome with <0.37% difference across all 24 samples. Mean read quality was slightly lower for the cortex than hippocampus or testes but did not affect precision.

Vertebrate methylation

Methylation varies with tissue and by species. To determine whether skimming is sufficient to capture primate-specific levels of global DNA methylation, I sequenced buffy coat DNA from five primates to a high depth and down-sampled the reads to a 0.01× equivalent, or 30 Mb (Fig. 4). The global levels are similar to what is seen in human blood.

Across vertebrates, methylation varies widely. Muscle tissue from 34 species was sequenced at a depth ranging from 3 to 200 Mb, median 80 Mb (Fig. 5A). Genome size correlated with percentage of methylation with an R^2 of 0.39, likely owing to increased repetitive element content and associated silencing (Fig. 5B). Summary statistics are shown in Table 3; complete data are available in Supplemental File S1.

Transposons

Genome skimming is unbiased by nature; however, some regions of DNA are enriched compared with single-copy genes; for example, repetitive elements and transposons are present in multiple copies in the nuclear genome and therefore are more likely to be sequenced by skims than single-copy regions. Importantly, DNA methylation plays a strong role in suppressing transposon mobilization and genomic dysregulation. I reasoned that with ~40% of a typical mammalian genome consisting of repetitive elements,

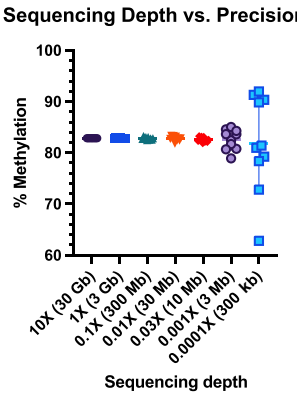


Figure 1. Sequencing depth versus precision. Reads are subsampled from an 11.2× coverage chimpanzee genome at coverage levels from 0.0001× to 10×. DNA methylation is calculated for 10 bootstrapped subsamples.

Table 2. Biological and technical replication of DNA methylation in mouse at low coverage

Tissue	Methylation (%)	SD (%)	Replicates	Median (Mb)	Mean % mapped	Mean quality
Hippocampus	78.62	1.64	5	28.3	98.81	30
Cortex	78.26	0.89	5	13.5	98.97	29.1
Testes	80.07	0.61	5	48.5	98.83	30.2
Low control	1.26	0.15	5	99.2	99.18	26
High control	99.43	0.11	4	51.0	98.96	30.1

genome skimming can assess methylation at the transposon family level by averaging methylation across all copies of a family.

To determine the minimum depth necessary for precise quantitation, I used the high-depth chimpanzee data set down-sampled from 10× to 0.0001×, quantified DNA methylation, and bootstrapped 10 replicates (Fig. 6; Table 4). For *Alu* elements, which number more than 1 million copies per cell, even 0.001× coverage, just 3 Mb sequenced, is sufficient for sub-1% accuracy. LINE-1 elements are present in 100,000 copies per cell, although most are truncated. Methylation at LINE-1 at the 0.01× coverage level, 30 Mb sequenced, obtained sub-1% precision. *Alu* mean methylation was very high at 94.69%, higher than both LINE-1 (90.16%) and the overall genome (82.87%). These values correlate with transposon age and CpG density, where *Alus* are younger and more CpG dense than LINE-1s, and both are under selection for increased DNA methylation to suppress mobilization.

The mouse genome skims provide validation and biological replication for transposon methylation assessment. I selected the highest copy number families in mouse, with >1% genomic content threshold, and compared their variability within and across tissues (Fig. 7; Table 5). Above the threshold are SINE families, *Alu*-B1, B2, and B4; ERV families, class II and ERV-MaLR; and LINE family L1. As in primates, all transposons had higher methylation than the global average of 78.26%–80.07%, depending on tissue. *Alu*/B1 and L1 elements were stable across tissues, whereas B2s, B4s, and ERVL-MaLRs were significantly increased in testes. ERV class II elements were the only family with a significant decrease in the testes, which is biologically significant as the most frequently active transposon in mice is the intracisternal A particle (IAP), an ERV class II element. These findings have significant implications on transposon based methylation assays.

By comparing the methylation of all repeat families across species, patterns emerge (Fig. 8A,B). The mouse repeat complement shows highest methylation in LTRs, SINEs (*Alu*/B1, B2, B4), ERVs, and LINE-1, which are the most active transposons in the mouse genome. The most prevalent, highest copy number transposons are highlighted in red and tend toward the highest levels of methylation. Similarly in the chimpanzee, SINE *Alu* and LINE-1 are among the most highly methylated and are the most prevalent in the primate genome. Chimpanzee ERVs remain highly methylated despite having lost mobility in the primate lineage. In both the mouse and chimpanzee, nontransposon repeats have the lowest levels of methylation, for example, low complexity repeats, tRNAs, and simple repeats. Note that the mouse is represented by five replicates taken from hippocampus, whereas the chimpanzee is represented by buffy coat DNA taken from a single individual.

Mitochondria

Genome skimming is frequently used to recover mitochondrial reads for species identification, mitogenome assembly, or other purposes; however, mitochondria are largely unmethylated. To confirm levels seen in other methods, I selected out mitochondria in the mouse samples. I filtered the reads assigned to "chrM" for both mouse tissues and in blood for the chimpanzee. Methylation analysis revealed no significant differences from the low controls (average 0.20%) for each of the three mouse tissues, whereas the high controls averaged 97.34% methylation. The

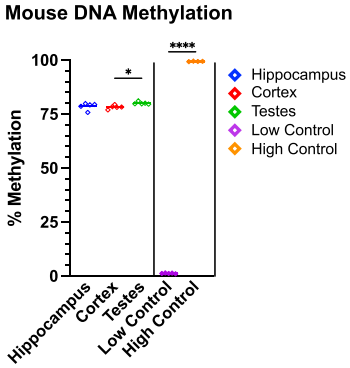


Figure 2. Replicate measures of DNA methylation in the mouse. Hippocampus, cortex, and testes are biological replicates. Control samples are technical replicates. Significance is compared within tissues and controls but not between them. (*) $P \leq 0.05$, (****) $P \leq 0.0001$.

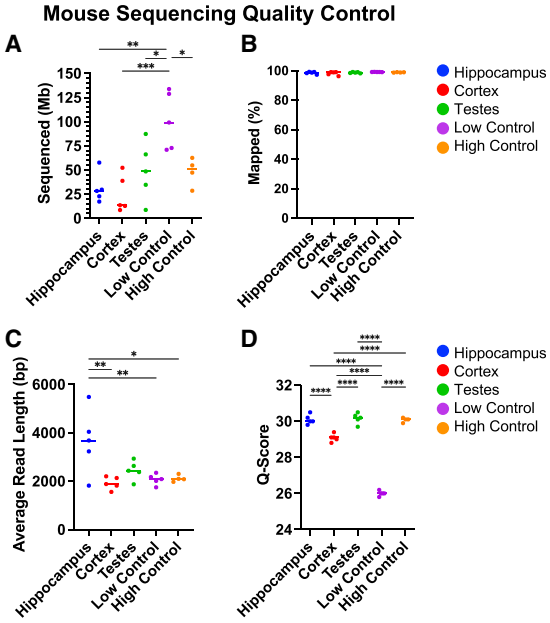


Figure 3. Mouse sequencing quality control. (*A*) Number bases sequenced in megabases. (*B*) Percentage of bases mapped to mm39 reference genome. (*C*) Average read length in base pairs. (*D*) Average read quality scores. (*) $P \leq 0.05$, (**) $P \leq 0.01$, (***) $P \leq 0.001$, (****) $P \leq 0.0001$.

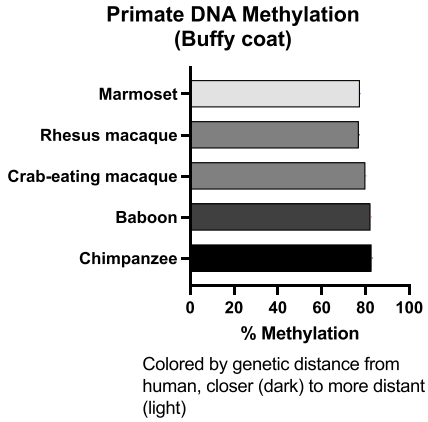


Figure 4. DNA methylation from primate buffy coat. The bars are colored by genetic distance from human, dark to light.

mouse mitochondrial depth ranged from 0.39× to 374× (average 12×) across all tissues, in line with sequencing depth. Because the average mouse sample was sequenced to only 35 Mb total, by increasing sequencing to the 300-Mb level, a coverage of 100× per mitogenome could be generated, enabling efficient assembly of mitogenomes. The mitochondrial-to-nuclear DNA ratio is typically reported as either mitochondria copy number per cell (mtDNA-CN) or as a natural log ratio (Fig. 9). The hippocampus and cortex averaged an mtDNA-CN of 1544 and 1738, respectively. The testes averaged an order of magnitude less at 185 copy number per nuclear genome.

Discussion

Genome skimming, also called low-pass or low-coverage sequencing, is defined as shallow sequencing down to 0.05× coverage of a genome (Li et al. 2021). Here I determined that just 0.01× coverage is sufficient to achieve high precision in global DNA methylation assessment in vertebrate genomes. Genome skimming has been used for diverse purposes, including plastid genome assembly, parasite identification, and evolutionary biology in extinct species, and is gaining software tools for phylogenetic analyses (Grewe et al. 2021; Ferrari et al. 2022; Odago et al. 2022; Reginato 2022; Papaiakovou et al. 2023). Classical genome skimming focuses on mitochondrial genome recovery for species or individual identification but does not typically use the nuclear portion of the skimmed reads or their DNA methylation. There are only six reports of nanopore sequencing used for genome skimming in animals at the time of publication, only one of which examines the epigenome (Euskirchen et al. 2017; Gan et al. 2019; Johri et al. 2019; Franco-Sierra and Díaz-Nieto 2020; De Vivo et al. 2022; Kneubehl et al. 2022). In only one case has genome skimming been used to reconstruct mitochondria with a methylation-aware base caller, but they did not assess either mitogenome or nuclear DNA methylation (Franco-Sierra and Díaz-Nieto 2020). The generation of long reads is crucial to improve alignment over traditional bisulfite-based short reads (Rachtman et al. 2020). The further use of nanopore sequencers for genome skimming will enable the acquisition of DNA methylation in parallel to the DNA sequence data for no additional cost.

Table 3. Summary statistics for vertebrates by class

Class	Mean genome size (Gb)	Mean methylation (%)	Median (Mb)	Mean % mapped	Mean quality
Amphibian	6.2 or 28.2	90.84	38.1	98.0	28.8
Fish	3.2	83.81	54.1	98.5	29.8
Reptile	1.4	81.20	19.8	96.1	26.4
Mammal	2.5	74.65	85.2	97.4	29.0
Bird	1.1	66.01	132.2	97.8	29.9

Genome skimming for DNA methylation is applicable to any method that generates nanopore sequence data, even in small quantities such as field sequencing. For instance, single-cell skimming is possible because only 30 Mb per sample is needed. With low cost enabled by sample multiplexing, genome skimming becomes ideal for epidemiological exposure monitoring for small-magnitude global changes in DNA methylation in a population. For this experiment, I skimmed 72 samples, yielding 12 Gb of sequencing. A typical minion flow cell can generate 15 Gb of data, so 96 samples could be multiplexed on a single flow cell. Bioinformatic pipelines are maturing, and global methylation is expected to become a commonly reported metric for every type of nanopore run in the future.

To validate the level of precision necessary, I sequenced five biological replicates of three mouse tissues at skimming coverage levels (0.0025× to 0.016×). Detection range and accuracy are confirmed by repeated measures of the technical replicate control samples. Biological replication is validated by highly consistent interindividual measurements

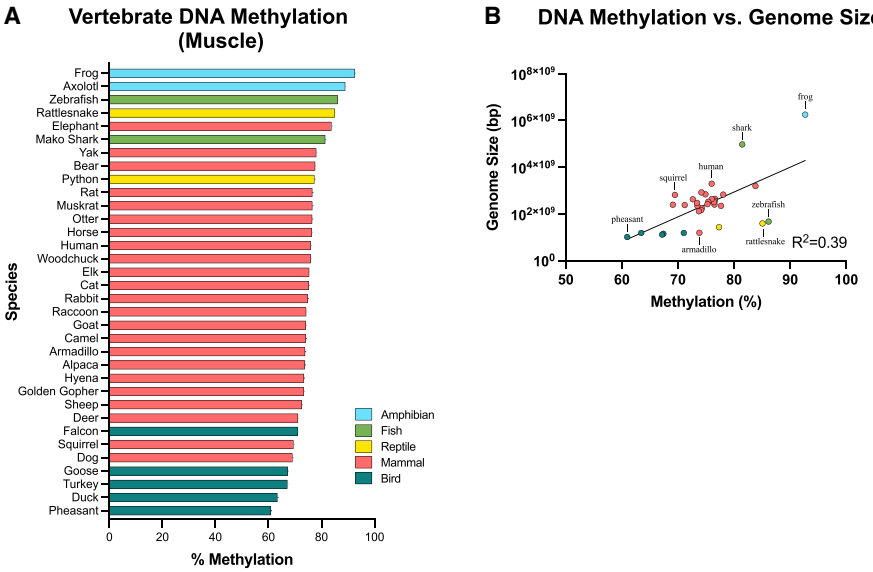


Figure 5. Vertebrate DNA methylation. (*A*) Global DNA methylation from skeletal muscle in vertebrates. (*B*) Genome methylation percentage versus genome size in base pairs. Axolotl was removed as an outlier, as it is an order of magnitude larger than other genomes.

Table 4. Chimpanzee *Alu* and LINE-1 methylation with 10 bootstrap replicates

	Alu		LINE-1	
Coverage	Mean (%)	SD (%)	Mean (%)	SD (%)
10× (30 Gb)	94.69	0.00	90.16	0.01
1× (3 Gb)	94.70	0.04	90.15	0.03
0.1× (300 Mb)	94.70	0.04	90.17	0.10
0.01× (30 Mb)	94.77	0.13	90.19	0.28
0.001× (3 Mb)	94.39	0.68	90.11	1.15
0.0001× (300 kb)	94.26	1.35	90.34	3.04

among multiple tissues in the mouse biological replicates. Compared with other methods of quantifying DNA methylation, genome skimming is quite accurate and reproducible. Although other methods exist to quantify global DNA methylation, this nanopore-based method has an attractive combination of ease of use, low cost, low barrier to entry, and high precision.

Currently the most common method for global DNA methylation uses a colorimetric antibody-based commercial ELISA kit, which suffers from poor resolution with small-magnitude differences and poor repeatability overall (Cantrell et al. 2020). Further, it only measures the methylated cytosine–to–total cytosine ratio, making its relation to cytosines in CpG context difficult to interpret.

Liquid chromatography uses native DNA degraded to single bases, losing positional information, and suffers from similar challenges as the ELISA method. A recent study using liquid chromatography across the tree of life assessed DNA modifications and reported a large variation in biological replicates. By the nature of the assay, it also reports only the methylated cytosine–to–total cytosine ratio and is difficult to extrapolate to cytosines in CpG context (Varma et al. 2022).

Global methods relying upon the Qiagen Pyromark line of pyrosequencers were formerly popular such as the luminometric assay (LUMA) and the LINE-1, *Alu*, and mouse IAP amplicon assays (Faulk et al. 2013, 2015). The LUMA assay relies upon isoschizomer methylation-specific enzymatic digestion followed by pyrosequencer quantitation (Karimi et al. 2006). The LINE-1 and *Alu* assays rely upon bisulfite conversion and determine methylation solely within transposon context, obscuring changes outside transposon regions (Poulin et al. 2018).

I skimmed 40 different vertebrate species across fish, amphibians, reptiles, mammals, and primates, generating global DNA

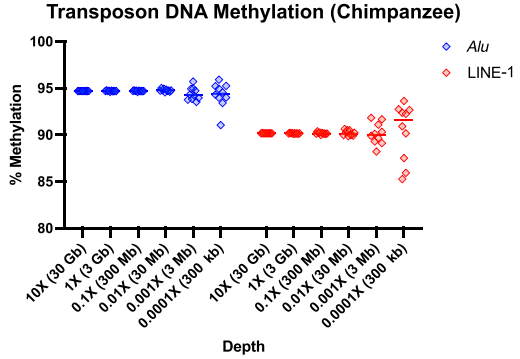


Figure 6. Transposon DNA methylation from chimpanzee buffy coat. From a single 11.72× coverage chimpanzee genome, 10 subsamples of the genome were taken according to coverage level. *Alu* and LINE-1 methylation are shown for the 10 bootstrap replicates.

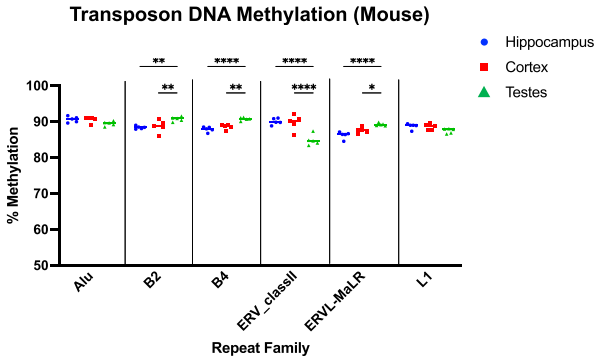


Figure 7. Transposon DNA methylation from mouse. Transposon families above a threshold of 2% genomic content are shown. DNA methylation is calculated from biological replicates of mouse tissues with coverage ranging from 0.0025× to 0.016×. Transposon families with significant differences are shown with ANOVA followed by Tukey's multiple comparison test. (*) $P \leq 0.05$, (**) $P \leq 0.01$, (****) $P \leq 0.0001$.

methylation values and transposon methylation. A study using WGBS in multiple vertebrate dermal fibroblasts showed global methylation that recapitulates the pattern of results very well (Al Adhami et al. 2022). The most comprehensive previous study of animal methylation assayed 580 species and matches the global patterns presented here (Klughammer et al. 2023). However, this impressive catalog was created using RRBS and is subject to the typical challenges caused by bisulfite conversion, as well as enrichment for CpG dense regions and lack of accurate transposon identification owing to short reads and loss of mapping ability.

I found that genome size directly correlates to methylation, explaining 39% of the variance, with a clear phylogenetic component. Clade-specific methylation results replicate the phyloepigenetic tree built by Haghani et al. (2021) that was made with a custom 40k epigenetic microarray used on 176 species.

Genome skimming is effective for transposon identification because repeats are present in many thousands of copies per nuclear genome and are known to be methylated at rates near 80% (Crary-Dooley et al. 2017). Transposon methylation in the LINE-1 element is often used as a proxy for global methylation in epidemiological studies (Perera et al. 2019); however, repeat copies are not truly identical and are divided by diagnostic mutations down to the subfamily and individual insertion level. Long reads allow precise identification of repeat type and often identify their exact genomic coordinates. Prior methods to determine DNA methylation of transposons included pyrosequencing, which amplifies a random mixture of transposons of a particular family and quantitates methylation at few specific CpG sites that are of unknown conservation in the bisulfite amplicon, leading to error and poor replication. Even at 0.001× coverage, equating to just 3

Table 5. Mouse repeat DNA methylation by family and tissue

Repeat family	Hippocampus (%)	SD (%)	Cortex (%)	SD (%)	Testes (%)	SD (%)
I-B1	90.60	0.83	90.54	0.79	87.99	2.91
B2	88.36	0.43	88.32	1.39	90.58	0.65
B4	87.87	0.70	84.82	8.22	90.80	0.46
ERV class II	90.03	0.89	89.75	2.14	85.43	1.66
ERVL-MaLR	86.22	1.00	76.68	24.25	89.16	0.38
L1	88.71	0.81	87.87	2.17	88.76	3.24

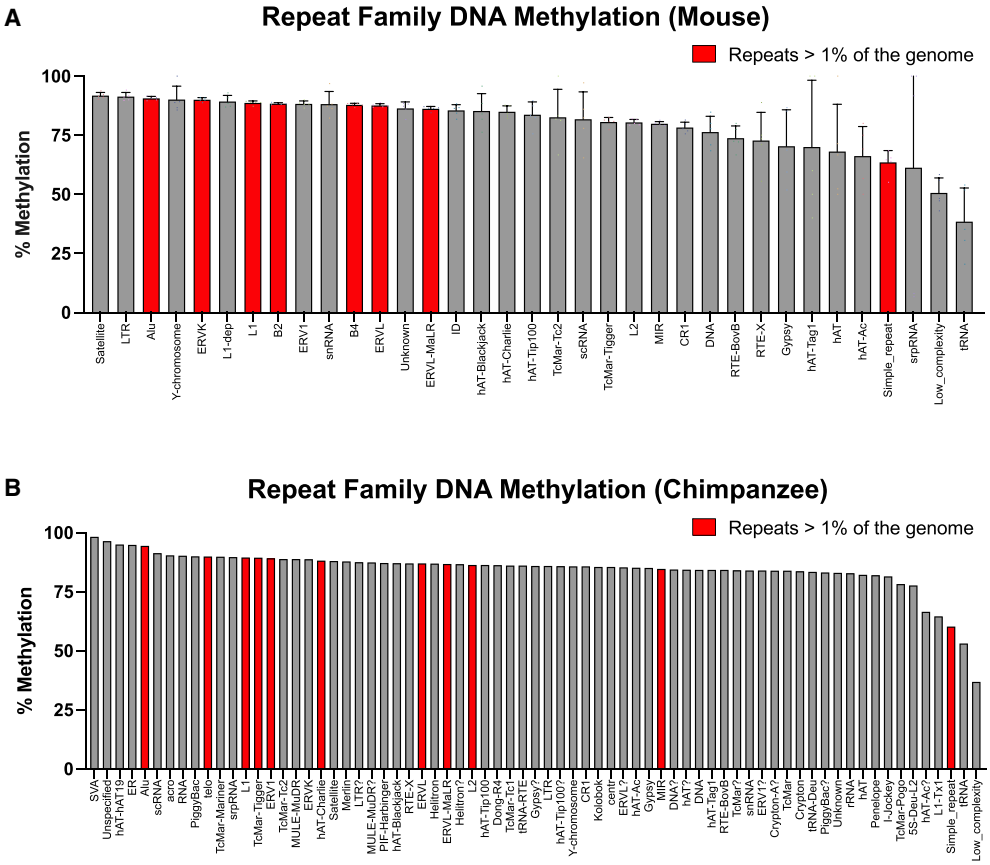


Figure 8. All repeat family DNA methylation. (*A*) All repeat families from mouse are shown with replication error indicated by error bars. (*B*) DNA methylation from chimpanzee is shown by repeat family, represented by a single individual. Transposons shown in red make up >1% of the genome of their respective species.

Mb of sequence, I was able to accurately determine *Alu* transposon methylation in the mouse genome, with <1% deviation between biological replicates. Genome skimming has been used in plants to identify and classify the extent of transposon content, but until now, it has not been used to assess DNA methylation of transposons (Choudhury et al. 2017).

I found biologically meaningful results in the mouse transposome, showing that ERV class II elements are substantially hypomethylated in mouse testes compared with other tissues. This is relevant because the most mutagenic transposon in the mouse genome, the IAP element, is an ERV class II, and its hypomethylation in testes suggests germ-line mobilization (Barbot 2002; Faulk et al. 2013).

Genome skimming is often used to sequence mitochondria owing to their naturally high copy number compared with the nuclear genome (Furtwängler et al. 2018). Typical uses are in species identification, assembly of mitogenomes, and ancient DNA haplotyping. In this study, the extraction and library preparation methods are not optimized for mtDNA, and therefore, counts should not be used as absolute values; however, they fall broadly in line with mtDNA: nuclear DNA ratios seen by others (Furtwängler et al. 2018; Longchamps et al. 2020). Even in the tissues with the highest mitochondrial copy number, approaching 1800:1, the mtDNA did not account for >6% of the DNA by weight (or base count), so depletion is not necessary. Intratissue variation in copy number was low, showing consistency and significant differences between tissues. DNA methylation of the mitochondria was extremely low, indistinguishable from the low controls, in line with other studies (Stoccoro and Coppedè 2021).

Regarding nonmodel organisms, good reference genomes are required, which may not exist for a given species. FASTmC is an alternative method for reference-free assessment of methylation using WGBS or inexpensive RRBS (Bewick et al. 2016). Also a new method of enzymatic methyl-seq has become available that mimics bisulfite conversion without the high damage and short reads typical of bisulfite treatment. For bisulfite methods in general, Bismark and other pipelines are well established. Regarding species' differences in global methylation, other variables beyond genome size could affect methylation, for example, differences in CpG island length and frequency, gene number, and, especially, transposon complement.

Genome skimming with nanopore sequencers is highly effective for global DNA methylation and transposon measurement. Multiplexing along with the ability to pause sequencing allows flexible experimental design and mitigates cost concerns. With improved base calling software and bioinformatic pipelines, global DNA methylation will likely be a basic quality metric for any type of nanopore run. Nanopore sequencing is revitalizing genome skimming in ecology with the detection of cryptic hybridization of threatened primates and in noninvasive methods to assemble mitogenomes (Malukiewicz et al. 2021; Wanner et al. 2021). Extending its utility by adding DNA methylation capability

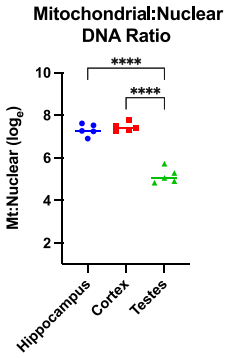


Figure 9. Mitochondrial-to-nuclear (mt:nuc) DNA. The mitochondrial-to-nuclear ratio is significantly reduced in mouse testes. (****) $P < 0.0001$.

will aid the use of genome skimming in epidemiology, conservation, and comparative genomics.

Methods

DNA extraction

DNA from primate buffy coat samples was extracted Qiagen DNA mini kit (cat. 51306) by the Southwest National Primate Center (SNPRC) and treated with Qiagen RNase A. DNA from mouse tissues and all vertebrate muscle samples was extracted using a Zymo quick-DNA miniprep plus kit (cat. D4068) by the author. Vertebrate skeletal muscle tissues were sourced from commercial meat suppliers or donated by research laboratories. The human muscle sample was obtained commercially from BioChemed Services and was deidentified before purchase. Mice in this study were of the agouti strain, wild-type allele sharing at least 93% similarity with C57BL/6 (Weinhouse et al. 2014). All were males 6–8 wk of age. Mice in this study were maintained in accordance with the guidelines for the care and use of laboratory animals and were treated humanely and with regard for alleviation of suffering. The study protocol was approved by the University of Minnesota institutional animal care and use committee (IACUC). Primate samples were acquired from the SNPRC from incidental blood collection during routine veterinary care and are under the aegis of SNPRC's IACUC approval.

Control DNA was generated by treatment of mouse liver gDNA extracted via a Zymo quick-DNA miniprep plus kit. For the low methylated control, 100 ng of liver gDNA was used as input to a Qiagen REPLI-g mini kit (150023) for whole-genome amplification, following the manufacturer's protocol. The resulting DNA was cleaned using a Axygen AxyPrep MAG PCR clean-up kit (Corning MAG-PCR-CL-50). DNA was incubated with 2× volume of magnetic bead solution on a shaker for 5 min, washed twice with 70% EtOH, and eluted with 100 µL of elution buffer. The yield was 10 µg per reaction. The highly methylated control DNA was generated using 10 µg of liver-derived gDNA as input in separate reactions with a Zymo CpG methylase kit (E2010). Output DNA was cleaned with Axygen beads, and the yield was 70% of input. Control DNA was then used for library prep and sequencing like any other sample extract.

Library preparation and run conditions

DNA from four primates was barcoded and prepped for sequencing using the native barcoding kit SQK-NBD114.24 following the manufacturer's instructions from Oxford Nanopore Technologies. The marmoset sample was prepared with the EXP-NBD104 barcoding kit and the SQK-LSK109 ligation sequencing kit. Primate DNA was sequenced at the University of Wisconsin Biotechnology Center, Sequencing Core on a PromethION 24 instrument. Two samples were multiplexed per PromethION flow cell.

All tissues from mouse and vertebrate muscle were library prepped using the rapid barcoding kit SQK-RBK114.24 following the manufacturer's instructions by the author at the University of Minnesota. Mouse tissue DNA biological replicates along with control DNA were run in a 24-sample multiplex along with technical replicates of five low and four high methylation individually barcoded controls from the same control reaction aliquots on a single MinION R10.4.1 flow cell for 24 h. Vertebrate tissue DNA was run in a 24-sample multiplex on a second MinION R10.4.1 flow cell for 24 h. All MinION runs were performed in the author's laboratory.

Alignment and methylation assessment

The complete bioinformatic pipeline is available in the Supplemental Methods and at GitHub (https://faulk-lab.github.io/skimming/). Briefly, Guppy v6.3.8 was used to call bases for all species with the super accuracy model "dna_r10.4.1_e8.2_400bps_modbases_5mc_cg_sup," which also calls 5 mC in CpG context. Guppy generates mapped or unmapped BAM files, depending on reference given. Here the reads were aligned post hoc with minimap2 (Li 2018), retaining modification flags by converting the BAM files to FASTQ with SAMtools (Li et al. 2009) v1.16.1 and mapping with minimap2 to create a mapped BAM file. Mapping efficiency was calculated using bamUtils v1.0.16 (https://genome.sph.umich.edu/wiki/BamUtil). Next, the BAM files were converted to BED format with modkit v0.1.4 (https://github.com/nanoporetech/modkit). Summary methylation was calculated with an awk script available in the detailed Supplemental Methods. Read and genome summary statistics were calculated with SeqKit (Shen et al. 2016). Differences in mouse tissue methylation were calculated with an ordinary one-way ANOVA followed by a Tukey's multiple comparison test. Differences in low versus high control methylation were calculated with an unpaired *t*-test. All mouse sequencing quality tests were compared with ordinary one-way ANOVAs with a Tukey's multiple comparison test.

All software was run on a single computer running Ubuntu v22.04 on an AMD 7950 Ryzen 32 thread CPU with 128-GB memory, 4-TB SSD, and an Nvidia 3080 Ti GPU for accelerated base calling. All statistics and figures were created using GraphPad Prism v9.4.1.

Genomes used in this study are listed in Table 6.

Vertebrate statistics

Vertebrate methylation was calculated similarly to that of mouse and primates, with alignment to the reference genome for each species downloaded from NCBI with accession numbers available as Supplemental File S1. Genome size versus methylation excluded axolotl as an outlier because its genome, at 28 Gb, is an order of magnitude larger than any other vertebrate. However, fitting with the pattern, it has the highest methylation percentage of any vertebrate sequenced.

Transposon analysis

Mouse (mm39) and chimpanzee (panTro6) repeatmasked tracks were downloaded from UCSC Genome Browser's Table Browser function with the following parameters: "Table Browser → Repeats → Table "rmsk" → All fields from selected table → repeat-

Table 6. Genome assemblies

Binomial	Genome
Rana catesbeiana	RCv2.1
Ambystoma mexicanum	AmbMex60DD
Danio rerio	GRCz11
Crotalus atrox	Cadam_11369
Loxodonta africana	Loxafr3.0
Isurus oxyrinchus	ASM2677070v1
Bos grunniens	BosGru3.1
Ursus americanus	gsc_jax_bbear_1.0
Python bivittatus	Python_molurus_bivittatus-5.0.2
Rattus norvegicus	mRatBN7.2
Ondatra zibethicus	OndZib_v1_BIUU
Lontra canadensis	GSC_riverotter_1.0
Equus caballus	EquCab3.0
Homo sapiens	GRCh38.p14
Marmota monax	Marmota_monax_Labrador192_F-V1.1
Cervus canadensis	ASM1932006v1
Felis catus	F.catus_Fca126_mat1.0
Oryctolagus cuniculus	UM_NZW_1.0
Procyon lotor	Plotor_platanus500
Capra hircus	ARS1.2
Camelus dromedarius	CamDro3
Dasypus novemcinctus	Dasnov3.2
Vicugna pacos	VicPac3.2
Crocuta crocuta	BGI_CrCroc_1.0
Ictidomys tridecemlineatus	HiC_Itri_2
Ovis aries	ARS-UI_Ramb_v2.0
Odocoileus virginianus	Ovir.te_1.0
Falco columbarius	ASM2756420v1
Sciurus carolinensis	mSciCar1.2
Canis familiaris	ROS_Cfam_1.0
Branta canadensis	GSC_cangoose_1.0
Meleagris gallopavo	Turkey_5.1
Anas platyrhynchos	ZJU1.0
Phasianus colchicus	ASM414374v1
Mus musculus	mm39

table.tsv." The table was converted to BED format with a custom awk script. BEDTools was used to intersect the genome-wide methylation BED file generated with modkit versus the repeatmask track BED file. An R script in R v4.2.2 was used to summarize the methylation values by repeat family and statistics applied were applied in GraphPad Prism (R Core Team 2022). Mouse transposon family significance was calculated with a two-way ANOVA followed by a Tukey's multiple comparison test. All scripts are available in Supplemental Methods.

Mitochondria analysis

Mouse and chimpanzee BED files were filtered for CpG sites mapped to "chrM" in each species and assessed for global methylation.

Data access

All processed sequencing data generated in this study are stored as BAM files containing modified base calls and have been submitted to the NCBI BioProject database (https://www.ncbi.nlm.nih.gov/bioproject/) under accession number PRJNA927034.

Competing interest statement

The author declares no competing interests.

Acknowledgments

I thank Arthur Rand, developer of Modkit, for development and updates to improve the accuracy of DNA methylation calls in BAM-to-BED file conversion. This work was supported by National Institutes of Health Grant R21AG071908 (C.F.), Impetus Grant (Norn Foundation; C.F.), and U.S. Department of Agriculture–National Institute of Food and Agriculture Grant USDA-NIFA MIN-16-129 (C.F.).

Author contributions: C.F. conceived the study, performed the sequencing, analyzed the data, and wrote the manuscript. C.F. read and approved the manuscript before submission.

References

Al Adhami H, Bardet AF, Dumas M, Cleroux E, Guibert S, Fauque P, Acloque H, Weber M. 2022. A comparative methylome analysis reveals conservation and divergence of DNA methylation patterns and functions in vertebrates. *BMC Biol* **20:** 70. doi:10.1186/s12915-022-01270-x

Baeza JA, Molina-Quirós JL, Hernández-Muñoz S. 2021. Genome survey sequencing of an iconic "trophy" sportfish, the roosterfish *Nematistius pectoralis*: genome size, repetitive elements, nuclear RNA gene operon, and microsatellite discovery. *Genes (Basel)* **12:** 1710. doi:10.3390/genes12111710

Barbot W. 2002. Epigenetic regulation of an IAP retrotransposon in the aging mouse: progressive demethylation and de-silencing of the element by its repetitive induction. *Nucleic Acids Res* **30:** 2365–2373. doi:10.1093/nar/30.11.2365

Bewick AJ, Hofmeister BT, Lee K, Zhang X, Hall DW, Schmitz RJ. 2016. *FAST^{m}C*: a suite of predictive models for nonreference-based estimations of DNA methylation. *G3 (Bethesda)* **6:** 447–452. doi:10.1534/g3.115.025668

Breton CV, Marsit CJ, Faustman E, Nadeau K, Goodrich JM, Dolinoy DC, Herbstman J, Holland N, LaSalle JM, Schmidt R, et al. 2017. Small-magnitude effect sizes in epigenetic end points are important in children's environmental health studies: the Children's Environmental Health and Disease Prevention Research Center's Epigenetics Working Group. *Environ Health Perspect* **125:** 511–526. doi:10.1289/EHP595

Cai L, Zhang H, Davis CC. 2022. PhyloHerb: a high-throughput phylogenomic pipeline for processing genome skimming data. *Appl Plant Sci* **10:** e11475. doi:10.1002/aps3.11475

Cantrell B, Friedman S, Lachance H, Bernier C, Murdoch B, Frattini S, Talenti A, Crepaldi P, McKay S. 2020. A novel understanding of global DNA methylation in bobcat (*Lynx rufus*). *Genome* **63:** 125–130. doi:10.1139/gen-2019-0046

Choudhury RR, Neuhaus J-M, Parisod C. 2017. Resolving fine-grained dynamics of retrotransposons: comparative analysis of inferential methods and genomic resources. *Plant J Cell Mol Biol* **90:** 979–993. doi:10.1111/tpj.13524

Crary-Dooley FK, Tam ME, Dunaway KW, Hertz-Picciotto I, Schmidt RJ, LaSalle JM. 2017. A comparison of existing global DNA methylation assays to low-coverage whole-genome bisulfite sequencing for epidemiological studies. *Epigenetics* **12:** 206–214. doi:10.1080/15592294.2016.1276680

De Vivo M, Lee H-H, Huang Y-S, Dreyer N, Fong C-L, de Mattos FMG, Jain D, Wen Y-HV, Mwihaki JK, Wang T-Y, et al. 2022. Utilisation of Oxford Nanopore sequencing to generate six complete gastropod mitochondrial genomes as part of a biodiversity curriculum. *Sci Rep* **12:** 9973. doi:10.1038/s41598-022-14121-0

Euskirchen P, Bielle F, Labreche K, Kloosterman WP, Rosenberg S, Daniau M, Schmitt C, Masliah-Planchon J, Bourdeaut F, Dehais C, et al. 2017. Same-day genomic and epigenomic diagnosis of brain tumors using real-time nanopore sequencing. *Acta Neuropathol* **134:** 691–703. doi:10.1007/s00401-017-1743-5

Faulk C. 2022. *De novo* sequencing, diploid assembly, and annotation of the black carpenter ant, *Camponotus pennsylvanicus*, and its symbionts by one person for $1000, using nanopore sequencing. *Nucleic Acids Res* **51:** 17–28. doi:10.1093/nar/gkac510

Faulk C, Barks A, Dolinoy DC. 2013. Phylogenetic and DNA methylation analysis reveal novel regions of variable methylation in the mouse IAP class of transposons. *BMC Genomics* **14:** 48. doi:10.1186/1471-2164-14-48

Faulk C, Kim JH, Jones TR, McEachin RC, Nahar MS, Dolinoy DC, Sartor MA. 2015. Bisphenol A-associated alterations in genome-wide DNA methylation and gene expression patterns reveal sequence-dependent and non-monotonic effects in human fetal liver. *Environ Epigenetics* **1:** dvv006. doi:10.1093/eep/dvv006

Ferrari G, Atmore LM, Jentoft S, Jakobsen KS, Makowiecki D, Barrett JH, Star B. 2022. An accurate assignment test for extremely low-coverage whole-genome sequence data. *Mol Ecol Resour* **22:** 1330–1344. doi:10.1111/1755-0998.13551

Franco-Sierra ND, Díaz-Nieto JF. 2020. Rapid mitochondrial genome sequencing based on Oxford Nanopore Sequencing and a proxy for vertebrate species identification. *Ecol Evol* **10:** 3544–3560. doi:10.1002/ece3.6151

Furtwängler A, Reiter E, Neumann GU, Siebke I, Steuri N, Hafner A, Lösch S, Anthes N, Schuenemann VJ, Krause J. 2018. Ratio of mitochondrial to nuclear DNA affects contamination estimates in ancient DNA analysis. *Sci Rep* **8:** 14075. doi:10.1038/s41598-018-32083-0

Gan HM, Linton SM, Austin CM. 2019. Two reads to rule them all: nanopore long read-guided assembly of the iconic Christmas Island red crab, *Gecarcoidea natalis* (Pocock, 1888), mitochondrial genome and the challenges of AT-rich mitogenomes. *Mar Genomics* **45:** 64–71. doi:10.1016/j.margen.2019.02.002

Grewe F, Kronforst MR, Pierce NE, Moreau CS. 2021. Museum genomics reveals the Xerces blue butterfly (*Glaucopsyche xerces*) was a distinct species driven to extinction. *Biol Lett* **17:** 20210123. doi:10.1098/rsbl.2021.0123

Haghani A, Lu AT, Li CZ, Robeck TR, Belov K, Breeze CE, Brooke RT, Clarke S, Faulkes CG, Fei Z, et al. 2021. DNA methylation networks underlying mammalian traits. bioRxiv doi:10.1101/2021.03.16.435708

Jahan NA, Lindsey LL, Kipp EJ, Reinschmidt A, Heins BJ, Runck AM, Larsen PA. 2021. Nanopore-based surveillance of zoonotic bacterial pathogens in farm-dwelling peridomestic rodents. *Pathogens* **10:** 1183. doi:10.3390/pathogens10091183

Johri S, Solanki J, Cantu VA, Fellows SR, Edwards RA, Moreno I, Vyas A, Dinsdale EA. 2019. "Genome skimming" with the MinION hand-held sequencer identifies CITES-listed shark species in India's exports market. *Sci Rep* **9:** 4476. doi:10.1038/s41598-019-40940-9

Karimi M, Johansson S, Ekström TJ. 2006. Using LUMA: a luminometric-based assay for global DNA-methylation. *Epigenetics* **1:** 45–48. doi:10.4161/epi.1.1.2587

Klughammer J, Romanovskaia D, Nemc A, Posautz A, Seid CA, Schuster LC, Keinath MC, Lugo Ramos JS, Kosack L, Evankow A, et al. 2023. Comparative analysis of genome-scale, base-resolution DNA methylation profiles across 580 animal species. *Nat Commun* **14:** 232. doi:10.1038/s41467-022-34828-y

Kneubehl AR, Muñoz-Leal S, Filatov S, de Klerk DG, Pienaar R, Lohmeyer KH, Bermúdez SE, Suriyamongkol T, Mali I, Kanduma E, et al. 2022. Amplification and sequencing of entire tick mitochondrial genomes for a phylogenomic analysis. *Sci Rep* **12:** 19310. doi:10.1038/s41598-022-23393-5

Laubach ZM, Perng W, Dolinoy DC, Faulk CD, Holekamp KE, Getty T. 2018. Epigenetics and the maintenance of developmental plasticity: extending the signalling theory framework. *Biol Rev Camb Philos Soc* **93:** 1323–1338. doi:10.1111/brv.12396

Li H. 2018. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34:** 3094–3100. doi:10.1093/bioinformatics/bty191

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, 1000 Genome Project Data Processing Subgroup. 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25:** 2078–2079. doi:10.1093/bioinformatics/btp352

Li JH, Mazur CA, Berisa T, Pickrell JK. 2021. Low-pass sequencing increases the power of GWAS and decreases measurement error of polygenic risk scores compared to genotyping arrays. *Genome Res* **31:** 529–537. doi:10.1101/gr.266486.120

Longchamps RJ, Castellani CA, Yang SY, Newcomb CE, Sumpter JA, Lane J, Grove ML, Guallar E, Pankratz N, Taylor KD, et al. 2020. Evaluation of mitochondrial DNA copy number estimation techniques. *PLoS One* **15:** e0228166. doi:10.1371/journal.pone.0228166

Malukiewicz J, Cartwright RA, Dergam JA, Igayara CS, Nicola PA, Pereira LMC, Ruiz-Miranda CR, Stone AC, Silva DL, da Silva FDFR, et al. 2021. Genomic skimming and nanopore sequencing uncover cryptic hybridization in one of world's most threatened primates. *Sci Rep* **11:** 17279. doi:10.1038/s41598-021-96404-6

Nurk S, Koren S, Rhie A, Rautiainen M, Bzikadze AV, Mikheenko A, Vollger MR, Altemose N, Uralsky L, Gershman A, et al. 2022. The complete sequence of a human genome. *Science* **376:** 44–53. doi:10.1126/science.abj6987

Odago WO, Waswa EN, Nanjala C, Mutinda ES, Wanga VO, Mkala EM, Oulo MA, Wang Y, Zhang C-F, Hu G-W, et al. 2022. Analysis of the complete plastomes of 31 species of *Hoya* group: insights into their comparative genomics and phylogenetic relationships. *Front Plant Sci* **12:** 814833. doi:10.3389/fpls.2021.814833

Papaiakovou M, Fraija-Fernández N, James K, Briscoe AG, Hall A, Jenkins TP, Dunn J, Levecke B, Mekonnen Z, Cools P, et al. 2023. Evaluation of genome skimming to detect and characterise human and livestock helminths. *Int J Parasitol* **53:** 69–79. doi:10.1016/j.ijpara.2022.12.002

Perera BPU, Faulk C, Svoboda LK, Goodrich JM, Dolinoy DC. 2019. The role of environmental exposures and the epigenome in health and disease. *Environ Mol Mutagen* **61:** 176–192. doi:10.1002/em.22311

Poulin M, Zhou JY, Yan L, Shioda T. 2018. Pyrosequencing methylation analysis. *Methods Mol Biol* **1856:** 283–296. doi:10.1007/978-1-4939-8751-1_17

Rachtman E, Balaban M, Bafna V, Mirarab S. 2020. The impact of contaminants on the accuracy of genome skimming and the effectiveness of exclusion read filters. *Mol Ecol Resour* **20:** 649–661. doi:10.1111/1755-0998.13135

R Core Team. 2022. *R: a language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna. https://www.R-project.org/.

Reginato M. 2022. A pipeline for assembling low copy nuclear markers from plant genome skimming data for phylogenetic use. *PeerJ* **10:** e14525. doi:10.7717/peerj.14525

Sarkar A, Lanciano S, Cristofari G. 2023. Targeted nanopore resequencing and methylation analysis of LINE-1 retrotransposons. *Methods Mol Biol* **2607:** 173–198. doi:10.1007/978-1-0716-2883-6_10

Shen W, Le S, Li Y, Hu F. 2016. SeqKit: a cross-platform and ultrafast toolkit for FASTA/Q file manipulation. *PLoS One* **11:** e0163962. doi:10.1371/journal.pone.0163962

Stoccoro A, Coppedè F. 2021. Mitochondrial DNA methylation and human diseases. *Int J Mol Sci* **22:** 4594. doi:10.3390/ijms22094594

Twyford AD, Ness RW. 2017. Strategies for complete plastid genome sequencing. *Mol Ecol Resour* **17:** 858–868. doi:10.1111/1755-0998.12626

Varma SJ, Calvani E, Grüning N-M, Messner CB, Grayson N, Capuano F, Mülleder M, Ralser M. 2022. Global analysis of cytosine and adenine DNA modifications across the tree of life. *eLife* **11:** e81002. doi:10.7554/eLife.81002

Wanner NM, Faulk C. 2021. Suggested absence of horizontal transfer of retrotransposons between humans and domestic mammal species. *Genes (Basel)* **12:** 1223. doi:10.3390/genes12081223

Wanner N, Larsen PA, McLain A, Faulk C. 2021. The mitochondrial genome and epigenome of the golden lion tamarin from fecal DNA using nanopore adaptive sequencing. *BMC Genomics* **22:** 726. doi:10.1186/s12864-021-08046-7

Weinhouse C, Anderson OS, Bergin IL, Vandenbergh DJ, Gyekis JP, Dingman MA, Yang J, Dolinoy DC. 2014. Dose-dependent incidence of hepatic tumors in adult mice following perinatal exposure to bisphenol A. *Environ Health Perspect* **122:** 485–491. doi:10.1289/ehp.1307449

Zhang L, Huang Y-W, Huang J-L, Ya J-D, Zhe M-Q, Zeng C-X, Zhang Z-R, Zhang S-B, Li D-Z, Li H-T, et al. 2023. DNA barcoding of *Cymbidium* by genome skimming: call for next-generation nuclear barcodes. *Mol Ecol Resour* **23:** 424–439. doi:10.1111/1755-0998.13719

Received January 27, 2023; accepted in revised form June 7, 2023.