A self-tuning sliding window method for detecting phenotype-linked regional poly-methylation architecture in sparse wildlife methylomes

Thomas Stocker[1]*

1. **University of Sydney, Sydney School of Veterinary Science, 2250 NSW, Australia. ORCID ID: https://orcid.org/0009-0006-3567-3338**

*** Corresponding author: Thomas.stocker@syndey.edu.au**

Highlights:

- **Methylation of an extreme phenotype**
- **Diving physiology of pinnipeds**
- **Methodological analysis of methylation**
- **Phenomics. Methylome and associations**

Abstract:

Despite featuring extreme physiological adaptations integration of wildlife species into the modern 'omics' frameworks are limited due to the sparsity in the data. To address the sparsity limitation a self-tuning sliding-window framework was developed for the identification of the regional poly-CpG methylation architecture associated with phenotypic traits. Under the framework the iteratively expanded window aligns against each chromosome for two constraints: (i) the window must increase its penalised R^2 statistic derived from univariate (ADVI-EWAS) and multivariate (Horseshoe) CpG-phenotype models and (ii) both models but agree upon the direction of the methylome-phenotype association. Any windows not passing the two conditions were discarded if they lacked sufficient CpG density, otherwise were retained but ceased to grow. Application of the framework to the blood methylome of 69 California Sea Lions (*Zalophus californianus*) detected three 'trusted' windows across the entire genome. Each of the windows were uniquely mapped to a single gene, these genes were GCSH, DNMT1, LMX1A. All three genes are mechanistically linked to the phenotypic trait and are consistent with the extreme phenotype of marine mammals. The results were independent of the age of the individuals tested. The results were not predictive of phenotype when applying epigenome wide association studies (EWAS); however, application of the poly-CpG architecture provided a phenotypically linked signal that was biologically interpretable. The sliding window method is computationally efficient, isolates short, directionally stable genome regions and outperforms single-site analyses. The generalisable statistical framework provided within the study could extract phenotype-linked molecular architecture from sparse wildlife methylomes.

Key words – Physiology, Phenotype, Methylation, Sliding-window

1. Introduction:

All 'omics' data is inherently sparse, often requiring large sample sizes for inferencing (1, 2). The application of 'omics' to wildlife species is even more pertinent, with smaller sample sizes afforded due to the opportunistic nature of sampling and greater variability present (3, 4). Wildlife species are frequently neglected within the 'omics' (3). Methylation studies are just one example of the phenomenon – with the inherent environmental stressors further complicating the consideration of the phenotype (5).

Unlike with genomics, methylation studies face greater variability within the models as a result of the phenotypic influences upon the methylome and the methylomes influences upon the phenotype (5). As such, wildlife methylome studies are typically restricted to molecular aging rather than phenotype-methylation association studies (6, 7). Single CpG models – including epigenome wide association studies (EWAS) – are just as sparse, if not more, sparse than genome wide association studies (GWAS). Both frameworks are already sparse in wildlife and frequently return an absence of significance in well studies model organisms (1, 2, 8). However, the absence of a causal variable is not sufficient to declare the absence of evidence within biological models (1, 2). Improving the sample size will reduce sparsity and enable greater significance, this is not always achievable for wildlife die to limited sampling opportunities (1).

Regional poly-CpG architecture dictates that the cumulative effect of CpGs within a defined genomic region is more biologically meaningful than that of a single-CpG model of phenotypic aassociations (9). Under this hypothesis, the accumulation of methylated sites results in the refolding of the chromosome – further reducing the expression of the particular chromosomal regions (9). The framework must tackle distance and issues of capturing noise in the poly-CpG model.

Under the sliding window framework the genome is regionally scanned across individual chromosomes and for each CpG iteratively expanded within the chromosome for the optimisation of a penalised R^2 parameter. The expansion of the window will continue until the either the R^2 penalised value doesn't increase sufficiently for each inclusion of a CpG and the agreement of a biologically meaningful value for the window's collective CpG scores and their estimated effects on the phenotype. Any windows failing to meet these conditions will be terminated and those under the threshold of CpG inclusion number will be permanently discarded. Application of a sliding window model reduces the computational burden and adds biological meaning (regionality) to the framework in a similar manner to GWAS (8).

Demonstration of the sliding window framework is necessary. The extreme physiology of the California Sea Lion (*Zalophus californianus*) renders the species suitable to study. Recently *Z. californianus* was molecularly aged through the methylome (6). Although these results found a strong association between the sex and chronological age upon the epigenetic age of the individuals was determined there was a lack of phenotypic inclusion in the model beyond these two variables. Thus, these samples demonstrate untapped potential. The maximum aerobic dive capacity was chosen as an ideal phenotype for the identification of the genes associative with a taxa-specific phenotype as the variable was a linear function of age – Sea Lions experience increasing physiological demands as they mature into independently feeding individuals – a process that can take years to complete (10, 11). It was hypothesised that the increasing physiological demands mid-dive is a driver for methylation changes – which will be measured through the sliding window framework (6).

The aim of the study is to apply a self-tuning sliding window framework to the analysis of the phenotype-linked poly-CpG structuring of the methylome of 69 individual blood samples of *Z. californianus.*

Methods:

2.1 Methods overview:

2.1.1 Introduction to the methods

An eight-stack pipeline was used to analyse the relationship between the methylome and the phenotype solely using Python-3.13. Statements of the full mathematical formulae to reproduce the statistics are listed in the Supplementary Methods. All code is available in the GitHub repository for the study (tsto3616/poly-CpG: The repository encodes the sliding window pipeline, accompanying statistics and data for a study on the California Sea Lion's epigenome).

All data included within the study was collected from open-access sources and the sources of the data both cite ethics approval (6, 10). No new animal procedures were conducted for these experiments.

2.1.2 Methylation data

Robeck, Haghani (6) studied the methylation of multiple pinniped species, of these samples 69 blood samples of *Z. californianus* were used for this study. The samples represented multiple age groups and the normalised methylation data was sourced from Gene Expression Omnibus (GEO accession: GSE227319). Methylation annotations were collected from the GitHub repository featured as part of the study by Robeck, Haghani (6). The genome/ gene annotation file for *Z. californianus* was collected from NCBI (GCA_009762295.2). Handling of the raw methylation data files is not discussed as it was not part of the study. Rather the normalised methylation data from GEO (GSE227319) were processed.

2.1.3 Phenotypic data

The oxygen carrying capabilities of *Z. californianus* were determined through the formula published by Weise and Costa (10). The formula is $MaxADC = 0.2357(years\ old) + 1.3571$, where the *MaxADC* represents the maximum aerobic diving capacity (in minutes) for the age of the animal. The maximum aerobic diving capacity is treated as a continuous phenotypic variable throughout downstream analyses.

2.1.4 Construction of the master file and basic statistics

The workflow involved the conversion of published β-values to M-scores using the standard logit transformation (Supplementary Methods 1.2) after removing the extreme values (close to 0 or 1) to prevent the production of infinite M-scores (6). The top 10,000 CpG sites for which there was a high rate of variance across the 69 samples were the only CpGs included in this study. Inclusion of greater numbers of CpG sites would create a significant computational burden with minimal benefit. Minimal variation would only weaken the signal to noise ratio and so these CpGs were removed.

2.1.5 Overview of the multi-stack modelling framework

Stack One clustered the methylation profiles of *Z. californianus* according to the M-scores of the methylation data by applying HDBSCAN.

Stack Two calculated the latent age for downstream analyses using the epigenomic calculated age and the chronological age. The latent age was used to determine the reliability of the clustering from Stack One.

Stack Three predicted the maximum aerobic dive capacity through a global Hierarchical Bayesian model – as determined by the methylation clusters and latent age reliability of the clusters (Stack One and Two respectively).

Stack Four predicted the relationship of CpGs through univariate epigenome wide association study (EWAS) sampled with Automatic Differentiation Variation Inference (ADVI). The ADVI model was informed by latent age (Stack Two) and the residuals from Stack Three as the outcome variable. The conversion of the results into the Bayesian probabilities and their standard deviation into measurable outcomes (Bayesian probability).

Stack Five involved the determination of a Horseshoe Shrinkage Bayesian model was fitted to effectively handle the sparse CpG data through a multivariate model.

Stack Six involved the isolation of CpGs predicted to be a significant predictor of maximum aerobic diving capacity for both the ADVI and Horseshoe Bayesian shrinkage model. This involved the mapping of the CpGs to the genomic coordinates.

Stack Seven evaluated the role of regional context in the poly-CpG scores for which a sliding window regression task was fitted. This model involved a moveable window for which the poly-CpG scores were considered for the maximisation of the correlation while limiting the sliding windows to the one chromosome to enable biologically relevant signal generation. The sliding windows were overlayed onto the genes of *Z. californianus* for the isolation of genes for which there is an association with the maximum aerobic dive capacity.

Stack Eight produced a comparison of the earlier models using trajectory analysis for the determination of the methylation profiling in response to the models and their variables. The pipeline involved dimension reduction through PCA, k-NN clustering, MST and DPT analyses for the determination of the pseudotime for which the methylation profiles differ.

2.2 Stack One — High-variance CpG selection, PCA reduction, and HDBSCAN clustering

Stack One performed clustering of the genome-wide methylation profiles for the identification of the methylation architectural structure across individual California Sea Lions.

The first 20 PCs were extracted to represent the major axes of methylation variance across the individuals and inputted into HDBSCAN (12).

2.3 Stack Two – Latent age and cluster reliability

Stack two produced a continuous variable called the latent age and a set of cluster reliability weights to quantify the effectiveness of the HDBSCAN clusters for the prediction of the age-related methylation structure. For the calculation of the latent age the chronological age and epigenetic age were combined into as single latent age value using the optimised weighting parameter using the 1D Fisher optimisation problem (13).

The final reliability metrics were averaged for production of a final reliability score for each cluster. These weights were used within Stack Three for the Hierarchical Bayesian model.

2.4 Stack Three – Bayesian Hierarchical Model

Stack Three was determined by a Bayesian Hierarchical model for the estimation of the relationship between latent age and maximum aerobic dive capacity by accounting for the cluster level structure and the reliability weights generated for each sample (Stack Two). The model was weighted by the cluster reliability score for the given animal within the cluster.

2.5 Stack Four – CpG-level EWAS via ADVI

Stack Four performed CpG EWAS with the Bayesian residuals from Stack Three as the outcome of the ADVI-EWAS model (14)

2.6 Stack Five – Horseshoe Bayesian shrinkage model:

Stack Five computed the singular-CpG score for the global-local shrinkage model (Horseshoe prior) and the estimation of methylation effects upon the maximum aerobic diving capacity. The Horseshoe model was selected for its ability to retain large CpG effects while

aggressively shrinking noise, which is appropriate for sparse high-dimensional methylation data (15).

2.7 Stack Six – Overlayed Manhattan and Volcano Plots:

Stack Seven incorporated the univariate CpG-specific EWAS results from Stack Four and the multivariate Horseshoe shrinkage estimates from Stack Five for the generation of comparative Manhattan and Volcano plots.

2.8 Stack Seven – Sliding Window CpG aggregation model and gene-phenotype association visualisation

Stack Seven represents the primary methodological contribution of the paper. It aggregated the CpGs within chromosomes in a computationally efficient manner (sliding windows) by self-tuning to identify the shortest genomic segments of neighbouring CpGs whose combined methylation signal showed a consistent and directionally stable association with maximum aerobic diving capacity. Consistent and directionally stable in this context refers to an agreed upon impact of methylation upon the aerobic diving capacity – be it hypo- or hypermethylation. CpGs were ordered by chromosome and genomic positioning. For each CpG the window was initiated at the CpG before being expanded iteratively to the downstream CpGs provided that: (1) both the ADVI-derived effects (γ_j) and the Horseshoe-derived effects (β_j) produced an increasing variance (R^2) explained by the window – with penalisation for the window size – and (2) both the ADVI model and the Horseshoe model must agree on the direction of association with the phenotype (methylation producing an increase or decrease in the maximum aerobic diving capacity). Windows failing to meet these criteria or not reaching the minimum CpG count were removed.

Each retained window was assigned a status: trusted, semi-trusted or weak. The status was determined by the joint explanation of variance (penalised R^2 support from the ADVI and

198 Horseshoe models. Only the trusted windows were deduplicated and mapped to the genes of
199 *Z. californianus* through the GREAT-style regulatory domain definition – for the GFF3 gene
200 annotation (16). Genes were only assigned to a window when the window's genomic span
201 overlapped the gene's regulatory domain.

202 The transformed trusted windows were converted into gene-level and window-level
203 methylation scores and visualised for their structure across animals and clusters. Oly trusted
204 windows that were unique were kept for downstream analyses to prevent the artificial inflation
205 of gene importance from duplicate gene-CpG associations. CpGs belonging to the gene-
206 assigned windows were aggregated into the per-animal scores using the logistic-scaled poly-
207 CpG transformation following the combing of the ADVI and Horseshoe effect sizes with the
208 following equation:

209

210 $$\theta_j = \frac{\gamma_j + \beta_j}{2}$$

211 The scores were further assembled into two matrices; animals x genes and animals x
212 windows. The corresponding cluster-level matrices were generated with the same process but
213 for the clusters rather than animals.

214 Gene-level directional effects were generated by computing the average combined effect
215 size ($\gamma_j + \beta_j$) across the CpGs belonging to each gene. The adaptive median-absolute-deviation
216 threshold was utilised to determine if the genes were contributing positively (Up), negatively
217 (Down) or not at all (Neutral). A positive contribution was translated as methylation increasing
218 the phenotypic trait. The per-animal gene-level effects were represented as a linear regression
219 model for the maximum aerobic dive capacity to visualise the phenotype-linked trends for the
220 selected genes. Sliding windows were used to capture regional methylation patterns, which

often exert biological effects through coordinated CpG behaviour rather than isolated CpG changes. Further detail is provided within Supplementary Methods 1.10.

To evaluate the performance of the sliding window framework under varying levels of noise and sparsity, a synthetic methylation dataset was generated across 3 x 3 grid of noise (0.5, 1.0, 2.0) and sparsity (0.2, 0.4, 0.8). Each dataset consisted of 200 CpGs with randomly assigned genomic positions for 120 simulated 'animals'. The portion of significant CpGs was determined by the sparsity parameter for the assignment of the true non-zero effects of both the ADVI-EWAS and the Horseshoe parameters. Methylation matrices were assumed to have a standard normal distribution with phenotypes resulting from the linear combination of the 'true' (simulated) ADVI-EWAS values plus the Gaussian noise. The estimated ADVI-EWAS and Horseshoe values were obtained by adding small Gaussian uncertainly to the 'true' effects. The posterior probabilities for both statistical frameworks were computed for the probit model and the CpGs were considered significant for both posteriors when exceeding 0.5 with a matching posterior direction. The sliding window algorithm was performed as described above. For each simulation the number of trusted windows was defined as above, the number of recovered windows was defined as at least one CpG overlapping with the window being significant, the precision, median window size and median R^2 was determined for every window.

2.9 Stack Eight - Methylation trajectory inferencing and phenotype alignment:

Stack Eight constructed a methylation-based developmental trajectory for the methylation profiles to visualise the association of individuals along a continuum. The continuum was aligned against the age, phenotype and gene-level methylation effects. The k-NN algorithm was used to cluster the data, the MST was used to construct a relative tree for the determination of association (17) and the DPT was used to provide a pseudotime-frame (18).

2.10 Overview of the computational processing:

All Stacks were conducted in python-3.13 (Python Software Foundation, Wilmington, DE, USA), on a laptop – except for Stack Four and Stack Five which were performed within Google Colab's python platform (Google LLC, Mountain View, CA, USA). The entire pipeline took under two days to execute with this combination. The Hierarchical Bayesian pipeline of Stack Three involved 8,000 tuning iterations and 16,000 draws. Stack Four processed 10,000 CpGs in batches of 1,000 CpGs for 20,000 iterations and 2,000 posterior samples. Stack Five involved 30,000 iterations and 2,000 draws, thus the entire process was highly computationally efficient.

The following python packages were used: pandas (3.0.3), numpy (2.2.6), matplotlib (3.10.8), seaborn (0.13.2), scipy (1.17.1), pymc (5.28.4), pytensor (2.38.2), arviz (0.22.0), statsmodels (0.14.6), hdbscan (0.8.44), scikit-learn (1.9.0) and network (3.6.1).

All accompanying code and data is available in the open-access GitHub repository associated with this study: tsto3616/poly-CpG: The repository encodes the sliding window pipeline, accompanying statistics and data for a study on the California Sea Lion's epigenome.

3. Results:

3.1 Blood methylation clusters of the *Zalophus californianus* are independent of age but predictive of aerobic diving capacity:

Four blood methylation profile clusters were formed, with the -1-cluster representing noise (Figure 1A). The clusters did not differ significantly for chronological or epigenetic age (Figure 1A; MANOVA, Roy's greatest root, $p < 0.05$; otherwise, *ns* for all other test statistics).

Despite age independence the clusters were significantly differentiated for the Hierarchical Bayesian and experimentally derived maximum aerobic dive capacity (Figure 1B; MANOVA, $p < 0.05$).

3.2 No significant CpGs explain the maximum aerobic diving capacity of *Zalophus californianus*:

Both univariate and multivariate CpG-phenotype scoring frameworks were applied to the study of the phenotype (maximum aerobic dive capacity) – both frameworks failed to identify any CpGs exceeding the significant threshold (Figure 2A, B). Effect sizes were small and consistent between the two frameworks; however, the probability of the CpG impacting upon the phenotype was significantly reduced for the multivariate CpG-phenotype scoring framework (Figure 2A, 2B). This justified the need for the construction of a poly-CpG-phenotype scoring framework.

3.3 Three genes inherently linked to the maximum aerobic diving capacity and the lactate dynamics of *Zalophus californianus*.

The sliding window framework identified three windows, which were uniquely mapped to the three genes: Glycine Cleavage System H (GCSH), DNA Methyltransferase 1 (DNMT1) and LIM Homeobox transcription factor 1-α (LMX1A). The self-tuning behaviour of the sliding window framework is illustrated for the three genes and the respective genomic order of the CpGs in Supplementary Figure 1.

All three genes were significant predictors of maximum aerobic diving capacity (Table 1) – consistent with either a negative or positive effect from methylation (Table 1; Figure 3A). The three genes each explained at least 40% of the respective variance (penalised R^2) for the maximum aerobic diving capacity (Table 1).

The GCSH gene was independent of the DNMT1 and LMX1A genes, whereas DNMT1 and LMX1A both retained coordinated methylation patterns (Figure 3B).

3.4 The sliding window framework is dependent on the noise: signal ratio for optimal performance.

The influence of noise and sparsity upon the performance of the sliding window framework was evaluated through Python simulated datasets for a grid of noise and sparsity values. At high signal levels (sparsity = 0.8), the model produced numerous trusted windows – but the exact number of windows varied with noise (Figure 4A). The sparsity values of 0.4 and 0.2 produced zero to two windows regardless of the noise (Figure 4A).

The difference in the window counts of the framework maintained a consistently low false-positive rate across the three noise to signal regimes (Figure 4B). The low noise-high sparsity condition produced four times as many windows as the intermediate condition – both conditions were determined to both present with perfect precision for the windows and an acceptable rate for the inclusion of insignificant, singular-CpGs (Figure 4B; Table 2). The summary statistics yielded the highest median R^2 value for the medium noise: signal regime – despite similar window size (Table 2).

The variability in the window detection was not attributed to the differences in the underlying γ or β estimates from the simulated ADVI-EWAS or Horseshoe models (Figure 4C, D). R^2 distributions for γ and β parameters did not differ in a meaningful manner between the two modelling approaches (Figure 4E, F). The dominant driver of window recovery was sparsity. Higher sparsity produced stronger and more coherent signals for window selection, whereas low sparsity reduced the ability of the framework to identify stable windows. This effect was noise agnostic (Figure 4E, F; Table 2).

3.5 Hierarchical analysis of the embedded methylation parameters not predictive of age or aerobic diving capacity.

Diffuse pseudotime revealed no genome-wide methylation trajectory association with the phenotype, but the maximum aerobic diving capacity was strongly associated with the three genes (Figure 4).

4. Discussion:

The aim of the paper was to develop a self-tuning window framework for the analysis of the environmental-methylome-phenotype axis of 69 *Z. californianus* individuals. To do so, blood samples were re-analyzed for the marine mammal specific phenotypic performance – the maximum aerobic diving capacity (6, 10). The results were consistent of no phenotype-methylation association when considering individual CpGs for both the univariate ADVI-EWAS and the multivariate Bayesian Horseshoe shrinkage model (14, 15). Presumably the results were due to the sparsity of the data, combined with the polygenic nature of the complex phenotypic trait assessed and the relatively small sample size (n = 69). However, the application of the sliding window framework provided vital insights into the environmental-methylome-phenotype axis by identifying three key genes associated with the phenotype. Thus, challenging the dominant assumption of methylome studies in mammals – the methylome is more complex than predominantly being a function of age (19, 20).

As Sea Lions age the development from a maternally dependent pup to an independent adult involves the ontogenic adaption towards the development of the physiological extremes needed for prolonged diving (10, 11). However, the age of the individual is said to be the primary predictor of this evolutionary dynamic. The California Sea Lion's methylome is a strong predictor of aging (6, 10). A strong mechanistic explanation of these physiological changes can be explained through the methylation of the three genes identified. The GCSH

gene is consistent with the development of post-diving hyperglycemia in pinnipeds (21-23). Deficiency of the GCSH gene in humans results in non-ketogenic hyperglycemia due to the accumulation of the gluconeogenic amino acid, glycine (24, 25). The excess glucose would drive a metabolic shift towards lactate production in the presence of anaerobic respiration at the end of the dive – resulting in potentially harmful physiological effects (26).

The lactate accumulation due to hypoxia may result in the phenomena known as lactylation. Lactylation is a well described pathway of gene expression influenced by lactate dynamics. Central to the regulation of methylation is the DNMT1 gene – lactate upregulates the expression of DNMT1 due to the hypermethylation of a gene with DNMT1-suppressive properties (27). DNMT1 also has protective factors against the induction of hypoxic induced apoptosis (28). DNMT1 reduces the methylation rate of LMX1 – explaining the interdependence of the poly-CpG-phenotype association disclosed by the sliding window framework (29). LMX1A knock-out results in dysregulation of mitochondria, producing excessive reactive oxygen species (ROS), which can be cytotoxic (30). Taken together, the three genes are mechanistically linked to the metabolic changes experienced in response to the physiological demands of diving.

The sparsity of the methylome reduced the power effect of the study, which was evident in the lack of singular CpGs associated with the phenotype (31). However, the inclusion of windows aggregated the regional CpG scores, shifting the analysis from noise abundant and numerous CpGs to a coordinated regional analysis approach. The process improved the reliability penalized R^2 value and directionality of the effect. As part of the model, the self-tuning sliding window analysis avoids arbitrary definitions of windows by limiting the windows to real chromosomes and imposing a penalization on the R^2 for the window length to prevent the detection of CpGs over vast distances within the one window. The end-result is phenotype-associated windows that are short, stable and directional in their predicted effect

upon the phenotype. Although the study was applied to one marine mammal species only, the method is broadly applicable due to its ability to handle sparse data and polygenic traits (32).

The simulation experiments demonstrated that the sliding window framework behaved predictably despite the varying noise to signal regimes. The high sparsity value produced the most trusted windows, the intermediate level of sparsity generated few windows, and the low level of sparsity generated no windows regardless of the noise level. For all settings generating a window, there was perfect window level precision; however, false positives were detected among the CpGs of windows. Despite this the framework is highly resistant to false-positive calls despite substantial noise. The ADVI-EWAS and or Horseshoe model parameters and their estimated R^2 distributions were not determinates of the number of windows selected, nor the variability within a window. Rather sparsity was the dominant determinant of window recovery – with denser signals capturing windows at a higher rate – as was expected given the penalisation of the window size. The penalisation of the window is inherently restricted by the density function (sparsity) of the CpG data. Future fine tuning of the model could incorporate a hyperparameter for the weighting of the CpG sparsity of the genome. This would be beneficial in the context of ecological studies, as the methylome sequencing coverage directly impacts upon the sparsity of the methylome (33). Wildlife studies are frequently limited by suboptimal sample preservation due to situations in the field – a model capable of accounting for the sparsity of the CpG rate per genomic nucleotide would offer a novel approach to counteract the coverage loss resulting from suboptimal preservation of DNA in the field (34).

Despite numerous limitations throughout the study, the sliding window framework still managed to produce a mechanistic explanation for the ontogenetic establishment of the extreme physiological shifts towards adulthood in the California Sea Lion. The limited sample size, does reduce the inferencing capabilities of the study, but it establishes the ability

386 of the sliding window framework to perform under both data and statistical sparsity (31). The
387 limitation of tissue sampling to the blood of pinnipeds is also a significant limitation.
388 Pinniped red blood cells lack nuclei and thus the epigenome is limited to cells that primarily
389 do not participate in the aerobic diving capacity – through oxygen reservation – rather the
390 effects are more likely secondary to physiological stressors rather than age (35). Examination
391 of other tissues – such as muscle – would benefit in the detection of genes associated with the
392 aerobic diving capacity of *Z. californianus*, but it would also provide stronger effect sizes
393 (10). The stronger effect sizes do not necessarily reflect the sparse nature of the data. Blood
394 samples are also one of the most accessible tissues for wildlife DNA sampling – allowing for
395 more meaningful comparisons for a pan-wildlife methylation tool (6, 34). However, blood is
396 not a universal window into whole animal methylation – tissue dependent methylation is well
397 described (36).

398 Limitations that detract from this study's applicability include the determination of the
399 aerobic dive capacity as a function of age and the cross-sectional nature of the data recycled
400 in the study. Given that the age of *Z. californianus* is a strong predictor of the aerobic dive
401 capacity and the molecular clock of individual pinnipeds, the methylation profiles predictive
402 of the aerobic dive capacity are inherently associated with the age of the individuals (6, 10).
403 This is further exacerbated by the estimation of the maximum aerobic dive capacity of *Z.*
404 *californianus* being calculated as a linear function of the age of the individuals sampled for
405 the methylome sequencing (10). The clusters identified were age independent despite the
406 genes' effect sizes being determined in part by cluster. Further research is a unique
407 opportunity to apply the framework to more diverse phenotypic traits that are (i) documented
408 as part of the sampling framework, (ii) are diverse within a species and (iii) are applicable
409 across species. Through future research a more mechanistic understanding of conservation
410 physiology and genetics can be established.

5. Conclusions:

The aim of the study was to develop a framework for poly-CpG regional analysis to improve the understanding of methylation-phenotype associations in wildlife. Through a sliding window framework, the regional poly-CpG effects were characterized for California Sea Lions – three windows were isolated producing three explanatory genes. The applicability of the method is widespread and can be applied to various wildlife species.

6. Acknowledgements:

No funding was provided. The Author would like to thank the academic advisors who did not contribute to the study, but provided much needed mentorship over the years.

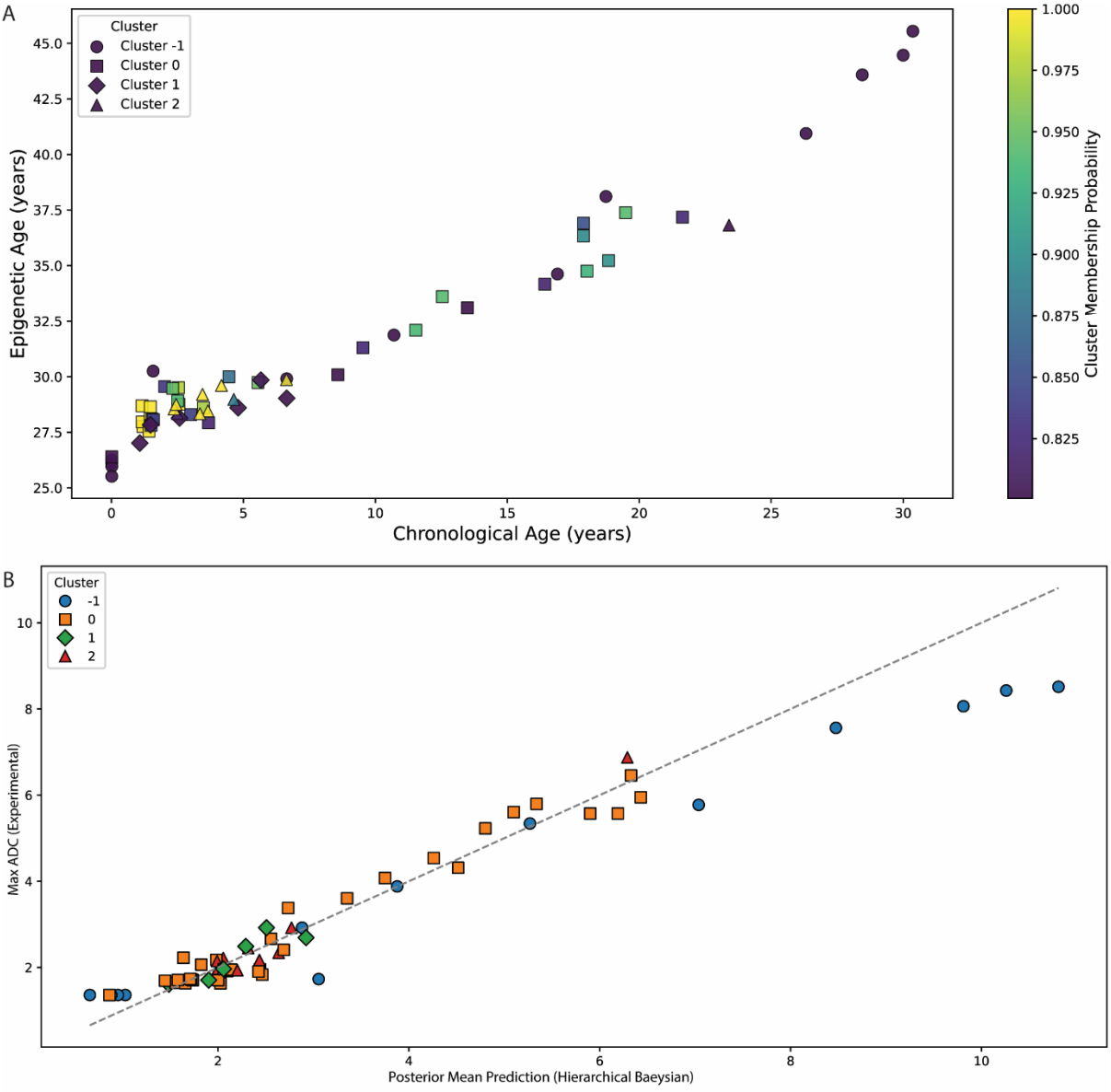


Figure 1: Hierarchical Bayesian prediction of the maximum aerobic diving capacity defined as a function of the experimentally derived maximum aerobic diving capacity. A. The chronological age (x axis) is compared to the epigenetic age (y axis), with the cluster (shape) and probability of assignment to a cluster (colour) featured. B. The Maximum experimental aerobic dive capacity (Max ADC – y axis) is defined for the mean posterior Hierarchical Bayesian prediction of the ADC (x axis). The dashed line represents the perfect prediction, not the linear regression line.

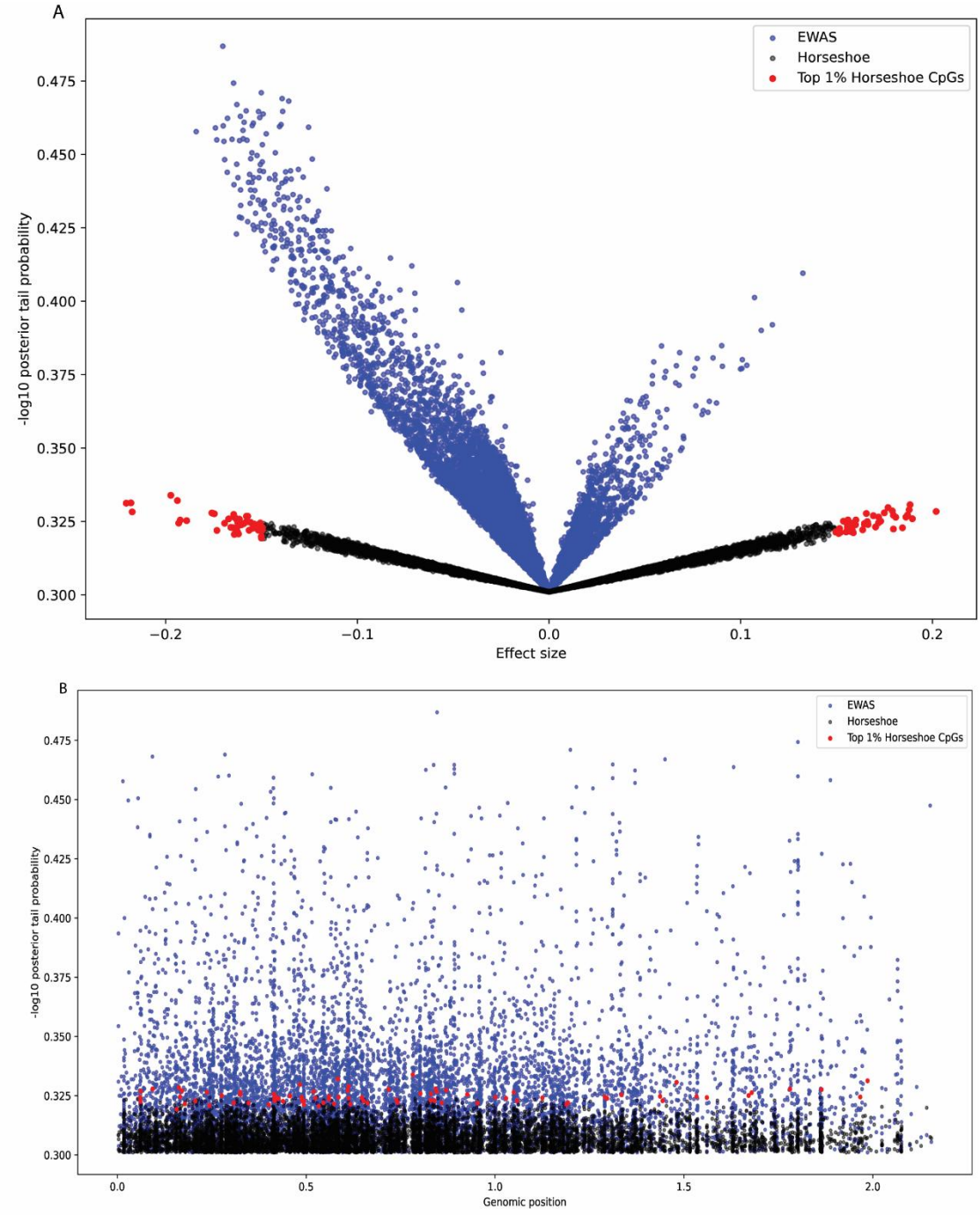


Figure 2: CpG scores are not a significant predictor for the aerobic diving capacity of *Zalophus californianus*. For both A and B, the y axis features the -$\log_{10}$ posterior tail of probability and the same colour coding system is retained. The black dots are the bottom 99% horseshoe values, the red dots the top 1% values and the blue dots are the ADVI-EWAS values. A. The Volcano plot of two contrasting analyses is provided for the ADVI-EWAS and the Horseshoe shrinkage model. The x-axis features the effect size. B. The ADVI and Horseshoe

437 CpG values were compared within a Manhattan plot with the x axis featuring the genomic
438 position.

439 **Table 1: The GCSH, DNMT1 and LMX1A genes are all significant predictors of**
440 **maximum aerobic diving capacity.** The summary statistics for the linear regression used in
441 Figure 3A are summarised. The slope and intercept predict the per-animal, gene effect.

Statistics:	**GCSH**	**DNMT1**	**LMX1A**
Slope	0.1880	-0.1812	-0.1247
Intercept	-1.3511	1.3657	1.6327
R^2	0.5537	0.5969	0.4148
p values	<0.0001	<0.0001	<0.0001
Standard error	0.0218	0.0192	0.0191

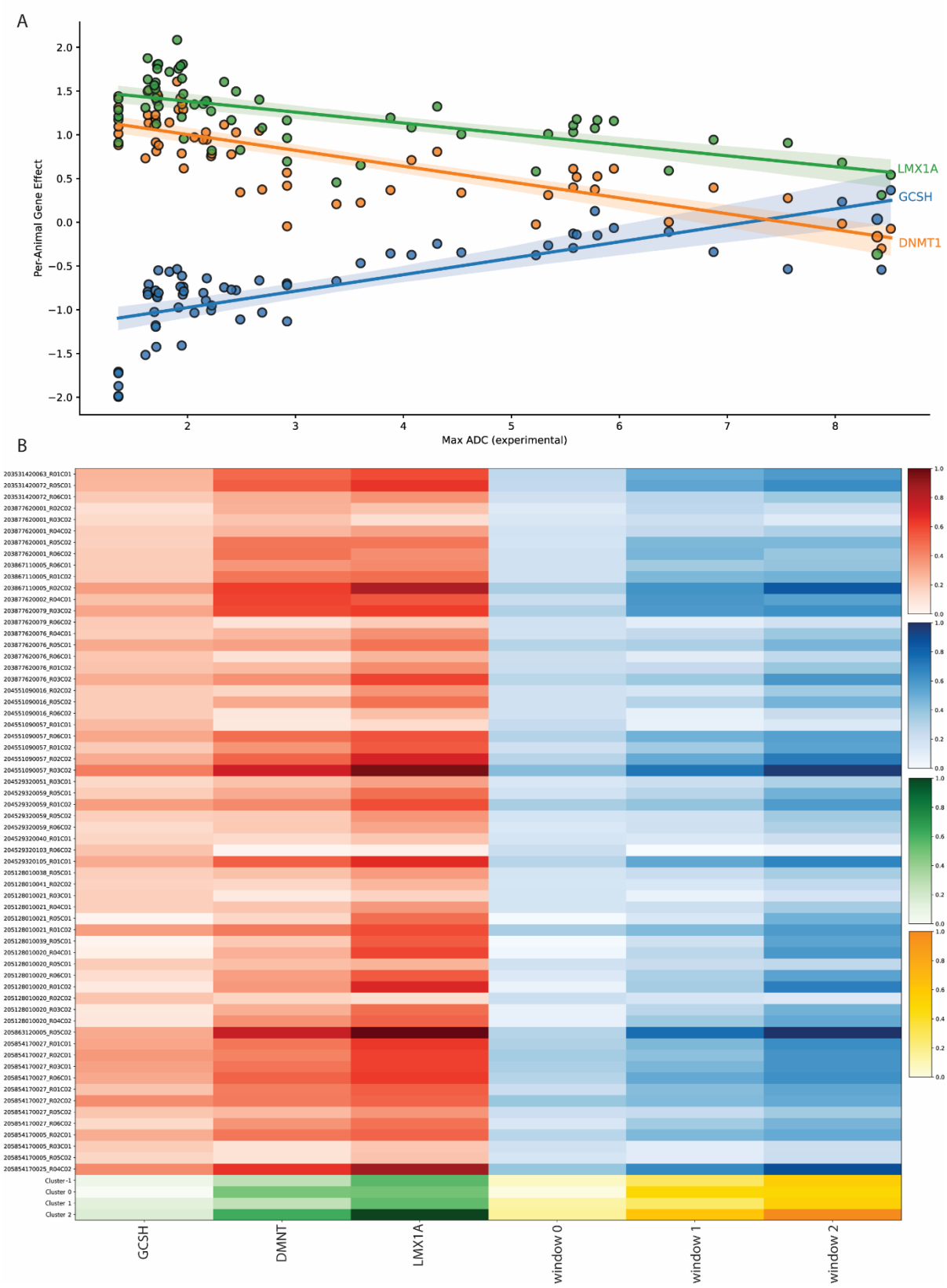


Figure 3: Three genes strongly associated with maximum aerobic diving capacity. A. The effect sizes of the genes were mapped through linear regression for the maximum aerobic

capacity – Max ADC. B. A heat map was generated for the relative effect of each gene upon animals (top left quad - red), window level methylation scores for the animals (top right quad - blue), gene upon HDBSCAN clusters (bottom left quad - green) and window upon the HDBSCAN clusters (bottom right quad - yellow).

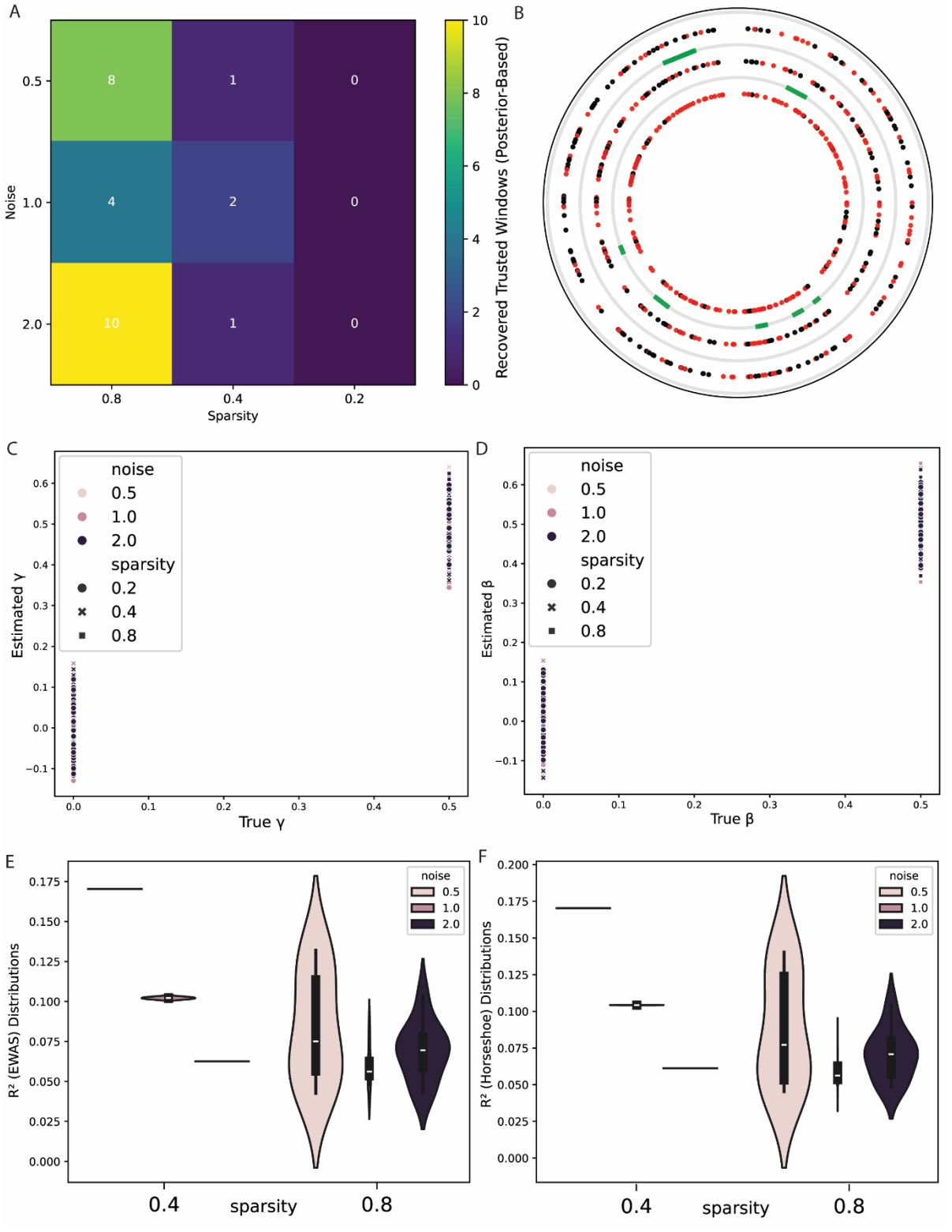


Figure 4: Simulations of the sliding window analysis reveal an absence of window identification at high noise to signal ratios and hallucinations at low noise to signal ratios. A. The heatmap shown is coloured for the number of trusted windows discovered for three tiers

of sparsity (0.8, 0.4, 0.2) and noise (0.5, 1.0, 2.0) – where the lower values of sparsity were predictive of a high noise: signal ratio, higher values the opposite. For the noise values the higher the value the greater the noise: signal ratio. The numbers inside the tiles are the number of windows identified. B. The circos plot was fitted for three combinations of the sparsity and the noise were selected for a small noise: signal ratio (0.8 sparsity and 0.5 noise), medium noise: signal (0.4 sparsity and 1.0 noise) and high noise: signal ratio (0.2 sparsity and 2.0 noise). The inner circles of red and black dots and the green bars represent the small noise: signal ratio dataset, the middle set of rings were for the medium noise: signal ratio and the outer rings of dots and absence of green bars is the high noise: signal ratio (for which there were no identified windows). The red dots represent the signal (CpG scores) and the black dots represent the noise (CpG scores). The green bars are trusted windows. C. The γ values from the ADVI-EWAS significance scoring from the simulated inputs/ 'true' values were modelled on the x axis and the sliding window estimated γ values were plotted on the y axis. This plot was repeated for D. but with the estimated β (y axis) and 'true' β (x axis). In plot E. the R^2 ADVI-EWAS distribution (y axis) was determined for the two sparsity categories for which sliding windows were successful – the 0.4 and 0.8 sparsity values (x axis). This was repeated in plot F. for the R^2 Horseshoe distribution (y axis) mapped against the same sparsity values (x axis).

Table 2: The noise to signal ratio has a dramatic effect on the calling of windows and the variance explained by the windows. The high noise: signal ratio was for a sparsity value of 0.2 and a noise value of 2, the medium was 0.4 (sparsity) and 1 (noise), whereas the small noise: signal ratio had a sparsity value of 0.8 and a noise value of 0.5. Detailed explanations of the workflow can be found in the methods.

Statistics:	High noise: signal ratio	Medium noise: signal ratio	Small noise: signal ratio
Significant CpGs	83	101	168
Trusted windows	0	2	8
Recovered windows	0	2	8
Precision	0	1.0	1.0
Median window size	0	4.5	4.0
Median R^2	0	0.1021	0.0751

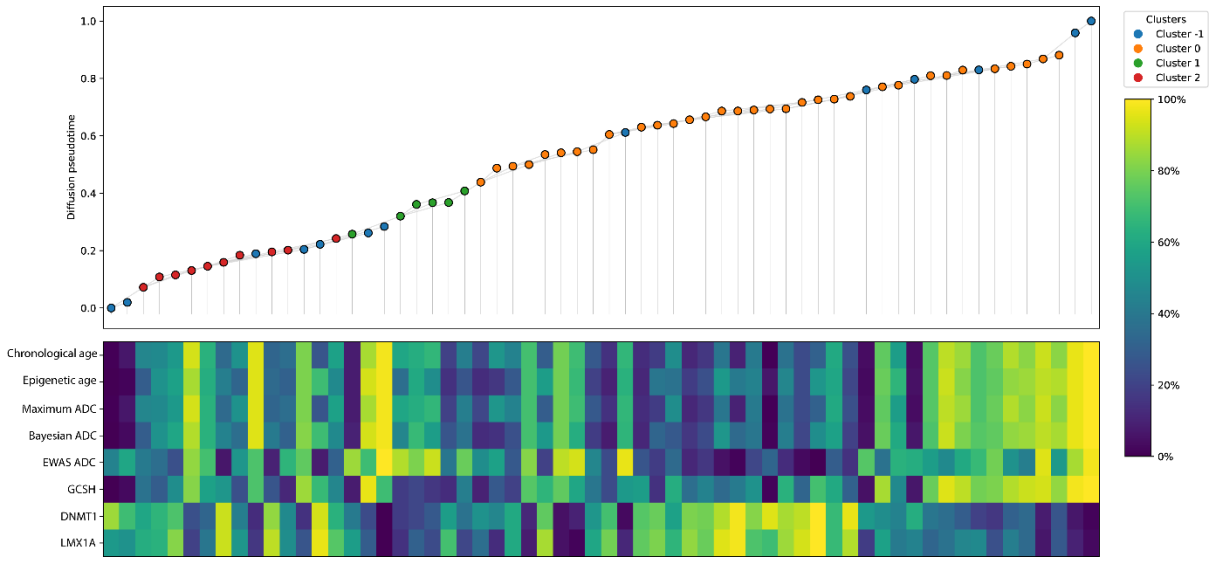


Figure 5: Hierarchical ordination of the methylation does not predict phenotype. The diffusion pseudotime (from the origin – 0 to the terminal individual – 1) was mapped for a k-NN, MST constructed hierarchical model for the aging and tracing of evolutionary methylation histories. The origin animal at the bottom of the graph was the root for the tree, it was the youngest chronological individual. The dot's colours in the top of sector of the graph represent the HDBSCAN clusters. The y axis of the top plot represents the pseudotime, the x axis was ordinally spaced out to accommodate for the heatmap below. The results were combined with

485 the phenotypic markers: chronological age, epigenomic age, maximum aerobic diving capaciry
486 (Maximum ADC), Bayesian ADC (the predicted aerobic dive capacity from the Hierarchical
487 Bayesian model), the EWAS ADC (generated from the ADVI-EWAS model) and the GCSH,
488 DNMT1 and LMX1A genes are displayed. The colour for the heatmap dictates the relative
489 proportional value for the given trait across all samples/ individuals.

490

Supplementary Methods:

1.1 Supplementary methods overview:

Removal of the excess CpGs was performed using the var() command from Python's numpy command. MANOVA was applied to understand the role of the chronological age and epigenetic age upon the formation of the HDBSCAN clusters.

The 20 PCs were inputted into HDBSCAN with a minimum cluster size of six individuals and minimum number of neighbouring samples of three (min_samples = 3) needed for the identification of a core sample within a cluster. The choice of the minimum cluster size and number of neighbouring samples which arbitrarily defined in this study were chosen on the basis of a small sample size, high dimensional PCA space. The pipeline is described in greater detail in Supplementary Methods 1.2

The weighted parameter (λ) was optimised using a 1D Fisher solution. Epigenetic ages were calculated through using the model provided by Robeck, Haghani (6) for all pinniped blood methylation clocks, as the California Sea Lion specific clock model was not freely available (Supplementary Methods 1.3). The weighting parameter was chosen to maximise the separation between clusters while minimised the within-cluster variability. Latent age values were assigned to each individual sample and/or animal (Supplementary Methods 1.4).

The stability and age-related relevance of each cluster was determined through three reliability metrics: (1) age coherence, (2) ordering reliability and (3) methylation tightness. Age coherence was calculated as how tightly chronological and epigenetic ages were grouped within a cluster. The ordering reliability was to determine if the clusters followed a monotonic progression in which the epigenetic age either always increased or decreased with the ordering of the hierarchical clusters. The methylation tightness quantified the within-cluster

514 homogeneity of the CpG M-values. The Hierarchical Bayesian model produced a divergence
515 rate of 0.075%. This was summarised in Supplementary Methods 1.5.

516 The maximum aerobic dive capacity was assumed to be a normally distributed outcome
517 variable with a mean provided as a linear prediction generated from an intercept, the latent age
518 effect and a random effect to account for HDBSCAN clustering. Individual weights were
519 included in the model as a scale for the impact of the cluster reliability for the given animal.
520 That is the more 'reliable' the cluster the animal belongs to, the more the animal contributes to
521 the scaling of the residual variance and thus contribute more strongly to the likelihood
522 prediction for the Bayesian model. For all informative priors, weak parameters were used to
523 prevent overfitting without over-constraining the model – a necessary consideration or the
524 limited sample size of the study (n=69). Posterior samples were estimated using Markov chain
525 Monte Carlo (MCMC). The posterior mean predictions were used to compute the residuals
526 needed for the downstream ADVI modelled EWAS in Stack Four. The formulae for the
527 Bayesian Hierarchical model are available in Supplementary Methods 1.6, with additional
528 information regarding the MCMC sampling.

529

530 The residual maximum aerobic dive capacity was determined as a function of latent age,
531 CpG-specific methylation, cluster-level random effects using the individual reliability weights
532 from Stack Two. To computationally manage the CpG EWAS (10,000 CpGs, 20,000 iterations
533 and 2,000 posterior samples) the process was conducted in batches of 1,000 CpGs with separate
534 regression fit for each CpG using ADVI rather than a MCMC posterior sampling. The ADVI
535 process provided a faster approximation of the posterior parameters. For each of the CpGs the
536 estimated CpG-specific effect size (γ) determines the effect of the methylation upon the residual
537 phenotype. The residual phenotype is a condition of the latent age and clustering associated

phenotype from Stack Three. The posterior means and standard deviations for the CpG-specific effect size were extracted and retained for all CpGs. These were used to compute the 95% credible intervals. These were used for the determination of the posterior probabilities of a positive benefit under the Gaussian variational posterior. The CpGs were classified as significant for the posterior probability of an effect being greater than 0.975 for either direction. The resultant CpG effect size data frame was used for downstream analyses. The statistical basis of Stack Four can be found in Supplementary Methods 1.7.

To achieve this the methylation matrix was standardised for the columns to ensure the effect sizes across the CpGs are comparable to the ADVI results of Stack Four. A Bayesian linear regression model with a horseshoe prior was fit to the standardised methylation matrix by applying ADVI for the computational efficiency.

The estimated model estimated a global shrinkage parameter (τ), singular-CpG-specific local shrinkage parameters (λ_j), and singular-CpG-specific regression coefficients (β_j). The Horseshoe prior enables strong reduction of the impact of noise (CpGs) while allowing for large CpG effects to remain unshrunk. Posterior samples were drawn from the fitted variational distribution with the posterior means and standard derivations of β_j were extracted every individual CpG. Additional resources for Stack Five can be found in Supplementary Methods 1.8.

Each CpG was aligned for its EWAS posterior mean effect size ($\hat{\gamma}_j$) and its comparable Bayesian tail probability for the Horseshoe model ($\hat{\beta}_j$) and posterior standard deviation. The two datasets were annotated on the CpG-level with the computation of the cumulative genomic positions for Manhattan-style plotting.

Both plots were developed by plotting the ADVI posterior, Bayesian tail probability for γ_j and the posterior tail of probability for the Horseshoe model (β_j). The CpGs from the

Horseshoe CpGs were plotted using the assumption that the posterior tail of probability (β_j) generated by the Horseshoe CpGs under variational approximation was comparable to the ADVI posterior Bayesian tail probability for γ_j. The Volcano plot visualised the direction of the significance for methylation while the Manhattan plot visualised the overall significance of methylation for the overlayed genomic positions. The top 1% most significant Horseshoe CpGs were highlighted in both plots. Comprehensive detail on the statistical foundations of these steps can be found in Supplementary Methods 1.9.

CpG methylation values were reduced using PCA and a k-nearest neighbour (k-NN) graph was constructed in PCA space with a minimum spanning tree (MST) extracted for the backbone of the methylation similarity score across individuals. Diffiusion pseudotime was computed for the y axis of the MST and rooted for the individual with the lowest chronological age. The production was a one-dimensional ordering of individuals that reflects the progressive changes in the methylation structure. The MST was visualised within the PCA space for clustering-based colouring (from Stack One) to highlight the relationship between the methylation-defined clusters and inferred trajectory.

For the integration of the phenotypic information the pseudotime ordering was used as a shared x-axis for the second panel. The second panel featured a tile-based heatmap of the ranked phenotypes with variables such as chronological age, epigenetic age, maximum aerobic diving capacity, the Bayesian residuals from Stack three, the ADVI derived poly-CpG scores from Stack Four result and the gene-level effects for the genes determined to be significant for the methylation-phenotype association. Each variable is percentile normalised and displayed as a horizontal tile row with precise alignment against the MST ordering. The results produced a unified figure for the comparisons of methylation trajectory structure against the physiological and gene-level phenotypes. Trajectory inference was used to identify latent methylation progression patterns that may reflect physiological ageing or adaptive diving

phenotypes. Further information detailing the Stack Eight pipeline can be found in Supplementary Methods 1.11.

1.2 Calculation of the M-scores:

The β-scores from the normalised methylation data were converted into M-scores for methylation using the formula:

$$M_{ij} = \log_2\left(\frac{\beta_{ij}}{1-\beta_{ij}}\right)$$

1.3 Estimation of the Epigenetic age:

The epigenetic age of the animals was predicted using a pinniped-wide methylation clock. For the calculation of the methylation age an elastic net epigenetic clock model was used whereby the methylation age (Age_i^{epi}) is a function of the intercept β_0 and the sum of the coefficients for the CpG (β_j) and the methylation scores (M-scores). The epigenetic age was calculated using the formula:

$$\mathrm{Age}_i^{\mathrm{epi}} = \beta_0 + \beta_1 M_{i1} + \beta_2 M_{i2} + \cdots + \beta_p M_{ip}$$

1.4 Latent age definition:

The latent age is then defined by the combination of the epigenetic age, the chronological age and age of the clusters from stack 1. The latent age was classified using the formula:

$$\mathrm{LatentAge}_i(\lambda) = \lambda\,\mathrm{Age}_i^{\mathrm{chrono}} + (1-\lambda)\,\widehat{\mathrm{Age}}_i^{\mathrm{epi}}, \qquad \lambda \in [0,1]$$

Where the λ is determined by the 1D Fisher optimisation formula:

$$\lambda^* = \arg\max_{\lambda\in[0,1]} J(\lambda)$$

608 $J(\lambda)$ is given by:

609

610 $$J(\lambda) = \frac{S_B(\lambda)}{S_W(\lambda)}$$

611 Where $S_B(\lambda)$ and $S_W(\lambda)$ – the between cluster and within cluster parameters respectively
612 – were given by:

613

614 $$S_B(\lambda) = \sum_k n_k \left(\overline{\text{LatentAge}}_k(\lambda) - \overline{\text{LatentAge}}(\lambda)\right)^2$$

615 $$S_W(\lambda) = \sum_k \sum_{i \in k} \left(\text{LatentAge}_i(\lambda) - \overline{\text{LatentAge}}_k(\lambda)\right)^2$$

616 The individual animal mean latent age was provided by the formula:

617

618 $$\overline{\text{LatentAge}}(\lambda) = \frac{1}{N} \sum_{i=1}^{N} \text{LatentAge}_i(\lambda)$$

619 The cluster mean latent age was provided by:

620 $$\overline{\text{LatentAge}}_k(\lambda) = \frac{1}{n_k} \sum_{i \in k} \text{LatentAge}_i(\lambda)$$

621 **1.5 Calculation of the reliability scores and weights:**

622 The formula for this reliability weighting is provided by formula below whereby the age
623 coherence (r_k^{coh}) is a function of the variance of the chronological age and epigenetic age. Age
624 coherence scores close to 1 indicate a tight age clustering for both the chronological and
625 epigenetic age, but a score of 0 indicates the opposite.

626

627 $$r_k^{\text{coh}} = \frac{1}{1 + \text{Var}\left(\text{Age}_i^{\text{chrono}} : i \in k\right) + \text{Var}\left(\widehat{\text{Age}}_i^{\text{epi}} : i \in k\right)}.$$

628 The order of the ages for the clusters is defined in two steps, first by the function:

629

630 $$\overline{\text{Age}}_k = \frac{1}{n_k} \sum_{i \in k} \widehat{\text{Age}}_i^{\text{epi}}$$

631 Then by the formula produces the reliability order score:

632

633 $$r_k^{\text{ord}} = \mathbb{I}\left(\overline{\text{Age}}_k < \overline{\text{Age}}_{k+1}\right)$$

634 The internal methylation tightness within the cluster is given by the formula:

635

636 $$r_k^{\text{tight}} = \frac{1}{1 + \text{Var}\left(M_{ij} : i \in k,\, j \in \mathcal{C}\right)}$$

637 Where the r_k^{tight} value is a function of the variation of M-score for the clusters for all
638 10,000 CpGs (where C is the CpGs) used for the clustering. If the clusters look similar in
639 methylation space the value is close to 1 but if not, it is close to 0. It is a molecular coherence
640 metric.

641 All the above cluster evaluation metrics are combined into a combined r_k for the
642 calculation of the weights for the Bayesian analysis.

643

644 $$r_k = \frac{r_k^{\text{coh}} + r_k^{\text{ord}} + r_k^{\text{tight}}}{3}$$

The reliability of the cluster (determined from r_k) is fed into the formula for the reliability of each individual animal. The reliability of the cluster and the predicted likelihood of the animal belonging to the cluster are combined to produce the metric:

$$w_i = p_i^{\text{cluster}} r_{c_i}$$

The w_i value is used as a Bayesian weight/ prior.

1.6 Determination of the Bayesian Hierarchical model:

The Bayesian Hierarchical model will be constructed using the formula:

$$Y_i \mid \mu_i, \sigma^2, w_i \sim \mathcal{N}\left(\mu_i, \frac{\sigma^2}{w_i}\right)$$

Where *i* denotes the animal, Y_i is the maximum aerobic diving capacity, the μ_i is the linear predictor defined below, the σ is residual normal variation, *N* indicates the normal distribution and w_i is the clipped weight (clipped at 10^{-6} to avoid zero variance issues) provided by the stack 2.

The σ value is defined as HalfNormal(5) whereby the residual noise is positive and moderate but allows noisy data to feature. This stabilises the model with weighted likelihoods.

The HalfNormal(5) equation dictates that the normal distribution has a prior ranging from 0 to ~10 with most values ranging from 0-5, values above ten are possible. It serves as a weakly informative prior for the prevention of unreasonable values in the absence of underfitting.

The formula for μ_i is provided below:

$$\mu_i = \alpha + \beta_{\text{age}} \, \text{LatentAge}_i + b_{c_i}$$

LatentAge$_i$ is output of the optimisation from stack 2. The α value is a weak global intercept prior that barely constrains the model. This prevents the drifting of the intercept due to the small sample size (n=69) and allows for an effect from the data. The β_{age} latent age prior is for the stabilisation of the regression model without biasing the magnitude of the effect or the direction of the age effect. The b_{ci} value is the cluster random effects defined by the cluster specific deviation of the cluster c_i for the animal i.

The α value is defined as $\alpha \sim N(0, 10^2)$, the β_{age} latent age as $\beta_{\text{age}} \sim N(0, 10^2)$, and the b_{ci} as $b_k|\tau^2 \sim N(0, \tau^2)$.

The τ value is provided by the formula: $\tau \sim \text{HalfNormal}(5)$ which allows for model cluster differences but also allows for extreme values if needed. It is a shrinkage mechanism to prevent the overinterpretation and or overfitting from clusters and or small sample size.

The posterior sampling was performed using No-U-Turn Sampler (NUTS) for which there were 4 chains, 2,000 tuning steps, 4,000 posterior draws, a target of acceptance of 0.9 and a fixed random seed for reproducibility.

The posterior means were used to calculate the residuals using the following formulae:

$$\hat{\mu}_i = \mathbb{E}[\alpha] + \mathbb{E}\left[\beta_{\text{age}}\right] \cdot \text{LatentAge}_i + \mathbb{E}\left[b_{c_i}\right]$$

Where the residuals are:

$$\text{Residual}_i = Y_i - \hat{\mu}_i$$

The residuals were included in the ADVI component of the EWAS for Stack Four.

1.7 Single-CpG EWAS model overview:

The ADVI process involved the optimising the parameters through the maximisation of the evidence lower bound (ELBO). The posterior summaries of the predicted effect of methylation upon the aerobic dive capacity residuals (γ_j) from Stack Three were generated by the sampling of the optimised variational distribution.

The model assumes a global normal and linear formula for the prediction of the residuals from stack 3 using the following equation:

$$r_{ij} = \alpha_j + \beta_{age,j} \cdot \text{LatentAge}_i + \gamma_j \cdot M_{ij} + b_{cij} + \varepsilon_{ij}, \qquad \varepsilon_{ij} \sim \mathcal{N}\left(0, \frac{\sigma_j^2}{w_i}\right)$$

The weight is derived from the stack 2, the $b_{cij} \sim N(0, \tau_b)$ for which $\tau_b \sim \text{HalfNormal}(5)$. The linear model has the following priors for all subscript *j,* each CpG:

$$\alpha_j \sim \mathcal{N}(0, 10^2), \qquad \beta_{age,j} \sim \mathcal{N}(0, 10^2), \qquad \gamma_j \sim \mathcal{N}(\mu_\gamma, \tau_\gamma^2)$$

The hyperpriors are defined as below:

$$\sigma_j \sim \text{HalfNormal}(5), \qquad \mu_\gamma \sim \mathcal{N}(0{,}10), \qquad \tau_\gamma \sim \text{HalfNormal}(5)$$

The ADVI method to estimate the γ_j values for the given age and M score of each CpG per animal (M_{ij}). The γ_j values are given by the Gaussian factorised approximation formula, where d_j is the number of parameters:

$$q(\theta_j) = \prod_{k=1}^{d_j} \mathcal{N}\left(\theta_{jk} \mid \mu_{jk}, s_{jk}^2\right)$$

Whereby the θ_j is defined as the following:

$$\theta_j = (\alpha_j,\ \beta_{\text{age},j},\ \gamma_j,\ \sigma_j,\ b_{1,j},\ b_{2,j},\ \ldots,\ b_{K,j})$$

The variation parameters (μ_{jk} and s_{jk}) are optimised using the ELBO method – as described:

$$\text{ELBO}(\phi_j) = \mathbb{E}_{q_{\phi_j}}\left[\log p\left(r \mid \theta_j\right) + \log p\left(\theta_j\right) - \log q_{\phi_j}\left(\theta_j\right)\right]$$

The summaries of the posteriors for the γ_j are obtained by sampling the variation approximations with the following equation:

$$\gamma_j^{(s)} \sim q(\gamma_j \mid \phi_j), \qquad s = 1, \ldots, S$$

The conversion of the Bayesian ADVI posterior values for each singular-CpG was used to calculate the mean and standard deviation of the predictions. These two metrics were fed into a Bayesian analogue of the p value to predict the significance of each CpG upon the aerobic dive capacity. This assumed a Gaussian variational posterior and produces a directional 'p value'. Significance was defined as $(p > 0.975) \mid (p < 0.025)$ – for the 95% confidence intervals as defined as:

$$\gamma_{j,\text{lower}} = \hat{\gamma}_j - 1.96 \cdot \text{sd}(\gamma_j), \qquad \gamma_{j,\text{upper}} = \hat{\gamma}_j + 1.96 \cdot \text{sd}(\gamma_j)$$

The equation for the conversion to the *p* values is provided below:

731 $$p(\gamma_j > 0) = 1 - \Phi\left(\frac{0 - \hat{\gamma}_j}{\mathrm{sd}(\gamma_j)}\right)$$

732 Where Φ is the cumulative distribution means μ and standard deviation σ.

733 **1.8 Horseshoe Bayesian shrinkage model:**

734 The model relied on the standardisation of the M-score matrix to stabilise the shrinkage
735 behaviour of the horseshoe prior. This was performed through:

736

737 $$M^*_{ij} = \frac{M_{ij} - \bar{M}_{.j}}{\mathrm{sd}(M_{.j})}$$

738 The Bayesian linear regression model with global-local shrinkage prior for the estimation
739 of the singular-CpG-specific methylation effects on the observed maximum aerobic diving
740 capacity is given for the individual i, and the single-CpG j:

741

742 $$y_i = \alpha + \sum_{j=1}^{J} \beta_j M^*_{ij} + \varepsilon_i, \qquad \varepsilon_i \sim \mathcal{N}(0, \sigma^2)$$

743 With the shrinkage structure:

744

745 $$\beta_j \sim \mathcal{N}(0, \tau\lambda_j)$$

746 Where:

747

748 $$\tau \sim \mathrm{HalfCauchy}(1), \qquad \lambda_j \sim \mathrm{HalfCauchy}(1)$$

749 With the additional priors:

$$\alpha \sim \mathcal{N}(0, 10^2), \qquad \sigma \sim \text{HalfNormal}(5)$$

Posterior inferencing used mean-field ADVI with the variational posterior assumed to be of a fully factorised Gaussian form:

$$q(\theta) = \prod_{k=1}^{d} \mathcal{N}\left(\theta_k \mid \mu_k, s_k^2\right)$$

Where the θ is:

$$\theta = \left(\alpha, \beta_1, \ldots, \beta_J, \sigma\right)$$

The posterior summaries were generated from the 30,000 iterations of ADVI and 2,000 draws to form the optimised variational distributions. Posterior summaries were generated for the estimate of β_j which was determined through the formula:

$$\hat{\beta}_j = \mathbb{E}\left[\beta_j \mid \text{data}\right], \qquad \text{sd}\left(\beta_j\right) = \sqrt{\text{Var}\left(\beta_j \mid \text{data}\right)}, \qquad \left|\hat{\beta}_j\right| = \left|\hat{\beta}_j\right|$$

1.9 Manhattan and Volcano plotting:

For each CpG the EWAS directional Bayesian tail of probability was computed under the Gaussian variational approximation:

$$p_{\gamma,j} = P\left(\gamma_j > 0 \mid \text{data}\right) = 1 - \Phi\left(0;\, \hat{\gamma}_j,\, \text{sd}\left(\gamma_j\right)\right)$$

Where the $\hat{\gamma}_j$ and $sd(\gamma_j)$ were denote the posterior mean and standard deviation respectively. The Volcano statistics were generated through:

$$\text{volcano}_{\gamma,j} = \begin{cases} -\log_{10}(1 - p_{\gamma,j}), & \hat{\gamma}_j \geq 0, \\ -\log_{10}(p_{\gamma,j}), & \hat{\gamma}_j < 0 \end{cases}$$

Genomic coordinates of the CpGs were cumulatively considered for the Manhattan plot as:

$$\text{position}_{ij} = \text{Start}_{ij} + \min(\text{Start}_{.j})$$

For the Horseshoe inputs the posterior probability of a positive effect are generated from the Gaussian variational posterior – computed as:

$$p_{\beta,j} = P(\beta_j > 0 \mid \text{data}) = 1 - \Phi\left(0; \hat{\beta}_j, \text{sd}(\beta_j)\right)$$

The Horseshoe volcano plots were generated through:

$$\text{volcano}_{\beta,j} = \begin{cases} -\log_{10}(1 - p_{\beta,j}), & \hat{\beta}_j \geq 0, \\ -\log_{10}(p_{\beta,j}), & \hat{\beta}_j < 0 \end{cases}$$

The top 1% of the Horseshoe CpGs were isolated as a function of the magnitude of the posterior mean:

$$|\hat{\beta}_j| > \text{percentile}_{99}(|\hat{\beta}|)$$

The Horseshoe genomic coordinates for the Manhattan plot were as described above for ADVI.

1.10 Sliding window and gene-phenotype visualisation analysis:

The CpG position was determined computationally using the following notation to align the CpGs – this was the basis of the sliding window model – so as to not confound biologically independent data from different strands of DNA.

$$\text{cpg(pos)}_j = \begin{cases} \text{start}_j, & \text{strand} = +, \\ \text{end}_j, & \text{strand} = - \end{cases}$$

The methylation matrix was constructed with the CpGs sorted by the Chromosome and genomic position. The CpGs, j, were computed within a window, w, through the following formula to determine the 'signal', where the M denotes the methylation matrix, the γ_w and the β_w denote the ADVI and Horseshoe values for each window:

$$S_\gamma = M_{\cdot,W}\,\gamma_W, \qquad S_\beta = M_{\cdot,W}\,\beta_W.$$

The signal was used to calculate the portion of the variability that was explained by the model. The formula for this process was:

$$R^2 = \left(\frac{\langle S - \bar{S},\, y - \bar{y} \rangle}{\|S - \bar{S}\|\, \|y - \bar{y}\|} \right)^2, \qquad \text{sign} = \text{sgn}(\langle S - \bar{S},\, y - \bar{y} \rangle)$$

A penalised score of the R^2 was generated through the formula:

$$\text{score} = R^2 - \lambda_{\text{penalty}} \cdot |W|$$

This score was iteratively calculated for the formula $score_{new} - score_{best} > minimum\text{-}R^2\text{-}increase$. In addition, both models must agree on the direction of the score – as determined by:

$$\text{sign}_{\gamma} = \text{sign}_{\beta} \neq 0$$

If either the R^2 growth condition or the direction of the score condition failed then the growth of the window ceased. Windows smaller than the predefined minimum number of CpGs (four – can be altered) then the window was not retained. For a window to receive a trusted status then the R^2 value for both the ADVI and Horseshoe sliding window models must be greater than 0.5. For semi-trusted window status, the R^2 value must be greater than 0.5 for either the ADVI or the Horseshoe sliding window model.

The windows were then assigned to genes using the transcription start site (TSS) – as defined by the strand. The basal domain was assigned as the TSS – 5,000bp to the TSS + 1,000bp. Domains were extended to the midpoints between adjacent genes – with the closest gene being assigned the CpG for intergenic regions. A window with the genomic span of s_w and e_w (start and end of the window) were assigned to the gene, g, only when:

$$\text{regulatory(start)}_g \leq e_w \quad \text{and} \quad \text{regulatory(end)}_g \geq s_w$$

The combined effect of the CpG was used for downstream calculation of the gene and or window effects on the phenotype of *Z. californianus*. For the calculation fo the combined effects the formula below was used:

$$\theta_j = \frac{\gamma_j + \beta_j}{2}$$

The logistic transformation was applied to each animal's methylation measurement:

$$\mathrm{poly}_{ij} = \sigma(\theta_j M_{ij}), \qquad \sigma(x) = \frac{1}{1+e^{-x}}$$

The combined effect per gene was defined as:

$$\mathrm{Effect}_g = \frac{1}{|W_g|}\sum_{j \in W_g}(\gamma_j + \beta_j)$$

The median absolute deviation (MAD) was used to classify the direction of impact – be it negative, positive or absent. For this calculation the following formulae were used:

$$\mathrm{Effect}_g = \frac{1}{|W_g|}\sum_{j \in W_g}(\gamma_j + \beta_j), \qquad \mathrm{MAD} = \mathrm{median}\left(\left|\mathrm{Effect}_g - \mathrm{median}(\mathrm{Effect})\right|\right)$$

For which the direction is determined by $\mathrm{Effect_g}$ being greater than 0.5(MAD) producing an Up direction, an $\mathrm{Effect_g}$ less than 0.5(MAD) producing a Down direction, otherwise producing a Neutral direction.

The linear regression model for the effect of the genes upon the maximum aerobic diving capacity (ADC) is predicted by the following equation:

$$E_{ig} = \alpha_g + \beta_g \,\mathrm{ADC}_i + \varepsilon_{ig}$$

For the slope and intercept values:

$$(\hat{\alpha}_g, \hat{\beta}_g) = \arg\min_{\alpha,\beta} \sum_{i=1}^{n} \left(E_{ig} - \alpha - \beta\ \mathrm{ADC}_i\right)^{2.}$$

The $\mathcal{E}_{ig}$ value in the global linear model represents the residual error term. The regression model was fitted for each individual gene.

The simulation for the sliding windows involved the mapping of artificial CpGs to a 100,000 nucleotide-long chromosome (the 'true' effects). For each sparsity level (0.2, 0.4, 0.8) the proportion of the CpGs determined to be significant was given by the formula:

$$\gamma_{\mathrm{true},i} = \beta_{\mathrm{true},i} = \begin{cases} 0.5 & \text{if CpG } i \text{ is selected as causal,} \\ 0 & \text{otherwise} \end{cases}$$

Where the γ and β values are substitutes for the ADVI-EWAS and Horseshoe parameters respectively. Methylation values were drawn from 120 individual simulated samples for which there was a presumption of a standard normal distribution: $M_{n\times p} \sim \mathcal{N}(0,1)$. Phenotypes were generated as a linear combination of the ADVI-EWAS effect with the inclusion of Gaussian noise – as defined by:

$$y = M\gamma_{\mathrm{true}} + \varepsilon, \qquad \varepsilon \sim \mathcal{N}(0, \sigma^2)$$

Where σ was 0.5, 1.0, 2.0 – the values of the noise quantification. The estimation of the parameters was carried out through the adding of Gaussian uncertainty to the 'true' effects:

$$\gamma_{\mathrm{est}} = \gamma_{\mathrm{true}} + \mathcal{N}(0,0.05), \qquad \beta_{\mathrm{est}} = \beta_{\mathrm{true}} + \mathcal{N}(0,0.05)$$

The posterior probabilities were determined with the following equations:

$$P(\gamma > 0) = \Phi\left(\frac{\gamma_{\text{est}}}{0.05}\right), \qquad P(\beta > 0) = \Phi\left(\frac{\beta_{\text{est}}}{0.05}\right)$$

Where Φ is the standard normal cumulative distribution function. CpGs were considered significant when both posterior probabilities were greater than 0.5 and both signs of the γ_{est} and β_{est} agreed. The sliding window approach was performed as described above, only without consideration of the chromosome since only one 'chromosome' was processed.

1.11 Trajectory based methylation profiling and phenotype association:

Principal component analysis was applied to the methylation matrix, M:

$$M_{\text{PCA}} = \text{PCA}_{20}(M)$$

Only the first two components were analysed further. A k-NN graph was constructed using Euclidean distances for the methylome. For each sample, *i*, the edges were added to its k=10 nearest neighbours and the minimum spanning tree (MST) generated. This provided a sparse backbone for the analysis of the methylome across individuals:

$$G_{\text{MST}} = \text{MST}(G_{\text{k-NN}})$$

The root node was chosen as:

$$i_{\text{root}} = \arg\min_i \text{Age}_i^{\text{chrono}}$$

Where the $\text{Age}_i^{\text{chrono}}$ is the chronological age. The diffusion pseudotime was computed along the shortest path distance of the MST:

$$\mathrm{DPT}_i = d(i_{\mathrm{root}}, i)$$

The results were normalised to [0, 1]:

$$\mathrm{DPT}_i^{\mathrm{norm}} = \frac{\mathrm{DPT}_i - \min_j \mathrm{DPT}_j}{\max_j \mathrm{DPT}_j - \min_j \mathrm{DPT}_j}$$

The PCA-MST embedding was generated by the plotting of the animals in the PC1, PC2 plane. The edges (*u, v*) were drawn as line segments and the nodes coloured by cluster labels from Stack One. Individuals were ordered according to increasing methylome pseudotime with the provided formula:

$$\pi = \mathrm{argsort}(\mathrm{DPT}^{\mathrm{norm}})$$

For each phenotype the values were rank normalised as described below:

$$\tilde{v}_i = 100 \times \frac{\mathrm{rank}(v_i) - 1}{n - 1}$$

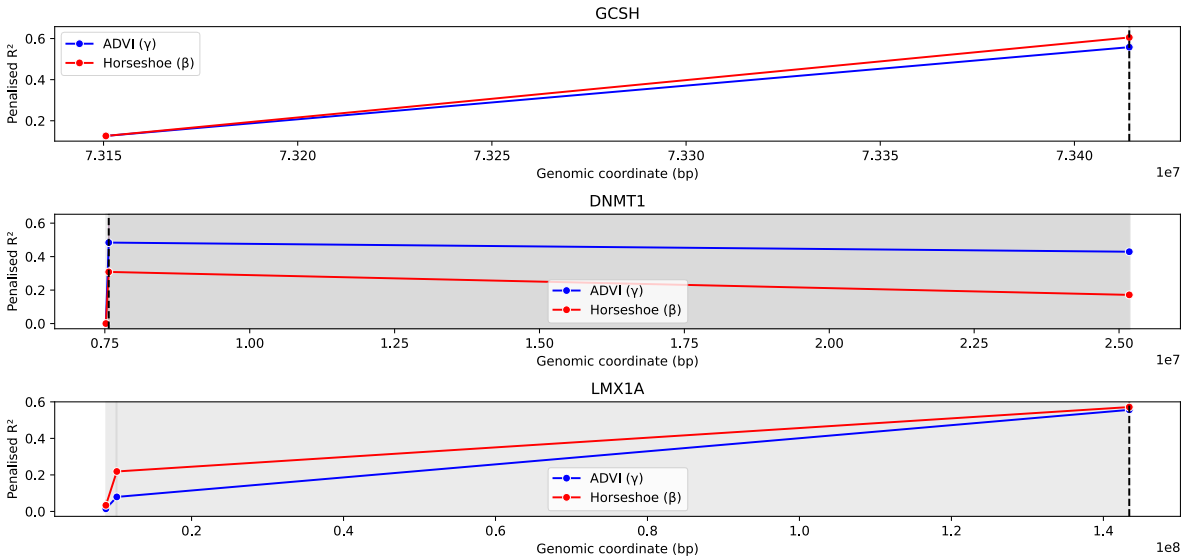
The heatmap was generated with rows corresponding to the individuals ordered by pseudotime:

$$H_{k,i} = \tilde{v}^{(k)}_{\pi(i)}$$

1.12 Data and code availability:

All data and code is available in the study's repository: …

Supplementary Figure 1: The sliding window self-tuning mechanism was illustrated to the three genes determined by four-CpG windows. The three genes were mapped for the genomic coordinates of *Z. californianus*. The coloured dots represent the individual CpGs, the black dashed line represents the termination of the biological signal (difference in directionality of the effect size). The grey shading represents the regions for which there is a consistent change in direction for both the ADVI and Horseshoe models. The genomic coordinates/ their spans are illustrated on the x axis. The genomic ranges for the genes are: 25Kb (GCSH), 31.7Kb (DNMT1) and 5.46Kb (LMX1A).



References:

1. Verplaetse N, Passemiers A, Arany A, Moreau Y, Raimondi D. Large sample size and nonlinear sparse models outline epistatic effects in inflammatory bowel disease. Genome Biol. 2023;24(1):224.

2. Kirpich A, Ainsworth EA, Wedow JM, Newman JRB, Michailidis G, McIntyre LM. Variable selection in omics data: A practical evaluation of small sample sizes. PLoS One. 2018;13(6):e0197910.

927 3. Flesch EP, Rotella JJ, Thomson JM, Graves TA, Garrott RA. Evaluating sample size
928 to estimate genetic management metrics in the genomics era. Molecular Ecology
929 Resources. 2018;18(5):1077-91.
930 4. Vicente-Santos A, Sandoval-Herrera N, Czirják GÁ, Neely BA, Becker DJ.
931 Proteomics Approaches to Ecoimmunology: New Insights into Wildlife Immunity and
932 Disease. Integrative and Comparative Biology. 2025;65(6):1744-54.
933 5. Bogan SN, Yi SV. Potential Role of DNA Methylation as a Driver of Plastic
934 Responses to the Environment Across Cells, Organisms, and Populations. Genome
935 Biology and Evolution. 2024;16(2):evae022.
936 6. Robeck TR, Haghani A, Fei Z, Lindemann DM, Russell J, Herrick KES, et al. Multi-
937 tissue DNA methylation aging clocks for sea lions, walruses and seals. Commun Biol.
938 2023;6(1):359.
939 7. Photopoulou T, Durbach I, Pirotta E, Barratclough A, Schwacke LH, Takeshita R, et
940 al. Methods for analysing wildlife DNA methylation data. Conserv Physiol.
941 2026;14(1):coaf091.
942 8. Uffelmann E, Huang QQ, Munung NS, de Vries J, Okada Y, Martin AR, et al.
943 Genome-wide association studies. Nature Reviews Methods Primers. 2021;1(1):59.
944 9. Pei Y, Li G. Unraveling the interplay of DNA methylation and chromosome
945 organization. Biochem Soc Trans. 2025;53(6):1433-44.
946 10. Weise MJ, Costa DP. Total body oxygen stores and physiological diving capacity of
947 California sea lions as a function of sex and age. J Exp Biol. 2007;210(Pt 2):278-89.
948 11. Fowler SL, Costa DP, Arnould JP, Gales NJ, Kuhn CE. Ontogeny of diving behaviour
949 in the Australian sea lion: trials of adolescence in a late bloomer. J Anim Ecol.
950 2006;75(2):358-67.

12. McInnes L, Healy J, Astels S. hdbscan: Hierarchical density based clustering. The Journal of Open Source Software. 2017;2:205.

13. Fisher RA. THE USE OF MULTIPLE MEASUREMENTS IN TAXONOMIC PROBLEMS. Annals of Eugenics. 1936;7(2):179-88.

14. Kucukelbir A, Tran D, Ranganath R, Gelman A, Blei D. Automatic Differentiation Variational Inference. 2016;18.

15. Carvalho CM, Polson NG, Scott JG. The horseshoe estimator for sparse signals. Biometrika. 2010;97(2):465-80.

16. McLean CY, Bristor D, Hiller M, Clarke SL, Schaar BT, Lowe CB, et al. GREAT improves functional interpretation of cis-regulatory regions. Nat Biotechnol. 2010;28(5):495-501.

17. Kruskal JB. On the shortest spanning subtree of a graph and the traveling salesman problem. Proceedings of the American Mathematical society. 1956;7(1):48-50.

18. Haghverdi L, Buttner M, Wolf FA, Buettner F, Theis FJ. Diffusion pseudotime robustly reconstructs lineage branching. Nat Methods. 2016;13(10):845-8.

19. Thomas K, Harvey JT, Goldstein T, Barakos J, Gulland F. Movement, dive behavior, and survival of California sea lions (Zalophus californianus) posttreatment for domoic acid toxicosis. Marine Mammal Science. 2010;26(1):36-52.

20. Crofts SJC, Latorre-Crespo E, Chandra T. DNA methylation rates scale with maximum lifespan across mammals. Nat Aging. 2024;4(1):27-32.

21. Robin ED, Ensinck J, Hance AJ, Newman A, Lewiston N, Cornell L, et al. Glucoregulation and simulated diving in the harbor seal Phoca vitulina. Am J Physiol. 1981;241(5):R293-300.

974 22. Hance AJ, Robin ED, Halter JB, Lewiston N, Robin DA, Cornell L, et al. Hormonal
975 changes and enforced diving in the harbor seal Phoca vitulina. II. Plasma
976 catecholamines. Am J Physiol. 1982;242(5):R528-32.
977 23. Davis RW. Lactate and glucose metabolism in the resting and diving harbor seal
978 (Phoca vitulina). Journal of comparative physiology. 1983;153(2):275-88.
979 24. Arribas-Carreira L, Dallabona C, Swanson MA, Farris J, Ostergaard E, Tsiakas K, et
980 al. Pathogenic variants in GCSH encoding the moonlighting H-protein cause combined
981 nonketotic hyperglycinemia and lipoate deficiency. Hum Mol Genet. 2023;32(6):917-33.
982 25. Majethia P, Somashekar PH, Hebbar M, Kadavigere R, Praveen BK, Girisha KM, et
983 al. Biallelic start loss variant, c.1A > G in GCSH is associated with variant nonketotic
984 hyperglycinemia. Clin Genet. 2021;100(2):201-5.
985 26. Marinkovich M, Wack RF, Field CL, Whoriskey ST, Kass PH, Gjeltema J. Evaluation
986 of Serial Blood Lactate and the Use of a Point-of-Care Lactate Meter in Live-Stranded
987 Pinnipeds. J Zoo Wildl Med. 2019;50(1):137-46.
988 27. Huang T, Zhou X, Mao X, Yu C, Zhang Z, Yang J, et al. Lactate-fueled oxidative
989 metabolism drives DNA methyltransferase 1-mediated transcriptional co-activator with
990 PDZ binding domain protein activation. Cancer Sci. 2020;111(1):186-99.
991 28. Xu R, Sun Y, Chen Z, Yao Y, Ma G. Hypoxic Preconditioning Inhibits Hypoxia-
992 induced Apoptosis of Cardiac Progenitor Cells via the PI3K/Akt-DNMT1-p53 Pathway. Sci
993 Rep. 2016;6:30922.
994 29. Paz MF, Wei S, Cigudosa JC, Rodriguez-Perales S, Peinado MA, Huang TH, et al.
995 Genetic unmasking of epigenetically silenced tumor suppressor genes in colon cancer
996 cells deficient in DNA methyltransferases. Hum Mol Genet. 2003;12(17):2209-19.

997 30. Doucet-Beaupre H, Gilbert C, Profes MS, Chabrat A, Pacelli C, Giguere N, et al.
998 Lmx1a and Lmx1b regulate mitochondrial functions and survival of adult midbrain
999 dopaminergic neurons. Proc Natl Acad Sci U S A. 2016;113(30):E4387-96.
1000 31. Xie J, Cai TT, Li H. Sample size and power analysis for sparse signal recovery in
1001 genome-wide association studies. Biometrika. 2011;98(2):273-90.
1002 32. Forsman A. Effects of genotypic and phenotypic variation on establishment are
1003 important for conservation, invasion, and infection biology. Proc Natl Acad Sci U S A.
1004 2014;111(1):302-7.
1005 33. Goldberg DC, Fu H, Atkins D, Moyer E, Lee CN, Deng Y, et al. KnowYourCG:
1006 Facilitating base-level sparse methylome interpretation. Sci Adv. 2025;11(43):eadw3027.
1007 34. Love Stowell SM, Bentley EG, Gagne RB, Gustafson KD, Rutledge LY, Ernest HB.
1008 Optimal DNA extractions from blood on preservation paper limits conservation genomic
1009 but not conservation genetic applications. Journal for Nature Conservation. 2018;46:89-
1010 96.
1011 35. Esipova P, Suvorova I, Yachmen V, Pushchin I. Blood Morphology and Hematology
1012 of Adult Baikal Seals (Pusa sibirica Gmelin, 1788) Under Professional Care. Animals
1013 (Basel). 2025;15(2).
1014 36. Mizoguchi BA, Valenzuela N. A Cautionary Tale of Sexing by Methylation: Hybrid
1015 Bisulfite-Conversion Sequencing of Immunoprecipitated Methylated DNA in Chrysemys
1016 picta Turtles with Temperature-Dependent Sex Determination Reveals Contrasting
1017 Patterns of Somatic and Gonadal Methylation, but No Unobtrusive Sex Diagnostic.
1018 Animals (Basel). 2022;13(1).

1019