Methylation-associated mutagenesis underlies variation in the mutation spectrum across eukaryotes.

Fabián Ramos-Almodóvar[1,2], Ziyue Gao[1], Benjamin F. Voight[1,3,4,*#], Iain Mathieson[1,*#]

1. Department of Genetics, Perelman School of Medicine, University of Pennsylvania
2. Genomics and Computational Biology Graduate Group, Perelman School of Medicine, University of Pennsylvania
3. Department of Systems Pharmacology and Translational Therapeutics, Perelman School of Medicine, University of Pennsylvania
4. Institute for Translational Medicine and Therapeutics, Perelman School of Medicine, University of Pennsylvania

* These authors jointly supervised this work
# Correspondence to BFV (bvoight@upenn.edu) or IM (mathi@pennmedicine.upenn.edu)

Abstract

Mutation spectra vary across genetic and environmental contexts, leading to differences between and within species. Most research on mutation spectrum has focused on the trinucleotide (3-mer) mutation types in mammals, limiting the breadth and depth of variation surveyed. In this study, we use whole-genome resequencing data across 108 eukaryotic species – including mammals, fish, plants, and invertebrates – to characterize pentanucleotide (5-mer) non-coding mutation spectra using a Bayesian approach. Our findings reveal cytosine transition mutability at CpG and (among plants) at CHG sites as the main drivers of variation in mutation spectra across eukaryotes, correlating strongly with genomic CpG and CHG depletion. However, despite the influence of methylation on CpG mutability, genome-wide average CpG methylation levels do not predict CpG transition rates across species and CHG methylation does not predict CHG transition rate, indicating unknown genetic or environmental factors influencing mutation rates at methylated cytosines. Together, our results illustrate the pivotal role of mutagenesis in shaping genome composition across eukaryotes and highlight a gap in knowledge about the mechanisms governing mutation rates.

Introduction

Mutation is a random process, but it is not uniform. DNA sequence context influences where, what type, and how frequently mutations occur[1,2]. One example is the hypermutability of methylated cytosines at CpG dinucleotides in vertebrate genomes, resulting from spontaneous deamination of 5-methylcytosine to thymine[3,4]. Recognizing the importance of sequence context, recent work has focused on inference of context-specific mutation spectra, which quantify relative mutability based on the flanking nucleotides of a mutated base[5–9]. Most studies of mutation spectra have focused on trinucleotide (3-mer) contexts, where the mutation of interest is analyzed along with one adjacent nucleotide on each side. Consideration of the trinucleotide context has facilitated the identification of mutational signatures in cancer genomes as well as characterization of heterogeneity in polymorphism spectrum across human populations[5–7,10–13]. However, extending the analysis to longer sequence contexts (e.g., 5-mer) provides a higher-resolution view and better captures differences in mutation processes[9,14,15]. We recently developed *Baymer*[8], a Bayesian hierarchical tree approach that facilitates accurate and robust inference of mutation spectra from polymorphism data. The number of events observed in polymorphism data is much larger (~100M) relative to *de novo* (~0.1M) datasets[16], justifying the use of polymorphism data to infer the mutational spectrum, at least in non-coding regions where the effect of selection is expected to be minimal and not context-specific.

Mutation rates differ across species due to genetic and environmental factors, yet the evolutionary forces shaping context-specific mutation rates remain largely unexplored outside of humans, primates and a few vertebrates[6,17–19]. With the expansion of large-scale sequencing initiatives and conservation genomics efforts, population-level polymorphism data are now available for diverse eukaryotic taxa beyond mammals[20–22]. These resources provide an opportunity to investigate mutation spectra across a wider phylogenetic range. For example, while cytosine methylation is almost exclusive to CpG contexts in vertebrates, it is found extensively in non-CpG contexts in plants and other species where we expect to discover distinct mutational spectra. To this end, we leverage whole-genome resequencing data from 108 eukaryotic species, including mammals, birds, fish, plants, and invertebrates, to characterize variation in mutation spectra in extended pentanucleotide (5-mer) sequence context. Using *Baymer*[8], we infer context-dependent mutation rates from non-coding polymorphisms and assess the biological and evolutionary factors driving mutation spectrum variation.

Results

A catalog of polymorphisms in the non-coding genomes of 108 eukaryotic species

The availability of population polymorphism data in diverse species allows us to study the evolution of mutational mechanisms across many previously unexplored lineages[20,21]. We collected publicly available polymorphism data from whole-genome shotgun sequencing for 113 eukaryotic species[23–122], including 35 mammals, 5 birds, 15 fish, 33 plants, and 12 invertebrates (**Fig. 1A, Supplemental Table 1**). To reduce issues related to genome assembly quality, variant callability, and the effects of selection on coding regions, we masked repetitive elements, low complexity regions, and exons from each species assembly, with the resulting regions subsequently referred to as the accessible non-coding genome. This resulted in polymorphism datasets with 10^5 – 10^9 SNPs per species (**Fig. 1B**).

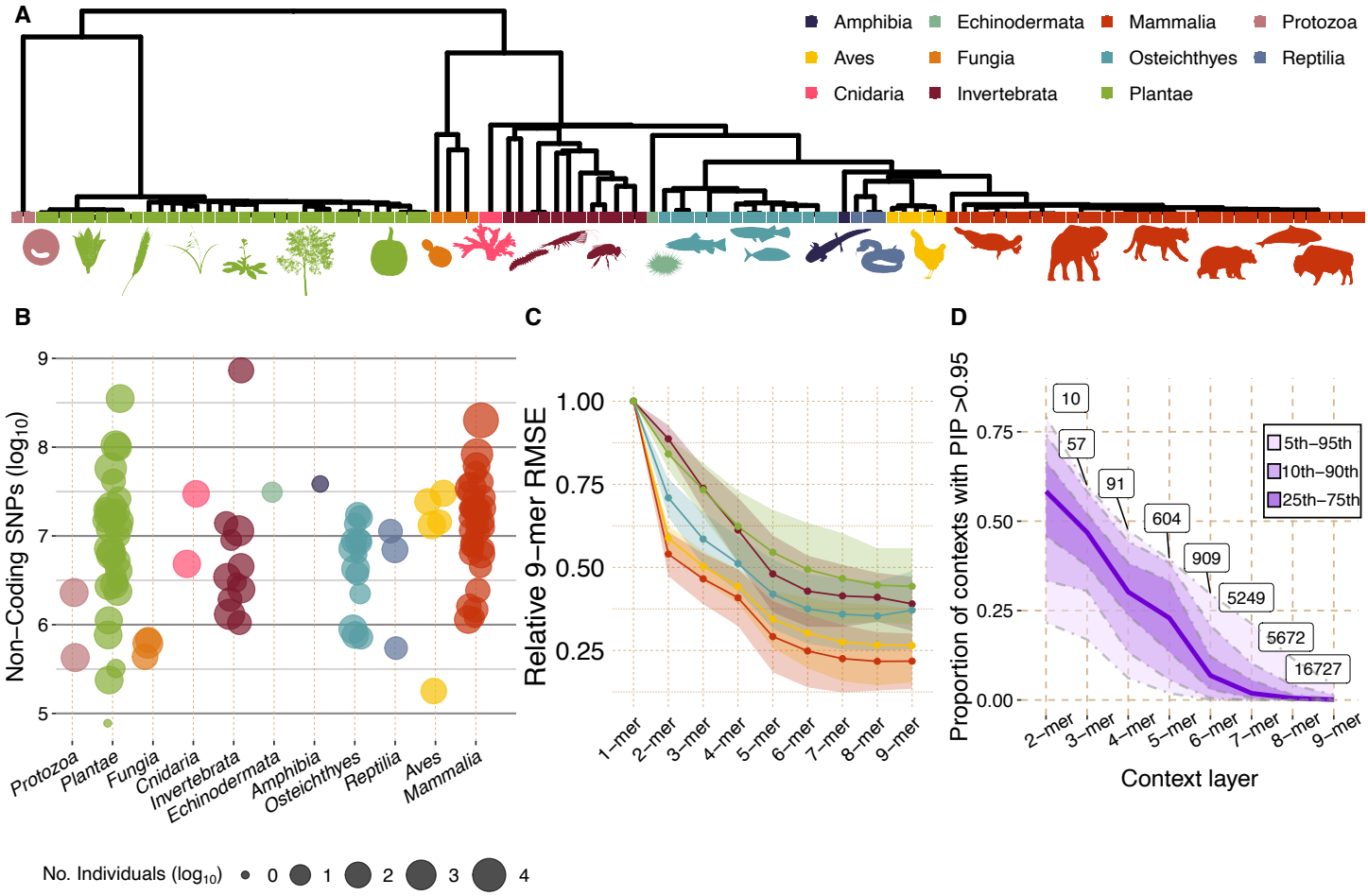

Fig. 1 | Data summary. **A**) Phylogenetic tree of collected species, generated with TimeTree, with assigned clades labeled at tips. **B**) Filtered single-nucleotide polymorphism (SNP) count ($\log_{10}$) per species is shown as individual points, colored by clade and sized by number of individuals ($\log_{10}$) in polymorphism dataset. **C**) Root mean square error (RMSE) between a k-mer mutation model trained on odd-numbered basepairs versus a 9-mer mutation model trained on even-numbered basepairs. Points show mean RMSE across species grouped by clade (only clades with n >=5 are shown). Shaded areas show one standard deviation from the mean. **D**) Proportions of contexts with posterior inclusion probability (PIP) greater than 0.95 in each context layer, for every species. Dark line represents the median proportion across species; shaded areas represent labeled quantiles. Number of contexts in 95th percentile is labeled for each layer.

Baymer captures variation in mutation spectra across species

We used *Baymer*[8] to infer sequence context windows (up to 4 flanking nucleotides, i.e., '9-mers') separately in each of the 113 species. *Baymer* infers the probability of polymorphism–an approximation to the relative mutation rate of each mutation type–by fitting a hierarchical Bayesian model to the number of observed polymorphisms and the corresponding frequency of the mutable context. A spike-and-slab prior allows contexts to have zero effect on mutation rate, preventing overfitting and allowing the model to adapt to species with very different numbers of observations. We use the default prior and MCMC settings as described in Adams, et al. (2023).

To evaluate the performance of our mutation models within species, we computed the root mean squared error (RMSE) between the mutation models generated from odd- and even-basepairs of the accessible non-coding regions (**Fig. S1**). A higher RMSE suggests higher uncertainty in estimated rates due to technical noise or data sparsity. There is a large drop in RMSE between 1-mer and 2-mer models for mammals and birds due to the inclusion of CpG contexts. For all species, RMSE decreases for larger contexts, but the decrease is substantially slower above 5-mer contexts (**Fig. 1C**). This reflects the fact that the average posterior inclusion probabilities (PIP)–the proportion of contexts with a non-zero effect–drops from 25% at the 5-mer level to 2% at the 7-mer level (**Fig. 1D**). We noted five species with near zero PIPs for every mutation context across every context layer (**Fig. S3**)–a pattern expected in random noise or sparse data that are not informative of mutational processes. We therefore excluded these five species from further analysis and focused on 5-mer contexts for downstream analyses in the 108 remaining species.

Cytosine transitions at methylation target sites explain most of the variation in mutation spectra

To investigate the variation of context-specific mutation rates in eukaryotes, we first performed principal component analysis (PCA) on the unscaled fitted 5-mer context-specific mutation spectra (**Fig. 2A**), noting that PCA applied to 7-mer (but not 3-mer) contexts produced qualitatively similar results (**Fig. S4**). We found that the first two principal components cluster mutation spectra by clade and capture 84% (PC1: 79.5%, PC2: 4.56%) of the variance in 5-mer mutation spectra across eukaryotes (**Fig. S5A**). The first and second principal components are characterized by variation in NNCGN > NNTGN and NNCHG > NNTHG mutation rates, respectively (where N is any nucleotide and H is any nucleotide other than guanine; **Fig. S5B**). We ruled out saturation of CpG sites due to sequencing of large sample sizes as the source of the captured variation by computing the proportion of CpG sites that are polymorphic for each species (**Fig. S6**).

While the PCA on unscaled mutation spectra reveals variation in absolute contributions to the mutation spectra, it can obscure relative differences due to the disproportionate contribution CpG transitions to the PCA loadings. To assess relative differences in mutation spectra, we performed PCA on scaled 5-mer mutation spectra, where the rates of each mutation type were standardized to have unit variance across species. When examining this relative variation, we found that the first two principal components capture 53.72% (PC1: 43.11%, PC2: 10.61%) of the variance in 5-mer mutation spectra across eukaryotes, and the species still cluster by clade and form a similar phylogenetic cline (**Fig. 2B**). The PC1 of scaled mutation spectra is characterized by variation in NNCGN > NNTGN mutation rate versus other mutation types and PC2 by variation in non-CpG transitions versus other mutation types (**Fig. 2C**). Mammals, birds, and reptiles have the highest weights in PC1, consistent with higher relative CpG mutation rates. The correlation between PCA results and context-specificity of cytosine methylation indicates that most of the variation in context-specific mutation rates across eukaryotes is driven by variation in the rate of cytosine transitions at CpG sites. These results underscore transitions at cytosine methylation target sites as the main driver of context-specific mutation rate variation across eukaryotes.

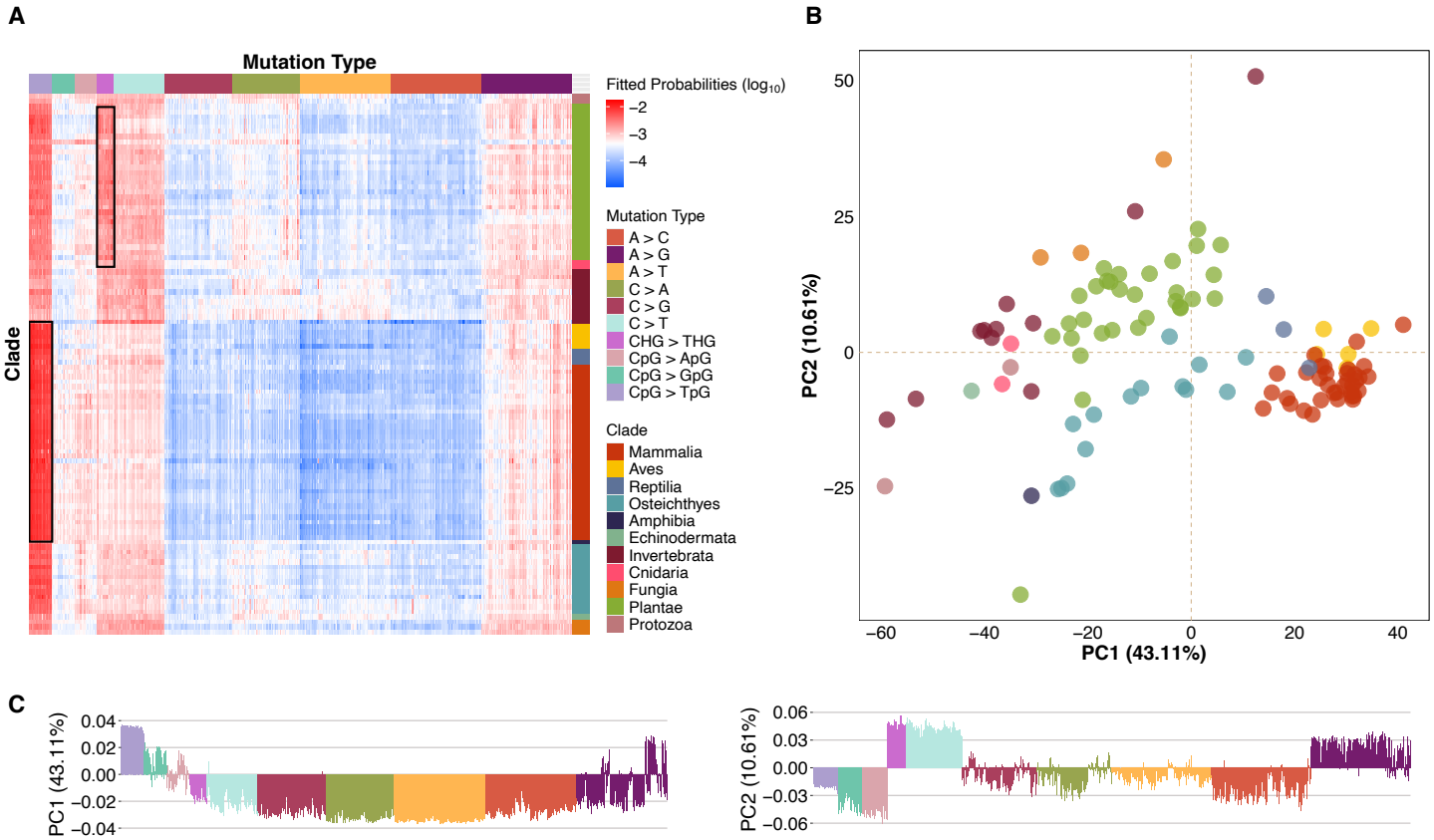

Fig. 2 | Principal component analysis (PCA) of scaled fitted 5-mer polymorphism probabilities. **A**) Heatmap of raw 5-mer context fitted probabilities; CpG > T and CHG > THG mutations are higlighted in boxes for tetrapods and plants, respectively. **B**) First two principal components from PCA on scaled polymorphism probabilities. Points show each species, colored by clade. **C**) First two principal component loadings, showing individual bars for each mutation type, colored by mutation type.

CpG transition rates predict genomic CpG depletion but are not predicted by methylation levels

We used the ratio of CpG>T and CpH>T polymorphism probabilities to measure the change in mutation rate due to the CpG context and refer to this as the CpG>T mutation rate ratio. Consistent with the mutagenic effect of methylation, this ratio is largest (5- to 20-fold) in vertebrates, which have high levels of CpG methylation, lower (1.5 to 4.5-fold) in plants and less than 1.5-fold in species without substantial methylation (**Fig. 3A**).

However, a closer analysis based on genome-wide CpG methylation levels derived from whole-genome bisulfite sequencing[24,123–133] reveals a more complicated pattern. Phylogenetic linear regression of CpG>T mutation rate ratio on average CpG methylation is not significant (P=0.21; **Fig. 3B**). This lack of relationship is clearly visible among vertebrates, where fish have high methylation levels but low mutation rate ratios, and birds vice versa. We note that the methylation data we collected is derived from somatic tissues. For seven vertebrates where germline methylation data were available[134–139], we find a Spearman correlation of 0.595 between somatic and germline methylation (**Fig. S7A**) and a relationship that is consistent with the somatic mutation data (i.e. higher methylation but lower CpG mutation in fish relative to mammals; **Fig. S7B**). Thus, presence of CpG methylation entails a higher CpG>T mutation rate ratios, but genome average CpG methylation levels do not predict the magnitude of the ratio. This suggests that some other genetic or environmental factors shape CpG mutation rates across eukaryotes by modifying deamination, repair, or replication-driven error rates at methylated cytosine.

In addition to CpG methylation, many plants have methylation in CHG and CHH contexts. We therefore corrected our analysis of CpG mutability by regressing the CpG>T / CHH>T mutation rate ratio on the difference between CpG and CHH genome-wide average methylation levels in plants (**Fig. 3C**) and observed consistent results (phylogenetic regression P = 0.97).

We next evaluated the effect of the CpG>T mutation rate ratio on CpG content of the accessible non-coding genome, as measured by the ratio of observed CpG fraction among all dinucleotides and the expected CpG fraction based on GC content (CpG O/E ratio). We found that the CpG>T mutation rate ratio strongly

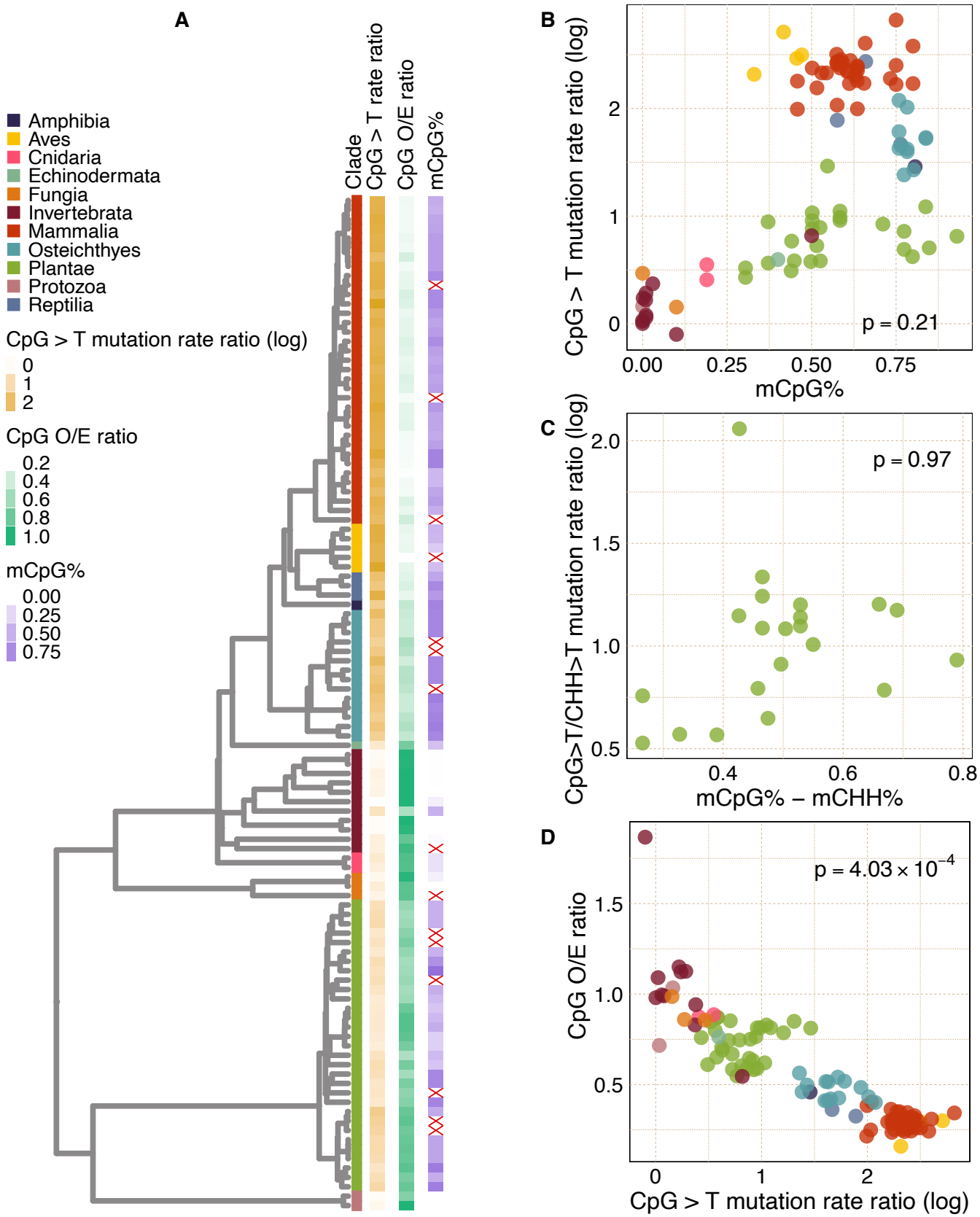

Fig. 3 | CpG methylation level, mutation rate, and genome composition. **A**) Phylogenetic tree with CpG-related data shown at tips (red crosses denote missing values). **B**) CpG > T mutation rate ratio vs genome-wide average CpG methylation. **C**) Same as B) but restricted to plants and removing the effect of CHG sites on CpG mutation rate ratio and methylation level. **D**) Observed/Expected CpG non-coding genome composition vs CpG > T mutation rate ratio. **B-D**) P-values correspond to phylogenetic regression results.

predicts the magnitude of CpG depletion (phylogenetic regression P = 4.03×10^{-4}) (**Fig. 3D**), with no depletion in species without elevated CpG>T mutation rate ratio; within-clade analysis is reported in **Table S1**. This aligns with the expectation of lower equilibrium genome representation of highly mutable contexts and suggests that CpG content in the non-coding genome is largely determined by the mutation spectrum with minimal influence of selective constraint on CpG content. The one outlier is the honeybee *Apis mellifera*, which has previously been noted to have an exceptionally high genomic CpG content[140]. To provide additional evidence, we examined changes along terminal branches, which offer a cleaner test of this relationship by focusing on recent evolutionary changes. We inferred the CpG O/E and relative mutation rate at the internal nodes of the tree using fastAnc in R and quantified the changes along terminal branches (**Fig S8**). We observed a significant overall negative correlation between shifts in CpG O/E ratio and relative mutation rate, consistent with the expectation that species experiencing recent increases in CpG mutation rates are moving towards lower CpG composition (and vice versa). Although this correlation reaches significance in vertebrates and invertebrates, driven predominantly by honeybee and krill, the negative trend is consistent across other clades (see results in **Table S2**).

In plants, CHG transition rates predict CHG depletion but are not predicted by methylation levels

To test the relationship between CHG methylation, mutation rate, and genome composition in plants, we extracted genome-wide average CHG methylation levels from whole-genome bisulfite sequencing[123] and computed the CHG>T / CHH>T mutation rate ratio and the CHG observed-to-expected (O/E) ratio for each species. Analogous to the findings for CpG methylation across eukaryotes, we found that the CHG>T mutation rate ratio is not predicted by CHG methylation (phylogenetic regression P = 0.57, **Fig. 4B**) but does predict the CHG O/E ratio (phylogenetic regression P = 6.75×10^{-7}, **Fig. S9**). To investigate whether the same or different factors underlie variation in mutability of CpG and CHG contexts, we regressed our species' mutation rate ratios on their methylation levels separately for CpG and CHG contexts and found that the residuals were highly correlated (Pearson's coefficient = 0.829; p = 3.3×10^{-6}) (**Fig. 4C**). This suggest that the same factors underlying the unexplained variation in CpG transition rates also underlie variation at CHG sites.

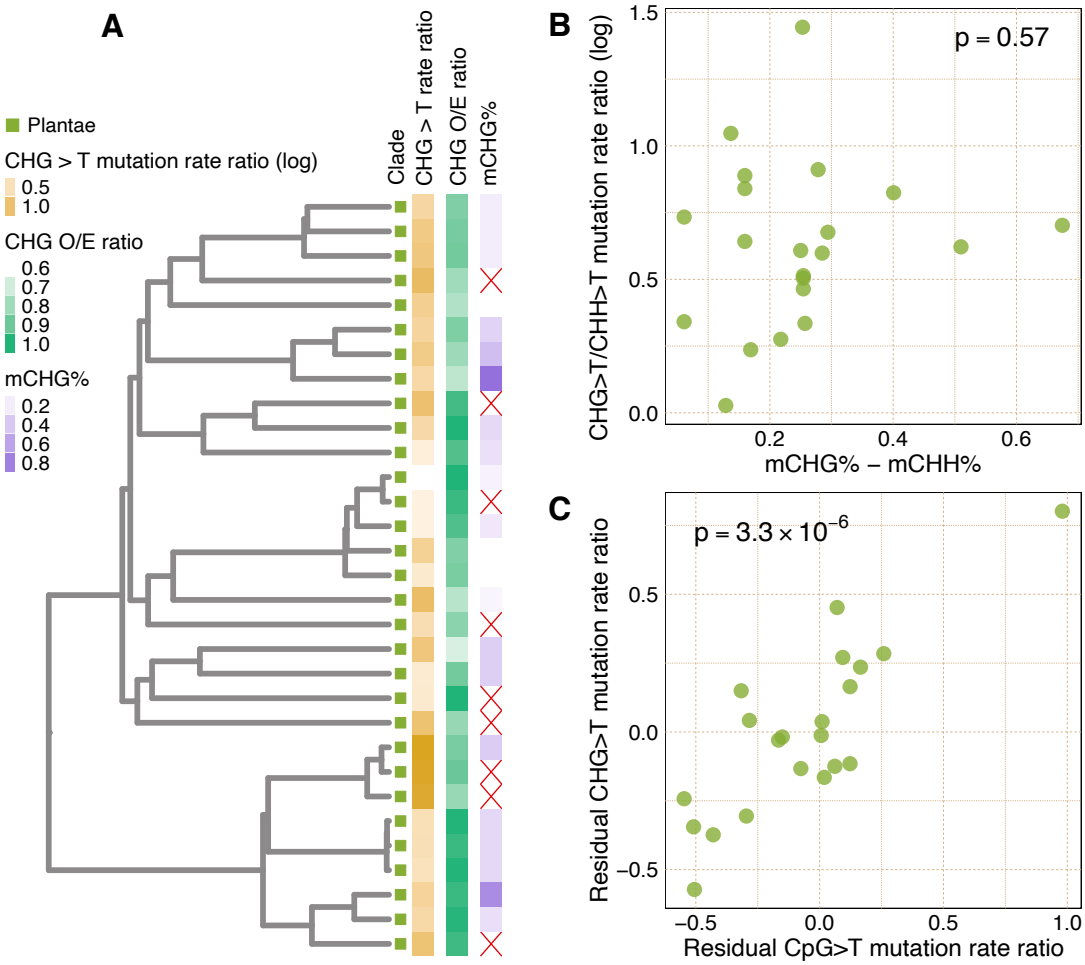

Fig. 4 | Plant CHG methylation and mutation rate. **A**) Phylogenetic tree with CHG data shown on tips (red crosses denote missing values). **B**) CHG > T mutation rate ratio vs genome-wide average CHG methylation. **C**) Plant CpG and CHG mutation rate ratio residuals from phylogenetic regressions of CpG>T mutation rate ratio on CpG methylation versus those from CHG>T mutation rate ratio regressed on CHG methylation. Plant species are shown as individual points (Pearson's r = 0.829).

Mutation signature analysis confirms transitions at methylation target sites as the main driver of mutation spectrum variation in eukaryotes

To corroborate our findings from Baymer-inferred context-specific mutation rates, we used SigFit[141] to model the observed 5-mer polymorphism spectrum in each species as a linear combination of *k* mutational signatures weighted by differential exposures in each species. Following Beichman et al.[142], we computed the cosine similarity between the observed and reconstructed mutation spectrum vectors to evaluate model fit to the data. After analyzing the reconstruction performance for *k*=2 to 10, we determined that four mutational signatures had optimal performance and interpretability (**Fig. S10**). These four signatures reflect a CpG transition signature, a CpG and CHG transition signature, and two background signatures (**Fig. 5A**). Consistent with our previous observations, the CpG transition signature was most active in mammals, birds, and reptiles while the signature of CpG and CHG methylation was most active but variable in plants (**Fig. 5B**); the two background signatures are present in all clades.

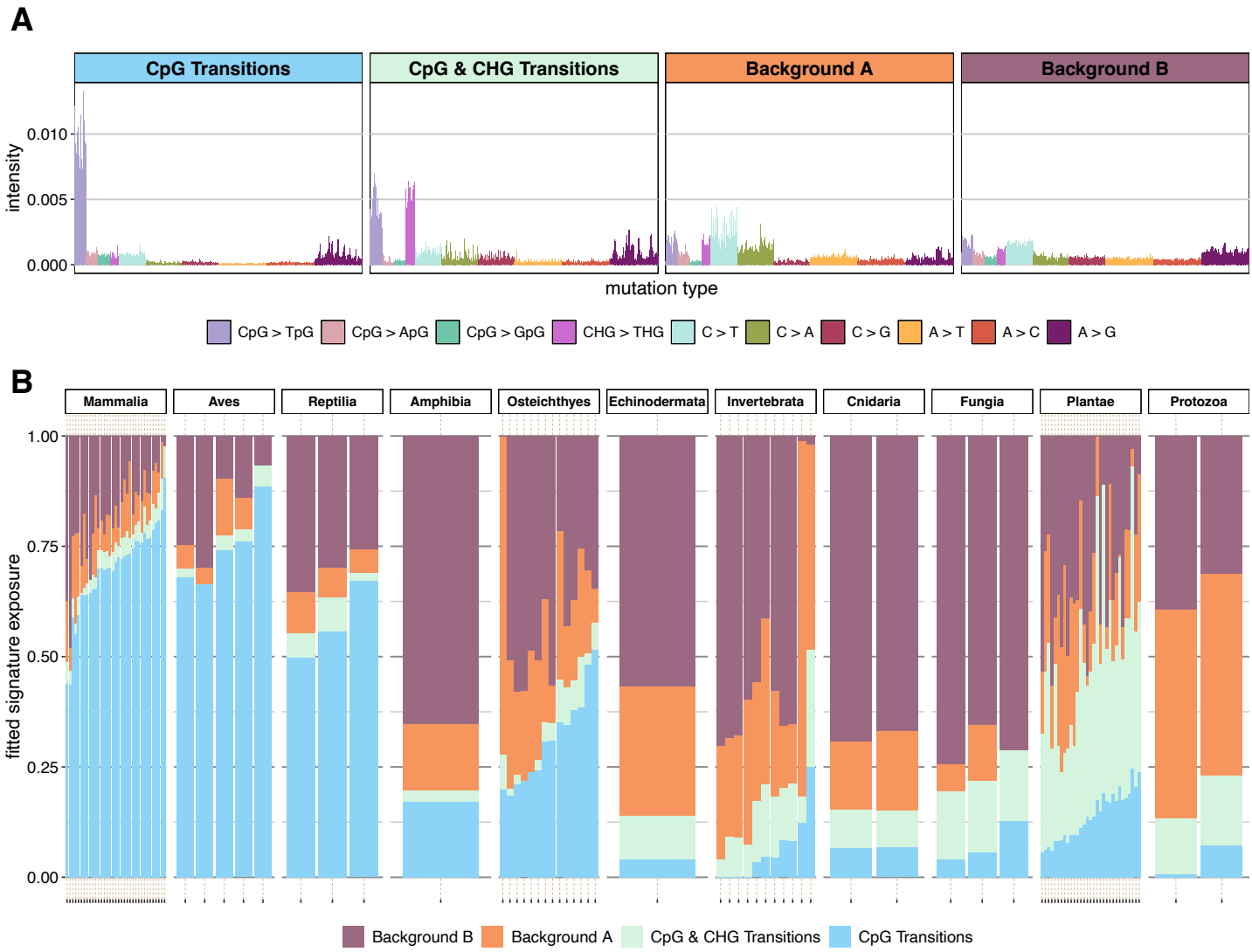

Fig. 5 | Mutational signature analysis in 5-mer polymorphism spectra. **A**) Extracted signature mutation spectra. **B**) Estimated contributions from each inferred signature to each species' polymorphism counts.

Discussion

Our study provides a comprehensive analysis of context-specific mutation spectra across 108 eukaryotic species, leveraging a phylogenetically diverse collection of polymorphism data from whole-genome sequencing. By extending mutation spectrum analysis to 5-mer contexts and looking across a wider range of species, we discover a new axis of mutation spectrum variation, corresponding to transitions at CHG contexts, in addition to the known role of CpG transitions. This is particularly important in plants, where methylation in CHG contexts is common and CHG > T mutations explain a large proportion of mutation spectrum variation.

While our findings reinforce the predominant role of transitions at methylation target sites on mutation spectrum variation across eukaryotic species, they also reveal that methylation is not the only factor at play, as genome-wide average methylation levels do not predict mutation rates at these sites across

species after correcting for phylogenetic non-independence. Therefore, while methylation is necessary for high transition rates at CpG and CHG sites, the variation in mutation rates is strongly modulated by other unknown genetic and environmental factors that differ across species and modify rates of deamination, replication error, or repair. One limitation of our current analysis is that it is based on somatic methylation data which might imperfectly reflect methylation patterns in the germline. Consequently, a poor correlation between methylation levels and mutation rates may be observed even if the former strongly determines the latter in the germline. If that is the case, our observations could reflect uncoupling between somatic and germline methylation levels across species. The relationship between mutation and methylation should be reanalyzed when germline methylation data are available for a broader range of eukaryotic species. In contrast, we found that genomic CpG and CHG depletion is highly correlated with cytosine transition rates at CpG and CHG sites, suggesting that these aspects of genomic composition are largely determined by mutational pressures rather than selection, biased gene conversion, or other forces.

In summary, this study advances our understanding of mutation spectrum variation across eukaryotes by integrating 5-mer context mutation spectrum models over a wide range of species. We show that polymorphism data serve as a reliable proxy for mutation data, although absolute mutation rate cannot be assessed and fine-scale variation may be missed due to data artefacts or the lack of polarization of ancestral alleles. Future work should aim to identify genetic or environmental correlates of CpG and CHG transition rates after accounting for methylation, investigate variation across genomic compartments, and sample more broadly from clades with extreme mutational spectra or genome compositions.

Methods

Data collection

Polymorphism data: We obtained species polymorphism datasets, in the form of .vcf files, from public repositories or directly from authors. We carried out this process by searching through the literature and limiting our selected species to those with polymorphisms called from whole-genome sequencing data, with at least 2x coverage and five individuals (excluding humans, sample size: median=81; mean=259). We also required that the reference genomes used for variant calling had coding region annotations. The full list of species, along with data sources, can be found in **Supplemental Table 1**.

For *Ornithorhynchus anatinus* (Platypus), we aligned 49 individual genomes (in FASTQ format) and called variants using DeepVariant[143] with its standard protocol (**Supplemental Data**).

Reference genome assembly: Corresponding reference assemblies for each species were downloaded from NCBI or publication repository, along with coding and RepeatMasker annotations. **Supplemental Table 1** contains list of sources for polymorphism data and reference genome assembly ID.

Data preparation

Non-coding genome: For each species, we limited the analysis to the accessible non-coding genome by masking out exons, repetitive elements, and low complexity regions. For coding sequences, we extracted genome coordinates (in the form of .bed files) for exons from coding sequence annotation files (downloaded from NCBI or publication repository in the form of .gtf or .gff files) generated from RNA sequencing, for the vast majority of our species (**Supplementary Table 1)**. Repetitive element coordinates were extracted from RepeatMasker annotations. When RepeatMasker annotation was not available, we used *RepeatMasker-4.1.5* with default parameters and specified clade repeat libraries to generate repetitive element annotations. Low-complexity DNA regions were masked using the NCBI tool, *DustMasker-1.0.0* (with -window = 32; default is 64). Our final genomic region for analysis was defined using *bedtools2-2.30.0 getfasta* function by excluding genome coordinates that overlapped exonic, repetitive, and low-complexity coordinates from the above.

Polymorphism data: We retained only single-nucleotide polymorphisms (SNPs) from the variant call format (vcf) files with complete 9-mer contexts in the accessible non-coding genome (defined above). Multiallelic sites were treated as independent mutations at the same site. Since the ancestral genome assembly is unavailable for most species, we used reference alleles for polarizing the polymorphisms, assuming the reference allele is the ancestral allele and non-reference allele the mutated allele. Reverse complementary mutation types were combined. To mitigate concerns of high error rate in singletons, we restricted our analysis to SNPs present in at least two individuals, when individual genotypes were available; otherwise, we removed singletons based on non-reference allele count. Because our polymorphism data derive from studies with different sequencing and variant calling strategies, we checked that the PCs of the mutation spectrum were uncorrelated with technical features including SNP count and reference genome N50 (**Fig. S8**).

Calculation of CpG and CHG observed/expected (O/E) ratio

The observed-over-expected CpG content ratio ($CpG_{O/E}$) was calculated as:

$$\boldsymbol{CpG_{O/E}} = \frac{[\boldsymbol{CpG}]_f}{[\boldsymbol{C}]_f[\boldsymbol{G}]_f}, \quad (1)$$

where the observed CpG frequency, $[\boldsymbol{CpG}]_f$, was calculated as the fraction of dinucleotide counts that are CpGs and the expected CpG frequency the product of the observed cytosine and guanine single-nucleotide frequencies, $[\boldsymbol{C}]_f$ and $[\boldsymbol{G}]_f$, in the accessible non-coding genome.

The observed-over-expected CHG content ratio ($CHG_{O/E}$) was calculated similarly as:

$$\boldsymbol{CHG_{O/E}} = \frac{[\boldsymbol{CHG}]_f}{[\boldsymbol{C}]_f[\boldsymbol{H}]_f[\boldsymbol{G}]_f}, \quad (2)$$

where the observed CHG (where H is any nucleotide other than guanine) frequency, $[\boldsymbol{CHG}]_f$, was calculated as the fraction of trinucleotide counts that are CHGs and the expected frequency the product of the observed frequencies of cytosine, non-guanine, and guanine single-nucleotide, $[\boldsymbol{C}]_f$, $[\boldsymbol{H}]_f$,and $[\boldsymbol{G}]_f$, in the accessible non-coding genome.

Methylation data collection

Cytosine methylation data: We collected averaged genome-wide cytosine methylation levels at CpG sites in each species and at CHG sites in plants from multiple sources (each species' methylation state and source is indicated in **Supplemental Table 2**) from whole-genome bisulfite sequencing (WGBS) studies[123,124]. Additionally, we collected averaged genome-wide cytosine methylation levels at CpG sites from testes WGBS for seven vertebrate species[134–139] (listed in **Supplemental Table 2**).

Context-specific mutation probability models using Baymer

To construct our mutation probability models, we used Baymer[8], a Bayesian hierarchical tree model approach that estimates context-specific mutation probabilities for increasing context layers, iteratively. Baymer generates regularized mutation probability estimates and provides a measure of uncertainty for each multiplicative shift in mutation probability leading up to the final k-mer mutation context. The context-specific mutation rates for any given context window are calculated as the product of the multiplicative shifts leading to the final mutation context. To generate context-specific mutation probability models from polymorphism data, Baymer requires a table with DNA context counts and polymorphic sites for the largest context size desired for the models (in our case, 9-mers). For each species, we first generated a table of 9-mer context counts within the accessible non-coding genome (defined above). Then, we identified and

counted polymorphic sites with complete 9-mer contexts in the same regions. Lower context size counts (e.g. 5-mers) were derived from the 9-mer context size counts.

Assessing robustness of the mutation model

We applied Baymer to these inputs to generate three models, each using one of three different partitions of the non-coding genome, which we refer to as ALL, ODD, and EVEN. The ALL model consisted of every context and polymorphism in the accessible non-coding genome, while the ODD and EVEN models were only generated using odd or even genome coordinates in the accessible non-coding genome, respectively. The ODD and EVEN models were generated for cross-validation of model performance in each species. We assessed within-species model robustness by computing the root mean square error (RMSE) between the ODD and EVEN polymorphism probability estimates; calculated as:

$$\boldsymbol{RMSE_{Odd/Even}} = \frac{\mathbf{1}}{\boldsymbol{K}}\sqrt{\sum_{\boldsymbol{i}=\mathbf{1}}^{\boldsymbol{K}}(\boldsymbol{log}(\boldsymbol{O_i}) - \boldsymbol{log}(\boldsymbol{E_i}))^{\mathbf{2}}} \tag{3}$$

Where **O** and **E** are the fitted context-specific rate vectors, of length K, from the ODD and EVEN models, respectively. **Fig. S1** shows the $RMSE_{Odd/Even}$ for the 5-mer mutation models. Fig. 1C shows the RMSE between 9-mer EVEN mutation models versus k-mer (k = 1 to 9) ODD mutation models, normalized by the k=1 RMSE.

Mutation rate ratios

Our mutation models infer content-specific relative mutation rates rather than absolute mutation rates. For this reason, we base our mutability analyses on relative differences in context-specific mutation rates, which we refer to as 'mutation rate ratios'. We estimated these ratios using 5-mer context-specific mutation rates.

CpG>T mutation rate ratio was defined for each species as the ratio of the average C>T mutation rate in all 5-mer CpG contexts over that of CpH contexts (H = A,C,T):

$$\boldsymbol{CpG > T\ Mutation\ Rate\ Ratio} = \frac{\frac{\mathbf{1}}{\boldsymbol{K}}\sum_{\boldsymbol{i}=\mathbf{1}}^{\boldsymbol{K}}[\boldsymbol{NNCGN}>\boldsymbol{NNTGN}]}{\frac{\mathbf{1}}{\boldsymbol{N}}\sum_{\boldsymbol{j}=\mathbf{1}}^{\boldsymbol{N}}[\boldsymbol{NNCHN}>\boldsymbol{NNTHN}]}, \tag{4}$$

where the summation is across all 5-mer contexts matching the specified patterns.

We note that plant species also experience extensive cytosine methylation at non-CpG contexts. Because of this, the CpG>T mutation rate ratio above may be a biased representation of the relative context-specific CpG>T mutation rate in these plants. To account for this, we defined the CpG>T / CHH>T mutation rate ratio for each plant species as the ratio of the average C>T mutation rate in CpG contexts over that of CHH contexts, removing impacts of methylation in CHG contexts:

$$\frac{CpG>T}{CHH>T}\ Mutation\ Rate\ Ratio = \frac{\frac{1}{K}\sum_{i=1}^{K}[NNCGN>NNTGN]}{\frac{1}{N}\sum_{j=1}^{N}[NNCHH>NNTHH]} \quad (5)$$

To analyze the relationship between CHG transition rates, methylation, and genome composition in plants, we defined the CHG>T / CHH>T mutation rate ratio for each plant species as the ratio of the average C>T mutation rate in 5-mer CHG contexts over that of CHH contexts:

$$\frac{CHG>T}{CHH>T}\ Mutation\ Rate\ Ratio = \frac{\frac{1}{K}\sum_{i=1}^{K}[NNCHG>NNTHG]}{\frac{1}{N}\sum_{j=1}^{N}[NNCHH>NNTHH]} \quad (6)$$

Phylogenetic tree

Our phylogenetic tree (in Newick file format) was constructed using estimated divergence times from TimeTree (timetree.org). **Supplemental Table 1** contains the list of species names used for TimeTree.

Principal component analysis (PCA) on raw and scaled fitted mutation probabilities

We performed principal component analysis (PCA) on the raw 3-mer, 5-mer, and 7-mer context size fitted mutation probability models, separately. We scaled the fitted mutation probabilities for each species so that they added up to one and used the prcomp(scale = FALSE, center = TRUE) R (v4.2) function to perform PCA based on single-value decomposition for supplementary figures. For the main text in **Fig. 2**, we performed PCA on scaled 5-mer context mutation probabilities, by standardizing each mutation type to have mean = 0 and standard deviation = 1 across species before PCA; prcomp(scale = TRUE, center = TRUE).

Phylogenetic linear regressions

To perform analysis accounting for the phylogenic tree structure given the species included in our study, we used the *phylolm-2.6.2* R package phylolm()[144] function with Pagel's lambda as a phylogenetic scaling factor and 1000 bootstraps for our regression models. The phylolm() function is an implementation of the phylogenetic generalized least squares regression model:

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{V}(\boldsymbol{\lambda})^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{V}(\boldsymbol{\lambda})^{-1}\boldsymbol{y} \quad (7)$$

where for n species and k independent variables, $\mathbf{y}$ ($n \times 1$) is the continuous outcome vector, $\boldsymbol{\beta}$ ($n \times 1$) is the vector of coefficients for the variable matrix $\mathbf{X}$ ($n \times k$), and $\mathbf{V}$ ($n \times n$) is the variance-covariance matrix based on the phylogenetic relatedness between the species. Pagel's lambda is then used to transform $\mathbf{V}$ into a

new matrix, **V(λ)** ($n \times n$), by scaling the off-diagonal elements of **V** using the scalar **λ**, which is an estimate of phylogenetic signal for the regressed trait. **λ** ranges between 0 and 1, where 1 indicates that the covariance structure is exactly given by the phylogenetic relatedness structure, and 0 indicates no phylogenetic signal for the regressed trait. We report two-tailed P-values.

Mutational signature analysis with SigFit

We extracted mutational signatures from the observed 5-mer polymorphism spectra using a non-negative matrix factorization approach, implemented in the *SigFit-2.2*[141] R package. SigFit uses a Bayesian multinomial model to fit specified and/or inferred mutational signatures to observed mutation spectra, an approach equivalent to non-negative matrix factorization. For **G** genomes and **M** mutation contexts, SigFit takes as input a mutation counts matrix (**G** by **M**) and a mutation opportunities matrix (**G** by **M**); the mutation opportunities matrix contains the number of mutable contexts for each given 5-mer mutation type in the species' genome. For our SigFit analysis, we used similar mutation and context counts as inputs for Baymer inference. We used the extract_signatures() function from SigFit, with iter = 30000 and nsignatures = k (for k = 2,3,...,10). For each SigFit model, we reconstructed the mutation spectra using the fitted mutational signatures and their estimated activity values in each species. We then computed the cosine similarity between the observed and reconstructed 5-mer mutation spectra vectors to measure each model's performance in each species.

Data and code availability

Code to reproduce the analysis in this paper is available at https://github.com/framos99/MutationSpectraEvolutionAcrossEukaryotes-Project. Fitted mutation spectra for each species are available at https://doi.org/10.5281/zenodo.15464759. Original sources for data are given in Supplementary Table 1.

Acknowledgements

F.RA. is grateful for support of the work from the University of Pennsylvania's Presidential PhD fellowship. This work was supported by NIGMS R35GM133708 (I.M.), R35GM146810 (Z.G.), NIEHS P30ES013508 (B.F.V.), and a Research Fellowship (FG-2021-15702) from the Alfred P. Sloan Foundation (Z.G). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

References

1. Hwang, D. G. & Green, P. Bayesian Markov chain Monte Carlo sequence analysis reveals varying neutral substitution patterns in mammalian evolution. *Proceedings of the National Academy of Sciences* **101**, 13994–14001 (2004).
2. Hodgkinson, A. & Eyre-Walker, A. Variation in the mutation rate across mammalian genomes. *Nat Rev Genet* **12**, 756–766 (2011).
3. Bird, A. P. DNA methylation and the frequency of CpG in animal DNA. *Nucleic Acids Research* **8**, 1499–1504 (1980).
4. Fryxell, K. J. & Zuckerkandl, E. Cytosine Deamination Plays a Primary Role in the Evolution of Mammalian Isochores. *Molecular Biology and Evolution* **17**, 1371–1383 (2000).
5. Harris, K. Evidence for recent, population-specific evolution of the human mutation rate. *Proceedings of the National Academy of Sciences* **112**, 3439–3444 (2015).
6. Harris, K. & Pritchard, J. K. Rapid evolution of the human mutation spectrum. *eLife* **6**, e24284 (2017).
7. Mathieson, I. & Reich, D. Differences in the rare variant spectrum among human populations. *PLOS Genetics* **13**, e1006581 (2017).
8. Adams, C. J. *et al*. Regularized sequence-context mutational trees capture variation in mutation rates across the human genome. *PLOS Genetics* **19**, e1010807 (2023).
9. Aggarwala, V. & Voight, B. F. An expanded sequence context model broadly explains variability in polymorphism levels across the human genome. *Nat Genet* **48**, 349–355 (2016).
10. Alexandrov, L. B. *et al*. Signatures of mutational processes in human cancer. *Nature* **500**, 415–421 (2013).
11. Alexandrov, L. B., Nik-Zainal, S., Wedge, D. C., Campbell, P. J. & Stratton, M. R. Deciphering Signatures of Mutational Processes Operative in Human Cancer. *Cell Reports* **3**, 246–259 (2013).
12. Alexandrov, L. B. *et al*. The repertoire of mutational signatures in human cancer. *Nature* **578**, 94–101 (2020).

13. Tate, J. G. *et al.* COSMIC: the Catalogue Of Somatic Mutations In Cancer. *Nucleic Acids Research* **47**, D941–D947 (2019).

14. Carlson, J. *et al.* Extremely rare variants reveal patterns of germline mutation rate heterogeneity in humans. *Nat Commun* **9**, 3753 (2018).

15. Seplyarskiy, V. *et al.* A mutation rate model at the basepair resolution identifies the mutagenic effect of polymerase III transcription. *Nat Genet* **55**, 2235–2242 (2023).

16. Bergeron, L. A. *et al.* The Mutationathon highlights the importance of reaching standardization in estimates of pedigree-based germline mutation rates. *eLife* **11**, e73577 (2022).

17. Beichman, A. C. *et al.* Evolution of the Mutation Spectrum Across a Mammalian Phylogeny. *Molecular Biology and Evolution* **40**, msad213 (2023).

18. Dumont, B. L. Significant Strain Variation in the Mutation Spectra of Inbred Laboratory Mice. *Molecular Biology and Evolution* **36**, 865–874 (2019).

19. Goldberg, M. E. & Harris, K. Mutational Signatures of Replication Timing and Epigenetic Modification Persist through the Global Divergence of Mutation Spectra across the Great Ape Phylogeny. *Genome Biology and Evolution* **14**, evab104 (2022).

20. Christmas, M. J. *et al.* Evolutionary constraint and innovation across hundreds of placental mammals. *Science* **380**, eabn3943 (2023).

21. Kuderna, L. F. K. *et al.* A global catalog of whole-genome diversity from 233 primate species. *Science* **380**, 906–913 (2023).

22. Lewin, H. A. *et al.* Earth BioGenome Project: Sequencing life for the future of life. *Proceedings of the National Academy of Sciences* **115**, 4325–4333 (2018).

23. Meadows, J. R. S. *et al.* Genome sequencing of 2000 canids by the Dog10K consortium advances the understanding of demography, genome function and architecture. *Genome Biology* **24**, 187 (2023).

24. Unneberg, P. *et al.* Ecological genomics in the Northern krill uncovers loci for local adaptation across ocean basins. *Nat Commun* **15**, 6297 (2024).

25. Cooke, I. *et al.* Genomic signatures in the coral holobiont reveal host adaptations driven by Holocene climate change and reef specific symbionts. *Science Advances* **6**, eabc6318 (2020).

26. Fuller, Z. L. *et al.* Population genetics of the coral Acropora millepora: Toward genomic prediction of bleaching. *Science* **369**, eaba4674 (2020).

27. 1,135 Genomes Reveal the Global Pattern of Polymorphism in Arabidopsis thaliana. *Cell* **166**, 481–491 (2016).

28. Rogivue, A. *et al.* Genome-wide variation in nucleotides and retrotransposons in alpine populations of Arabis alpina (Brassicaceae). *Molecular Ecology Resources* **19**, 773–787 (2019).

29. Parejo, M. *et al.* AmelHap: Leveraging drone whole-genome sequence data to create a honey bee HapMap. *Sci Data* **10**, 198 (2023).

30. Ouyang, J. *et al.* Chromosome-level genome and population genomics reveal evolutionary characteristics and conservation status of Chinese indigenous geese. *Commun Biol* **5**, 1–12 (2022).

31. Love, R. R. *et al.* Chromosomal inversions and ecotypic differentiation in Anopheles gambiae: the perspective from whole-genome sequencing. *Molecular Ecology* **25**, 5889–5906 (2016).

32. Lee, D. *et al.* Population analysis of the Korean native duck using whole-genome sequencing data. *BMC Genomics* **21**, 554 (2020).

33. Hilali, S. E. *et al.* Chromosome-scale genome assembly and gene annotation of the hydrothermal vent annelid Alvinella pompejana yield insight into animal evolution in extreme environments. 2024.06.25.600561 Preprint at https://doi.org/10.1101/2024.06.25.600561 (2024).

34. Guo, N. *et al.* Construction and Application of an F1-Derived Doubled-Haploid Population and High-Density Genetic Map for Ornamental Kale Breeding. *Genes* **14**, 2104 (2023).

35. Stroupe, S. *et al.* Chromosome-level reference genome for North American bison (Bison bison) and variant database aids in identifying albino mutation. *G3 Genes|Genomes|Genetics* **13**, jkad156 (2023).

36. Hayes, B. J. & Daetwyler, H. D. 1000 Bull Genomes Project to Map Simple and Complex Genetic Traits in Cattle: Applications and Outcomes. *Annual Review of Animal Biosciences* **7**, 89–102 (2019).

37. Thomas, C. G. *et al.* Full-genome evolutionary histories of selfing, splitting, and selection in Caenorhabditis. *Genome Research* **25**, 667 (2015).

38. Crombie, T. A. *et al.* Deep sampling of Hawaiian Caenorhabditis elegans reveals high genetic diversity and admixture with global populations. *eLife* **8**, e50465 (2019).

39. Lei, Y. *et al.* Whole-genome resequencing reveals the origin of tea in Lincang. *Front. Plant Sci.* **13**, (2022).

40. Wolf, M. *et al.* The genome of the pygmy right whale illuminates the evolution of rorquals. *BMC Biology* **21**, 79 (2023).

41. Patiranage, D. S. *et al.* Genome-wide association study in quinoa reveals selection pattern typical for crops with a short breeding history. *eLife* **11**, e66873 (2022).

42. Svardal, H. *et al.* Ancient hybridization and strong adaptation to viruses across African vervet monkey populations. *Nat Genet* **49**, 1705–1713 (2017).

43. Martinez Barrio, A. *et al.* The genetic basis for ecological adaptation of the Atlantic herring revealed by genome sequencing. *eLife* **5**, e12081 (2016).

44. Gao, B. *et al.* Chromosome genome assembly and whole genome sequencing of 110 individuals of Conogethes punctiferalis (Guenée). *Sci Data* **10**, 805 (2023).

45. Huang, W. *et al.* Natural variation in genome architecture among 205 Drosophila melanogaster Genetic Reference Panel lines. *Genome Research* **24**, 1193 (2014).

46. Signor, S. A., New, F. N. & Nuzhdin, S. A Large Panel of Drosophila simulans Reveals an Abundance of Common Variants. *Genome Biology and Evolution* **10**, 189–206 (2018).

47. Todd, E. T. *et al.* The genomic history and global expansion of domestic donkeys. *Science* **377**, 1172–1180 (2022).

48. Jagannathan, V. *et al.* Comprehensive characterization of horse genome variation by whole-genome sequencing of 88 horses. *Animal Genetics* **50**, 74–77 (2019).

49. Stern, D. B. & Lee, C. E. Evolutionary origins of genomic adaptations in an invasive copepod. *Nat Ecol Evol* **4**, 1084–1094 (2020).

50. Gheyas, A. *et al.* Whole genome sequences of 234 indigenous African chickens from Ethiopia. *Sci Data* **9**, 53 (2022).

51. Happ, M. M. *et al.* Comparing a Mixed Model Approach to Traditional Stability Estimators for Mapping Genotype by Environment Interactions and Yield Stability in Soybean [Glycine max (L.) Merr.]. *Front. Plant Sci.* **12**, (2021).

52. Hansen, C. C. R. *et al.* Genomic diversity and differentiation between island and mainland populations of white-tailed eagles (Haliaeetus albicilla). *Molecular Ecology* **32**, 1925–1942 (2023).

53. Svardal, H. *et al.* Ancestral Hybridization Facilitated Species Diversification in the Lake Malawi Cichlid Fish Adaptive Radiation. *Molecular Biology and Evolution* **37**, 1100–1113 (2020).

54. Das, P. *et al.* De novo Assembly and Genome-Wide SNP Discovery in Rohu Carp, Labeo rohita. *Front. Genet.* **11**, (2020).

55. Wang, L., Liu, S., Yang, Y., Meng, Z. & Zhuang, Z. Linked selection, differential introgression and recombination rate variation promote heterogeneous divergence in a pair of yellow croakers. *Mol Ecol* **31**, 5729–5744 (2022).

56. Tollis, M. *et al.* Elephant Genomes Reveal Accelerated Evolution in Mechanisms Underlying Disease Defenses. *Molecular Biology and Evolution* **38**, 3606–3620 (2021).

57. Stankowski, S. *et al.* Widespread selection and gene flow shape the genomic landscape during a radiation of monkeyflowers. *PLOS Biology* **17**, e3000391 (2019).

58. Hoelzel, A. R. *et al.* Genomics of post-bottleneck recovery in the northern elephant seal. *Nat Ecol Evol* **8**, 686–694 (2024).

59. Euclide, P. T. *et al.* Conserved islands of divergence associated with adaptive variation in sockeye salmon are maintained by multiple mechanisms. *Molecular Ecology* **33**, e17126 (2024).

60. Cádiz, M. I. *et al.* Whole genome re-sequencing reveals recent signatures of selection in three strains of farmed Nile tilapia (Oreochromis niloticus). *Sci Rep* **10**, 11514 (2020).

61. Choi, J. Y. *et al.* The complex geography of domestication of the African rice Oryza glaberrima. *PLOS Genetics* **15**, e1007414 (2019).

62. Pan, Z. *et al.* Whole-genome sequences of 89 Chinese sheep suggest role of RXFP2 in the development of unique horn phenotype as response to semi-feralization. *GigaScience* **7**, giy019 (2018).

63. Fair, B. J. *et al.* Gene expression variability in human and chimpanzee populations share common determinants. *eLife* **9**, e59929 (2020).

64. Hoge, C. *et al.* Patterns of recombination in snakes reveal a tug-of-war between PRDM9 and promoter-like features. *Science* **383**, eadj7026 (2024).

65. Rogers, J. *et al.* The comparative genomics and complex population history of Papio baboons. *Science Advances* **5**, eaau6947 (2019).

66. Lucius, M. D. *et al.* Genomic variation in captive deer mouse (Peromyscus maniculatus) populations. *BMC Genomics* **22**, 662 (2021).

67. Celemín, E. *et al.* Evolutionary history and seascape genomics of Harbour porpoises (Phocoena phocoena) across environmental gradients in the North Atlantic and adjacent waters. *Molecular Ecology Resources* **n/a**,.

68. Guo, J. *et al.* An integrated peach genome structural variation map uncovers genes associated with fruit traits. *Genome Biology* **21**, 258 (2020).

69. Saremi, N. F. *et al.* Puma genomes from North and South America provide insights into the genomic consequences of inbreeding. *Nat Commun* **10**, 4769 (2019).

70. Harpak, A. *et al.* Genetic Adaptation in New York City Rats. *Genome Biol Evol* **13**, evaa247 (2020).

71. Peter, J. *et al.* Genome evolution across 1,011 Saccharomyces cerevisiae isolates. *Nature* **556**, 339–344 (2018).

72. Xia, W. *et al.* Population genomics reveals structure at the individual, host-tree scale and persistence of genotypic variants of the undomesticated yeast Saccharomyces paradoxus in a natural woodland. *Molecular Ecology* **26**, 995–1007 (2017).

73. Gao, G. *et al.* A New Single Nucleotide Polymorphism Database for North American Atlantic Salmon Generated Through Whole Genome Resequencing. *Front Genet* **11**, 85 (2020).

74. Saha, A. *et al.* Whole-genome resequencing confirms reproductive isolation between sympatric demes of brown trout (Salmo trutta) detected with allozymes. *Molecular Ecology* **31**, 498–511 (2022).

75. de los Ríos-Pérez, L. *et al.* An ultra-high density SNP-based linkage map for enhancing the pikeperch (Sander lucioperca) genome assembly to chromosome-scale. *Sci Rep* **10**, 22335 (2020).

76. Poelstra, J. W., Richards, E. J. & Martin, C. H. Speciation in sympatry with ongoing secondary gene flow and a potential olfactory trigger in a radiation of Cameroon cichlids. *Molecular Ecology* **27**, 4270–4288 (2018).

77. Mathur, S., Mason, A. J., Bradburd, G. S. & Gibbs, H. L. Functional genomic diversity is correlated with neutral genomic diversity in populations of an endangered rattlesnake. *Proceedings of the National Academy of Sciences* **120**, e2303043120 (2023).

78. Causse, M. *et al.* Whole genome resequencing in tomato reveals variation associated with introgression and breeding events. *BMC Genomics* **14**, 791 (2013).

79. Boatwright, J. L. *et al.* Sorghum Association Panel whole-genome sequencing establishes cornerstone resource for dissecting genomic diversity. *Plant J* **111**, 888–904 (2022).

80. Gimenez, S. *et al.* Adaptation by copy number variation increases insecticide resistance in the fall armyworm. *Commun Biol* **3**, 1–10 (2020).

81. Gaurav, K. *et al.* Population genomic analysis of Aegilops tauschii identifies targets for bread wheat improvement. *Nat Biotechnol* **40**, 422–431 (2022).

82. Delorean, E. *et al.* High molecular weight glutenin gene diversity in Aegilops tauschii demonstrates unique origin of superior wheat quality. *Commun Biol* **4**, 1–9 (2021).

83. Zhao, S. *et al.* Whole-genome sequencing of giant pandas provides insights into demographic history and local adaptation. *Nat Genet* **45**, 67–71 (2013).

84. Timoshevskaya, N., Voss, S. R., Labianca, C. N., High, C. R. & Smith, J. J. Large-scale variation in single nucleotide polymorphism density within the laboratory axolotl (Ambystoma mexicanum). *Developmental Dynamics* **250**, 822–837 (2021).

85. Fitak, R. R. *et al.* Genomic signatures of domestication in Old World camels. *Commun Biol* **3**, 1–10 (2020).

86. Pérez-Moro, C. *et al.* Discovery of variation in genes related to agronomic traits by sequencing the genome of Cucurbita pepo varieties. *BMC Genomics* **26**, 335 (2025).

87. Sallam, A. H. *et al.* Genome-Wide Association Mapping of Stem Rust Resistance in Hordeum vulgare subsp. spontaneum. *G3 Genes|Genomes|Genetics* **7**, 3491–3507 (2017).

88. Michell, C. T., Pohjoismäki, J. L. O., Spong, G. & Thulin, C.-G. Mountain- and brown hare genetic polymorphisms to survey local adaptations and conservation status of the heath hare (Lepus timidus sylvaticus, Nilsson 1831). *Sci Data* **9**, 667 (2022).

89. Foote, A. D. *et al.* Genome-culture coevolution promotes rapid divergence of killer whale ecotypes. *Nat Commun* **7**, 11693 (2016).

90. Kajiya-Kanegae, H. *et al.* OryzaGenome2.1: Database of Diverse Genotypes in Wild Oryza Species. *Rice* **14**, 24 (2021).

91. The 3, 000 rice genomes project. The 3,000 rice genomes project. *GigaScience* **3**, 2047-217X-3–7 (2014).

92. Wei, W. *et al.* Metabolome-Based Genome-Wide Association Study Provides Genetic Insights Into the Natural Variation of Foxtail Millet. *Front Plant Sci* **12**, 665530 (2021).

93. MalariaGEN *et al.* Pf7: an open dataset of Plasmodium falciparum genome variation in 20,000 worldwide samples. *Wellcome Open Res* **8**, 22 (2023).

94. Kuhnert, P., Loosli, N., Brodard, I., Lindtke, D. & Jores, J. Resistance of zebu cattle (*Bos indicus*) to colonization by major ruminant hoof pathogens. *Veterinary Microbiology* **296**, 110184 (2024).

95. NextGen project variation for Capra aegagrus. European Variation Archive: PRJEB5978; Retreived from https://www.ebi.ac.uk/ena/browser/view/PRJEB5978?show=analyses (2021).

96. NextGen project variation for Capra hircus. European Variation Archive: PRJEB6057; Retrieved from https://www.ebi.ac.uk/ena/browser/view/PRJEB6057 (2021).

97. Resequencing of a pair of chickpea near-isogenic lines with contrasting flowering time. European Variation Archive: PRJEB73790; Retrieved from https://www.ebi.ac.uk/ena/browser/view/PRJEB73790 (2024).

98. Butler, M. G. *et al.* SNPfisher: tools for probing genetic variation in laboratory-reared zebrafish. *Development* **142**, 1542–1552 (2015).

99. Louis, M. *et al.* Selection on ancestral genetic variation fuels repeated ecotype formation in bottlenose dolphins. *Science Advances* **7**, eabg1245 (2021).

100. Wang, C. *et al.* Population structure and genetic diversity in Eucalyptus pellita based on SNP markers. *Front Plant Sci* **14**, 1278427 (2023).

101. Han, F. *et al.* Gene flow, ancient polymorphism, and ecological adaptation shape the genomic landscape of divergence among Darwin's finches. *Genome Res.* **27**, 1004–1015 (2017).

102. Parallel evolution in the clingfish genus Gouania. European Variation Archive: PRJEB60875; Retrieved from https://www.ebi.ac.uk/ena/browser/view/PRJEB60875 (2024).

103. Chen, S. *et al.* A genomic mutational constraint map using variation in 76,156 human genomes. *Nature* **625**, 92–100 (2024).

104. Spivakov, M. *et al.* Genomic and Phenotypic Characterization of a Wild Medaka Population: Towards the Establishment of an Isogenic Population Genetic Resource in Fish. *G3 Genes|Genomes|Genetics* **4**, 433–445 (2014).

105. NextGen project variation for Ovis orientalis. European Variation Archive: PRJEB6495; Retrieved from https://www.ebi.ac.uk/ena/browser/view/PRJEB6495 (2021).

106. Ozerov, M. *et al.* Genomics of humic adaptation in Eurasian perch (Perca fluviatilis): SNP genotypes of 32 perch individuals, supplementary figures and tables. 233736280 bytes Dryad https://doi.org/10.5061/DRYAD.M0CFXPP4T (2022).

107. Sequencing of common bean genotypes provides comprehensive resources for genetic studies and molecular breeding. European Variation Archive: PRJEB18671; Retrieved from https://www.ebi.ac.uk/ena/browser/view/PRJEB18671 (2021).

108. De Meulenaere, K., Cuypers, B., Gamboa, D., Laukens, K. & Rosanas-Urgell, A. A new Plasmodium vivax reference genome for South American isolates. *BMC Genomics* **24**, 606 (2023).

109. Draft genome of wild Prunus yedoensis var. nudiflora (King cherry flower). European Variation Archive: PRJEB28064; Retrieved from https://www.ebi.ac.uk/ena/browser/view/PRJEB28064 (2021).

110. Whole genome resequencing of the human parasite Schistosoma mansoni reveals population history and effects of selection. European Variation Archive: PRJEB13625; Retrieved from https://www.ebi.ac.uk/ena/browser/view/PRJEB13625 (2021).

111. Greater amberjack genetic variants. European Variation Archive: PRJEB24756; Retrieved from https://www.ebi.ac.uk/ena/browser/view/PRJEB24756 (2021).

112. Glasenapp, M. R. & Pogson, G. H. Extensive introgression among strongylocentrotid sea urchins revealed by phylogenomics. *Ecology and Evolution* **13**, e10446 (2023).

113. Bovo, S. *et al.* Whole-genome sequencing of European autochthonous and commercial pig breeds allows the detection of signatures of selection for adaptation of genetic resources to different breeding and production systems. *Genetics Selection Evolution* **52**, 33 (2020).

114. Cornejo, O. E. *et al.* Population genomic analyses of the chocolate tree, Theobroma cacao L., provide insights into its domestication process. *Commun Biol* **1**, 1–12 (2018).

115. Kinneberg, V. B., Lü, D. S., Peris, D., Ravinet, M. & Skrede, I. Introgression between highly divergent fungal sister species. *Journal of Evolutionary Biology* **36**, 1133–1149 (2023).

116. Single-nucleotide polymorphism matrices for a large diversity panel of wheat (Triticum aestivum L.). European Variation Archive: PRJEB52759; Retrieved from https://www.ebi.ac.uk/ena/browser/view/PRJEB52759 (2022).

117. de Jong, M. J. *et al.* Range-wide whole-genome resequencing of the brown bear reveals drivers of intraspecies divergence. *Commun Biol* **6**, 1–16 (2023).

118. Iannucci, A. *et al.* Population structure, genomic diversity and demographic history of Komodo dragons inferred from whole-genome sequencing. *Molecular Ecology* **30**, 6309–6324 (2021).

119. Genome-wide scan for runs of homozygosity in South American Camelids. European Variation Archive: PRJEB61878; Retrieved from https://www.ebi.ac.uk/ena/browser/view/PRJEB61878 (2023).

120. Bukowski, R. *et al.* Construction of the third-generation Zea mays haplotype map. *Gigascience* **7**, 1–12 (2018).

121. Genome sequencing of Aspergillus oryzae/flavus. European Variation Archive: PRJEB79400; Retrieved from https://www.ebi.ac.uk/ena/browser/view/PRJEB79400 (2024).

122. Li, C. *et al.* Genome Variation Map: a worldwide collection of genome variations across multiple species. *Nucleic Acids Res* **49**, D1186–D1191 (2021).

123. Niederhuth, C. E. *et al.* Widespread natural variation of DNA methylation within angiosperms. *Genome Biology* **17**, 194 (2016).

124. Klughammer, J. *et al.* Comparative analysis of genome-scale, base-resolution DNA methylation profiles across 580 animal species. *Nat Commun* **14**, 232 (2023).

125. Jones, C. M., Lim, K. S., Chapman, J. W. & Bass, C. Genome-Wide Characterization of DNA Methylation in an Invasive Lepidopteran Pest, the Cotton Bollworm Helicoverpa armigera. *G3 (Bethesda)* **8**, 779–787 (2018).

126. Deshmukh, S., Ponnaluri, V. C., Dai, N., Pradhan, S. & Deobagkar, D. Levels of DNA cytosine methylation in the Drosophila genome. *PeerJ* **6**, e5119 (2018).

127. Ponts, N. *et al.* Genome-wide mapping of DNA methylation in the human malaria parasite Plasmodium falciparum. *Cell Host Microbe* **14**, 696–706 (2013).

128. Pandey, G., Yadav, C. B., Sahu, P. P., Muthamilarasan, M. & Prasad, M. Salinity induced differential methylation patterns in contrasting cultivars of foxtail millet (Setaria italica L.). *Plant Cell Rep* **36**, 759–772 (2017).

129. Hernando-Herraez, I., Garcia-Perez, R., Sharp, A. J. & Marques-Bonet, T. DNA Methylation: Insights into Human Evolution. *PLoS Genet* **11**, e1005661 (2015).

130. Bhatia, H., Khemka, N., Jain, M. & Garg, R. Genome-wide bisulphite-sequencing reveals organ-specific methylation patterns in chickpea. *Sci Rep* **8**, 9704 (2018).

131. Kong, W. *et al.* 5mC DNA methylation modification-mediated regulation in tissue functional differentiation and important flavor substance synthesis of tea plant (Camellia sinensis L.). *Hortic Res* **10**, uhad126 (2023).

132. Kronforst, M. R., Gilley, D. C., Strassmann, J. E. & Queller, D. C. DNA methylation is widespread across social Hymenoptera. *Current Biology* **18**, R287–R288 (2008).

133. Dimond, J. L. & Roberts, S. B. Convergence of DNA Methylation Profiles of the Reef Coral Porites astreoides in a Novel Environment. *Front. Mar. Sci.* **6**, (2020).

134. Molaro, A. *et al.* Sperm Methylation Profiles Reveal Features of Epigenetic Inheritance and Evolution in Primates. *Cell* **146**, 1029–1041 (2011).

135. Hossain, M. N. *et al.* Cold exposure impacts DNA methylation patterns in cattle sperm. *Front Genet* **15**, 1346150 (2024).

136. Chen, S. *et al.* Comparative Analyses of Sperm DNA Methylomes Among Three Commercial Pig Breeds Reveal Vital Hypomethylated Regions Associated With Spermatogenesis and Embryonic Development. *Front. Genet.* **12**, (2021).

137. Chen, L. *et al.* WGBS of embryonic gonads revealed that long non-coding RNAs in the MHM region might be involved in cell autonomous sex identity and female gonadal development in chickens. *Epigenetics* **19**, 2283657.

138. Jiang, L. *et al.* Sperm, but Not Oocyte, DNA Methylome Is Inherited by Zebrafish Early Embryos. *Cell* **153**, 773–784 (2013).

139. Wellband, K., Roth, D., Linnansaari, T., Curry, R. A. & Bernatchez, L. Environment-driven reprogramming of gamete DNA methylation occurs during maturation and is transmitted intergenerationally in Atlantic Salmon. *G3 (Bethesda)* **11**, jkab353 (2021).

140. Weinstock, G. M. *et al.* Insights into social insects from the genome of the honeybee Apis mellifera. *Nature* **443**, 931–949 (2006).

141. Gori, K. & Baez-Ortega, A. sigfit: flexible Bayesian inference of mutational signatures. 372896 Preprint at https://doi.org/10.1101/372896 (2020).

142. Beichman, A. C. *et al.* “Evolution of the mutation spectrum across a mammalian phylogeny”. 2023.05.31.543114 Preprint at https://doi.org/10.1101/2023.05.31.543114 (2023).

143. Poplin, R. *et al.* A universal SNP and small-indel variant caller using deep neural networks. *Nat Biotechnol* **36**, 983–987 (2018).

144. Ho, L. si T. & Ané, C. A linear-time algorithm for Gaussian and non-Gaussian trait evolution models. *Syst Biol* **63**, 397–408 (2014).