OXFORD

Briefings in Bioinformatics, 2023, **24(1)**, 1–12
https://doi.org/10.1093/bib/bbac617
Problem Solving Protocol

Comparative epigenome analysis using Infinium DNA methylation BeadChips

Wubin Ding, Diljeet Kaur, Steve Horvath and Wanding Zhou
Corresponding author. Wanding Zhou, E-mail: wanding.zhou@pennmedicine.upenn.edu

Abstract

The arrival of the Infinium DNA methylation BeadChips for mice and other nonhuman mammalian species has outpaced the development of the informatics that supports their use for epigenetics study in model organisms. Here, we present informatics infrastructure and methods to allow easy DNA methylation analysis on multiple species, including domesticated animals and inbred laboratory mice (in *SeSAMe* version 1.16.0+). First, we developed a data-driven analysis pipeline covering species inference, genome-specific data preprocessing and regression modeling. We targeted genomes of 310 species and 37 inbred mouse strains and showed that genome-specific preprocessing prevents artifacts and yields more accurate measurements than generic pipelines. Second, we uncovered the dynamics of the epigenome evolution in different genomic territories and tissue types through comparative analysis. We identified a catalog of inbred mouse strain-specific methylation differences, some of which are linked to the strains' immune, metabolic and neurological phenotypes. By streamlining DNA methylation array analysis for undesigned genomes, our methods extend epigenome research to broad species contexts.

Keywords: DNA methylation, comparative epigenetics, mouse epigenetics

Introduction

DNA methylation, an ancient biochemical innovation traced back to a viral defense mechanism in bacteria [1], is a classical epigenetic mark in higher order eukaryotes [2]. It is extensively implicated in transcriptional regulation [3], cell differentiation [4], organismal development [5] and human diseases [6]. The evolution of genome-wide DNA methylation plays a significant role in regulating transposable element expansion in the host genomes [7], shaping the host genome composition [8] and driving genome evolution [9]. Nonhuman model organisms, e.g. inbred laboratory mouse, and comparative approaches have played a significant role in the discovery of methylation biology and the evolution of epigenetic mechanisms in higher order eukaryotes [10–14]. Recently, DNA methylation was also linked to traits and phenotypes in domesticated animals [15, 16] and postulated to serve biomarkers in livestock improvement programs [17, 18].

The Infinium DNA methylation microarray technologies (e.g. EPIC and HM450 arrays) are among the most widely used DNA methylation assay technologies for human DNA [19]. Despite its success in human epigenetic research, this technology had limited applications in non-human species until recently. This lack of adoption is partly attributable to the need to predetermine the target CpGs during the array design and the human array's limited coverage of the non-human genomes. Several attempts have been made to co-opt the Infinium array to primate and mammalian species using probes targeting evolutionarily conserved sequences [20–26]. Recently, Infinium arrays for non-human species were developed and have become available, best represented by the MM285 array, which targets 284 860 CpGs in the C57BL/6 J mouse genome, and the HorvathMammalMethylChip40 (Mammal40) array, a mammalian array that targets around 37 488 CpGs relatively conserved across mammalian species [27]. These two arrays were designed for multiple mouse strains or mammalian species. Despite the progress in assay development, using one array for multiple genomes is met with technical challenges in analyzing the array data. Notably, one must accommodate more flexible probe usage depending on the input DNA. For example, the array signal detection success is usually determined by comparing the probe signal intensity to the background. One can derive these background signals from the built-in control probes [28] and the out-of-band signals of Infinium-I chemistry [29]. However, both approaches depend on assumed probe alignment and may not faithfully represent the actual signal background when those assumptions fail to hold in arbitrary genomes. Similarly, these probe mapping assumptions affect background subtraction [30], dye bias correction [29] and other data normalization techniques that assume the detection success of most designed probes [31] or correct annotation of the control probes in every target species [32].

To streamline the use of Infinium arrays in model organisms and farm animals and their use for comparative epigenetic studies, we present a suite of informatics solutions to allow

Wubin Ding is a postdoctoral researcher at the Children's Hospital of Philadelphia, his research interest includes bioinformatic data analysis and software or algorithm development.
Diljeet Kaur is a Ph.D. student from the University of Pennsylvania. Her research focuses on developmental and computational epigenetics.
Steve Horvath is a professor in Human Genetics and Biostatistics at the University of California, his methodological research area lies at the intersection of biostatistics, bioinformatics, computational biology, cancer research, genetics, epidemiology, machine learning and systems biology.
Wanding Zhou is an assistant professor at the University of Pennsylvania and Children's Hospital of Philadelphia. His research interests include DNA methylation drift and inference of cell-type-specific epigenetic signals.
Received: October 12, 2022. **Revised:** December 5, 2022. **Accepted:** December 15, 2022

DNA methylation analysis on multiple species and mouse strains. These solutions are implemented in our tool SeSAMe [29] (Version 1.16.0+), which features novel methods and *ab initio* characterization of the EPICv2, EPIC, MM285 and Mammal40 array probes in over 300 species. These methods include (1) inference of the source species based on signal intensities, (2) determination of the probe utility space, including MM285 array probes whose methylation readings are potentially influenced by the presence of inbred mouse strain-specific SNPs, (3) methods and data infrastructure for Infinium-I color channel specification, (4) support of replicate probe design and daughter-strand probe designs, and (5) genome-specific data preprocessing, including signal background subtraction, dye bias correction and detection *P*-value calculation. Applying the new computational pipeline to query inbred mouse strain-associated methylation differences, we identified epigenetic links to metabolic and neurological phenotypes, including differential methylation involved in neuron and liver tissue-specific transcription factoring binding and micro-imprinting. We also uncovered principles of epigenetic regulation at various evolutionary scales in diverse mammalian species. By providing probe annotation and informatics for genome-dependent preprocessing, SeSAMe broadens the application scope of the Infinium DNA methylation BeadChip to arbitrary genomes, facilitating comparative epigenome studies at the population scales.

Results

Utility of Infinium BeadChips on vertebrate genomes

To analyze Infinium BeadChip from arbitrary genome sources, we developed an informatics infrastructure composed of (1) computational methods for online annotation of probe masking and color channel (Figure 1A), (2) a database of probe utility space on candidate genomes (Figure 1B), (3) inference methods of species and mouse strains from probe signal intensities and (4) adapted genome-aware background subtraction, dye bias correction and other data preprocessing methods. Our new data processing workflow (openSesame) allows for customizable end-to-end species- and strain-specific data preprocessing using one line of R code (Figure 1A). Targeting 310 (305 vertebrate) species and 37 inbred mouse strains, we defined the probe utility spaces for each candidate genome. This space, maximized at the designed species/-mouse strain, shrinks in distantly related genomes (Figure 1B). The shrinkage occurs more rapidly in the MM285 and the EPIC arrays than in the Mammal40 array, which targets more conserved genomic territory by design (Figure 1B, Figure S1A). Genic region probes, particularly at the promoter and 3′-end of genes, are more retained (Figure 1C, Figure S1B), and the retention rate tracks overall sequence conservation (Figure 1D).

It is established that adjacent SNPs influence the methylation reading on Infinium BeadChips [33, 34], partially defining probe utility. A naïve multiple regression of an MM285 data set revealed that 34.85% of mouse strain-specific methylation reading is under direct influences of SNPs (Figure S1C, odds ratio = 3.18, $P < 2.2e\text{-}16$). As a positive control, explicitly designed strain-specific SNP probes (rs probes) are highly enriched for strain-specific readings (Figure S1D). We classified SNP influences into six groups (Figure S1E) based on the SNP location, allele information and the probe design, as illustrated in Figures 1E and S1F. In brief, SNPs that alter target CpGs or impact probe hybridization/extension can create artificially high, low or intermediate readings. In part, such suboptimal hybridization/extension is caused by mutated probe target sequences (Figure 1E) and the extension base that flips the measurement color channel (Figure S1G). We reported 63 339 probes whose reading can be influenced by neighboring genetic variants and 15 002 probes on which the SNPs are present with no influence. We validated the prediction of all the putative SNP-influenced probes in a dataset of 191 inbred mouse tissue samples from 20 strains and 6 tissue types. The beta values for artificial high and low methylation reading group probes are close to 1 and 0, respectively. In contrast, the beta values for the probes in the suboptimal hybridization group approach 0.5 (Figure 1F) as both the methylated and the unmethylated allele signals approximate signal background. The SNP-influenced probes are most prevalent in three wild-derived strains—CAST/EiJ, MOLF/EiJ and PWK/PhJ (Figure S1H), highlighting the critical need to consider these artifacts in studying these wild mouse strains. We developed an automated pipeline that annotates SNP influence for other genomes not included in our collection.

Species inference and genome-specific Infinium data preprocessing

Towards a fully automated data-driven workflow, we sought to explore whether the target species could be inferred entirely from the methylation array data. We reasoned that probes with mappable sequences are more likely to emit a signal and be successfully detected. As such, probes with genome-specific mapping differences may carry the power of genome discrimination. Indeed, clustering probes by total signal intensity grouped samples by species origin and to a lesser degree, by tissue type (Figures 2A and S2A). Based on this intuition, we developed the SPIRAL (SPecies Inference fRom Alignment) method based on probe alignment scores and their signal detection *P*-values (Figure 2B, see section MATERIALS AND METHODS). These two metrics both reflect probe hybridization success and are inversely associated (Figure S2B). In brief, we first selected probes with the strongest and weakest signal detection as determined by their detection *P*-values. These positive and negative probes were then compared to their alignment scores in the candidate genomes. The species with the maximum area under the curve (AUC) is returned as the predicted species (see Figure S2C for an example). For the designed genome, the number of negative probes can be lacking, leading to genome discrimination primarily driven by other thermodynamic properties of probe hybridization rather than sequence dissimilarity (Figure S2D and E). We empirically determine that the sample has come from the designed genome based on the probe success rate (see section MATERIALS AND METHODS, Figure S2F).

We applied SPIRAL to 211 public EPIC/HM450 array data generated on non-human species, including primates and mice. In all test cases, including 23 samples adapted to query 5-hydroxymethylation, SPIRAL correctly identified the species' origin (Figure 2C). We then applied SPIRAL to a combined Mammal40 dataset of 883 samples from 22 species (Figure 2D) and an MM285 dataset of 30 human, rat and hamster DNA samples (Figure S2G). On the mammal40 array dataset, SPIRAL made accurate predictions in 7 families and only made mistakes between closely related candidate species (Figure 2D). Notable mistakes include flying fox (Pteropodidae) samples being confused for greater horseshoe bats, a species very close to the large flying fox, in 58 of 235 cases (Figure 2D). Similarly, from the MM285 dataset, SPIRAL can correctly identify human, rat and hamster samples (Figure S2G). The SPIRAL algorithm provides inference of the samples' origin of species from data, which is key to genome-specific preprocessing and normalization (below).

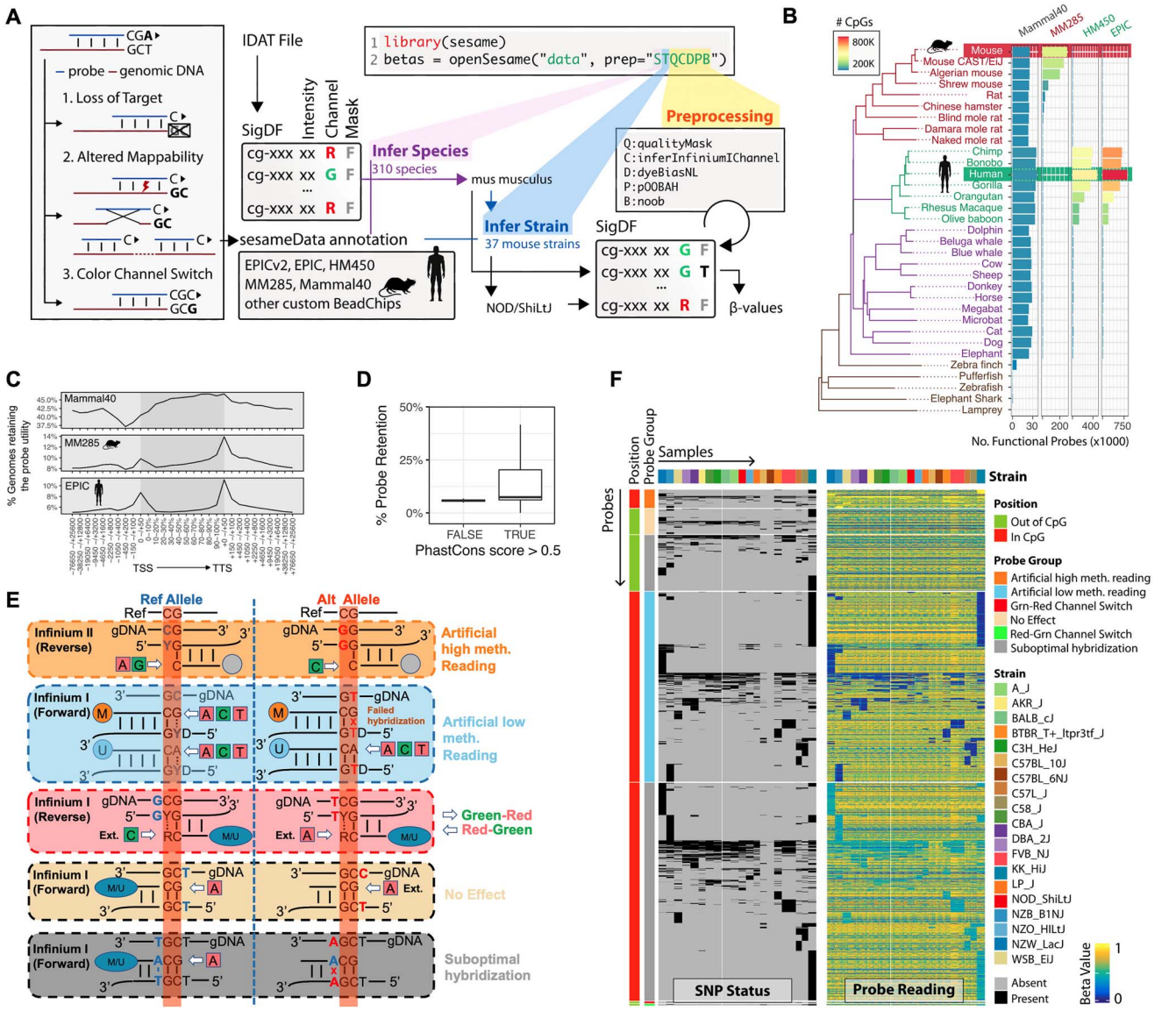


Figure 1. Utility of Infinium BeadChips on vertebrate genomes. (**A**) SeSAMe workflow of species and strain inference and genome-specific preprocessing. (**B**) Annotation of four Infinium BeadChip probes on the number of functional probes across 310 genomes (showing 33 representative species). (**C**) Genic view of probe retention of three generations of Infinium BeadChips. (**D**) Evolutionarily conserved versus non-conserved regions in the cross-species retention rate of probes on the MM285 array. (**E**) A schematic illustration of six groups of SNP influences on DNA methylation reading (R = A,G; Y=C,T,U; D = A,G,T). (**F**) Validation of strain-specific SNP influence on DNA methylation readings in 191 inbred mouse strain samples. Three wild-derived strains (CAST_EiJ,PWK_PhJ and MOLF_EiJ) were shown separately due to their higher SNP number.

Previous Infinium array data processing methods are established mainly under the assumption that the array has been run on the designed species. Therefore, the chip's global performance should evaluate the success of the whole experiment. In contrast, when the array is run on an arbitrary genome, one may expect detection success from a subset of the probes. We modified existing Infinium BeadChip data processing methods, e.g. background subtraction, dye bias correction and detection *P*-value calculation, to work in a genome-specific manner. For example, we use genome-specific out-of-band channel specification to parameterize signal background as is needed by the noob [30] and the pOOBAH algorithms [29]. Similarly, a substantial number of probes are expected to fail in a sample from a non-designed species. Focusing on probes with predicted utility in the target species effectively removed artifactual intermediate methylation readings (Figure S2H) associated with low signal intensity (Figure S2I), and improved the detection rate estimation (Figure S2J). Our pipeline collectively corrects the skewed methylation readings in a calibration experiment running MM285 (Figure 2E and F) on rat DNA of titrated methylation levels. The pipeline is reduced to the generic processing in performance on the designed genome (Figure S2H). Most of the probe replicates targeting the same cytosines or CpGs but with different designs [35] generate largely concordant DNA methylation measurements after preprocessing (Figure 2G), suggesting its robustness.

SeSAMe associates mouse strain-specific methylations to phenotypes

With proper probe masking and signal normalization, we explored the DNA methylation diversity among inbred mouse strains. We analyzed a public data set of 238 liver, spleen, tail and frontal

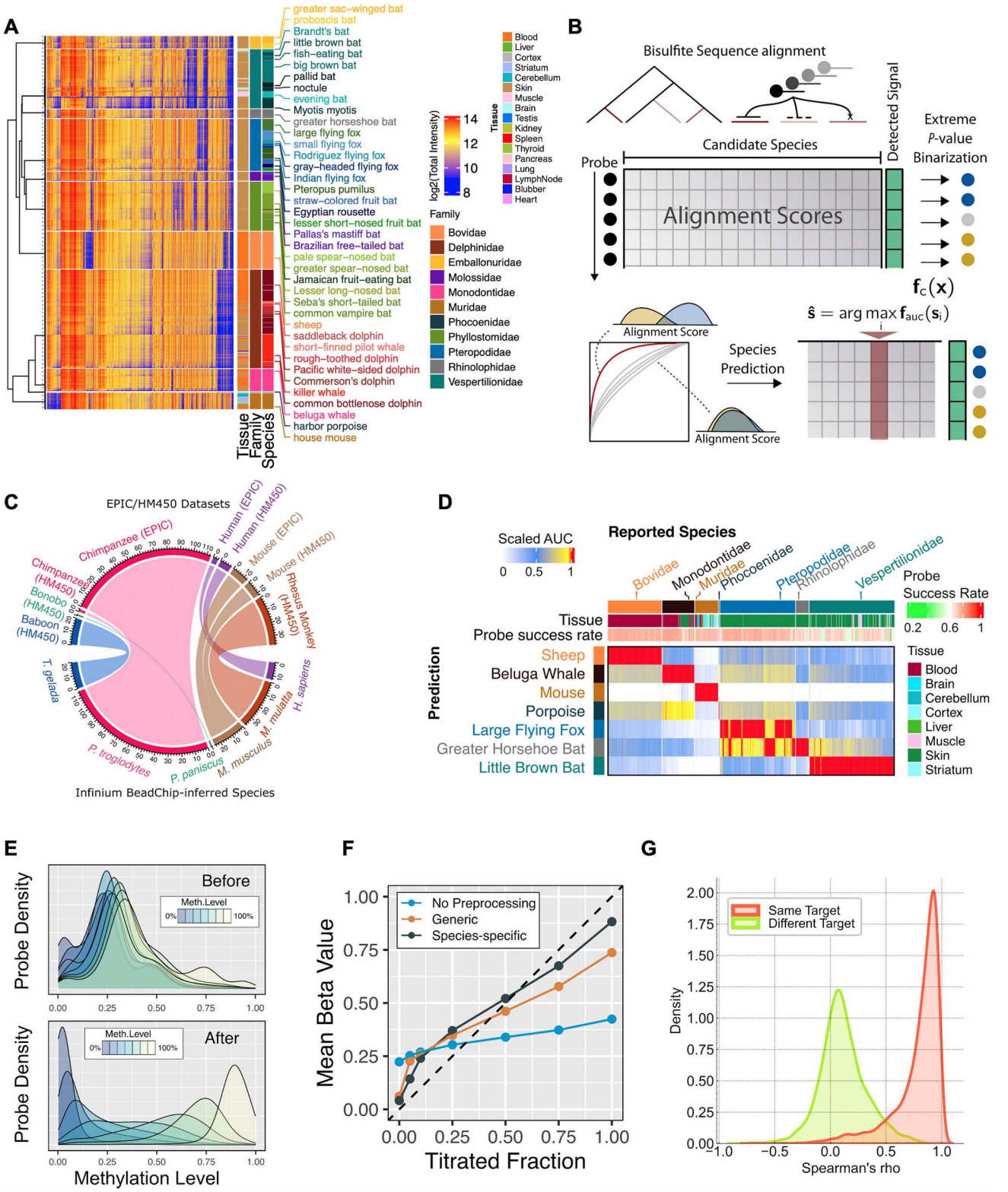


Figure 2. Species inference and genome-specific Infinium data preprocessing. (**A**) Heatmap showing the total probe signal intensity across 1644 Mammal40 datasets clustered both row and column-wise. We selected top 200 probes (ranked by AS) with AS >47 in each species, combined them to obtain 4569 unique probes. (**B**) The species inference workflow. (**C**) Comparison of species prediction accuracy on human (HM450 and EPIC) array datasets. (**D**) Heatmap of the species prediction AUC in a Mammal40 dataset comparing reported species (column) and predicted species (row). Rows and columns color indicates the family of each species. Only candidate species predicted by at least one sample are shown. (**E**) Methylation titration comparison on MM285 rat DNA datasets (GSE184410) before and after SeSAMe processing. (**F**) Scatter plot comparing the mean beta value of each sample to the titrated fraction with generic, species-specific data preprocess and without data preprocessing on the rat DNA titration data. Generic processing does not include species inference and SigDF updates but still relies on detection *P*-value to perform probe masking. (**G**) Comparison of correlations between replicate probes and randomly selected probes targeting different loci. We randomly selected 10 000 records from each group for visualization.

lobe brain tissue samples from 25 mouse strains. The DNA methylomes are primarily grouped by tissue types (Figure 3A). Within each tissue type, the samples are further sub-clustered by mouse strains (Figure 3A), suggesting the prevalence of strain-specific methylation variation. By performing a multiple regression of DNA methylation on tissue, strain and sex, we identified 18 822 strain-specific differentially methylated CpGs (SDMCs, Methods). Most SDMCs strongly depend on tissue type (Figure 3B), with effect sizes highly intertwined (Figure S3A). In contrast, the SNP probes, which measure variant allele frequencies instead of cytosine methylation levels, were predominantly tissue independent (Figure S3B). Among the four tissue types we studied, the spleen has the most SDMCs, followed by the liver, tail and brain, suggesting that methylation variation may be linked to immune and metabolic phenotypes among these mouse strains.

We classified the SDMCs based on whether their methylation states depend on the tissue types. Tissue-independent hypermethylated SDMCs were enriched in CTCF binding sites, intermediately methylated CpGs and variably methylated-IAPs (VM-IAPs) (Figure 3C). VM-IAPs were found to have altered methylation levels across mouse individuals [36] (Figure 3C). Our findings support a previous study showing that VM-IAPs are found in multiple tissue types and are frequently adjacent to CTCF binding sites in the genome [36]. In contrast, tissue-independent hypomethylated SDMCs are slightly enriched at quiescent regions, and pseudogene transcription start sites (Figure S3D).

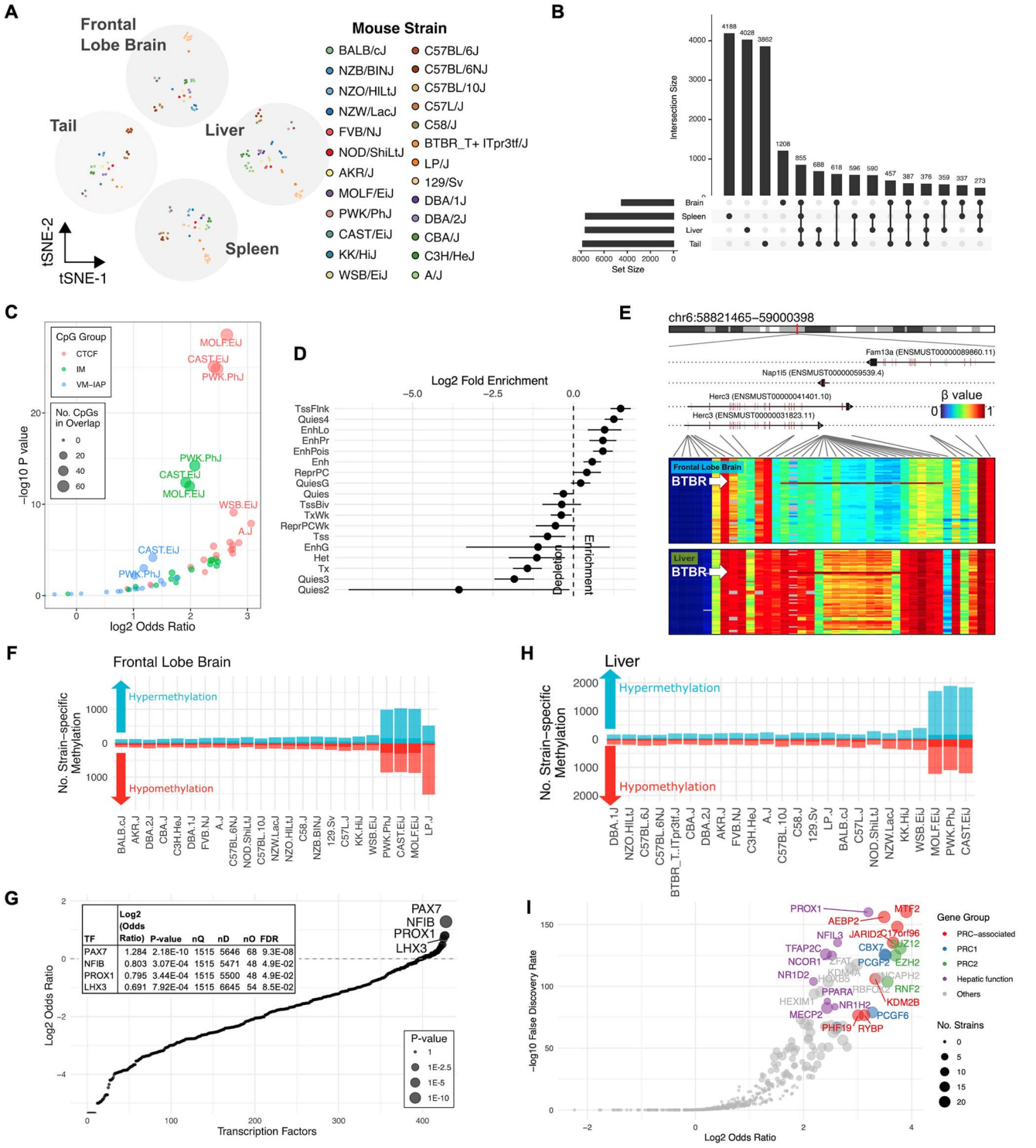
Studying tissue-specific SDMCs, on the other hand, revealed an enrichment at enhancer regions (e.g. ChromHMM states TssFlnk, EnhLo, EnhPr, EnhPois and Enh) but depletion at heterochromatic regions (Het), promoters (Tss) and gene bodies (Tx) (Figure 3D). Figure 3E shows an example of a hypermethylated brain-specific SDMC at the Nap1l5 locus in the BTBR mice. Despite being present in all four tissues studied, this methylation gain spans the most extended genomic region in the frontal lobe brain tissue. In contrast, the hypermethylation occurs at only a few CpGs in the liver tissue (Figure 3E). Nap1l5 is a paternally expressed, micro-imprinted gene nested within a non-imprinted host gene Herc3. Nap1l5 and Herc3 are highly expressed in cortical neurons [37]. The brain-specific, imprinting-associated DMR in BTBR mice may be implicated in their known loss of corpus callosum neurons and autism-related phenotypes [38].

Comparing frontal lobe brain tissue in different strains suggests more hypomethylation in LP/J mice (Figure 3F) with a lower global methylation average (Figure S3E). Investigating hypomethylated CpGs in LP/J mice revealed that these SDMCs are frequently bound by the paired box protein PAX7, NFI gene complex member NFIB and homeobox transcription factors, e.g. LHX3 (Figure 3G). Notably, PAX7 is key to lineage specification of early neural crest development in mammals [39] and non-mammalian vertebrates [40]. LP/J mice have a reduced number of neural crest-derived melanocytes in the coat and choroid layer of the eye. This difference in neural crest development, likely attributed to the Ednrb gene mutation [41], leads to dark eyes in the LP/J mice. The impaired neurogenesis may also be implicated in the increased propensity for audiogenic seizures in LP/J [42]. Similarly, the LIM factor LHX3 plays a crucial role in motor neuron differentiation [43], and NFIB cooperates with NFIA in coordinating fetal mouse forebrain development [44]. Further comparison of LP/J-specific hypomethylated sites with a mouse cell type-specific methylation atlas [45] suggests that this hypomethylation pattern might be due to an enrichment of hippocampus-specific excitatory neurons, e.g. at the dentate gyrus (DG) and hippocampal subfield CA1–3, as well as adult neural precursors (ANP) (Figure S3F).

Comparing the liver methylation profiles in different strains revealed an association with body and fat weight. The three wild-derived strains, CAST/EiJ, MOLF/EiJ, PWK/PhJ, tend to have more hypermethylated than hypomethylated CpGs in the liver (Figure 3H). Most hypermethylations are of small effect sizes (delta≤0.5). These hypermethylated sites in the liver may be associated with the smaller body weight and fat composition in these wild-derived mice [46]. KK/HiJ, a type 2 noninsulin-dependent diabetes and obesity model, has the most liver methylation variation only next to the four wild-derived strains [46], suggesting impaired epigenetic regulation of metabolism. Methylation change in the liver is disproportionately found at the binding sites of Polycomb repressive complexes, including PRC2, PRC1 and other PRC-associated co-factors (Figure 3I). The Polycomb repressive complex is known to pre-mark unmethylated DNA that gains methylation during cell proliferation [47–49]. Hence differential hepatic cell proliferation may explain the strain-specific methylation differences in the liver. Furthermore, transcription factors that govern hepatic cell differentiation and function are also differentially methylated (Figure 3I). Notably, HNF4A, PROX1 and FOXAs are among the TFs most frequently associated with hypermethylation alongside CTCF/cohesion complexes (Figure S3G). HNF4A and FOXAs are central canonical TFs for hepatocyte development and function [50]. PROX1 is an early marker of the developing mouse liver, controlling hepatocyte migration during liver morphogenesis [51]. NFIL3 is highly expressed in the liver and regulates hepatic gluconeogenesis [52]. The binding sites of several metabolic nuclear receptors, such as NR1D2, NR1H2 (Liver X receptor beta) and PPARA [53], were also enriched at strain-specific methylation differences. Intriguingly, the DNA methylation reader MECP2, best known for its role in neurogenesis and causing Rett syndrome when mutated, also regulates lipid metabolism in the liver in coordination with the NCOR1 corepressor complex [54]. NCOR1 and MECP2 are both bound to differential methylated sites across strains in the liver (Figure 3I). In contrast, we only identified two liver hypomethylation-associated TFs: CBX5 and ZFP57. CBX5/HP1A is a heterochromatin protein [55], consistent with methylation loss most occurring at heavily methylated heterochromatic regions.

SeSAMe analysis reveals principles of the epigenome evolution

Changes in DNA methylation contribute to the evolution of transcriptional regulation and phenotypic variation at a level beyond that of the gene sequence evolution [56]. We applied SeSAMe to analyze the methylomes of 857 blood samples from 16 species. First, the global blood DNA methylation levels averaged over different chromatin states are stable from species to species. Active transcription start sites and gene bodies display the least methylation variation featuring the lowest and the highest average methylation levels across ChromHMM states, respectively (Figure 4A). This genic stability (Figure S4A) is consistent with the evolutionary conservation of gene expression patterns across different mammalian species [57]. In contrast, enhancers are more variably methylated than promoters across species, in agreement with these elements having evolved more rapidly [58] and contributed to the gene expression robustness, possibly through partial redundancy [59]. Blood samples of the same species share more significant similarities in the DNA methylome than those of different species (Figure S4B). Within each species group, samples are further segregated by sex (Figure S4B, right



TF	Log2 (Odds Ratio)	P-value	nQ	nD	nO	FDR
PAX7	1.284	2.18E-10	1515	5646	68	9.3E-08
NFIB	0.803	3.07E-04	1515	5471	48	4.9E-02
PROX1	0.795	3.44E-04	1515	5500	48	4.9E-02
LHX3	0.691	7.92E-04	1515	6645	54	8.5E-02

Figure 3. SeSAMe associates mouse strain-specific methylations to phenotypes. (**A**) t-SNE embedding of methylomes of tissue samples from 25 different strains. SNP-influenced probes are masked. (**B**) An UpSet plot of strain-specific differential methylated CpGs (SDMCs) showing strong dependence on tissue types. (**C**) Enrichment of SDMCs shared across tissues in CTCF binding sites (red), consistent intermediate methylation (IM, green) and variably methylated regions (VMR, blue). (**D**) Enrichment of tissue-specific SDMCs in enhancer regions among different chromatin states (from ENCODE ChromHMM). SDMCs from all strains are merged to perform enrichment test. (**E**) Example differentially methylated regions at the Nap1l5/Herc3 locus, specific to the BTBR strain. (**F**) The number of frontal lobe brain SDMCs in different strains. Only strains with both sexes represented are shown. Frontal lobe brain SDMCs with LP/J showed the most methylation loss. (**G**) Transcription factor binding site (TFBS) enrichment of LP/J-specific hypomethylation. nQ, nD and nO represent the size of query, database and overlap CpGs set. (**H**) The number of liver SDMCs in different strains. Liver SDMCs with wild-derived strains showed more hyper- than hypomethylation. (**I**) Enrichment of liver-specific SDMCs at Polycomb repressive complex targets and hepatocyte-associated transcription factors.

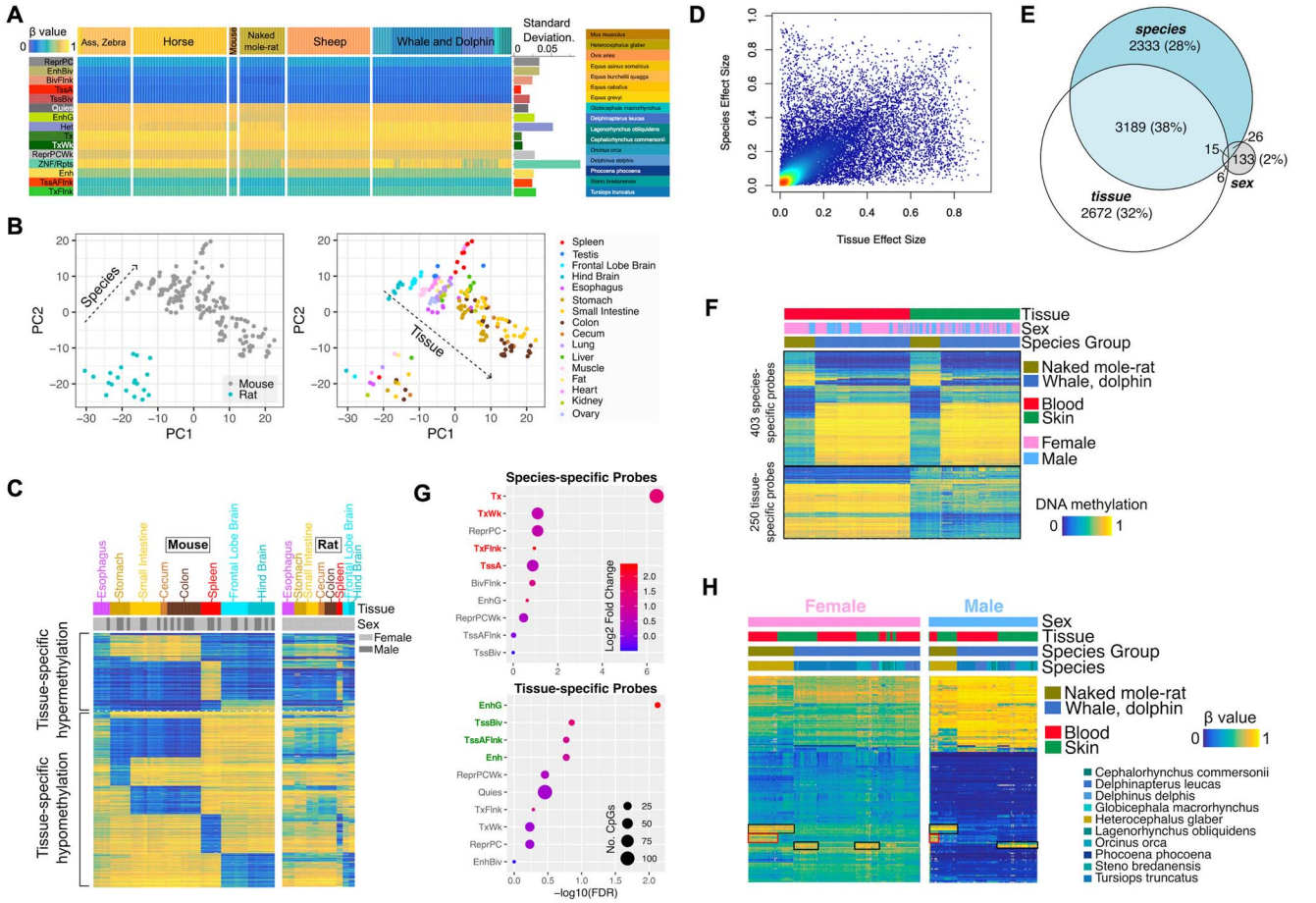


Figure 4. SeSAMe analysis reveals principles of the epigenome evolution. (**A**) Comparison of chromatin state methylation average across 857 blood methylomes from 16 species. (**B**) Principal component analysis of 188 methylomes profiled from the mouse and rat DNA. (**C**) Tissue-specific hyper- and hypomethylation signatures comparing mouse and rat tissues. (**D**) The effect sizes of tissue- and species-specific methylation in multiple regression modeling of 703 naked mole-rat and cetaceans Mammal40 datasets. (**E**) Overlap of species-, tissue- and sex-specific methylation in the 703-sample analysis in 4D. (**F**) Strictly tissue- and species-specific methylation in the 703-sample analysis in 4E. (**G**) ChromHMM state enrichment of strict species- and tissue-specific methylation. (**H**) Comparison of two species groups (naked mole-rat and cetaceans) and two tissue types (blood and skin) on the methylations of 158 X-linked CpGs.

panel). Further investigation of a data set of diverse mouse and rat tissue types revealed a joint DNA methylation determination by tissue and species, spanning the two leading principal components of the methylome data (Figure 4B). The tissue-type orders along the tissue-type axis were mainly conserved between mouse and rat samples, pointing to the relative stability of tissue biology across rodents (Figure 4B, right panel). To further evaluate the conservation of tissue-specific methylation, we identified tissue-specific hypo and hypermethylation of eight different tissue types (cecum, colon, esophagus, frontal lobe brain, hindbrain, small intestine, spleen and stomach) in mice. We studied matched tissues in the rat (Figure 4C). Focusing on the MM285 probes compatible with the rat and the mouse genome, we uncovered the conservation of these tissue-specific methylations, with brain-specific methylations most conserved, followed by the spleen. This methylome conservation may implicate transcriptional conservation. For example, the preservation of brain tissue methylation is consistent with a slower transcriptional change in brain tissue across mammals [57,60]. Compared to the brain-specific methylations, the tissue-specific methylation pattern is less conserved in the gastrointestinal (GI) tract, suggesting more rapid evolution of GI-specific methylation signature in the rodent clade.

To validate the interplay of tissue type and species in more remotely related species, we compared two Mammal40 datasets of blood and skin tissues of mole rats and cetaceans (whales and dolphins) (see section MATERIALS AND METHODS). We performed a multiple regression analysis to model the roles of tissue type, species group and sex in governing the DNA methylome dynamics. Again, we found highly intertwined tissue-specific and species-specific methylation (Figure 4D). About, 38% of the methylation variation depends on both tissue and species (Methods), while 32%, 28% and 2% of the methylations vary by tissue, species and sex alone (Figure 4E). Despite the extensive statistical interaction between tissue and species, we identified 403 strictly species-specific and 250 strictly tissue-specific probes with an effect size of the corresponding predictor greater than 0.3 but the effect sizes of other predictors smaller than 0.1 (Figure 4F). Querying the genomic distribution of species- and tissue-specific methylation revealed that strictly tissue-specific methylation is slightly enriched in enhancer regions (ChromHMM code Enh, TssAFlnk). In contrast, species-specific methylation differences are strongly enriched in gene bodies (ChromHMM code Tx, TxWk, Figure 4G, see section MATERIALS AND METHODS). This enrichment reflects the role of enhancer regions in lineage specification and its evolutionary conservation across species.

On the other hand, species-specific methylation at gene bodies reflects more the difference in gene expression states.

Finally, from the regression, we identified 180 sex-specific methylations from the above Mammal40 datasets, and these CpG sites are all X-linked on the human genome. X-linked gene promoters can be mono-allelically methylated in the female somatic cells (Figure 4H). The monoallelic methylation is associated with the X chromosome inactivation and is largely transcriptionally silenced. Comparing cetaceans and naked mole rats, we found that the inactive X-associated DNA methylation is mainly conserved, with most X-linked promoter probes displaying an intermediate methylation level (Figure 4H). Some CpGs lose intermediate methylation patterns in either of the two species groups and become fully methylated (black boxes in Figure 4H), likely due to gene silencing on both the active and inactive X chromosomes. Some tissue dependencies have also been noted, including CpGs that lost intermediate methylation only in blood but not skin tissue (red boxes in Figure 4H). Further study of these X-linked probes in a more extensive data set of 2141 samples from 44 species revealed that bats have more intermediate methylation loss than cetaceans (Figure S4C). Collectively, we observed preservation of the sex dimorphism associated with X chromosome inactivation, while most male samples are consistently unmethylated at these sites.

Discussion

The Infinium DNA Methylation BeadChip has been one of the most used assay technologies to profile human genome-wide DNA methylation. We presented genome inference methods and adapted genome-specific data preprocessing techniques to streamline the study of DNA methylation from arbitrary genomes using a single platform. Notably, we compared the methylomes of different inbred mouse strains to reveal extensive and intricate strain-specific methylation differences, highly intertwined with sex, tissue type and organismal age as shown previously [35]. We demonstrated that some of these differences might be linked to the corresponding mouse phenotype differences in a related tissue context. For example, we showed that LP/J mice, known to have impaired neural crest development, carry differential methylation in the binding sites of transcription factors involved in neurogenesis. BTBR strain, an autism disease model with documented loss of corpus callosum neurons, possesses unique methylation patterns in genes specifically expressed in neurons. Despite these associations, it is possible that many methylation differences merely reflect the genetic distance without simple phenotypic manifestation. This would be supported by a higher number of SDMCs in wild-derived strains. Schilling compared C57BL/6 J and BALB/cJ and found that most allele-specific DNA methylation is mainly determined by cis-acting sequences [61], highlighting a complex interplay of the genotype, phenotype and DNA methylation in mammalian cells. To enhance the sensitivity of detecting such association and its interaction with other covariates such as age and tissue, one could investigate more diverse mouse populations, such as the collaborative cross and the diverse outbred populations. Our strain-specific methylation catalog represents the epigenetic heterogeneity of the founder strains in these more diverse mouse cohorts.

The evolution of whole genome DNA methylation, which first emerged in early vertebrates [13] and prevailed in most human tissues, is not fully understood. We studied DNA methylation in different species contexts across a diverse group of mammalian species. By studying chromatin-aggregated methylation, we had not observed global methylation increase from other mammals, such as sheep, whales and horses, to humans, indicating that the human-like DNA methylome has appeared before the mammalian evolution and likely appeared more abruptly than progressively, possibly aligned with the evolution of DNA methyltransferases [62].

Our analysis sheds light on the principles of future generations of multi-species array design. We showed that most probes from species-specific arrays, such as the human EPIC and mouse MM285 array, are restricted to a small phylogenetic clade. The number of usable probes decays fast as the target species move away from the designed species, posing challenges to both inter-species comparison and inter-array normalization. The Mammal40 array retains a high fraction of probe utility within the mammalian species, meeting its design objective [27]. However, the Mammal40 array only optimally covers the human genome, and the number of usable probes still decreases as the target genome becomes less closely related to the designed genome. In the Mammal40 array, the support for different alleles was partly achieved by the wobble base design [27], which may expand the species scope of these platforms without increasing the probe number. Multiple versions of the same probe may be included to account for this bias in coverage and allow a fair comparison of DNA methylation on a dinucleotide resolution.

Materials and methods

Species inference

Probes sequences were mapped to the 310 candidate genomes with BISCUIT (https://github.com/zhou-lab/biscuit), and the alignment score of Infinium-I probes were calculated by taking the maximum of the alignment scores of the two alleles. The probe success rate is the number of probes with a detected *P*-value <0.05 over the total number of probes. The probe detection *P*-value was calculated with the pOOBAH method as implemented in the SeSAMe package [29]. The extreme probes were projected into positive (detected *P*-value <0.01) and negative group (*P*-value >0.1) to reflect whether the probe hybridization was successful.

$$f_c\left(b_k\right) = \begin{cases} 1 & \text{if } P\text{-value of probe k} \le 0.01, \\ 0 & \text{if } P\text{-value of probe k} \ge 0.1 \end{cases}$$

where b is a Boolean vector. We then calculate the area under the curve (AUC) for each candidate species, comparing the alignment scores of the positive and negative probes:

$$f_{auc}\left(s_i\right) = \left(\sum\nolimits_k R_i^k \times b_k - n_1 \times \left(n_1 + 1\right)/2\right) / \left(n_1 \times n_2\right)$$

where R_i^k is the rank of alignment score for each probe k in candidate species i. n is the length of b, $n_1 = \sum_{i=1}^{n} b_i$ is the number of positive probes and $n_2 = n - n_1$ is the number of negative probes. The candidate species with the maximum AUC was assigned as the predicted species to the given sample (Figure 2A). If the whole-array success rate is over 0.95 or 0.8 and the maximum AUC is under 0.5, we directly assign the designed reference genome as the prediction for human and mouse arrays respectively. To keep the same number of positive and negative probes, we balanced the number of positive and negative probes when calculating AUC. Species taxonomy was downloaded from the NCBI taxonomy database [63].

Species- and strain-specific preprocessing

We adapted three preprocessing components, i.e. dye bias correction, detection *P*-value masking and background subtraction, chained in the *openSesame* function in the SeSAMe package (Version 1.16.0+). The adaptation includes (1) the out-of-band signals were combined with the internal negative control probes to parameterize the signal background. (2) The SPIRAL species inference was employed before other signal preprocessing components. All the preprocessing uses an updated, genome-specific color channel in the SigDF object. (3) The order of normalization steps defaults to channel inference, dye bias correction, detection *P*-value masking and background subtraction but can be customized to suit the user's preference. (4) Infinium-I probes that were non-uniquely mapped were excluded from the out-of-band signal pool. (5) Masked probes are excluded in functions that explicitly equalize the beta value distributions of probes of different design types, such as BMIQ [31]. (6) A color channel inference component was added to supplement alignment-based designation to detect residual color channel switches from data. (7) Quality control-based masking is used before all the other processing components. (8) Detection *P*-value calculation is done before background subtraction, which will modify the signal and may affect the out-of-band signal assumption used on pOOBAH. Raw detection rate is computed by the number of probes with significant detection *P*-values divided by the total number of probes. The species-specific detection rate is computed by the number of functional probes with a significant signal divided by the species' total number of putative functional probes.

Predicting the influence of strain-specific variants

The genetic variants of inbred mouse strains were downloaded from Mouse Genomes Project (ftp://ftp-mouse.sanger.ac.uk/) [64]. We used BEDTools [65] to find strain-specific variants (SNP and Indel) located within 5 bp from the 3′-end of each probe's extension base. To get reliable variants, we used the following conditions to filter variants: (1) FILTER = PASS, (2) Genotype (GT) was homozygous (1/1, 2/2, 3/3, 4/4 or 5/5), (3) GQ > 20 and (4) DP > 8 [66]. Groups were assigned to each probe based on the SNP position, the SNP type, the probe type, the probe directionality and the strand of bisulfite conversion (Figure S1E). The effect of SNP on probe methylation reading was classified into the following six groups: (1) no effect, (2) artificial low methylation reading, (3) artificial high methylation reading, (4) suboptimal hybridization, (5) G-R (green-to-red channel switch) and (5) R-G (red-to-green channel switch). To illustrate the color channel switch caused by SNP, only CpGs grouped to G-R or R-G were selected, and the U and M intensity for the green and red channels were calculated by SeSAMe *readIDATpair* function using the mouse array MM285 array dataset.

DNA methylation BeadChip data

Raw IDAT files for the Infinium array data were downloaded from Gene Expression Omnibus using the following accessions. Mammal40 data were downloaded from GSE169218 [67], GSE164127 [68], GSE147004 [69] and GSE173330 [70]. MM285 data were downloaded from GSE184410. One hundred and thirteen chimpanzee data were retrieved from GSE136296; 11 mouse data were downloaded from GSE110600; 61 HM450 samples were downloaded from GSE49177 (including bonobo, chimpanzee, human, mouse and rhesus samples); 23 samples interrogate 5hmCs using TET-mediated cytosine oxidation (GSE49177) and 20 baboon samples were downloaded from GSE101733. IDAT files were preprocessed to the total intensity and the DNA methylation beta value matrices using the *openSesame* workflow implemented in the SeSAMe package with default options [29].

Strain-specific differential methylation analysis

Raw IDAT files for 238 mouse samples were retrieved from our prior study (GSE184410) before being processed using the openSesame pipeline with the default parameters. In total, 3000 most variable CpGs were selected before using the Rtsne package (https://github.com/jkrijthe/Rtsne) to create a 2D embedding. We performed a multiple regression of strain and sex for each tissue type to identify the strain-specific differential methylated CpGs (SDMCs). The median DNA methylation level across all strains was used as the reference level. Sex-specific probes (with modeled beta difference ≥ 0.01 between the two sexes) were excluded. Probes with an absolute value of slope coefficient ≥ 0.2 are considered SDMCs. Excluding all probes with variants, we calculated the effect sizes of strain and tissue-specific methylation differences (between the maximum and the minimum coefficient). For each strain, we define tissue independent SDMCs as CpGs whose methylation slope coefficients ≥0.2 in at least three of the four tissue types. Tissue-dependent hypermethylated SDMCs are defined as CpG probes that are hypermethylated in only one of the four tissue types. We visualized the Herc3/Nap1l5 region (from 10 kb upstream to 80 kb downstream of the Herc3 gene) using the *visualizeGene* function in the SeSAMe package. The global methylation average calculated for different strains was compared to C57BL/6 J. Male and female samples were separated.

Single-cell brain whole-genome bisulfite sequencing data

The processed methylation level data from a published dataset [45] was downloaded from GEO with accession ID GSE132489. The methylation beta value in each CpG site was merged with *biscuit mergecg* (https://github.com/huishenlab/biscuit), and the average beta values were calculated for each cell type. We plotted 2966 probes with LP/J methylation levels lower than C57BL/6 J than 0.35.

Species-specific differential methylation analysis

ChromHMM averages were calculated using the *dbStat* function in the SeSAMe package. Mammal40 ChromHMM state overlap was calculated using the consensus human chromatin state as a surrogate [71]. Metagene plot was calculated using the *KYCG_plotMeta* function in the SeSAMe package. For clarity, we randomly sampled three samples from each combination of tissue type and species group, leading to 204 datasets. Horse sexes were missing and inferred from the fraction of intermediate X-linked CpGs (beta value average between 0.3 and 0.7) with a cutoff of 0.35. The principal component analysis was performed using the *prcomp* function in the R stats package. To select tissue methylation signatures, we performed a nonparametric test comparing beta values of the target tissue with those of the non-target tissues. We only kept CpGs whose methylation can fully separate the two mouse sample groups. We further require the delta beta value of the two sample groups to be greater than 0.3 and the fraction of the missing beta value to be under 30%. The top 50 CpGs with the greatest absolute value of delta beta were kept for visualization. We performed multiple regression modeling tissue, species, sex and tissue–species interaction. Probes with an effect size ≥0.2 and *F*-test *P*-value ≤0.05 are considered as significantly methylated CpG specific to that variable. Tissue-specific probes are defined

as ones with tissue effect size ≥0.3 and species effect size ≤0.1. Species-specific probes are defined as ones with species effect size ≥0.3 and tissue effect size ≤0.1. We excluded probes with more than 20 samples showing NA in the methylation reading.

CpG enrichment analysis

The enrichment test for strain, tissue, species-specific methylated CpGs were performed using the testEnrichment function in the SeSAMe package. In brief, we manually curated our database sets using diverse sources of publicly available data. To engineer genomic features, we used BEDTools to intersect the Infinium BeadChip manifest with genomic coordinates from ChromHMM [72], GENCODE [73], ReMap [74], and various other genomic and epigenomic annotations [34]. We used Fisher's exact test to evaluate the statistical significance of set overlaps and report the Benjamini–Hochberg adjusted *P*-value each curated CpGs set.

Key Points

- This study introduced novel computational methods to process the methylation BeadChip data on arbitrary genomes.
- Meta-analysis of strain-specific DNA methylome atlas across 25 inbred mouse strains.
- Mouse strain-specific DNA methylome profiles are associated with immune, metabolic and neurological disease phenotypes.

Funding

The NIH/NIGMS (grant R35GM146978 awarded to W.Z.). It was also supported by W.Z.'s startup fund at Children's Hospital of Philadelphia and research sponsorship by FOXO Bioscience.

Availability

The SeSAMe R package (version 1.16.0+) and documentation are available on the Bioconductor home page at https://bioconductor.org/packages/release/bioc/html/sesame.html. The HM450, EPIC, EPICv2, MM285 and Mammal40 probe annotations are available at http://zwdzwd.github.io/InfiniumAnnotation, https://github.com/zhou-lab/InfiniumAnnotationV1. The mouse strain-specific SNP maskings are available at https://github.com/zhou-lab/InfiniumAnnotationV1/raw/main/Anno/MM285/MM285.mm10.mask.tsv.gz. Python script to annotate the influence of SNPs on DNA methylation reading for future array generations and new genome assemblies is available at https://github.com/zhou-lab/InfiniumManifestAnnotator.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Data availability

All data generated or analyzed during this study are included in this published article and its supplementary information files.

Competing interests

W.Z. receives research funding from FOXO Bioscience.

Authors' contributions

W.Z. conceived the study, W.Z. and W.D. implemented the algorithm, D.K. and S.H. helped with the data analysis, W.Z. and W.D. wrote the manuscript, S.H. revised the manuscript, all authors read and approved the final version.

References

1. Bestor TH. DNA methylation: evolution of a bacterial immune function into a regulator of gene expression and genome structure in higher eukaryotes. *Philos Trans R Soc Lond B Biol Sci* 1990;**326**:179–87.
2. Li E, Zhang Y. DNA methylation in mammals. *Cold Spring Harb Perspect Biol* 2014;**6**:a019133.
3. Jaenisch R, Bird A. Epigenetic regulation of gene expression: how the genome integrates intrinsic and environmental signals. *Nat Genet* 2003;**33**(Suppl):245–54.
4. Riggs AD. X inactivation, differentiation, and DNA methylation. *Cytogenet Cell Genet* 1975;**14**:9–25.
5. Greenberg MVC, Bourc'his D. The diverse roles of DNA methylation in mammalian development and disease. *Nat Rev Mol Cell Biol* 2019;**20**:590–607.
6. Robertson KD. DNA methylation and human disease. *Nat Rev Genet* 2005;**6**:597–610.
7. Deniz Ö, Frost JM, Branco MR. Regulation of transposable elements by DNA modifications. *Nat Rev Genet* 2019;**20**:417–31.
8. Cohen NM, Kenigsberg E, Tanay A. Primate CpG islands are maintained by heterogeneous evolutionary regimes involving minimal selection. *Cell* 2011;**145**:773–86.
9. Zhou W, Liang G, Molloy PL, *et al.* DNA methylation enables transposable element-driven genome expansion. *Proc Natl Acad Sci U S A* 2020;**117**:19359–66.
10. Zhong X. Comparative epigenomics: a powerful tool to understand the evolution of DNA methylation. *New Phytol* 2016;**210**:76–80.
11. Tweedie S, Charlton J, Clark V, *et al.* Methylation of genomes and genes at the invertebrate-vertebrate boundary. *Mol Cell Biol* 1997;**17**:1469–75.
12. Zemach A, Zilberman D. Evolution of eukaryotic DNA methylation and the pursuit of safer sex. *Curr Biol* 2010;**20**:R780–5.
13. Zemach A, McDaniel IE, Silva P, *et al.* Genome-wide evolutionary analysis of eukaryotic DNA methylation. *Science* 2010;**328**:916–9.
14. Yue F, Cheng Y, Breschi A, *et al.* A comparative encyclopedia of DNA elements in the mouse genome. *Nature* 2014;**515**:355–64.
15. Mi S, Chen S, Li W, *et al.* Effects of sperm DNA methylation on domesticated animal performance and perspectives on cross-species epigenetics in animal breeding. *Anim Front* 2021;**11**:39–47.
16. de Souza MM, Niciura SCM, Rocha MIP, *et al.* DNA methylation may affect beef tenderness through signal transduction in Bos indicus. *Epigenetics Chromatin* 2022;**15**:15.
17. Ibeagha-Awemu EM, Zhao X. Epigenetic marks: regulators of livestock phenotypes and conceivable sources of missing variation in livestock improvement programs. *Front Genet* 2015;**6**:302.
18. Nery da Silva A, Silva Araujo M, Pértille F, *et al.* How epigenetics can enhance pig welfare? *Animals (Basel)* 2021;12.
19. Bibikova M, Barnes B, Tsan C, *et al.* High density DNA methylation array with single CpG site resolution. *Genomics* 2011;**98**:288–95.

20. Wong NC, Ng J, Hall NE, *et al.* Exploring the utility of human DNA methylation arrays for profiling mouse genomic DNA. *Genomics* 2013;**102**:38–46.
21. Heyn H, Moran S, Hernando-Herraez I, *et al.* DNA methylation contributes to natural human variation. *Genome Res* 2013;**23**: 1363–72.
22. Chopra P, Papale LA, White ATJ, *et al.* Array-based assay detects genome-wide 5-mC and 5-hmC in the brains of humans, non-human primates, and mice. *BMC Genomics* 2014;**15**:131.
23. Needhamsen M, Ewing E, Lund H, *et al.* Usability of human Infinium MethylationEPIC BeadChip for mouse DNA methylation studies. *BMC Bioinformatics* 2017;**18**:486.
24. Gujar H, Liang JW, Wong NC, *et al.* Profiling DNA methylation differences between inbred mouse strains on the Illumina human Infinium MethylationEPIC microarray. *PLoS ONE* 2018;**13**:e0193496.
25. Housman G, Havill LM, Quillen EE, *et al.* Assessment of DNA methylation patterns in the bone and cartilage of a nonhuman primate model of osteoarthritis. *Cartilage* 2019;**10**:335–45.
26. Guevara EE, Lawler RR, Staes N, *et al.* Age-associated epigenetic change in chimpanzees and humans. *Philos Trans R Soc Lond B Biol Sci* 2020;**375**:20190616.
27. Arneson A, Haghani A, Thompson MJ, *et al.* A mammalian methylation array for profiling methylation levels at conserved sequences. *Nat Commun* 2022;**13**:783.
28. Lehne B, Drong AW, Loh M, *et al.* A coherent approach for analysis of the Illumina HumanMethylation450 BeadChip improves data quality and performance in epigenome-wide association studies. *Genome Biol* 2015;**16**:37.
29. Zhou W, Triche TJ, Laird PW, *et al.* SeSAMe: reducing artifactual detection of DNA methylation by Infinium BeadChips in genomic deletions. *Nucleic Acids Res* 2018;**46**:e123.
30. Triche TJ, Weisenberger DJ, Van Den Berg D, *et al.* Low-level processing of Illumina Infinium DNA methylation BeadArrays. *Nucleic Acids Res* 2013;**41**:e90.
31. Teschendorff AE, Marabita F, Lechner M, *et al.* A beta-mixture quantile normalization method for correcting probe design bias in Illumina Infinium 450 k DNA methylation data. *Bioinformatics* 2013;**29**:189–96.
32. Fortin J-P, Labbe A, Lemire M, *et al.* Functional normalization of 450k methylation array data improves replication in large cancer studies. *Genome Biol* 2014;**15**:503.
33. Chen Y, Lemire M, Choufani S, *et al.* Discovery of cross-reactive probes and polymorphic CpGs in the Illumina Infinium HumanMethylation450 microarray. *Epigenetics* 2013;**8**:203–9.
34. Zhou W, Laird PW, Shen H. Comprehensive characterization, annotation and innovative use of Infinium DNA methylation BeadChip probes. *Nucleic Acids Res* 2017;**45**:e22.
35. Zhou W, Hinoue T, Barnes B, *et al.* DNA methylation dynamics and dysregulation delineated by high-throughput profiling in the mouse. *Cell Genomics* 2022;**2**(7).
36. Kazachenka A, Bertozzi TM, Sjoberg-Herrera MK, *et al.* Identification, characterization, and heritability of murine metastable Epialleles: implications for non-genetic inheritance. *Cell* 2018;**175**:1259–1271.e13.
37. Davies W, Smith RJ, Kelsey G, *et al.* Expression patterns of the novel imprinted genes Nap1l5 and Peg13 and their non-imprinted host genes in the adult mouse brain. *Gene Expr Patterns* 2004;**4**:741–7.
38. Wahlsten D, Metten P, Crabbe JC. Survey of 21 inbred mouse strains in two laboratories reveals that BTBR T/+ tf/tf has severely reduced hippocampal commissure and absent corpus callosum. *Brain Res* 2003;**971**:47–54.
39. Murdoch B, DelConte C, García-Castro MI. Pax7 lineage contributions to the mammalian neural crest. *PLoS ONE* 2012;**7**:e41089.
40. Basch ML, Bronner-Fraser M, García-Castro MI. Specification of the neural crest occurs during gastrulation and requires Pax7. *Nature* 2006;**441**:218–22.
41. Carrasquillo MM, McCallion AS, Puffenberger EG, *et al.* Genome-wide association study and mouse model identify interaction between RET and EDNRB pathways in Hirschsprung disease. *Nat Genet* 2002;**32**:237–44.
42. Fuller JL, Sjursen FH. Audiogenic seizures in eleven mouse strains. *J Hered* 1967;**58**:135–40.
43. Thaler JP, Lee S-K, Jurata LW, *et al.* LIM factor Lhx3 contributes to the specification of motor neuron and interneuron identity through cell-type-specific protein-protein interactions. *Cell* 2002;**110**:237–49.
44. Steele-Perkins G, Plachez C, Butz KG, *et al.* The transcription factor gene Nfib is essential for both lung maturation and brain development. *Mol Cell Biol* 2005;**25**:685–98.
45. Liu H, Zhou J, Tian W, *et al.* DNA methylation atlas of the mouse brain at single-cell resolution. *Nature* 2021;**598**:120–8.
46. Ikeda H. KK mouse. *Diabetes Res Clin Pract* 1994;**24**(Suppl):S313–6.
47. Gal-Yam EN, Egger G, Iniguez L, *et al.* Frequent switching of Polycomb repressive marks and DNA hypermethylation in the PC3 prostate cancer cell line. *Proc Natl Acad Sci U S A* 2008;**105**: 12979–84.
48. Schlesinger Y, Straussman R, Keshet I, *et al.* Polycomb-mediated methylation on Lys27 of histone H3 pre-marks genes for de novo methylation in cancer. *Nat Genet* 2007;**39**:232–6.
49. Widschwendter M, Fiegl H, Egle D, *et al.* Epigenetic stem cell signature in cancer. *Nat Genet* 2007;**39**:157–8.
50. Watt AJ, Garrison WD, Duncan SA. HNF4: a central regulator of hepatocyte differentiation and function. *Hepatology* 2003;**37**: 1249–53.
51. Burke Z, Oliver G. Prox1 is an early specific marker for the developing liver and pancreas in the mammalian foregut endoderm. *Mech Dev* 2002;**118**:147–55.
52. Kang G, Han H-S, Koo S-H. NFIL3 is a negative regulator of hepatic gluconeogenesis. *Metab Clin Exp* 2017;**77**:13–22.
53. Wang Y, Nakajima T, Gonzalez FJ, *et al.* PPARs as metabolic regulators in the liver: lessons from liver-specific PPAR-null mice. *Int J Mol Sci* 2020;**21**(6):2061.
54. Kyle SM, Saha PK, Brown HM, *et al.* MeCP2 co-ordinates liver lipid metabolism with the NCoR1/HDAC3 corepressor complex. *Hum Mol Genet* 2016;**25**:3029–41.
55. Lomberk G, Wallrath L, Urrutia R. The Heterochromatin Protein 1 family. *Genome Biol*. 2006; **7**:228
56. King MC, Wilson AC. Evolution at two levels in humans and chimpanzees. *Science* 1975;**188**:107–16.
57. Brawand D, Soumillon M, Necsulea A, *et al.* The evolution of gene expression levels in mammalian organs. *Nature* 2011;**478**:343–8.
58. Villar D, Berthelot C, Aldridge S, *et al.* Enhancer evolution across 20 mammalian species. *Cell* 2015;**160**:554–66.
59. Berthelot C, Villar D, Horvath JE, *et al.* Complexity and conservation of regulatory landscapes underlie evolutionary resilience of mammalian gene expression. *Nat Ecol Evol* 2018;**2**:152–63.
60. Cardoso-Moreira M, Halbert J, Valloton D, *et al.* Gene expression across mammalian organ development. *Nature* 2019;**571**:505–9.
61. Schilling E, El Chartouni C, Rehli M. Allele-specific DNA methylation in mouse strains is mainly determined by cis-acting sequences. *Genome Res* 2009;**19**:2028–35.
62. de Mendoza A, Poppe D, Buckberry S, *et al.* The emergence of the brain non-CpG methylation system in vertebrates. *Nat Ecol Evol* 2021;**5**:369–78.

63. Schoch CL, Ciufo S, Domrachev M, *et al*. NCBI taxonomy: a comprehensive update on curation, resources and tools. *Database (Oxford)* 2020;**2020**:2020.
64. Keane TM, Goodstadt L, Danecek P, *et al*. Mouse genomic variation and its effect on phenotypes and gene regulation. *Nature* 2011;**477**:289–94.
65. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* 2010;**26**:841–2.
66. Carson AR, Smith EN, Matsui H, *et al*. Effective filtering strategies to improve data quality from population-based whole exome sequencing studies. *BMC Bioinformatics* 2014;**15**:125.
67. Seiler Vellame D, Castanho I, Dahir A, *et al*. Characterizing the properties of bisulfite sequencing data: maximizing power and sensitivity to identify between-group differences in DNA methylation. *BMC Genomics* 2021;**22**:446.
68. Wilkinson GS, Adams DM, Haghani A, *et al*. DNA methylation predicts age and provides insight into exceptional longevity of bats. *Nat Commun* 2021;**12**:1615.
69. Lu AT, Narayan P, Grant MJ, *et al*. DNA methylation study of Huntington's disease and motor progression in patients and in animal models. *Nat Commun* 2020;**11**:4529.
70. Robeck TR, Fei Z, Lu AT, *et al*. Multi-species and multi-tissue methylation clocks for age estimation in toothed whales and dolphins. *Commun Biol* 2021;**4**:642.
71. Ernst J, Kellis M. ChromHMM: automating chromatin-state discovery and characterization. *Nat Methods* 2012;**9**:215–6.
72. van der Velde A, Fan K, Tsuji J, *et al*. Annotation of chromatin states in 66 complete mouse epigenomes during development. *Commun Biol* 2021;**4**:239.
73. Frankish A, Diekhans M, Ferreira A-M, *et al*. GENCODE reference annotation for the human and mouse genomes. *Nucleic Acids Res* 2019;**47**:D766–73.
74. Hammal F, de Langen P, Bergon A, *et al*. ReMap 2022: a database of human, mouse, drosophila and arabidopsis regulatory regions from an integrative analysis of DNA-binding sequencing experiments. *Nucleic Acids Res* 2022;**50**:D316–25.