Mohammadmahdi Hajimoradkhani

Intergenic Co-evolution in Primates: Co-ordinated mutational patterns in the Genetic Architecture of Lifespan and Body Mass

MASTER'S THESIS

Supervisors:

Prof. Hafid Laayouni (Universitat Pompeu Fabra)

Prof. Arcadi Navarro Cuartiellas (Universitat Pompeu Fabra)

Tutor:

Prof. Sergio Gómez Jiménez (Universitat Rovira i Virgili)

Master's Degree in Health Data Science

Table of contents

1. Introduction

1.1. Background and motivation

1.1.1. The genetic architecture of complex traits, epistasis and evolution

The genetic architecture of complex traits[1], such longevity, is a fundamental topic in modern biology with broad implications for evolutionary theory, precision medicine and agriculture. Comparative studies within populations and between species allow the identification of genes and genetic variants that correlate with trait variation[2], [3]. However, identifying the alleles involved is only part of the picture, equally important is understanding their mode of action. Alleles can influence phenotypes in an additive manner -where their effects are independent of other genetic loci, or through *epistatic* interactions such that the effect of a particular allele is dependent on the identity of the alleles at other genomic loci[1], [4]. Epistasis introduces non-linear effects into the genotype-phenotype relationship and is regarded as an emergent property of biological networks, particularly biological pathways and gene regulatory networks. In such networks, genes may interact directly (physically) or indirectly as part of the same functional pathways influencing cellular processes (e.g. senescence) or higher level phenotypes (e.g. longevity).

There is an ongoing debate about the relative importance of additive versus epistatic effects in shaping complex phenotypes[5], [6], [6], [7], [8]. While substantial evidence has arisen from evolutionary studies[8], [9], [10], [11], [12] and studies involving model organisms[13] suggesting the significant role of *intragenic* and *intergenic* epistasis, despite most statistical models linking genetic variation to phenotypic variation in human populations under an assumption of additive gene action[2]. While additive gene action may explain more of the variation in complex phenotypes in human populations according to some simulations and models- such studies suffer from insufficient sample sizes and unfavourable allele frequencies to quantify epistatic effects. In fact, pointing to evolution and the tree of life itself, the identity of alleles in a given genetic context has been demonstrated to be a significant factor in molecular evolution of proteins, and also in genetic incompatibilities between species, such that measures of selection and molecular phenotypes reflect this in large studies with high throughput data and large numbers of species[8], [14], [15], [16], [17].

When studying the impact of mutations, especially in applied clinical settings and drug discovery, it should be noted that evolution provides powerful grounds for understanding such effects under a given genetic context- many loci that are fixed in most of the human population but undergo mutations in disease are mutational steps away from variants in other species, especially closely related species such as primates[18], [19], [20], [21]. Looking at differences in effects in different genetic contexts not only potentially consolidates the role of epistasis in the genetic architecture of complex traits, it also allows us to use evolution as the biggest experiment undertaken by nature to facilitate precision medicine[22]; given that evolution optimises the fitness of the genetic systems of populations of species through acting as a filter for genetic variants, the comparison between problematic allelic combinations found in humans and what has been observed, or not observed,

across primate evolution could further our understanding of the biological system perturbed by a disease or trait state as a function of the variants[23].

An evolutionary approach further avoids a common limitation of population studies such as GWAS which is statistical power, as genotype-phenotype associations across evolutionary timescales are not subject to the same constraints in obtaining a statistically significant signal, as sample sizes are often not sufficient in GWAS[1], [13] studies for the p-values of certain variant effects to survive the multiple testing correction, a case that is more typical for allelic variants in low frequencies. This problem is enormously more pronounced when considering 2-way interaction terms (to consider epistasis) in the GWAS linear model measuring effect sizes of variants towards a phenotype,and the issue grows in scale when considering higher order interactions between loci.

1.1.2. Evolution, genetic networks and co-evolution: co-ordinated mutational patterns

Given that genes work in complex networks that determine phenotypes in tandem with environmental influence[24], it should be noted that evolution does not act on individual genes alone but on networks of genes[25]. Therefore, mutations do not occur in isolation, but in co-ordinated patterns- a notion that is demonstrated in subsequent sections of this work. The concept of co-ordinated mutational patterns is closely related with epistasis, where the effect of a variant is dependent on the genetic context, or in other words, the identity of the variants at other loci; in this study we consider strictly *intergenic* loci. In the context of evolution, the dependency of the effect of a variant towards a phenotype translates to the idea that the effect of a mutation is dependent on other loci with meaningful interactions at the genetic network scale (not to be confused with strictly physically interacting loci), which is the principle behind the assumption of co-ordinated mutational patterns- the basis of the co-evolutionary perspective we maintain in our analysis for included genes.

In recent years, high throughput and sophisticated genomics assays have produced high quality genomes across the tree of life, spanning eukaryotic, mammalian and primate taxa. We focus specifically on the primate phylogeny where high quality genomes are available for 230 species and a modern state of the art phylogenetic tree to depict the evolutionary relationships between the given primates[26]. The particular focus on primates is due to a myriad of reasons spanning statistical methodology, allometric considerations, data availability and the similarity of primate genetic systems to that of humans for translatable research. Through the lens of systems genetics perspective that inspires us to use co-evolution as a means of exploring the genetic architecture of life history traits such as longevity and body mass, looking at primate genomes in relation to the human genome has the advantage where a large degree of similarity exists between humans and other primates, especially those with most recent common ancestors such as chimpanzees, where this similarity allows changes with adaptive and co-ordinated mutational significance to stand out at the statistical level, and easier to validate in downstream experimental analyses to extract translatable value from our work as a contribution towards precision medicine and understanding of the *Human* genetic architecture of life history traits. The aforementioned is made possible by assays that have produced a wealth of primate phenotypic data from primates in zoos and natural habitats, such that the integrated consideration of the

phenomes and genomes allows the establishment of the genotype-phenotype maps necessary to conduct co-evolutionary analyses.

Indeed, it has been well studied that allometric predictors of longevity, such as body size, underestimate the lifespan of primates[27]. Specifically, *metabolic scaling* states that larger species tend to live longer due to lower metabolic rates leading to less oxidative damage per unit time[28]. This biochemical foundation has led to the development of mathematical models that scale the expected lifespan of a species proportional to the body size[29]. Primates generally overperform in their lifespans based on this relationship however, suggesting that the underlying molecular evolution in the genetic systems of primates that enabled adaptation to increase survival rates must have had an overlap with the genetic systems of primates that determine primate longevity. This *pleiotropic* effect is representative of the evolutionary trade-offs between particular traits and fitness that is a hallmark of evolution.

Given the potential biomedical and evolutionary significance of such changes, it becomes crucial to identify and understand the co-occurring mutations that may have contributed to these outcomes. By tracing these coordinated changes over millions of years of primate evolution, we hope to shed light on the genetic systems that enabled increased lifespan, and in doing so, to provide insights into the genetic basis of longevity in humans.

1.1.3. Phylogenetic comparative methods: Building genotype-phenotype maps using divergence data

The co-evolutionary approach, which takes into account a system of genes and the co-ordinated mutational patterns between intergenic loci across primate evolution, requires identifying the system of genes that are associated with a particular phenotype. In the context of this study the focal traits are longevity (represented by maximum lifespan across primates) and body mass. Modern Phylogenetic Comparative Methods (PCMs) such as RERConverge[30] and CAASTools[31] have been developed to identify associations between genes, or loci towards a particular phenotype while controlling for phylogenetic structure. These PCMs take as input multiple sequence alignments (MSAs), a phylogenetic tree representing the more accurate depiction of the relationships between the species present in the multiple sequence alignments, and the phenotypic values of the focal trait for the species present in the analysis- this integrated approach allows the establishment of genotype-phenotype maps based on molecular and phenotypic patterns observed across evolution. Such methods have been increasing in relevance and sophistication in recent years thanks to powerful efforts by the community to gather phenomes to facilitate integrated analysis in conjunction with the large number of genomes assembled across the tree of life.

RERConverge is a PCM that operates at the *gene* level where it considers shifts in the relatively evolutionary rate (RER) of a particular gene across a phylogeny, and identifies whether there is an association between shifts in the values of the focal trait and shifts in the RER of the gene. In brief, if a consistent signal persists across the phylogeny demonstrating that there is an association between shifts in the RER of the gene and

shifts in the phenotype, while controlling for phylogenetic structure, then the gene can be inferred to have been involved in the genetic architecture of the trait across evolution. The method does not however infer any particular positions or specific states in the sequences that were associated with these shifts, therefore it does not directly infer loci that could be used for an analysis with our philosophy.

CAASTools operates at both the *gene* level and the *position* level where, given it takes as input the MSA, phylogeny and trait values, but it also takes into consideration two groups defined by the user- the *tops* corresponding to the highest phenotypic extremes for the trait, and the *bottoms* corresponding to the lowest phenotypic extremes for the trait. This categorisation of the species is then used to detect whether there are positions in the MSA of the gene that show a pattern of molecular convergence coinciding with phenotypic convergence- do all of the top species have a particular amino acid (AA) while all of the bottom species have a particular amino acid in the focal position? This is the framework to identify convergent amino acid substitutions (CAAS). In the context of longevity, identifying CAAS in each gene across the primate phylogeny means identifying *loci* in genes that display a statistically significant convergent molecular pattern controlled for phylogenetic structure- independently of how closely related they are, the longest lived primates have particular AA residues at a given locus, while the shortest lived primates have another AA residue at that position of a gene. This methodology allows the establishment of a genotype-phenotype map across primate evolution but crucially also provides candidate loci (i.e. gene positions) that evolution has acted upon.

With PCMs such as CAASTools providing genes and positions that present an association with our focal traits of longevity and body mass, it becomes possible to leverage powerful existing methods such as ancestral sequence reconstruction (ASR) to reconstruct the history of changes of candidate positions across the primate phylogeny. The framework developed in *Methodology* thus allows us to infer the history of changes of focal positions in each gene in combination with focal positions of other genes to identify whether an associative, co-ordinated pattern of mutations has been present in the genetic systems of primates consequential for longevity and body mass.

1.1.4. The gaps in knowledge

Since the advent of the Human Genome Project, a substantial amount of studies have focused on polymorphism data pertaining to human populations as a means of studying the genetic architecture of complex traits, especially so in the domains of disease and human longevity. While these studies have made important contributions, they suffer from sample sizes that require genetic variants at low frequencies within a population to have an effect large enough that could be detected. Further, such studies consider mutations and variants in the scope of single genes or segments of genetic sequences as opposed to a systems genetics view that looks at how combinations of variants across a genetic system are contributing to a phenotype. This is due to limitations in statistical power in considering interaction terms in the linear models measuring effect sizes of variants upon phenotypes, and tandem simplifying assumptions that neglect epistasis as a contributing factor to the genetic architecture of traits[1].

In evolutionary studies, significant work has been done in demonstrating epistasis as a factor in the mode and tempo of the molecular evolution of individual proteins, and also in the context of how epistasis has been a foundation upon which natural selection has acted across the evolution of a large number of genes in the eukaryotic tree of life. The drawback however, is that the focus was in the general theory of how evolution has behaved, and more of a focus on *intragenic* epistasis in both cases[8]. There is a lack of depth in assessing how networks of genes have evolved with special consideration of *intragenic* epistasis. Additionally, such theoretical studies have had a general focused as opposed to an application specific exploration of the primate phylogeny that we propose in this work, where we believe the systems genetics view at the heart of the co-evolutionary framework, in conjunction with the close evolutionary relationship between humans and other primate species provides grounds for research that could have more immediate impact in translatable domains such as precision medicine.

1.2. Objectives

The inquiry presented in introduction can be dissected into several stages of analysis, where in most cases the results of each sub-inquiry serve as the input to the next. To carry out the methodological work and interpretation attached to each stage, we use pre-existing work carried out in our lab which has established a catalogue of genes and constituent CAAS positions across primate orthologues with respect to maximum lifespan and body mass independently- for further discussion refer to methodology.

1.2.1. Reconstruction of the evolutionary history of CAAS

In the first stage of our work, we use ASR to reconstruct the history of changes for all genes implicated in the genetic architecture of longevity and body mass across primate evolution. To qualify as a gene associated in the genetic architecture of these focal traits, the gene must harbour at least one statistically significant CAAS position. This reconstruction maps the history of changes of each position of each gene across the primate phylogeny, revealing the most probable set of changes that have occurred across primate evolution to produce the sequences observed today. Naturally, this reconstruction also gives the catalogue of changes (i.e. evolutionary branches and the exact amino acid transitions) that the CAAS positions have undergone. Fulfillment of this objective serves as a precursor step to co-evolutionary analysis that will require the identity of the phylogenetic branches where each CAAS position has changed, and the exact AA transition inferred on the branch.This procedure, like all of the analysis in our workflow, is performed independently for the maximum lifespan (proxy for longevity) and body mass datasets- where each dataset contains the genes and CAAS positions identified in the genes.

1.2.2. The catalogue of co-occurring changes between intergenic CAAS positions: The magnitude and statistical significance of coevolution

Given the history of the changes obtained for each CAAS position, where we know the identity of the branches, number of changes and the exact transitions for each CAAS position in the corresponding dataset; we consider the genes inferred in the genetic architecture of each trait *across evolution* as a system to pave

the way for co-evolutionary analysis. To do so, we generate the set of all pairwise combinations of genes in each dataset, and for a given pair of genes we consider the set of all pairwise combinations of CAAS positions. We identify the number of times each intergenic CAAS pairs has undergone co-occurring mutations and the number of times they have changed individually. Then, we design and implement simulations to assess which intergenic CAAS pairs have co-evolved more than one would expect by random chance. This allows us to answer the fundamental questions- how much co-evolution is occurring between CAAS positions inferred to be part of the genetic architecture of our traits? Do we observe that the genes harbouring these positions are co-evolving as a system in a way that significantly deviates from what is expected under a random model of coevolution?

Answering these fundamental questions allows us to move on to exploring whether an associative structure of co-transition exists across the intergenic loci belonging to each system of genes, segregated by trait.

1.2.3. Patterns of co-transition: Do pairwise co-transitions in the genetic system of a trait follow an associative pattern

The inference of co-occurring mutations amongst intergenic CAAS pairs, and the subsequent simulations allow us to filter for intergenic CAAS pairs that have a magnitude of coevolution beyond what is expected from random chance. This allows considering the exact identity of the AA co-transitions in these co-occurring changes, where a co-transition equates to the AA transitions undergone by each CAAS of an intergenic CAAS pair on a branch of the primate phylogeny. Here we define as our goals the ability to represent these co-transitions, pooled together across the dataset, in a data structure that is quantitatively and qualitatively tractable in terms of performing statistical tests and mathematical operations to answer a fundamental question- across the intergenic CAAS loci in our dataset that are co-evolving far beyond what is expected under randomness, are the pattern of co-transitions associative?

An associative pattern of co-transition spanning all of the significant co-evolution observed across our focal loci would suggest a dependency between particular AA states at loci across evolution such that evolution has permitted only specific mutations to occur in consideration of the state at another loci- a significant result in understanding the genetic architecture of longevity and body mass, and one that would be consistent with evolutionary studies that have demonstrated epistasis created dependencies between mutations. Such a result would be profound in its ability to be translated into value for precision medicine in humans where diseases and senescence are being addressed using advanced knowledge of the genetic architecture of traits in tandem with multimodal data integration across the omics data spectrum.

1.2.4. Methods development

To explore the objectives, and fundamental questions posed by our line of inquiry, we develop methods integrating mathematics, code implementation and HPC compatibility. While this is not a primary scientific focus, at a technical level it is one of the objectives to produce a workflow that could be versatile enough to be used with other types of positions other than CAAS, re-usable by other researchers and reproducible for

broader applications. Thus, we aim not only to advance knowledge on the scientific problems discussed, but to also make technical contributions in the field to accelerate contributions in knowledge to understanding the genetic architecture of complex traits across evolution using a systems-genetics co-evolutionary framework.

2. Methodology

2.1. Data preparation, exploration and filtering

2.1.1. Data selection

The data is based on a catalogue of 233 primate genomes, and the phylogenetic tree was constructed using the ultra-conserved regions of these genomes to produce the current state of the art representation of the primate phylogeny. Our in-house data's starting point consists of the AA translated codon alignments of ~ 16000 1-to-1 primate orthologues. Data was gathered for ~ 98 traits, including maximum lifespan and body mass, for as many species as possible to create what we internally denote as the Primate Genome-Phenome Archive (PGA), which consists of the set of genomes and matching phenomes. Note that the sparsity was greater across the phenomes. Following the establishment of the PGA, a large scale computational experiment was performed using CAASTools for each of the ~ 16000 genes against the ~ 98 traits, with the goal of identifying positions in each gene, that are CAAS towards a particular trait- the position of each CAAS is computed based on the MSA. This data served as the starting point for this investigation. One issue however was that the number of species and the identity of species was different across the MSA, as not all genomes had the same level of quality and coverage and therefore some genes may have been missing, or some orthologues simply may not have existed in some species. As a consequence, the ~ 16000 genes could be partitioned into different sets, where each had a distinct number and set of species present in the MSA.

The ~ 16000 genes collapsed to ~ 3000 when selecting the set of genes that had the maximum number of identical species (230 species) present in the MSA. This criteria was important because any inference based on ASR would require the tree topology to be identical across different genes to facilitate the analysis of co-occurring changes between two intergenic CAAS positions. For instance, a gene with 220 species in the MSA would have a different topology for the phylogenetic tree than a gene with 230 species, causing issues in identifying which branches of the primate phylogeny harboured co-occurring changes between the CAAS positions of each gene. Resolving such an issue may require collapsing multiple branches of the larger tree into one, however this will predictably creates biases in statistical parameters used in simulations later on, where these biases will be against comparisons between trees with more numerous branches- consequently we controlled for this by simply selecting genes that allow working with an identical and fixed topology.

The set of ~ 3000 genes containing 230 identical species contained genes with at least one statistically significant CAAS for all of the ~ 98 traits. Our focal traits however, are longevity- represented by maximum lifespan, and body mass. We selected these two life history traits given the reasoning explored in the

Introduction, in addition to maximum lifespan in particular representing a maximum across a population and therefore being less susceptible to sampling errors as a function of population variance and biases generated from phenotypic measurements of primates that may have been kept in zoos. Out of the ~ 3000 genes, we selected 883 genes for maximum lifespan after filtering for genes that may have had an erroneous number of CAAS identified relative to the distribution of significant CAAS per gene. Body mass, being a life history trait that could show signals of genetic trade offs enforced by evolution to optimise fitness, also serves as a statistical outgroup with 213 genes passing the filters to be included in the analysis- comparing the signal of co-evolution and hypothesis of associative co-transition, across the maximum lifespan dataset with more CAAS positions distributed across 883 genes, compared to the body mass dataset with 213 genes, allowed utilising some of the trait data beyond maximum lifespan to compare results at different stages.

We suspect the genes which harboured erroneously large numbers of CAAS to have had misaligned MSAs, or to have been paralogues which were false positives in the 1-to-1 orthology test performed prior to our filtering of the data. These genes were all excluded and the analysis was performed with the trait gene sets mentioned above.

2.1.2. Alignment length and number of CAAS hits

As part of a post-hoc filtering step before implementation and execution of our pipeline, we computed the correlation between the length of a gene (length of the MSA) and the number of CAAS hits per gene. This was done to ensure that no genes hosting a large number of CAAS as a function of the length of their MSA were included in the trait datasets given our expectation that CAAS hits on a gene should be independent of the gene length.

	Maximum Lifespan	**Body Mass**
Pearson Correlation Coeff.	0.0139	-0.0467

Tab 1. Table showing the correlation between the length of the alignment and the number of CAAS per gene across each corresponding dataset

Based on the information shown in *Tab 1*, the number of CAAS inferred per gene is not a function of the length of the alignment of the corresponding gene, as in both datasets the correlation is quite close to 0.

2.1.3. Distribution of the number of CAAS per gene

Given that we aim to generate the set of all intergenic CAAS pairs across all pairwise combinations of genes, we explore the distribution of significant CAAS hits per gene. Exploring the distribution for each dataset is important given that some combinations of genes will simply have a larger set of pairwise intergenic CAAS pairs between them, which could be taken into account if the goal is to stratify the results by gene pairs which have displayed a stronger co-evolutionary pattern. For instance, if two genes have a large number of intergenic CAAS pairs between them as a function of each having a large number of CAAS hits, independently

of gene length as we explored in the previous correlation analysis, then these genes have a stronger signal of co-evolution if all of these pairs are co-evolving non-randomly when compared to a pair of genes with fewer numbers of possible intergenic pairs co-evolving significantly. Another useful application of this preliminary exploration was the elimination of genes with extreme numbers of CAAS hits relative to the distribution of CAAS hits across the genes of the dataset- for instance, the genes *PRAME* and *CES3* of the maximum lifespan dataset were removed before the analysis as they had upwards of 150 significant CAAS hits each.

Note that the maximum lifespan geneset consists of 885 genes and 1330 significant CAAS positions, while the body mass geneset consists of 213 genes harbouring 397 significant CAAS positions. This should be taken into account when interpreting the magnitude of the frequencies of the number of CAAS in the distributions plotted below.

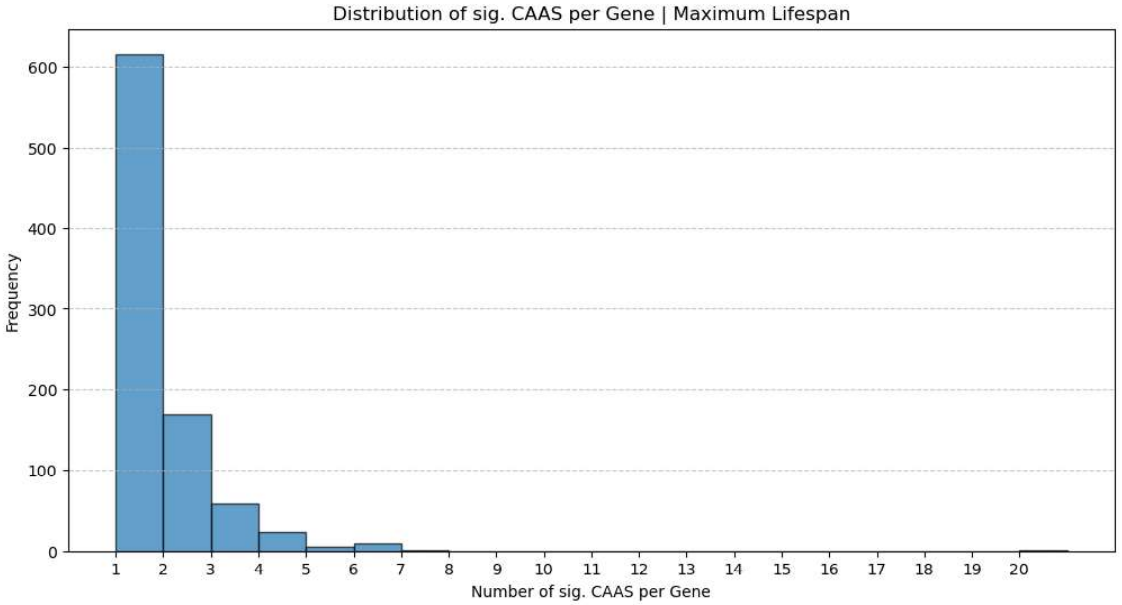


Fig 1. Distribution of the number of sig. CAAS per gene in the Maximum Lifespan dataset.

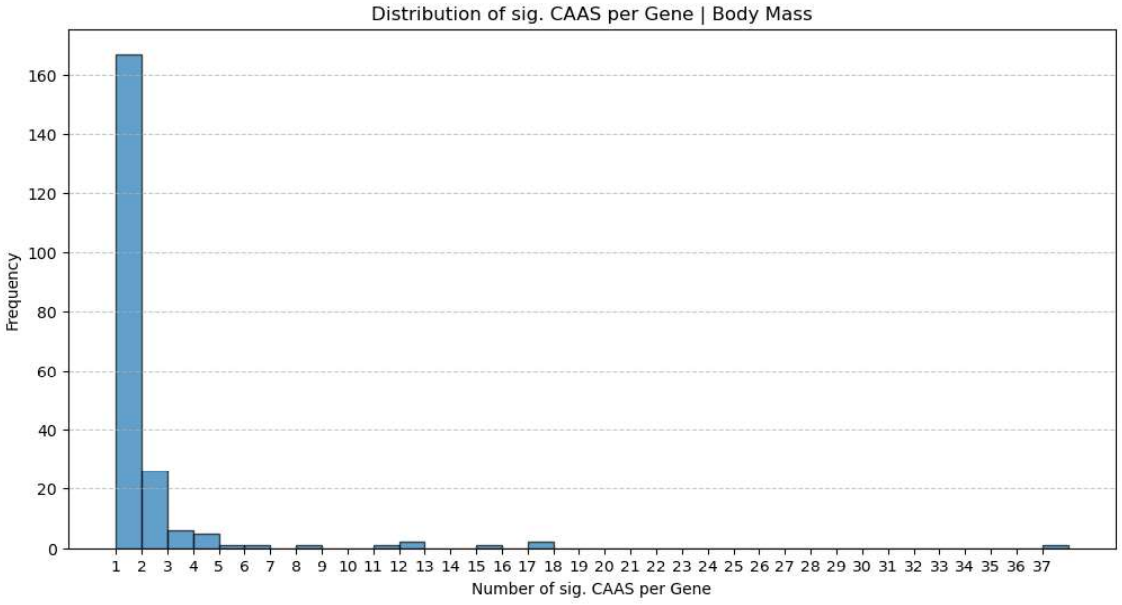


Fig 2. Distribution of the number of sig. CAAS per gene in the Body Mass dataset.

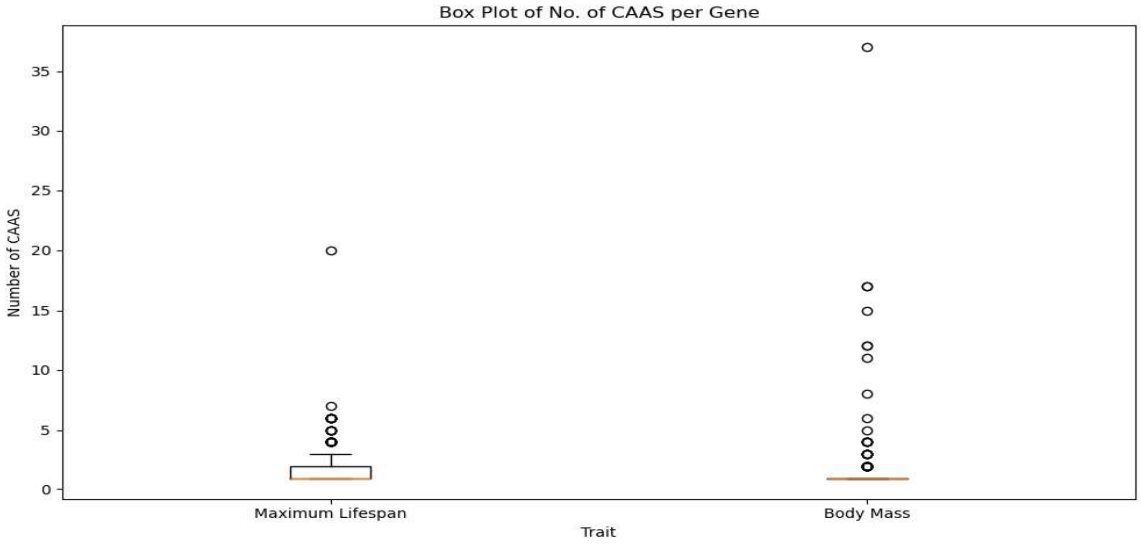


Fig 3. Boxplots merging the information contained in *Fig 2* and *Fig 3*.

2.1.4. Distribution of variability across the dataset and CAAS positions

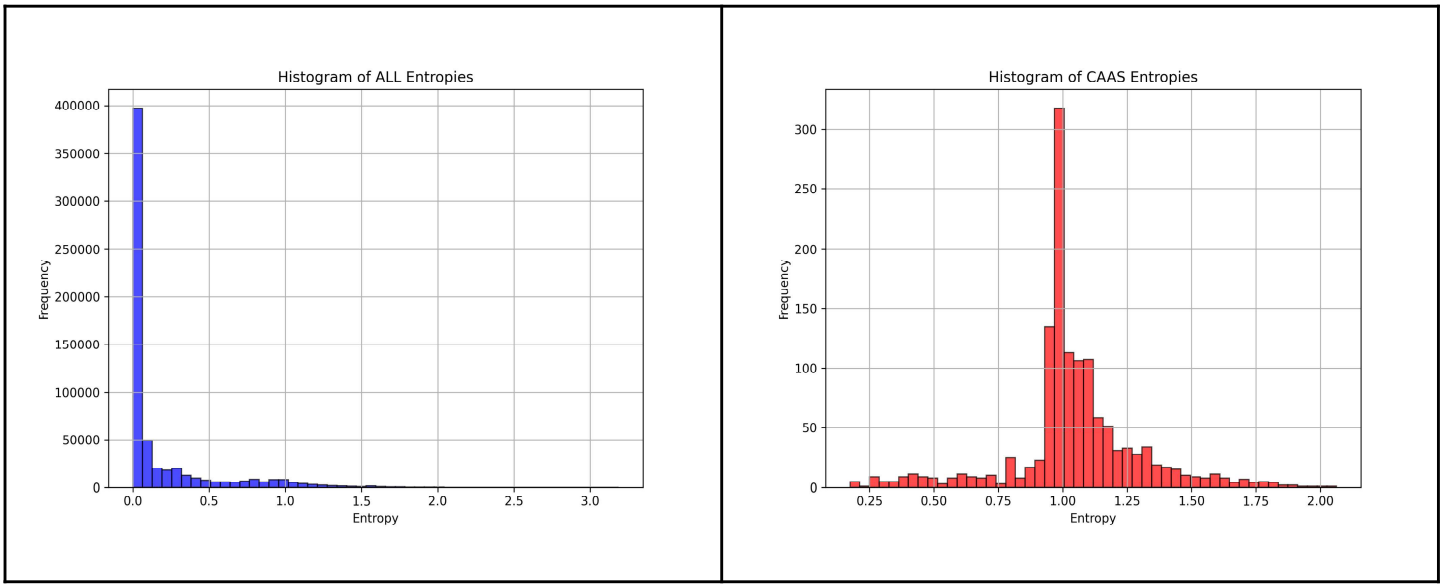
We evaluate the distribution of variability across CAAS and non-CAAS positions in the maximum lifespan and body mass datasets for two related reasons. The first is that once ancestral sequence reconstruction is performed, the higher the variability of the position across the multiple sequence alignment of the extant sequences, the more branches (i.e. transitions from one internal node to another) could be detected with changes for that position across the phylogenetic tree. Therefore the number of changes inferred across evolution, expressed in the number of branches hosting transitions for a position, is a (not necessarily linear) function of the variability of the position in extant species. The second reason, which expands on this, is that we assess the co-evolutionary strength between two intergenic positions based on three factors- the first is how many times *Position A* has changed, and the second being how many times *Position B* has changed and the third being how many times *Position A* and *Position B* have changed together. Consequently, if we consider a set of positions (CAAS and non-CAAS) that may be a part of the genetic architecture of a given trait, the statistical power to detect significant co-evolutionary patterns is greater for positions that have more variability.

To illustrate, let us consider two distinct cases. In the first case, we have an intergenic pair of positions, where both positions have changed many times across evolution, and many of these changes occurred in tandem with one another. In the second case, we consider an intergenic pair where each position has changed only a few times, and even if most of these changes were in tandem with each other. Naturally, the first case represents a much more convincing case of co-evolution that would be less probable to be a false-positive; the conclusion from exploring the two key aforementioned motivations to explore the distribution of variability across CAAS and non-CAAS is that it serves as a post-hoc sanity check, where aside from the genotype-phenotype correlation detected in CAAS positions, we also check whether there is an advantage to using this class of positions at a statistical level, increasing power to detect meaningful co-evolution and lowering potential false positives.

The measure we use to express the variability of a position is the Shannon entropy[32], a common metric in the genomics community used as a measure of how much information content is contained in a probability distribution. In the case of a position in an AA MSA, the position represents a probability distribution where each AA letter is an event with an attached probability. The maximum value (4.32) is reached if all AA letters of the 20 letter alphabet are equally probable in that position of the multiple sequence alignment. The minimum value (0) is reached if one AA state is fixed in all the species in that particular position.

$H(X) = -\sum_{i=1}^{n} P(x_i) \log_2 P(x_i)$	$H(site_j) = -\sum_{i=1}^{20} P(a_i) \log_2 P(a_i)$

Tab 2. The equation to the left is the general expression of the Shannon entropy while the expression on the right is the equation applied to a given site of a multiple sequence alignment computed over the 20-letter AA alphabet.



Tab 3. The figure to the left depicts the entropy distribution across all positions in the genes in the maximum lifespan dataset. The figure to the right depicts the entropy distribution across all CAAS positions in the same genes, showing a distinct difference in the variability required to infer a CAAS position as by definition a minimum amount of variability should to be present to identify AA states attached to long lived species and AA states attached to short lived species. CAAS positions make up 0.25% of all positions in the dataset. The power law distribution of entropy is a consequence of most positions being fixed, and the distribution of entropies across CAAS positions represents a departure from the entropy of non-CAAS loci.

2.2. Ancestral Sequence Reconstruction

The ancestral sequence reconstruction was performed for all genes across both the maximum lifespan and body mass datasets using PAML[33], which uses the marginal Empirical Bayes method to compute a probability distribution over the AA alphabet at each internal node in the given primate phylogenetic tree topology. The state selected at each node is the one with the highest probability, and the probability is also a direct measure of the certainty of the reconstruction. The term “marginal” means that the reconstruction of each node is done independently of all of the other nodes, where the alternative is the “joint” reconstruction,

meaning that a global solution is computed for all nodes at once. While the latter option is technically more correct, it suffers from an exponential increase in computational complexity as a function of the number of taxa in the multiple sequence alignment, and thus it was computationally infeasible. Aside from the distinction between joint and marginal ancestral state inference, the other alternative to the Empirical Bayes method is Maximum Likelihood estimation of ancestral nodes. While both methods have highly similar same accuracy and pitfalls- biasing ancestral protein sequences' physical properties toward extant species as ancestral states are computed as a function of extant sequences, in addition to not being able to infer more than one change on a given branch with either method, Maximum Likelihood estimation simply provides likelihoods of states at internal nodes, and thus cannot give a direct measure of the accuracy of the reconstruction, even if the AA state with the highest likelihood corresponds to the representative state at the ancestral node. The Empirical Bayes method goes one step further in giving the direct probabilities of the AA states at the given node, even if in most cases the representative AA inferred for a node is shared between both methods.

It should be noted that in our downstream analysis, even if some nodes are equiprobabilistic for all AA states, a change can be considered as having occurred on the branch formed by the equiprobable node and the following node(s). However, when performing analyses that directly take into account the identity of the amino acids involved in the transition, for instance in the co-transition matrix analysis, these events are filtered out. Thus, such (rare) cases are only used where we select two intergenic positions and evaluate the significance of the number of times they have changed on the same branches across primate evolution. While we may lack the resolution at a particular ancestral node to know the exact amino acid transition that occurred between that node and either of its two daughter nodes, a change must surely have happened on one of the two branches to produce the uncertainty, as if both daughter nodes shared the same amino acid state there would be no uncertainty in the AA state of the parent node. In the results for the maximum lifespan dataset that tabulate the intergenic CAAS pairs with at least one co-occurring change, only 0.08% of cases (cases being co-transitions considered between the pairs) had an unresolved AA transition involved. In the body mass dataset they correspond to 0.5% of cases. As mentioned previously, despite only being used to quantify the significance of co-evolutionary patterns, the structure of the pipeline easily allows removing these cases from the results should we choose to abandon them, and their extremely low frequency means their impact on the global significance of results is negligible.

Prior to the execution of the Empirical Bayes algorithm to reconstruct ancestral states, a maximum likelihood parameter estimation step is applied to certain model parameters which the Empirical Bayes algorithm uses to guide the reconstructions. This is not to be confused with maximum likelihood ancestral sequence reconstruction. Typically, this includes the estimation of an empirical amino acid substitution matrix, and optimisation of the branch lengths of the phylogenetic tree that captures relationships between the taxa present in the input MSA. In our execution, we have forced PAML to not rescale the branches based on information from the MSA, as the input (species) tree is the state of the art representative phylogenetic relationship between the taxa and rescaling this tree to a structure that is a hybrid between the gene tree and species tree would introduce overfitting. To further control for overfitting, an empirical amino acid substitution matrix was not used, and instead the widely used Jones-Taylor-Thornton (JTT) amino acid substitution matrix

was used. This general substitution matrix is a common input parameter in ancestral sequence reconstruction assays, and despite being derived three decades ago, it contains the key aspect of different rates of transitions as a function of biochemical properties. Ultimately, in our ancestral sequence reconstruction runs the only maximum likelihood parameter estimation done by PAML was in modelling the variation rate across different sites using a gamma distribution. This simply took into account heterogeneity in site variation. Naturally, this process is done for each gene and no parameter estimations are shared between any genes.

2.3. Co-occurring changes and co-transitions

From PAML's output for ancestral sequence reconstruction we have access to the reconstructed states for every position of each gene. From these positions, the CAAS positions were selected in accordance with each dataset (i.e. maximum lifespan or body mass). We record the following information for all of the CAAS positions:

1. The identity of the branches in which the position changed
2. The amino acid transition corresponding to each change

Following this procedure, we performed a simple computational experiment for each dataset where we created a library of all pairwise combinations of genes in the dataset, and generated a library of all pairwise combinations of CAAS positions between these genes. This allowed us to identify the branches in which two CAAS positions changed together, and the co-transitions that occurred on these branches i.e. the transition attached to each position on that branch. As a consequence, we were able to record the number of times the possible pairwise combinations of intergenic CAAS positions changed together as a precursor to the downstream analysis of quantifying the significance of the number of times two positions changed together as a measure of their co-evolutionary relationship.

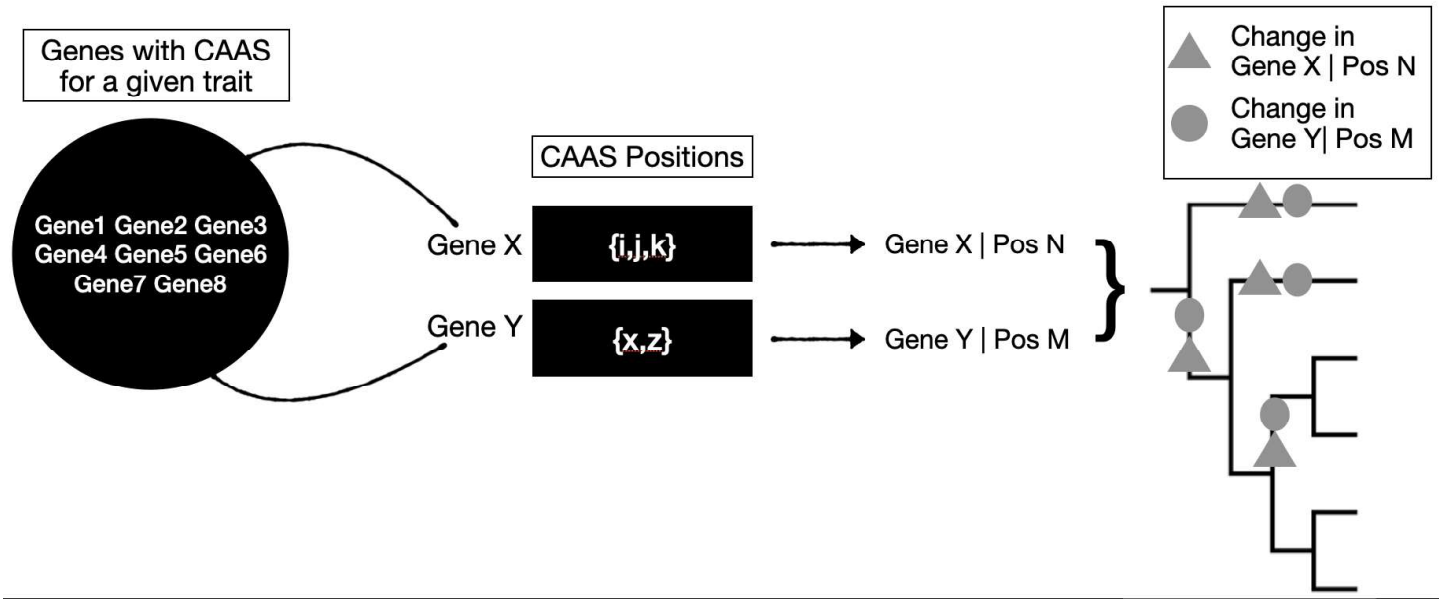


Fig 4. Scheme depicting a case where two genes are selected from all possible pairings of genes, and an intergenic CAAS pair is selected from these two genes, where the number and identity of branches in which these two positions together and individually are identified.

While we have not approached the problem from this angle yet, the setup of our data and pipeline allows us to also stratify by branch and note all the changes across all CAAS positions that occurred on that branch. This means that we could consider more than just pairwise changes but rather up to N positions at once on that branch with N being the number of CAAS positions that changed on the branch. Such an analysis would be interesting given that we have evidence that these positions and genes are representing an underlying co-evolving system driven by biological processes which we plan to identify, and thus we could select specific branches that bifurcate towards lineages of interest (for instance internal branches leading to long lived and short lived species) and study the relationship between the co-transition patterns and their impact on the underlying biology of the trait of interest. As an example to stimulate our imagination, distinct co-transitions in transcription factors in branches leading to long lived species may provide clues as to why the descending lineages are longer lived than other lineages as a function of their gene regulatory networks. Such analyses could contribute enormous to experimental assays designed to show epistasis as a factor in primate evolution and genetic architecture, as they would allow knowing combinations of alleles at a large number of loci in various internal branches that lead to primate lineages with certain trait characteristics (for instance short lived and long lived). The allelic combinations present in the primates themselves would further provide context on what alleles could be tested for.

2.4. Simulations approach for significance testing of co-occurring changes between intergenic CAAS pairs

Here we propose both a general description of the strategy undertaken to produce simulations that could quantify the statistical significance of the observed co-evolutionary patterns expressed in the number of co-occurring changes between intergenic CAAS pairs, and provide an additional mathematical overview as a complement.

2.4.1. General description

Following completion of the computational experiment of identifying the branches in which the intergenic pairs of CAAS positions changed together; a process in which the exact branches, co-transitions and the total number of co-occurring changes is identified, the natural goal is to quantify the statistical significance of the patterns of co-evolution. When we select two intergenic CAAS positions (x, y), we first take three key pieces of information into account:

1. The number of branches where x has changed (x_λ)
2. The number of branches where y has changed (y_λ)
3. The number of times (x, y) have changed together. $(x, y)_\lambda$

Using this information, we formulate a simulation strategy to realise our previously stated goal of quantifying the significance of the observed co-evolutionary patterns expressed in the number of co-occurring changes

between the intergenic CAAS pairs. The basic principle of this simulation strategy is the placement of mutations for each position randomly across the primate phylogenetic tree, and then counting the number of times the mutations occurred together. This process is repeated 10^5 times for each CAAS pair. This allows us to define a simulated distribution of co-occurring changes as a function of (x_λ) and (y_λ), and we can quantify the significance of $(x, y)_\lambda$ by counting the proportion of simulated cases where the number of co-occurring changes is greater than or equal to $(x, y)_\lambda$. Note that only one mutation could be placed on a branch, as in ancestral sequence reconstruction we could infer at most one change as having occurred on a branch, with this change resolving to the final difference between an ancestral node and a daughter node.

When placing the mutations, we make the initial assumption that the probability of a mutation occurring on a branch is proportional to the branch length. In fact, across both the maximum lifespan and body mass datasets, the correlation between the length of a branch and the total number of co-occurring changes (i.e. co-transitions) on that branch is 30%. While co-occurring changes are a subset of all the changes that have occurred on the given branches, the widely accepted proposition in evolutionary biology that the probability of a mutation on a branch is proportional to the branch length holds in our dataset. In fact a correlation of only 30% could be exploited to design a conservative random simulations strategy relative to what has been observed, as the branch length is not a *strong* predictor of a co-occurring change in *observation*; by using the branch lengths directly as a driver of placing mutations we increase the number of scenarios in our null model where mutations fall on longer branches, increasing the probability of both positions undergoing a mutation on these longer branches, ultimately leading to a higher frequencies of larger values of co-occurring changes in the null distributions generated. This can be regarded as an intentional bias created lower the significance of p-values of lower values of $(x, y)_\lambda$ through increasing the importance of branch lengths in the mutation probability of a position on a branch (and as a consequence the co-occurring mutations probability) relative to what has been observed in our inferred reality.

Given that the null distributions are created as a function of x_λ and y_λ, we note that many positions may share these values, in addition to the fact that the same particular position can participate in different pairwise combinations of intergenic CAAS pairs. This enables us to perform an optimisation trick where we do not need to create 10^5 scenarios of placing x_λ mutations for position x and y_λ mutations for position y from scratch each time (this would amount to 10^5 multiplied by the number of intergenic pairs of CAAS across the dataset) ; rather, we can create these scenarios independently $5(10^4)$ times for the interval of λ present across the dataset, and simply shuffle the scenarios and randomly draw a scenario of mutations for each position to count the number of co-occurring changes, repeating this 10^5 times to generate the null distribution of co-occurring changes between the two positions.

In other words, suppose that the interval of the number of individual changes of positions forming combinations of intergenic CAAS $\{(x_i, y_j)\}$ is $[2, 30]$ i.e. $\{x_{\lambda} \cap y_{\lambda}\} \epsilon [2, 30]$. For each $\boldsymbol{\lambda} \epsilon [2, 30]$, we simulate $5(10^4)$ mutation scenarios, which means placing mutations on branches through sampling branches without replacement $\boldsymbol{\lambda}$ times. Then, when we consider a pair (x, y), taking the corresponding set of simulations with $\boldsymbol{\lambda} = x_{\lambda}$, and $\boldsymbol{\lambda} = y_{\lambda}$, and in each iteration, we shuffle and randomly draw from the $5(10^4)$ mutation scenarios for $\boldsymbol{\lambda} = x_{\lambda}$, and $\boldsymbol{\lambda} = y_{\lambda}$ respectively. We then count the number of branches which have hosted a mutation for both positions and repeat this process up to 10^5 iterations to generate the null distribution to calculate the p-value of the observed $(x, y)_{\lambda}$.

2.4.2. Mathematical formulation and pseudocode

A. Quantifying the interval of the number of changes across all positions

Consider $S_{coc} = \{(x_i, x_j) \mid i, j \epsilon CAAS\}$, where this set represents all intergenic CAAS pair combinations with more than 0 co-occurring changes.

We can collapse this set into $S = \{x_i \mid i \epsilon S_{coc}\}$, which represents all CAAS positions having participated in an intergenic CAAS pair with more than 0 co-occurring changes.

Each position $x_i \epsilon S$, $\exists\, \boldsymbol{\lambda}(x_i)$, where $\boldsymbol{\lambda}(x_i)$ is the number of branches in which position x_i has changed.

We can define an interval for the range of values of $\boldsymbol{\lambda}$, $I = [a, b]$ where all $\boldsymbol{\lambda}(x_i) \epsilon I \; \forall \; x_i \epsilon S$.

Consequently, for all $i \epsilon I$ we perform the mutation placement strategy defined in the next step.

B. Mutation placement strategy and mutation scenario library generation

The goal of this step is to generate $5(10^4)$ scenarios (a “scenario” would resolve to a tree with a specified number of branches where mutations have been placed) for each $i \epsilon I$, which means that for the entire range of values of changes observed for each CAAS position belonging to the set of all intergenic CAAS pairs with more than one co-occurring change, we generate a library of mutation scenarios.

Building on the assumption that the probability of a mutation occurring on a branch is proportional to the branch length, we define the probability of a mutation occurring on a branch as:

$$P(B_i) = \frac{\text{Length}(B_i)}{\sum_{i=1}^{458} \text{Length}(B_i)}$$

The rationale for this is simply that in each mutation placement step, the probability of a mutation should be 1 as a mutation must be placed by definition, but which branch inherits the mutation? The selected branch is sampled based on its length relative to the entire length of the tree, meaning that a consistent mutation probability distribution is defined as per above by dividing each branch's length by the total length of the tree. It should be taken into account that the particular tree we are working with is ultrametric. Given that in ancestral sequence reconstruction we could infer at most one change on a branch, this is built into the mutation placement strategy.

As a consequence, once a branch j has been sampled, we enforce $P(B_j) = 0$, and rescale the remaining probabilities $P(B_i) \mid i \neq$ j by the length of the tree minus $Length(B_j)$. Thus when B_j is sampled:

$$P(B_j) = 0 \;\Rightarrow\; P(B_i) = \frac{\text{Length}(B_i)}{\sum_{i=1}^{458} \text{Length}(B_i) - \text{Length}(B_j)}$$

We continue sampling branches until we reach i, which is simply the number of changes that position has undergone, and this is the termination condition for one scenario, returning a tree with a given set of branches hosting mutations. The pseudocode follows to generate a scenario library:

*For **k**= 1 until **k** = 50 000:* (Generate up to 50 000 scenarios for the scenario library of a given $i \in I$)

*For **m**=1 until **m**=i:* (For one scenario, generate up to i mutations)

Sample branch B_k

Place mutation on branch B_k

Set $P(B_k) = 0$

Add B_k *to mutation-saturated set of branches* X

Re-normalise branch selection probabilities $P(B_c) \mid c \not\subset X$

continue

This process is repeated for all $i \in I$, generating a library of scenarios for the entire interval of individual changes observed across intergenic CAAS with non-zero co-occurring changes. The generation of these scenario libraries allows us to sample scenarios to generate null distributions of co-occurring changes, which in turn allows us to avoid the extremely expensive computations of performing this process repetitively when positions with the same number of changes are analysed for the co-occurrence statistical significance quantification.

C. Computing p-values for observed co-occurring changes across intergenic CAAS pairs

Once we have generated the mutation scenario libraries, we refer back to the three properties we introduced in the general description of the simulation strategy:

1. The number of branches where x has changed (x_λ)
2. The number of branches where y has changed (y_λ)
3. The number of times (x, y) have changed together. $(x, y)_\lambda$

Consider the set of all intergenic CAAS pairs with at least one inferred co-occurring change, S_{coc}. For all the pairs of positions in this set, let (x, y) be the selected pair for which we perform the following:

*For **k**= 1 until **k** = 100 000:*

*Shuffle mutation scenario library **i** =* x_λ (Randomly shuffle the order of the scenarios stored in the libraries)

*Shuffle mutation scenario library **i** =* y_λ

Randomly sample mutation scenario R_X

Randomly sample mutation scenario R_y

Record $(R_x, R_y)_\lambda$ ($(R_x, R_y)_\lambda$: Number of co-occurring changes between random mutation scenarios $R_X \cap R_y$)

The pseudocode above describes the process that generates the distribution of the number of co-occurring changes under our given model of random mutation placement as a function of the number of times each position has changed individually. The p-value of $(x, y)_\lambda$, the observed number of co-occurring changes for a given intergenic CAAS pair is obtained through counting the number of co-occurring changes under the null distribution greater than or equal to $(x, y)_\lambda$.

We have to note two limitations in this approach. One is the resolution at which we could compute the p-value; given the 10^5 instances of sampling mutation scenarios from the libraries (where each library is defined based on the number of mutations placed in the constituent scenarios), if a co-occurrence event has not been generated in this range of comparisons we can only assess that $p < 10^5$.

The second limitation is that $5(10^4)$ scenarios are generated for each library. This was done for computational efficiency, and in fact for scenarios, and perhaps some improbable scenarios of mutations could not be simulated entirely given the probabilistic approach of sampling branches to place mutations. Given the interval $I = [a, b]$ representing the range of positional changes, used to generate the mutation scenario libraries, for

smaller $i \in I$, we can cover a larger proportion of the space of all possible combinations of mutations on branches. For larger i, $5(10^4)$ scenarios would capture a fraction of all possible combinations of mutations placed on the constituent branches of the tree. Therefore, we lose resolution to define the full mutation combination space as the number of times each position has changed increases. While this is non-trivial, we ought to refer back to the massive increase in computational scale required to capture a larger proportion of possible scenarios in our simulations. Nonetheless, the bias we have created by using branch lengths as a *stronger* predictor of a branch being sampled for mutation placement can balance our decreasing resolution (when increasing number of changes for a position) to define the combinatorial space with respect to our goal of significance quantification; when sampling the mutational scenarios of two positions, mutations placed will have gravitated to longer branches for the mutation scenarios of both positions, increasing the likelihood of counting co-occurrences to generate a conservative null distribution of co-occurring changes, even if we have not explored the full space of possibilities.

2.5. Quantification of the magnitude of co-evolutionary signal across all intergenic CAAS pairs

The total number of possible co-occurring changes across a dataset is a function of the number of branches in the phylogeny M | $M = 458$, and the number of all intergenic CAAS pairs formed across the dataset. We can express the total number of pairwise combinations of intergenic CAAS between all the pairwise combinations of genes in the dataset as:

$$U = \sum_{i \in G} \sum_{\substack{j \in G \\ j \neq i}} (N_i \times N_j)$$

Where G is the set of all genes in the dataset, and N is the number of CAAS positions inferred in the corresponding gene and U is the number of all possible pairwise intergenic CAAS positions across all pairwise combinations of genes in the dataset. In a simplistic case, if we consider two positions, given a tree with M branches, the maximum number of times the two positions could have changed would be M, not taking into account additional information about the number of changes each position has undergone. Then, the total number of theoretically possible co-occurring changes across the dataset, C_T, would be computed by applying this logic to all possible intergenic CAAS pairs. This would mean $C_T = MU$. Given that $M = 458$ for our tree topology, we deduce:

$$C_t = MU = 458 \left(\sum_{i \in G} \sum_{\substack{j \in G \\ j \neq i}} (N_i \times N_j) \right)$$

If we take C_{obs} as the total number of observed co-occurring changes for a dataset, and C_0 as the number of times a pairwise intergenic CAAS pair has been evaluated to have 0 co-occurring changes, we can quantify C_0

as $C_0 = C_t - C_{obs}$, where C_{obs} is computed as the sum of all instances in which a branch harboured a co-occurring change across all intergenic CAAS pair combinations. We have to empirically compute C_T as the number of pairwise CAAS combinations between two genes can vary, as the number of CAAS per gene is specific to each gene, but as an exercise let us compute the theoretical lower bound of the total number of co-occurring changes across the dataset, $L(C_T) \mid L(C_T) < C_T$.

This is done through a simplifying assumption; the median of the skewed discrete distribution defining the number of CAAS per gene across the dataset is 1 (*Fig. 2, Fig. 3)*, meaning that independently of gene length (as the correlation between gene length expressed in amino acids and number of CAAS is ~ 0) We can typically expect each gene to contribute one position to the set of all intergenic CAAS pairs it forms with another gene, as the gene is expected to have a single CAAS position. This would simply the number of intergenic CAAS pairs being formed between each pair of genes to 1, and it simplifies the expression of $L(C_T)$ where $L(C_T)$ would simply resolve to the number of all pairwise combinations of genes in the dataset. We then propose:

$$L(C_t) = ML(U) = M\frac{N_g(N_g - 1)}{2}$$

Such that N_g is the number of genes with at least one CAAS position for the given trait. In the case of the maximum lifespan dataset, where $N_g = 883$, this would resolve to:

$$L(C_t) = ML(U) = 458\frac{883(883 - 1)}{2} = 178346574$$

Knowing how much of the total possible co-occurring changes have been observed in our data allows the quantification of the sparsity of the signal we have observed, and in comparative analyses where we sample non-CAAS within the same genes, or sample positions from random subsets of genes and perform the same analysis applied to the CAAS positions, it provides an angle to perform comparisons to assess differences in the magnitude of co-evolutionary patterns.

However, we have to take into account that it would be bizarre for a position to have changed on all 458 branches, let alone having undergone such an amount of co-occurring changes with another position. In reality, if we consider an intergenic CAAS pair (m, n), we denote (μ_m, μ_n) as the number of times each corresponding position has changed. Then, the maximum number of observed co-occurring changes possible for this intergenic CAAS pair is $min(\mu_m, \mu_n)$. This would modify the definition of C_T to:

$$C_t = \sum_{i \in G} \sum_{\substack{j \in G \\ j \neq i}} \sum_{m \in S_i} \sum_{n \in S_j} \min(\mu_m, \mu_n)$$

Where G is the set of all genes with a CAAS in the dataset, S_i is the set of CAAS positions of gene i, S_j is the set of CAAS positions of gene j, μ_m is the number of times CAAS position m has changed and μ_n is the number of times CAAS position n has changed. In this case, it is more ideal to compute C_T than $L(C_T)$ as the layer of complexity in estimating the error of $L(C_T)$ is not trivial due to the presence of the term $min(\mu_m, \mu_n)$. Once C_T has been computed, we simply need to compute C_{obs} as the total observed number of co-occurring changes- across all intergenic CAAS pairs in the dataset, and this allows us to evaluate the ratio C_{obs}/C_T as the fraction of all possible co-occurring changes observed across the intergenic CAAS pairs- a crude measure of how much of the total possible coevolution has been realised in our dataset.

2.6. Expectation of the magnitude of co-evolutionary signal under the random mutation simulation strategy

The C_{obs}/C_T ratio defined in the previous section can be compared to the total number of co-occurring changes across the datasets we would expect under a random mode of evolution (i.e. random mutation placement strategy), where we take into account the null distributions of $(x, y)_\lambda \mid n = 10^5$ defined for each possible x_λ and y_λ where $(x, y)_\lambda$ is the number of co-occurring changes observed by placing x_λ and y_λ mutations randomly based on the strategy described in 2.4.Simulations. For each intergenic CAAS pair's null distribution of co-occurring changes, we can compute the median and the average, and sum these values for each metric across the datasets to obtain the total number of expected number of co-occurring changes depending on the representative value taken for the distributions. This facilitates a comparison between the C_{obs}/C_T and the respective null expectations, $E(C|M_{H0})/C_T$ and $E(C|\mu_{H0})/C_T$- where M is the median, μ is the mean and $H0$ is the null distribution of co-occurring changes. Formally we can define $E(C|M_{H0})$ and $E(C|\mu_{H0})$ as:

$$\mathbb{E}(C \mid \mathcal{M}[H_0]) = \sum_{i \in G} \sum_{\substack{j \in G \\ j \neq i}} \sum_{m \in S_i} \sum_{n \in S_j} \mathcal{M}[H_0](\lambda_m, \lambda_n)$$

$$\mathbb{E}(C \mid \mu[H_0]) = \sum_{i \in G} \sum_{\substack{j \in G \\ j \neq i}} \sum_{m \in S_i} \sum_{n \in S_j} \mu[H_0](\lambda_m, \lambda_n)$$

2.7. Pairwise co-transition analysis: are co-transitions between significantly co-evolving CAAS pairs associative or independent?

2.7.1. Multigraph definition of transitions and co-transitions

Taking the set of all co-transitions and corresponding transitions aggregated from all the intergenic CAAS pairs in the co-occurring changes analysis, we model the system as an undirected multigraph, where each node is an amino acid transition. Each edge between two nodes represents one exact count of a co-transition, such that up to N edges between transitions V_i and V_i means N co-transitions have occurred across all intergenic CAAS pairs in the dataset that contain those two particular transitions- recalling that a co-transition is inferred if two CAAS positions from two distinctly different genes have been inferred to have changed on the same branch of the phylogenetic tree based on ASR. Naturally, this undirected multigraph allows self loops, where a self loop equates to a co-transition where the coinciding transitions between the loci had the same identity i.e. a self loop is a single count of $AA_{i->j} \cap AA_{i->j}$ defined on the vertex $V: AA_{i->j}$. Given 20 letter AA alphabet S, we have $|S^2|$ vertices corresponding to the $20 \cdot 20$ possible AA transitions, however transitions where $AA_{i->i}$ have degree of 0 by default, as they identify a state remaining the same on a particular branch, and we have not taken conservation of states into account in our study of co-occurring changes and coinciding transitions.Note that the undirected property is due to difficulties in ASR methodology determining which of the two transitions across the CAAS pair occurred first, making the establishment of directionality an exercise we have omitted.

The abstract representation of the transitions and co-transitions using an undirected multigraph with self loops serves two key primary purposes of our analysis, where the stated goal is investigating the hypothesis of associative co-transitions based on evidence in the literature that epistasis is a driving force of molecular evolution as the component of the genetic architecture upon which selection acts upon. The first purpose is that a symmetrical co-transition matrix could be defined based on the architecture of the transition multigraph. This allows us to both represent the idea and analysis in a generalisable and visualisable way, but also allows tractability in the implementation of matrix and vector based operations to conduct the statistical tests proposed for this analysis. The second purpose is for the generation of a null distribution to represent randomised co-transition matrices conditioned on an independent mode of co-transition across the dataset, where the multigraph approach allows using powerful techniques such as the configurational model to generate ensembles of randomised versions of the same graph, from which the co-transition matrices defined in the following section could be defined to perform our statistical operations.

2.7.2. Definition of the co-transition matrix

The detection of co-occurring changes across the set of all intergenic CAAS pairs, defined by selecting a position each from all pairwise gene combinations in the dataset, and the follow-up evaluation of the statistical significance of these CAAS positions allowed us to filter the set of all intergenic CAAS positions such that only pairs that were statistically significant were considered for the follow up analysis. Given that the number of

times such pairs changed together was not random through significant deviation from what was expected under the null distribution, generated from randomly placing mutations for based on the number of times each position had changed- and then counting the number of co-occurring changes, it then became of interest to test whether the pattern of co-transitions underlying the co-occurring changes exhibited properties that suggest changes are not independent from one another.

Another closely related goal was to represent co-transitions in a data structure that could be easily interpreted and communicated, and general statistical techniques to detect associations and non-independence could be applied to such a data structure. Accordingly, we implemented the novel idea of co-transition matrices, where the 20 letter amino acid alphabet produces a 400 by 400 matrix, such that the rows represent a particular transition from one amino acid to another, and the columns represent one possible transition from an amino acid to another, containing the full set of possible transitions across the amino acid alphabet.

More formally given the amino acid alphabet S with $|S|$ amino acids, we define the co-transition matrix as:

$$M \in \mathbb{R}^{|S|^2 \times |S|^2}$$

Where $|S^2|$ is the number of all possible amino acid transitions defined on the alphabet S, and each entry is the number of times a co-transition was observed, meaning the total number of branches harbouring that particular co-transition across all intergenic CAAS pairs analysed. We represent this as:

$$\forall i,j,\ M_{(i,j)} = f\left((\mathrm{AA}_x \to \mathrm{AA}_y), (\mathrm{AA}_m \to \mathrm{AA}_n)\right),\ \text{with}\ \mathrm{AA}_x, \mathrm{AA}_y, \mathrm{AA}_m, \mathrm{AA}_n \in S \times S$$

$$\text{s.t.}\ \forall x,y,m,n,\ (x = y \vee m = n) \Rightarrow M_{(i,j)} = 0, \quad \text{where}\ i = f(\mathrm{AA}_x \to \mathrm{AA}_y),\ j = f(\mathrm{AA}_m \to \mathrm{AA}_n)$$

Where the co-matrix has the value 0 for entries that have either of the co-transitions consisting of a transition from an amino acid state to itself, meaning conservation, (e.g. A->A). This is obvious given the fact that a co-transition means that on a particular branch of the phylogeny, two positions had a coinciding change; a co-transition could not be defined if only one position forming an intergenic CAAS pair changed.

In parallel to the co-transition matrix, a “traditional” transition matrix was defined where all the transitions involved in co-transitions were aggregated into a transition matrix T defined as:

$$T \in \mathbb{R}^{|S| \times |S|}$$

Where $T_{(i,j)}$ denotes a transition from amino acid j to amino acid i. On the 20 letter amino acid alphabet, this defines the 20 by 20 amino acid transition matrix we have constructed from de-coupling the co-transitions and simply counting each observed transition within this matrix. When computing expected values under an independent model of co-transition (no dependency between transitions) for the purpose of hypothesis testing, either the marginal sums of the co-transition matrix could be used, or the transition matrix itself as the two approaches are mathematically equivalent.

2.7.3. Hypothesis formulation

The co-transition matrices provide the setting to test whether the frequencies of particular co-transitions could be described under a model of independence, such that the frequency $M_{(i,j)}$ could be computed as a product of the independent probabilities of transitions i and j where $i \epsilon S \times S \ \wedge \ j \epsilon S \times S$; the competing hypothesis is that there is an association (deviation from independence) between transitions i and $j \ \forall \ i \epsilon S \times S \ \wedge \ j \epsilon S \times S$ which has produced the observed frequencies of co-transitions.

Generally, we are testing whether the global patterns of co-transitions, quantified by their frequencies, display an associative structure or are simply a function of an independent model of co-transition based on the frequencies of the constituent transitions that define the co-transitions.

2.7.4. Computation of the expectation under the null model of independence

There are two approaches to compute the expectation of each particular co-transition. First, we note that the independent probability of two co-occurring transitions i and j would be:

$$P(i \cap j) = P(i) \cdot P(j)$$

Then, if the total number of co-transitions is N_T, the expected number of times transitions i and j would have co-occurred to produce a particular co-transition under the assumption of independence is:

$$N_T \times P(i \cap j) = P(i) \cdot P(j) \times N_T$$

The first approach to quantify this value for all of the defined co-transitions is to use the co-transition matrix directly, where the probability of each transition is computed as the marginal sum divided by the total number of co-transitions.

$$\mathbb{E}(M_{(i,j)}) = \left(\frac{\sum_{j=1}^{|S|^2} M_{(i,j)}}{N_T} \times \frac{\sum_{i=1}^{|S|^2} M_{(i,j)}}{N_T} \right) \times N_T \quad | \quad N_T = \sum_{i=1}^{|S|^2} \sum_{j=1}^{|S|^2} M_{(i,j)}$$

Or alternatively, we can represent this equation as:

$$\mathbb{E}(M_{(i,j)}) = \left(\frac{\sum_{j \in S \times S} M_{(i,j)}}{N_T} \times \frac{\sum_{i \in S \times S} M_{(i,j)}}{N_T} \right) \times N_T \quad | \quad N_T = \sum_{i \in S \times S} \sum_{j \in S \times S} M_{(i,j)}$$

The previous approach is the one used in our computations, however, given that we have a transition matrix that describes the frequencies of the transitions, it is the second approach that could be undertaken to compute the expected values of the co-transitions under an independent model. We have empirically validated that this second approach produces expected values in agreement with the method of using the marginal sums of the co-transition matrix. Thus, if we represent transition i as $m \rightarrow n \mid m \epsilon S \ \wedge \ n \epsilon S$, and transition j as $v \rightarrow w \mid v \epsilon S \ \wedge \ w \epsilon S$, letting T represent the transition matrix such that $T_{(n,m)}$ equates to the frequency of the transition $m \rightarrow n$, we can compute the expected value of a co-transition under independence as:

$$\mathbb{E}(m \to n,\ v \to w) = \frac{T_{(n,m)}}{N_T} \times \frac{T_{(w,v)}}{N_T} \times N_T \quad | \quad N_T = \sum_{x=1}^{|S|} \sum_{y=1}^{|S|} T_{(x,y)}$$

Or alternatively, we can represent this as:

$$\mathbb{E}(m \to n,\ v \to w) = \frac{T_{(n,m)}}{N_T} \times \frac{T_{(w,v)}}{N_T} \times N_T \quad | \quad N_T = \sum_{x \in S} \sum_{y \in S} T_{(x,y)}$$

Noting that if $i = m \to n \wedge j = v \to w$ then:

$$T_{(n,m)} = \sum_{j=1}^{|S|^2} M_{(i,j)} = \sum_{j \in S \times S} M_{(i,j)}$$

Given this formulation of the expected co-transition frequencies under the model of independence, we used this to compute the expected co-transition frequencies for the observed co-transition matrix computed from the data, and for randomised matrices used to generate the null distribution of the test-statistic (the distribution of the the statistic representing the deviation between observed and expected values), which will be discussed in the following subsections detailing this analysis.

2.7.5. Quantification of the deviation between observed co-transition frequencies and expected values under the assumption of independence between co-transitions

To test the hypothesis of an independent or associative structure of co-transition frequencies defined in the previous section, we first compute the expected values for each co-transition under an independent model. The following step resolves to the computation of a test statistic to measure the deviation of the observed values from the frequencies expected under the null model. To do so, we can consider the co-transition matrix as a contingency table on a conceptual level, such that a co-transition is defined as the joint occurrence of the two categorical variables denoting transitions, where the levels of each variable are the transitions themselves. Given the two other prerequisites where all entries are non-negative integer entries, and that the marginal sums across the rows and columns are meaningful and descriptive, each cell of the matrix is then regarded as the joint frequencies of a particular level of each of the transitions, and the fundamental requirements of formulating this matrix as a contingency table are met. This is especially true considering that we are testing for the presence of an association or independent structure.

The most widely used measure of testing a contingency table is the chi-square test of independence, however, the matrix sparsity and proportion of counts $M_{(i,j)} < 5$ in this case violate fundamental conditions required to perform the test without inflating the significance level. Indeed if we consider the chi-square test statistic's formula:

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c} \frac{(O_{(i,j)} - E_{(i,j)})^2}{E_{(i,j)}}$$

We clearly see that low counts would inflate the value of the test statistic as the deviation from the expected values would be large. In the table below we represent the sparsity and percentage of cases where $M_{(i,j)} < 5$:

	Maximum Lifespan	**Body Mass**
$M_{(i,j)} = 0$	85.10%	88.14%
$M_{(i,j)} < 5$	92.06%	99.07%
$M_{(i,j)} = 0 \vert i \vee j \notin \{AA_x \rightarrow AA_x\}$	84.61%	87.65%
$M_{(i,j)} < 0 \vert i \vee j \notin \{AA_x \rightarrow AA_x\}$	91.57%	98.58%

Tab 4. Representation of the percentage of co-transitions defined over the amino acid alphabet where the frequency was 0 or less than 5; given that a row or column includes a transition from an amino acid state of itself the frequency is 0 by default; the percentage of sparse and low frequency cells are also provided while controlling for this value.

We observe that the co-transition matrices we are working with are very sparse, and a large proportion of defined co-transitions have very low frequencies. While this may suggest that a large number of co-transitions were not allowed by evolution, the lower sparsity and proportion of low-frequency cells observed for the maximum lifespan co-transition matrix, relative to the body mass co-transition matrix, suggests that the statistical power to explore the full space of possible co-transitions is dependent on the number of positions (and consequently position pairs) analysed, given that the maximum lifespan dataset contains approximately $9(10^5)$ possible pairs of intergenic CAAS positions and the body mass dataset contains $8(10^4)$ pairs of intergenic CAAS positions. The correction performed for the sparsity and low frequency metrics was a simple correction for the fact that by definition transitions from an amino state to itself will have 0 frequency. This means that any element $M_{(i,j)}$ where i or j is a transition from one amino acid state to itself will have 0 counts. Thus if we let any of the metrics be N, N_{adj} and consider that for the 20 letter amino acid alphabet S, $|S| = 20$, and $|S|^4$ is the number of possible co-transitions defined:

$$N_{\text{adj}} = N - \left(\frac{|S| \cdot |S| + |S| \cdot |S| - 20}{|S|^4}\right) \cdot 100 = N - \left(\frac{2|S|^2 - |S|}{|S|^4}\right) \cdot 100$$

$$N_{\text{adj}} = N - \left(\frac{2 \cdot 20^2 - 20}{20^4}\right) \cdot 100 = N - \left(\frac{780}{160{,}000}\right) \cdot 100 = N - 0.4875$$

The explicit reasoning for this formula is that $2|S|^2$ is the number of times where either a row or a column corresponds to a co-transition where there is a self-transition that is 0 by definition. We subtract 20 from this

value to account for double counting of diagonal elements. Then, the ratio of this value over the total number of possible co-transitions in the co-transition matrix (i.e. cells) is taken and then subtracted from the sparsity or low-frequency metric as a percentage, given that both metrics are presented as a percentage.

Given the prevalence of sparsity and low-frequency co-transitions, we undertook a dual approach where we used a test statistic that is less sensitive to the deviation between low observed values and expected values, while simulating randomised variants of the co-transition matrix to generate the null distribution of the test statistic, as a chi-square approximation of the asymptotic distribution of the test statistic could not be used due to the significant deviations from the properties required to perform such an inference in a statistically sound manner.

The test statistic used to measure the deviation between observed and expected values for our inferred co-transition matrix is the square root Tukey-Freeman correction of the chi-square test statistic[34], expressed as below for each (co-transition) cell of the co-transition matrix:

$$X^2_{\mathrm{TF}} = \sum_{i,j} \left(\sqrt{O_{ij}} - \sqrt{E_{ij}} \right)^2$$

This statistic stabilises the variance when the cell counts are low, avoids division by low expected values and consequently controls for skewness in the distribution of the cellwise differences. The major caveat however, is that this test statistic penalises larger deviations between observed and expected values in cases that are not a consequence of low observed frequencies relative to high expected values, meaning that there is a tradeoff between statistical power and controlling for inflated significance, where integration of “desirable” large deviations into the test statistic is penalised in turn for penalising large deviations caused by a mismatch between low observed frequencies and higher expected values. Given that this test statistic does not have a defined asymptotic distribution, we simulated random tables under given constraints (see next section) to approximate the p-value obtained, where we took the observed test statistic for the maximum lifespan and body mass co-transition matrices respectively, and counted the number of times a value equal to or greater was observed in the null distribution of the test statistic (generated from the random tables.)

2.7.6. Simulation of randomised co-transition matrices to generate the null distribution of the Tukey-Freeman correction of Chi-square test statistic

To generate the null distribution of the test statistic, we need to simulate randomised tables under the assumption of independent co-transitions. However, creating a symmetrical randomised matrix with the dimensions as our observed data is not a trivial task. This is because it would require either sophisticated monte-carlo methods with high computational cost to ensure the necessary constraints and properties of the co-transition matrix are met, or the use of algorithms such as the Patefield randomised contingency table algorithm[35], which despite satisfying marginal totals,do not ensure symmetry in the output.

To address this, we use a multigraph representation of the transitions and apply the multigraph configurational model[36] to generate 10^5 random multigraphs. These simulate the scenario of independent co-transition. Each of these multigraphs can then be encoded as a co-transition matrix and used in the statistical hypothesis testing described above. A key feature of the multigraphs generated using the configurational model is that they preserve the original degree sequence of the multigraph. This is equivalent to creating randomised co-transition matrices where the marginal sums(rows and columns totals) of the co-transition matrix remain constant across all randomized versions. In other words, the total number of transitions for each amino acid is fixed, but the pairing of co-transitions and their frequencies is randomised. The multigraph configurational model ensures that the random multigraphs are uniformly sampled from the space of all possible multigraphs with the specified degree sequence (i.e fixed transition counts per AA).

3. Results

3.1. Magnitude of the co-evolutionary signal across all intergenic CAAS pairs

	C_t	C_{obs}	C_{obs}/C_t
Body Mass	350 031	66 800	$19.7 \cdot 10^{-2}$
Maximum Lifespan	3 385 982	667 467	$19 \cdot 10^{-2}$

Tab 5a. Summary of the computed C_t values and observed C_{obs} values for the maximum lifespan and body mass datasets. The third column represents what fraction of total possible co-occurring changes has been fulfilled in the observed data, amounting to roughly ~20%.

	$E(C\|M_{H0})$	$E(C\|\mu_{H0})$	$E(C\|M_{H0})/C_t$	$E(C\|\mu_{H0})/C_t$	$C_{obs}/E(C\|\mu_{H0})$
Body Mass	510	8457	10^{-3}	$2 \cdot 10^{-2}$	9.85
Maximum Lifespan	2162	62861	$6 \cdot 10^{-4}$	$1.8 \cdot 10^{-2}$	10.5

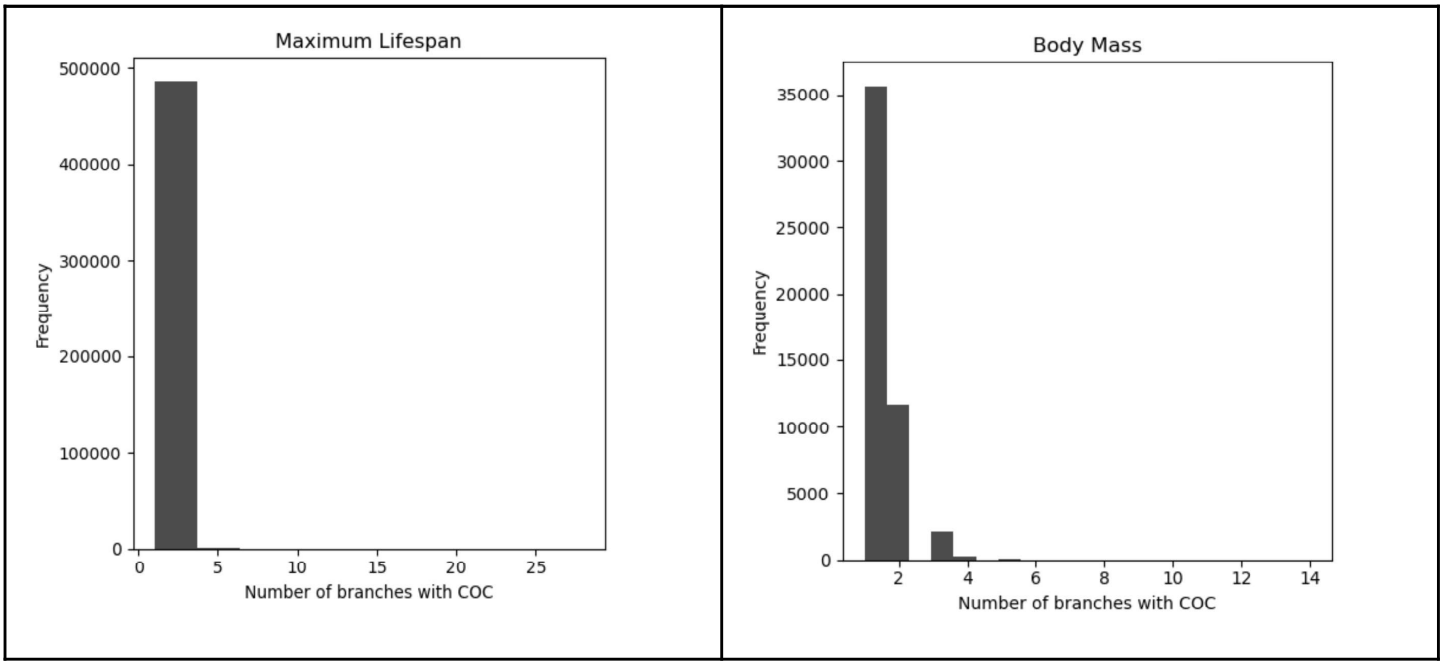
Tab 5b. Summary of the expected numbers of co-occurring changes across all intergenic CAAS pairs for the maximum lifespan and body mass datasets, created from summing the mean and the median from the set of all null distributions of co-occurring changes defined for each intergenic CAAS pair per the random mutations simulation strategy. We can observe that out of the total possible number of co-occurring changes estimated, across both datasets ~10 more co-occurring changes have been observed relative to the summed mean of the null distributions of co-occurring changes.

3.2. Significance of observed co-occurring changes between intergenic CAAS pairs

3.2.1. Atlas of results of significance tested co-occurring changes analysis results

Following the generation of the all pairwise combinations of genes, and the ensuing pairwise combination of gene pairs' CAAS positions (intergenic CAAS pairs), the number and identity of branches (co-occurring changes) in which positions changed together was obtained, with the transitions underlying the co-transitions also attached to this information. The simulations approach, which quantifies the significance of the number of branches in which positions co-evolved, assigned p-values (number of scenarios in which the observed number of co-occurring changes was present in the distribution of co-occurrence counts across simulated scenarios) to the intergenic CAAS pairs' observed number of co-occurring changes across our inferred trajectory of primate evolution. The full tables, one for the maximum lifespan dataset (*Supplementary Table 1)*, and one for the body mass dataset (*Supplementary Table 2)* are available with a data dictionary (*Supplementary Tables 1 and 2 Data Dictionary)* describing the variables in both tables, provided in the *Supplementary Materials*. These tables summarise information on all the intergenic CAAS pairs, across all pairwise combinations of genes in the corresponding datasets, where at least one co-occurring change was observed.

3.2.2. Statistical summary of observed results of co-occurring changes analysis



Tab 6. Distribution of the number of co-occurring changes across the maximum lifespan and body mass datasets. Each unit of observation is the number of co-occurring changes of a particular intergenic CAAS pair; the distribution is defined by the number of co-occurring changes across all intergenic CAAS pairs in the dataset with at least one observed co-occurring change.

From the distribution of the number of co-occurring changes computed across all distinct intergenic CAAS pairs in the dataset, we can observe a skewed distribution for both datasets where the majority of the number of the number of co-occurring changes are observed in the interval $[1, 4]$. By design, given that the simulations approach places mutations on branches using branch lengths, with a bias towards longer branches, the scenarios that led to lower counts of observed co-occurring changes are likely to be plentiful in the null distributions and as a consequence these scenarios would not have significant p-values. The exception to this however, is when for a given CAAS pair (m, n), $Obs_{COC} = min(\mu_m, \mu_n)$ where Obs_{COC} is the observed number of co-occurring changes for a given intergenic CAAS pair. For both the maximum lifespan and body mass datasets, $\sim 100\%$ of the previously mentioned cases had statistically significant simulation p-values under the Benjamini-Hochberg FDR correction. A subset of these results demonstrated perfect co-evolution, where $Obs_{COC} = \mu_m = \mu_n | (m, n)$. For the maximum lifespan dataset, this spans 1% of intergenic CAAS pairs with significant BH-FDR corrected p-values, and 0.4% of all intergenic CAAS pairs across the maximum lifespan dataset. For the body mass dataset, these cases were 0.2% of all intergenic CAAS pairs with significant BH-FDR corrected p-values, and 0.04% of all intergenic CAAS pairs analysed. We summarise these statistics below in *Tab 6.*

Statistic	Maximum Lifespan dataset	Body Mass dataset
Number of analysed intergenic CAAS pairs where: $Obs_{COC} > 1 \mid (m, n)$	488 128	49 844
Number of intergenic CAAS pairs with BH-FDR corrected significant p-values	162 882	8509
Number of intergenic CAAS pairs with Bonferroni corrected significant p-values	100 651	4444
Number of intergenic CAAS pairs with un-observed $Obs_{COC} \mid (m, n)$ across simulations i.e. p-value=0	100 651	4444
Number of intergenic CAAS pairs with $Obs_{COC} = min(\mu_m, \mu_n) \mid (m, n)$	34244	1461
Number of intergenic CAAS pairs with $Obs_{COC} = \mu_m = \mu_n \mid (m, n)$	2218	22

Tab 7. Various descriptive statistics computed from the results of the co-occurring changes analysis applied to the maximum lifespan and body mass datasets, such that (m, n) denotes an intergenic CAAS pair, and each count in the table denotes a single experiment where two intergenic CAAS are selected, the co-occurring changes are identified and consequently the simulation approach is integrated to quantify the number of times the observed value was present in random mutation scenarios.

To put into perspective the amount of intergenic CAAS pairs with a statistically significant number of co-occurring changes under our random mutations simulation strategy, or the other subcategories of significant intergenic CAAS pairs of relevance such as those where $Obs_{COC} = \mu_m = \mu_n$ or $Obs_{COC} = min(\mu_m, \mu_n)$,we have computed the number of all intergenic CAAS pairs present in each dataset, where this number is simply the number of combinations of CAAS pairs across pairwise combinations of all genes in each dataset. The equation below, used previously in the initial construction of the logic to quantify the magnitude of co-evolutionary signal present in the dataset, provides this number.

$$U = \sum_{i \in G} \sum_{\substack{j \in G \\ j \neq i}} (N_i \times N_j)$$

For the maximum lifespan dataset, this number is $U = 882902 \simeq 9(10^5)$, whereas for the body mass dataset it is $U = 77249 \simeq 8(10^4)$, noting that this number is smaller for the body mass dataset as there are less genes and as a result, there are less CAAS positions that make up the combination space; in the case of the maximum lifespan dataset, $1.6(10^5)$ out of $9(10^5)$ intergenic CAAS pairs demonstrating a significant co-evolutionary signal, in addition to $\sim 20\%$ of all possible co-occurring changes having taken place provides evidence that a common architecture of biological processes, pathways and interactions could be driving this observed signal.

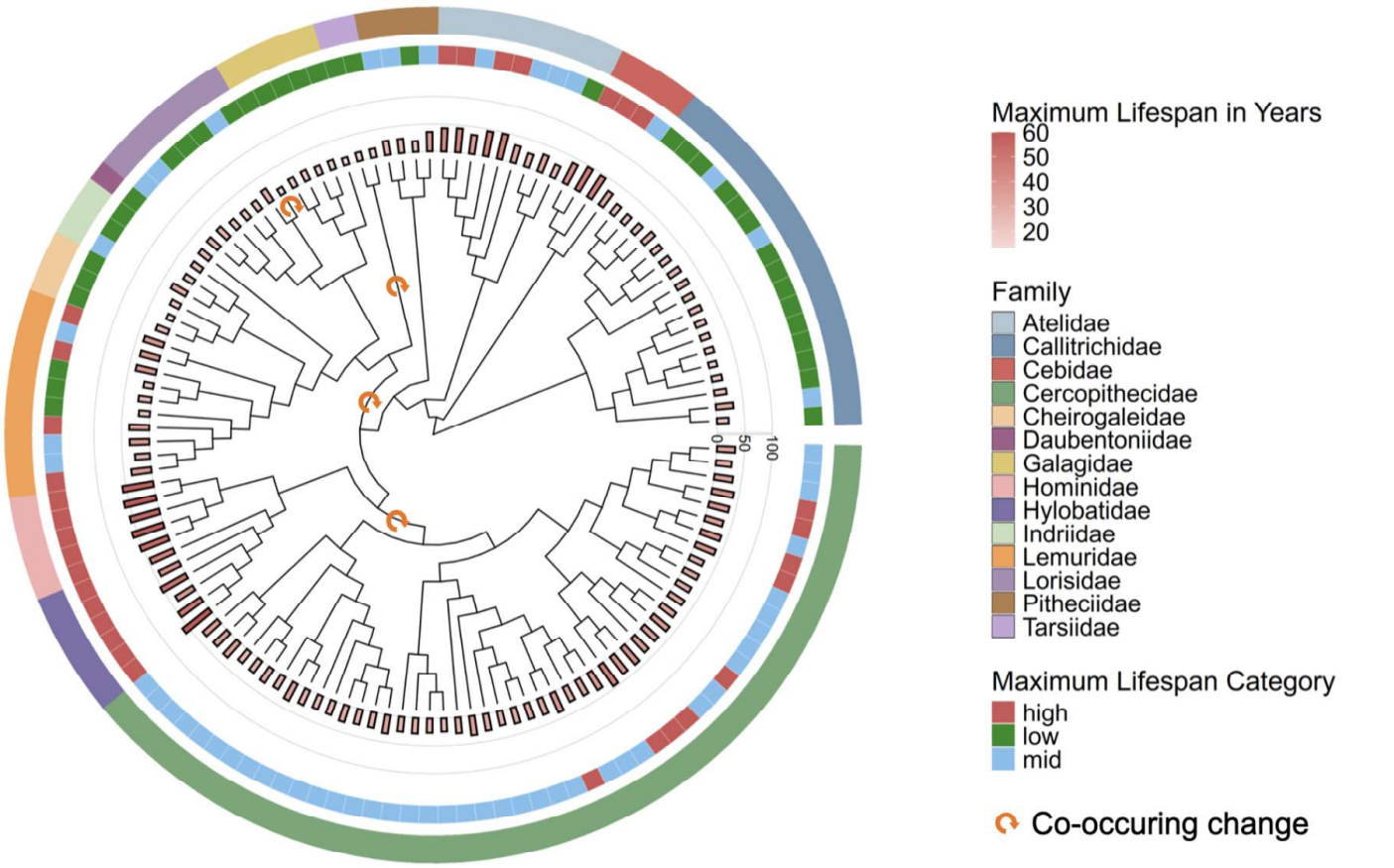


Fig 5. In the analysis performed between a pair of CAAS positions, one from the gene *ERCC6* and one from *TIMELESS,* both involved in DNA repair, we can see the four instances where these positions changed together- in fact the only times these positions were inferred to change occurred on the same branches, suggesting perfect co-evolution within the considered evolutionary time frame. The identity of the branches in which the changes occurred have been highlighted, and the exact amino acid transitions are known. Noting the maximum lifespan of the descending lineages from these branches, where the innermost shows barplots displaying the the magnitude of maximum lifespan for the corresponding species and the middle ring shows the relative category of maximum lifespan achieved across primates, we could trace such co-occurring changes to potentially consequential changes in the lifespan of descending primate lineages.

3.3. Hypothesis testing of a global associative structure of co-transitions across aggregated intergenic CAAS pairs

For both the maximum lifespan and body mass co-transition matrices, there were 0 Tukey-Freeman corrected chi-square test statistics from the null distribution that were equal or greater, indicating that the true p-value must be less than 10^5 , allowing us to reject the hypothesis that the observed frequencies of co-transitions were produced under a model of independent co-transition. This result, combined with the sparsity of the matrices where a large number of co-transitions across the intergenic CAAS pairs were not allowed across primate evolution, and the sizeable proportion of intergenic CAAS pairs displaying strong co-evolution across both datasets, suggests evidence in favour of the widely studied theories that epistasis is a major component of the genetic architecture of complex traits across long term evolution. However, there is room to expand our analysis further to demonstrate evidence of epistasis rigorously through biological and mathematical dissection of the data, such as implementing existing analyses and development of our own tools to further this strong narrative. We summarise the results of the hypothesis testing in the figures below.

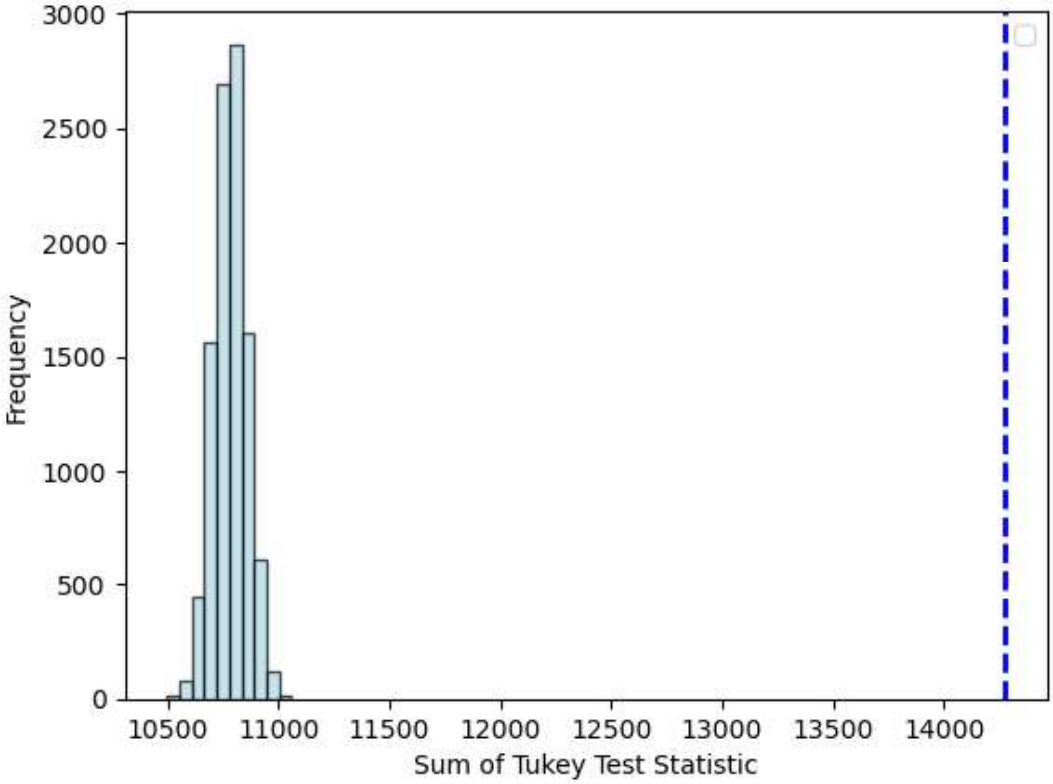


Fig 6. The null distribution of the Tukey Test statistic (Tukey-Freeman correction of the chi-square test statistic) is shown for a null distribution generated from the generation of 10^5 randomised tables using the configurational model. The test statistic is summed for all cells of the co-transition matrix for each given randomised table, and the blue dotted line represents the observed value of the test statistic for the maximum lifespan co-transition matrix

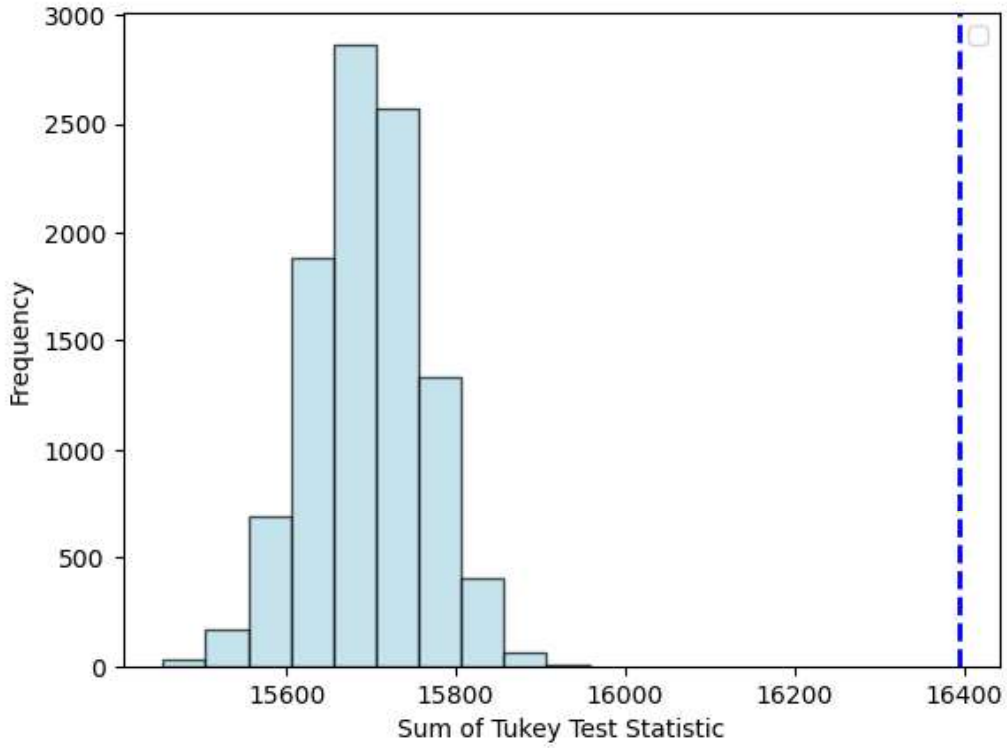


Fig 7. The null distribution of the Tukey Test statistic (Tukey-Freeman correction of the chi-square test statistic) is shown for a null distribution generated from the generation of 10^5 randomised tables using the configurational model. The test statistic is summed for all cells of the co-transition matrix for each given randomised table, and the blue dotted line represents the observed value of the test statistic for the body mass co-transition matrix.

Both the maximum lifespan (*Supplementary Table 3)* and the body mass co-transition matrices (*Supplementary Table 4)* have been provided in the *Supplementary Materials*.

4.Discussion

4.1. Summary of main findings

To explore the potential co-evolutionary structure of the genetic systems underlying longevity and body mass across primates, we combined ancestral sequence reconstruction with a high resolution catalogue of CASS positions. Using simulations that preserve branch wise mutation dynamics, we tested whether co-occurring changes between intergenic loci exceeded random expectation, and whether the resulting co-transition pattern followed a structured, associative model. This multi-layered approach allowed us to systematically assess the magnitude, statistical significance, and the organization of intergenic co-evolution within each trait.

In line with this framework, our findings revealed a widespread and significant signal of non random co-evolution. For the maximum life span dataset, out of ~ 488,000 intergenic CAAS pairs with at least 1 co-occurring change, ~ 163,000 (~33%) displayed a statistically significant pattern of co-occurrence after BH-FDR correction. For body mass, ~8,500 out of ~50,000 pairs (~17%) were similarly significant. Notably, in both datasets, a substantial number of intergenic pairs—100,651 in lifespan and 4,444 in body mass—had simulation p-values of 10-7 (0 events igual or more extreme than observed in 100,000 simulations) reinforcing the strength of the co-evolutionary signal. At a systems level, the number of observed co-occurring changes exceeded the expected number by approximately 10.5-fold in the lifespan dataset and 9.85 -fold in the body mass dataset. Beyond frequency alone, analysis of co-transition matrices revealed a global associative structure in the identity of amino acid changes: the observed transition patterns deviated so markedly from randomised configurations that none of the 100,000 simulated matrices produced a test statistic equal to or exceeding the observed value.

The CAAS positions underlying these results appear to operate within an interdependent system of genes that contributes to the architecture of complex traits. The emergent network of significantly co-evolving genes reflects a broader principle: genes function not in isolation but as part of coordinated molecular systems that have evolved under shared selective pressures. The observed magnitude of coevolution, together with the identity of the changes involved, strongly supports the view that evolutionary constraints have acted across genes, shaping the trajectories of lifespan and body mass in primates.

At the pairwise level, this is exemplified by the ~$1.6 \cdot 10^5$ significant intergenic CAAS pairs for lifespan(~ 33.4% of tested pairs) and ~$8.5 \cdot 10^3$ for body mass (~17.1%). These results imply that the loci involved in trait-associated convergence (CAAS) do not act independently, but rather co-evolve in statistically detectable patterns—likely reflecting functional coupling. The fact that the exact amino acid transitions were inferred via ancestral sequence reconstruction opens the door to future analyses of combinatorial effects across multiple

loci. For instance, subtrees of the primate phylogeny—such as those shown in Fig. 5—highlight how specific co-occurring mutations may align with shifts in trait values, though a more exhaustive mathematical and biological treatment would be required to capture their full implications.

Finally, the global pattern observed in the co-transition matrix—where particular amino acid transitions tend to pair preferentially with others—provides further evidence that molecular evolution in these traits is not governed solely by independent changes. Rather, these results point toward an epistatic layer in the genetic architecture of complex traits, not only within genes (as has been shown in prior work) but also between them. This epistasis at the intergenic level may represent a critical axis of constraint and opportunity in the evolution of life history traits.

4.2. Biological interpretation of the results

The signal of widespread intergenic co-evolution observed in our results provides compelling evidence that the genetic systems underlying maximum lifespan and body mass are governed by coordinated constraints that extend beyond single-locus effects. Particularly in the case of maximum lifespan, the high proportion of intergenic CAAS pairs exhibiting statistically significant co-occurrence (~33.4%) suggests that longevity may be under stronger evolutionary constraint than body mass, where the corresponding figure was ~17.1%. This discrepancy could reflect differences in the underlying genetic architectures of these traits, where body mass may involve a broader set of loci with greater environmental modulation, while longevity is more tightly coupled to a restricted set of pathways such as DNA repair, oxidative stress response, and cellular maintenance systems.

A notable example that supports this interpretation is the ERCC6–TIMELESS gene pair, both involved in DNA repair, where all observed changes occurred in perfect synchrony across the same branches of the primate phylogeny. This case, highlighted in Figure 5, exemplifies how co-evolutionary analysis at the intergenic level can uncover putative functionally coupled loci. While such perfect co-evolution is rare, its presence in core biological processes lends support to the idea that the coordinated evolution of life history traits reflects evolutionary dependencies between genes in shared pathways.

The CAAS positions used in this analysis not only represent sites of evolutionary convergence associated with phenotypic extremes but also show elevated Shannon entropy compared to non-CAAS sites. This distinction contributes both biological and statistical value to our framework. The higher variability observed in CAAS positions likely improves the detection power of our co-evolutionary metrics, while also reflecting the underlying evolutionary dynamics required for convergent adaptation. This dual role strengthens the rationale for focusing on CAAS in the context of systems-level evolutionary inference.

At a broader scale, when we looked at all the amino acid changes together in a co-transition matrix, the clear deviation from independence in the pairing of amino acid transitions, supported by a complete absence of null replicates with test statistics as extreme as those observed,indicates that not only do co-occurring mutations happen more often than expected - their specific transitions appear to be non randomly structured. This

pattern suggests the existence of a global associative structure in amino acid change dynamics, consistent with intergenic epistasis acting as a relevant layer across the evolution of complex traits. Together, these findings contribute to a growing body of evidence that epistasis is not only a driver of molecular evolution at intragenic level, but also plays a relevant role in shaping interactions between loci.

4.3. Limitations

While the analytical framework presented here provides robust evidence for non random co-evolution, several methodological limitations should be recognised. First, the resolution of the ancestral state reconstruction, -though generally correct- remains limited in certain internal nodes of primate phylogeny. In cases where the inferred amino acid states have the same probability, it becomes difficult to determine the exact identity of transitions, leading to an under-estimation of the co-occurring changes. These ambiguous cases were filtered out from the analysis introducing a small but systematic conservative bias.

Another limitation lies in the pairwise focus of the present study. While the pipeline is readily extensible to higher-order analyses, where three or more loci are considered simultaneously, such combinatorial analysis would have prohibitive computational costs. Future implementations of this framework could prioritise biologically relevant subsets — for example, examining clusters of co-evolving genes within known pathways — to investigate whether coordinated higher-order transitions shape the evolution of life history traits.

Finally, our null model incorporates in the simulations the branch length to weight for the mutation's probability, however, it does not account for lineage specific differences in mutation rate and selective pressures. Future simulation models can be refined to include these factors to increase the resolution of the analysis.

4.4. Future work

We aim to replicate the analysis we have done by selecting random genes, and generating random intergenic pairs of positions between these genes, where we can make comparisons between what we have obtained in our results to the magnitude of coevolution and the patterns of mutations that would be observed when our analytical workflow is applied to a randomised subset. We aim to replicate this analysis for two independent sets of random intergenic pairs of positions.First we will select at random positions that have at some minimal level of variability in the gene MSA column to so that ASR would be able to infer changes across the branches of the primate phylogeny. The second set of intergenic pairs of positions will be positions sampled where their entropy is similar to the entropy distribution of CAAS positions- this would allow comparing the results obtained for the CAAS positions to positions that may have not been identified as CAAS towards a focal trait, but would provide ASR with similar statistical power to detect changes (and co-occurring changes) across the primate phylogeny.

5.Bibliography

[1] T. F. C. Mackay and R. R. H. Anholt, “Pleiotropy, epistasis and the genetic architecture of quantitative traits,” *Nat. Rev. Genet.*, Apr. 2024, doi: 10.1038/s41576-024-00711-3.

[2] E. Uffelmann *et al.*, “Genome-wide association studies,” *Nat. Rev. Methods Primer*, vol. 1, no. 1, p. 59, Aug. 2021, doi: 10.1038/s43586-021-00056-9.

[3] E. E. Eichler, “Genetic Variation, Comparative Genomics, and the Diagnosis of Disease,” *N. Engl. J. Med.*, vol. 381, no. 1, pp. 64–74, Jul. 2019, doi: 10.1056/NEJMra1809315.

[4] T. Matsui *et al.*, “The interplay of additivity, dominance, and epistasis on fitness in a diploid yeast cross,” *Nat. Commun.*, vol. 13, no. 1, p. 1463, Mar. 2022, doi: 10.1038/s41467-022-29111-z.

[5] R. F. Campbell, P. T. McGrath, and A. B. Paaby, “Analysis of Epistasis in Natural Traits Using Model Organisms,” *Trends Genet.*, vol. 34, no. 11, pp. 883–898, Nov. 2018, doi: 10.1016/j.tig.2018.08.002.

[6] W. Huang *et al.*, “Epistasis dominates the genetic architecture of *Drosophila* quantitative traits,” *Proc. Natl. Acad. Sci.*, vol. 109, no. 39, pp. 15553–15559, Sep. 2012, doi: 10.1073/pnas.1213423109.

[7] W. G. Hill, M. E. Goddard, and P. M. Visscher, “Data and Theory Point to Mainly Additive Genetic Variance for Complex Traits,” *PLoS Genet.*, vol. 4, no. 2, p. e1000008, Feb. 2008, doi: 10.1371/journal.pgen.1000008.

[8] M. S. Breen, C. Kemena, P. K. Vlasov, C. Notredame, and F. A. Kondrashov, “Epistasis as the primary factor in molecular evolution,” *Nature*, vol. 490, no. 7421, pp. 535–538, Oct. 2012, doi: 10.1038/nature11510.

[9] J. Burch *et al.*, “Wright was right: leveraging old data and new methods to illustrate the critical role of epistasis in genetics and evolution,” *Evolution*, vol. 78, no. 4, pp. 624–634, Mar. 2024, doi: 10.1093/evolut/qpae003.

[10] L. Di Bari, M. Bisardi, S. Cotogno, M. Weigt, and F. Zamponi, “Emergent time scales of epistasis in protein evolution,” Mar. 15, 2024. doi: 10.1101/2024.03.14.585034.

[11] M. S. Johnson, G. Reddy, and M. M. Desai, “Epistasis and evolution: recent advances and an outlook for prediction,” *BMC Biol.*, vol. 21, no. 1, p. 120, May 2023, doi: 10.1186/s12915-023-01585-3.

[12] F. J. Poelwijk, M. Socolich, and R. Ranganathan, “Learning the pattern of epistasis linking genotype and phenotype in a protein,” *Nat. Commun.*, vol. 10, no. 1, p. 4213, Sep. 2019, doi: 10.1038/s41467-019-12130-8.

[13] T. F. C. Mackay, “Epistasis and quantitative traits: using model organisms to study gene–gene interactions,” *Nat. Rev. Genet.*, vol. 15, no. 1, pp. 22–33, Jan. 2014, doi: 10.1038/nrg3627.

[14] Annalisa Pastore and Marco Fantini, “Protein Structural Information and Evolutionary Landscape by In Vitro Evolution (DCA RELATED9”.

[15] T. N. Starr and J. W. Thornton, “Epistasis in protein evolution,” *Protein Sci.*, vol. 25, no. 7, pp. 1204–1218, Jul. 2016, doi: 10.1002/pro.2897.

[16] G. Pedruzzi, A. Barlukova, and I. M. Rouzine, “Evolutionary footprint of epistasis,” *PLOS Comput. Biol.*, vol. 14, no. 9, p. e1006426, Sep. 2018, doi: 10.1371/journal.pcbi.1006426.

[17] K. Buda, C. M. Miton, and N. Tokuriki, “Pervasive epistasis exposes intramolecular networks in adaptive enzyme evolution,” *Nat. Commun.*, vol. 14, no. 1, p. 8508, Dec. 2023, doi: 10.1038/s41467-023-44333-5.

[18] E. E. Eichler, “Genetic Variation, Comparative Genomics, and the Diagnosis of Disease,” *N. Engl. J. Med.*, vol. 381, no. 1, pp. 64–74, Jul. 2019, doi: 10.1056/NEJMra1809315.

[19] Y. Quan, Z. Wang, X. Chu, and H. Zhang, “Evolutionary and genetic features of drug targets,” *Med. Res. Rev.*, vol. 38, no. 5, pp. 1536–1549, Sep. 2018, doi: 10.1002/med.21487.

[20] R. C. Albertson, W. Cresko, H. W. Detrich, and J. H. Postlethwait, “Evolutionary mutant models for human disease,” *Trends Genet.*, vol. 25, no. 2, pp. 74–81, Feb. 2009, doi: 10.1016/j.tig.2008.11.006.

[21] M. L. Benton, A. Abraham, A. L. LaBella, P. Abbot, A. Rokas, and J. A. Capra, “The influence of evolutionary history on human health and disease,” *Nat. Rev. Genet.*, vol. 22, no. 5, pp. 269–283, May 2021, doi: 10.1038/s41576-020-00305-9.

[22] L. Quintana-Murci, “Understanding rare and common diseases in the context of human evolution,” *Genome Biol.*, vol. 17, no. 1, p. 225, Dec. 2016, doi: 10.1186/s13059-016-1093-y.

[23] G. Housman and J. Tung, “Next-generation primate genomics: New genome assemblies unlock new questions,” *Cell*, vol. 186, no. 25, pp. 5433–5437, Dec. 2023, doi: 10.1016/j.cell.2023.11.014.

[24] M. Civelek and A. J. Lusis, “Systems genetics approaches to understand complex traits,” *Nat. Rev. Genet.*, vol. 15, no. 1, pp. 34–48, Jan. 2014, doi: 10.1038/nrg3575.

[25] J. M. Cork and M. D. Purugganan, “The evolution of molecular genetic pathways and networks,” *BioEssays*, vol. 26, no. 5, pp. 479–484, May 2004, doi: 10.1002/bies.20026.

[26] L. F. K. Kuderna *et al.*, “A global catalog of whole-genome diversity from 233 primate species”.

[27] G. Muntané *et al.*, “Biological Processes Modulating Longevity across Primates: A Phylogenetic Genome-Phenome Analysis,” *Mol. Biol. Evol.*, vol. 35, no. 8, pp. 1990–2004, Aug. 2018, doi: 10.1093/molbev/msy105.

[28] J. R. Speakman, “Body size, energy metabolism and lifespan,” *J. Exp. Biol.*, vol. 208, no. 9, pp. 1717–1730, May 2005, doi: 10.1242/jeb.01556.

[29] J. P. D. Magalhães, J. Costa, and G. M. Church, “An Analysis of the Relationship Between Metabolism, Developmental Schedules, and Longevity Using Phylogenetic Independent Contrasts,” *J. Gerontol. Ser. A*, vol. 62, no. 2, pp. 149–160, Feb. 2007, doi: 10.1093/gerona/62.2.149.

[30] A. Kowalczyk, W. K. Meyer, R. Partha, W. Mao, N. L. Clark, and M. Chikina, “RERconverge: an R package for associating evolutionary rates with convergent traits,” *Bioinformatics*, vol. 35, no. 22, pp. 4815–4817, Nov. 2019, doi: 10.1093/bioinformatics/btz468.

[31] F. Barteri *et al.*, “CAAStools: a toolbox to identify and test Convergent Amino Acid Substitutions”.

[32] R. P. Simões, I. R. Wolf, B. A. Correa, and G. T. Valente, “Uncovering patterns of the evolution of genomic sequence entropy and complexity,” *Mol. Genet. Genomics*, vol. 296, no. 2, pp. 289–298, Mar. 2021, doi: 10.1007/s00438-020-01729-y.

[33] Z. Yang, “PAML 4: Phylogenetic Analysis by Maximum Likelihood,” *Mol. Biol. Evol.*, vol. 24, no. 8, pp. 1586–1591, Apr. 2007, doi: 10.1093/molbev/msm088.

[34] C. B. Read, “Freeman—Tukey chi-squared goodness-of-fit statistics,” *Stat. Probab. Lett.*, vol. 18, no. 4, pp. 271–278, Nov. 1993, doi: 10.1016/0167-7152(93)90015-B.

[35] W. M. Patefield, “Algorithm AS 159: An Efficient Method of Generating Random R × C Tables with Given Row and Column Totals,” *Appl. Stat.*, vol. 30, no. 1, p. 91, 1981, doi: 10.2307/2346669.

[36] G. Casiraghi and V. Nanumyan, “Configuration models as an urn problem,” *Sci. Rep.*, vol. 11, no. 1, p. 13416, Jun. 2021, doi: 10.1038/s41598-021-92519-y.