Supplementary Information

Recurrent erosion of *COA1/MITRAC15* exemplifies conditional gene dispensability in oxidative phosphorylation

Sagar Sharad Shinde, Sandhya Sharma, Lokdeep Teekas, Ashutosh Sharma, Nagarjun Vijay

Computational Evolutionary Genomics Lab, Department of Biological Sciences, IISER Bhopal, Bhauri, Madhya Pradesh, India

*Correspondence: nagarjun@iiserb.ac.in

S1 text: *COA1* is a distant homolog of *TIMM21*

We have tried to identify homologs of *COA1* and traced the history of this gene family. The rationale for determining the homologs of *COA1* is to evaluate the presence of potential paralogs that could compensate for the loss of the *COA1* gene. Our analysis finds that the closest homolog for *COA1* is *TIMM21*. Subsequently, we estimated when the *TIMM21* and *COA1* genes emerged through a duplication event. Demonstrating that both these genes are well conserved illustrates that the loss of *COA1* in various vertebrate species is striking and potentially functionally relevant. Distinct one-to-one orthologs of *COA1* and *TIMM21* can be identified based on sequence identity and gene order in mammals, birds, amphibians, lobe-finned fishes, and jawless fishes, which share the second round of whole-genome duplication (2R-WGD) and lack the third round of whole-genome-duplication (3R-WGD). Two copies of the *TIMM21* gene occur in several species of ray-finned fishes, which have undergone 3R-WGD. Pre-2R-WGD invertebrate species such as the fruit fly and oriental fly have homologs of *COA1* and *TIMM21*.

The presence of distinct copies of *COA1* and *TIMM21* in pre-2R-WGD species suggests that both genes existed in the LVCA (Last Vertebrate Common Ancestor) (**Supplementary Fig. S1**). Homologs of both *COA1* and *TIMM21* are present in *Schizosaccharomyces pombe* and *Saccharomyces cerevisiae.* Based on the presence of these homologs in fungi and animals, we can track the presence of these genes to the LOCA (Last Opisthokont Common Ancestor). A further search of *COA1* and *TIMM21* homologs revealed their presence in Amoebozoa (*Dictyostelium discoideum*), Chlorophyta (*Chlamydomonas reinhardtii*), Plantae (*Arabidopsis thaliana*), and Protista (**Supplementary Fig. S2**). The widespread presence of these homologs suggests both copies were present in the Last Eukaryotic Common Ancestor (LECA).

Identification of one-to-one orthologs relies on the conserved synteny of the region containing the focal gene. However, synteny is not conserved over very long (> 500 million years) timescales[1], and high confidence identification of one-to-one orthologs is not always feasible. In the case of such distantly related genes, the Pfam 34.0 database[2] of protein families provides a valuable resource. We found that *TIMM21* (PF08294) and *COA1* (PF08695) families are members of the MIM-OM_import (CL0455) clan, which is characterized as "mitochondrial membrane proteins with importing activities.". We used all proteins that form part of these protein families as input to clans to identify the clustering patterns. When both *TIMM21* (PF08294) and *COA1* (PF08695) families are concatenated and used as input for CLANS, we recover the clustering pattern (**Fig. 1A** and **Supplementary Fig. S3**) seen while running CLANS on the HHblits hits of the human *COA1* protein (**Supplementary Fig. S4**). The clustering pattern of *TIMM21* (PF08294) and *COA1* (PF08695) families suggests that all *TIMM21* proteins are much more similar to each other than *COA1* proteins. However, identical to *COA1*, we can see separate clusters for fungal, plant, and animal sequences within *TIMM21* proteins (**Fig. 1A** inset and **Supplementary Fig. S5**). Such independent clustering of *TIMM21* and *COA1* homologs suggests that lineage-specific duplications after the LECA are unlikely.

The orthology-based analysis and sequence clustering suggest that both *TIMM21* and *COA1* existed in the LECA. While we cannot conclusively exclude post-LECA lineage-specific duplications, the data does not seem to support this scenario. Earlier reports support the presence of *TIMM21* in the LECA[3,4]. However, the *COA1* gene and its origin are understudied. We can envisage the following simple scenarios that can result in the LECA containing both *TIMM21* and *COA1* homologs:

1. *TIMM21* from archaea + *COA1* from bacteria→FECA→LECA
2. *COA1* from archaea + *TIMM21* from bacteria→FECA→LECA
3. *COA1* from archaea →FECA—(Duplication event)→LECA
4. *COA1* from bacteria →FECA—(Duplication event)→LECA
5. *TIMM21* from archaea →FECA—(Duplication event)→LECA
6. *TIMM21* from bacteria →FECA—(Duplication event)→LECA

Many bacterial sequences belong to the *COA1* protein family in contrast to *TIMM21*, which has a minimal number of bacterial homologs. Hence, the data supports scenarios 1 and 4, which suggest that *COA1* has a bacterial origin. Scenarios 5 and 6, which indicate duplication of *TIMM21* within the FECA, will find it challenging to explain the large number of bacterial homologs of *COA1*. Scenarios 2 and 3 suggest that *COA1* has an archaeal origin and has minimal support as very few *COA1* sequences occur in archaea.

We used the human *COA1* sequence to perform an iterative profile-profile search of the uniport (UniRef30_2020_06) database using HHblits. The list of proteins identified as homologs of human *COA1* (**Supplementary File S1-S2**) and primate *COA1* orthologs (**Supplementary File S3**) contain several *TIMM21* like proteins. Iterative PSI-BLAST search identified *TIMM21* homologs from the second iteration onwards and found an increasing number of *TIMM21* hits in each subsequent iteration (**Supplementary File S4**). Out of the 500 top search results from HHblits, 59 have annotation as "Cytochrome C oxidase assembly factor" or "Cytochrome C oxidase assembly protein" or "*COA1*", and 120 as "*TIMM21*" homologs. The annotation of 13 proteins is "hypothetical", nine are "membrane" proteins, eight are "DUF1783 domain-containing" proteins, and 27 proteins are from diverse proteins. The remaining 264 of the 500 hits are "Uncharacterized". The large number of "Uncharacterized" proteins identified are challenging to interpret. Hence, to trace the relationships between the proteins identified as homologs of *COA1*, we investigated the sequence identity-based clusters established by CLANS (**Supplementary Fig. S6**). The large group of red dots consists of proteins annotated as *TIMM21*, and the collection of blue dots contains proteins annotated as *COA1*. Homologs of *COA1* from bacterial species form two clusters, a distinct light blue cluster consisting of predominantly Planctomycetes bacteria and a diffuse bunch of brown dots that consists of largely proteobacterial species. The group of orange dots consists of proteins annotated as *COA1* in fungal genomes. The *COA1* homologs in plants consist of a yellow cluster consisting of *Arabidopsis thaliana* homolog At2g20390 and the magenta cluster of *TIMM21*-like proteins containing *Arabidopsis thaliana* homolog At2g40800.

To further verify whether the database matches are homologous, we evaluated the biological function, secondary structure similarity, relationship among top hits, and occurrence of conserved motifs. To obtain secondary structure predictions for the proteins *COA1* and *TIMM21*, we used the PROTEUS (version 2.0) webserver[5]. The HeliQuest webserver[6] provided each predicted helix's physicochemical properties and amino acid compositions (**Supplementary Fig. S7-S8**). While the three-dimensional (3-D) structure of the *COA1* protein is not available yet, multiple structures of the *TIMM21* protein are available in the Protein Data Bank (PDB). It is possible to use comparative/homology modeling to predict the 3-D structure based on the protein structure of a related protein[7]. Hence, we used the comparative modeling approach implemented in Modeller (v10.0) software to model the structure of *COA1* based on the homologous structures available in PDB. The Phyre2[8] and Expasy Swiss-Model[9] webservers also predicted homologous 3-D structures of *COA1* (**Supplementary File S5**). All the top hits were from 3-D structures of the IMS (Inter Membrane Space) domain of *TIMM21* protein. The IMS domain of *TIMM21*, whose 3-D structures are available on PDB, contains only the part of the protein that occurs after the membrane-spanning helix. To model the structure of *COA1* using these existing 3-D structures, we used the *COA1* amino acid sequence that occurs after the membrane-spanning domain. We visualized (**Supplementary Fig. S9**) the structure of *TIMM21* and the predicted *COA1* structure using (UCSF Chimera v1.15) ChimeraX[10].

S2A text: *COA1* gene duplication, pseudogenisation, and exon reorganization

In species with functional and pseudogene copies of *COA1*, these copies have diverged considerably and formed distinct haplotypes. For example, the blastn search of sequencing short-read data from the human genome with a *COA1* gene sequence as a query results in two distinct haplotypes. One set of reads correspond to the intact *COA1* gene in humans, and the other set of reads are from the pseudogenic copy (**Supplementary Fig. S17**). Comparative analysis of Primate genome assemblies suggests that the pseudogenic copy results from a duplication of *COA1* within the primate lineage (**Supplementary Fig. S18**).

Independent duplication of *COA1* has occurred in carnivores (**Supplementary Fig. S19**). However, the duplicated copy has undergone pseudogenization and diverged from the functional gene sequence similar to Primates. For example, sequencing raw read data in the tiger consist of two distinct haplotypes corresponding to the intact and pseudogene copies (**Supplementary Fig. S20**). While the intact copy is located at a genomic region (*STK17A* & *HECW1* upstream and *BLVRA* & *VOPP1* downstream) with conserved synteny across other mammals, the pseudogene copy occurs adjacent to the *PRR32* gene. Outgroup species such as horse (*Equus caballus*) and pangolin (*Manis javanica*) have a single copy of the *COA1* gene with all raw reads supporting a single haplotype (**Supplementary Fig. S21-S22**). Both sub-orders (Caniformia and Feliformia) within Carnivora share this duplication of the *COA1* gene (**Supplementary Fig. S19**). The intact *COA1* copy is expressed in diverse transcriptomes among Caniformia species, while the pseudogene copy lacks expression. The first and second exons are orthologous; however, the genomic location of the transcribed third exon is different between Feliformia (cat-like-exon-3) and Caniformia species (dog-like-exon-3) (**Fig. 2** and **Supplementary Fig. S23**).

The final exon of the *COA1* gene in Feliformia extends to 163 base pairs (*Panthera tigris altaica, Panthera leo, Panthera pardus,* and *Lynx lynx*) and 160 base pairs (*Puma concolor* and *Felis catus*) compared to the 100 base pairs in Caniformia species. A single deletion event causes the difference of three base pairs between these two groups of Feliformia at the 24th base of exon-4 (**Supplementary Table S1**). The extended final exon shared by all Feliformia species results from a two-base frameshift deletion before the erstwhile stop codon in exon-4. Despite the extended last exon in Feliformia species, the full-length open reading frames of Feliformia (130/131 amino acids) and Caniformia (135 amino acids) are comparable.

The shorter reading frame in Feliformia results from most *COA1* transcripts skipping the dog-like-exon-3, whose inclusion results in premature stop codons in all the seven Feliformia species. The dog-like-exon-3 is present in all *COA1* transcripts of Caniformia species and does not contain gene-disrupting changes. A single base deletion in all Feliformia species changes the end phase of exon-2 to maintain an intact reading frame while skipping the dog-like-exon-3. Transcriptomes of the cat (*Felis catus*) from the spleen (**Supplementary Fig. S24-S34**) and puma (*Puma concolor*) from blood (**Supplementary Fig. S35-S42**) exhibit expression of a proto cat-like-exon-3 which gets spliced into some of the *COA1* transcripts. However, the majority of transcripts skip this proto cat-like-exon-3 which contains premature stop codons. These changes in exon splicing patterns between Caniformia and Feliformia species appear to result from changes in splice factor binding sites at the *COA1* locus (**Supplementary Fig. S43**).

In contrast to primates and carnivores, reads support multiple haplotypes of *COA1* only in the second exon of naked mole-rat (**Supplementary Fig. S44**). Hence, the duplicated copy of *COA1* in naked mole-rat appears to have mostly degraded. However, we cannot rule out the possibility that the reads from other haplotypes spanning the remaining three exons are missing due to high GC content. The sequencing reads support a single intact open reading frame in the red squirrel (**Supplementary Fig. S45**) and platypus (**Supplementary Fig. S46**). Although a single haplotype occurs in the raw read dataset of chicken, this haplotype has gene-disrupting changes (**Supplementary Fig. S47**). The gene-disrupting modifications identified in the chicken *COA1* gene were investigated further by screening long-read datasets, transcriptomes, and genomes of various galliform species.

After duplication of the *COA1* gene in Primates, an extension of the N-terminal region has occurred in Cercopithecidae and Catarrhini and is transcriptionally active (**Supplementary Fig. S48**). However, new world monkeys do not have this N-terminal extension denoted as exon-1a. Both Cercopithecidae and Catarrhini have an additional start codon in exon-1a upstream from the original start codon in the ancestral exon-1 denoted as exon-1b in species with N-terminal extension (**Supplementary Table S2)**. A striking difference between Cercopithecidae and Catarrhini is the lack of the internal start codon in Cercopithecidae, where Catarrhini has a start codon. Since proteome level data is not available for these species, we rely solely on the RNA-seq datasets and start and stop codons within the expressed transcripts to evaluate the exon/intron structure changes. Using these carefully

annotated Primate sequences of *COA1*, we verified a previous report[11] of positive selection in this gene among Primates (**Supplementary Table S3**).

S2B text: *COA1* gene duplication, pseudogenisation, and exon reorganization

RNA-seq data is available only from the skin tissue and one unknown tissue sample in the cheetah. Unfortunately, these RNA-seq datasets are of poor quality, evidenced by the lack of gene expression information for most genes (**Supplementary Fig. S55-S60)**. Nonetheless, to evaluate if the skin is a relevant tissue to screen for the expression level of *COA1*, we analyzed skin transcriptome datasets from closely related species. The *COA1* gene is robustly expressed in the skin tissue of cats and dogs (**Supplementary Fig. S61-S65 and S66-S67**). The presence or absence of evidence of transcription or translation is not conclusive evidence of gene loss. It is known that several pseudogenes are not only transcribed but are also translated. However, they may not form functional proteins as the full-length protein is not produced due to premature stop codons[12]. Hence, the presence of premature stop codons that disrupt the functional domains of the protein or truncate a significant fraction of the protein is considered more reliable evidence of gene loss.

Two versions of the cheetah genome (Aci_jub_2/ GCA_003709585.1 and aciJub1/ GCA_001443585.1) are currently available in the NCBI (National Center for Biotechnology Information) database. A third version of the cheetah genome generated by scaffolding the second version using Hi-C data is also available (aciJub1_HiC.fasta). While these multiple versions of the cheetah genome assembly have improved the contiguity (as seen by the increased N50), the actual base pair level sequence correctness cannot be assessed directly from the genome assembly. Some of the carnivore genome assemblies are known to have base-pair level errors that can affect the results of evolutionary analysis[13]. Hence, we rely upon the sequencing raw read datasets to ensure the validity of gene loss.

The *COA1* gene has undergone a duplication followed by pseudogenisation of one copy in the ancestor of all carnivore species (**Supplementary Fig. S19**). We used the *COA1* sequence of the puma (*Puma concolor*), which has an intact ORF, as a query to search for the *COA1* gene in cheetah. Based on the results of the blastn search of the sequencing raw read datasets, we found evidence of at least two haplotypes of *COA1* in each of the three cheetah individuals screened. Various scenarios that can result in an open reading frame among these haplotypes were evaluated to verify gene loss in the cheetah (**Supplementary Fig. S68-S70**). In the first haplotype, located in the syntenic region of *COA1*, one base change occurs in exon two position 27 (C → T), leading to a premature stop codon (TAG). The second haplotype (i.e., the duplicated copy of *COA1*) is located between PRR32 and DCAF12L2 (DDB1- and CUL4-associated factor 12-like protein 2 gene) on the scaffold NW_020836464.1. We found two single base insertions in exon two at positions 11 (T) and 89 (C), leading to an altered reading frame in haplotype two. A single base substitution (C → T) at position 59 in this new reading frame leads to a premature stop codon (TGA). Exon four has an insertion of CTT, TAT, AAACA, and A at positions 22, 82, 84, and 145, respectively, and it has two stop codons at positions 24-26 (TAA) and 84-86 (TAG).

In haplotype two, we found few reads that do not support T insertion at position 11 in exon two. This difference between the haplotype two reads could result from individual polymorphism (**Supplementary Fig. S71-S73**). However, even when the insertion at position 11 is missing in exon two, a premature stop codon occurs in exon four at position 31, where C → T substitution leads to a stop codon (TAA). Despite this, exon four contains two more stop codons at positions 79-81 (TGA) and 84-86 (TGA). For each of the three haplotypes, we tried different possible combinations of exons to evaluate if any combination produced an intact ORF. We found that none of the exon combinations lead to an intact ORF (**Supplementary Fig. S74-S77**). The conserved domain identified in each ORF is shown as a green line below the exons. In most cases, we can see that the truncated ORF covers only a fraction of the conserved domain. Functional evaluation of the short proteins would give a definitive answer about the relevance of these ORF's.

S3 text: *COA1* gene loss in galliform species

We find evidence of *COA1* gene loss in multiple galliform species. The loss of *COA1* could be verified through the presence of gene disrupting mutations, lack of gene expression in multiple tissues, and signatures of relaxed selection detected by molecular evolutionary analyses in galliform species consisting of the red-legged partridge (*Alectoris rufa*), Chinese bamboo partridge (*Bambusicola thoracicus*), common quail (*Coturnix coturnix*), Japanese quail (*Coturnix japonica*), lesser sage-grouse (*Centrocercus minimus*), chicken (*Gallus gallus*), rock ptarmigan (*Lagopus muta*), black grouse (*Lyrurus tetrix*), wild turkey (*Meleagris gallopavo*), helmeted guineafowl (*Numida meleagris*), marbled wood quail (*Odontophorus gujanensis*), Indian peafowl (*Pavo cristatus*), mikado pheasant (*Syrmaticus mikado*) and greater prairie chicken (*Tympanuchus cupido*). In addition to these species, the common pheasant (*Phasianus colchicus*) and golden pheasant (*Chrysolophus pictus*) may have acquired regulatory changes resulting in a lack of gene expression and relaxed selection. Although we see signatures of relaxed selection in wood duck or Carolina duck (*Aix sponsa*), Muscovy duck (*Cairina moschata*), and common crossbill (*Loxia curvirostra*), these species appear to have intact reading frames that are transcribed. However, despite being galliform birds, species such as Australian brushturkey (*Alectura lathami*), northern bobwhite (*Colinus virginianus*), scaled quail (*Callipepla squamata*), and white-crested guan (*Penelope pileata*) have intact gene sequences that are under purifying selection (see **Supplementary Table S11**). Retention of an intact *COA1* gene in these galliform species may result from a more ancestral duck-like muscle fiber distribution with a lower proportion of white fiber.

Several flight-degenerate species from diverse taxa[14] such as rifleman (*Acanthisitta chloris*), speckled mousebird (*Colius striatus*), sunbittern (*Eurypyga helias*), hoatzin (*Opisthocomus hoazin*), and South African ostrich (*Struthio camelus australis*) have retained intact *COA1* sequences. Various tinamou species such as the rufescent tinamou (*Crypturellus cinnamomeus*), little tinamou (*Crypturellus soui*), undulated tinamou (*Crypturellus undulatus*), martineta tinamou (*Eudromia elegans*), white-throated tinamou (*Tinamus guttatus*), tawny-breasted tinamou (*Nothocercus julius*), hooded tinamou (*Nothocercus*

nigrocapillus), ornate tinamou (*Nothoprocta ornata*), Andean tinamou (*Nothoprocta pentlandii*) and Chilean tinamou (*Nothoprocta perdicaria*) also have intact *COA1* gene sequences. Other Palaeognathae flightless birds such as the North Island kiwi (*Apteryx australis mantelli*), Okarito kiwi (*Apteryx rowi*), southern cassowary (*Casuarius casuarius*), emu (*Dromaius novaehollandiae*), and greater rhea (*Rhea americana*) also have intact *COA1* gene sequences. The flightless cormorant (*Nannopterum harrisi*), as well as other species of cormorants such as double-crested cormorant (*Phalacrocorax auritus*), Neotropic cormorant (*Phalacrocorax brasilianus*), great cormorant (*Phalacrocorax carbo*), and pelagic cormorant (*Phalacrocorax pelagicus*), also have intact *COA1* gene sequences. The ground-dwelling brown mesite (*Mesitornis unicolor*), kākāpō also called owl parrot (*Strigops habroptila*), emperor penguin (*Aptenodytes forsteri*), Adélie penguin (*Pygoscelis adeliae*), and African penguin (*Spheniscus demersus*) are also considered flightless but retain the *COA1* gene. The greater roadrunner (*Geococcyx californianus*) has a limited flying ability but is capable of very fast running. Similarly, the kagu (*Rhynochetus jubatus*) is almost flightless and spends most of its time on the ground. Both the greater roadrunner and kagu have intact *COA1* sequences.

Overall, we see that the *COA1* gene loss is found in species with a high proportion of white muscle fiber in the pectoralis and is not simply a consequence of degeneration of flight abilities. Increased use of hind limb muscles in flightless birds and a lower fraction of white muscle fiber in the pectoralis muscle seem sufficient to retain an intact *COA1* gene. Although quantitative information is not available, the red-winged blackbird (*Agelaius phoeniceus*), snowy owl (*Bubo scandiacus*), and Canada geese (Branta canadensis) have predominantly red muscle fiber and have intact *COA1* gene sequence. Similarly, the cockatiel (*Nymphicus hollandicus*), dark-eyed juncos (*Junco hyemalis*), and species closely related to the Smithsonian gull (*Larus smithsonianus*) are reported[15] to have only red muscle fiber, and all of these species have intact *COA1* gene sequence. Hence, based on currently available data, we find a significant negative correlation pattern between loss of *COA1* gene and predominance of white muscle fiber.

S4A text: *COA1* occurs in an evolutionary breakpoint region

While verifying gene loss, we need to consider two aspects. First, we need to confirm whether the raw read dataset supports the genome assembly sequence and identified gene disrupting mutations. Second, we need to evaluate whether the gene order is conserved in the region adjacent to the focal gene, allowing for accurate identification of the ortholog. Hence, the first criteria require high coverage data from a sequencing technology with a low error rate (for instance, Illumina sequencing data). The second criteria require evaluation of gene order and the genome assemblies need to be of high contiguity. Additionally, long-range sequence information should be available from either long-read sequencing technologies like PacBio or complementary methods such as Hi-C or synthetic long reads. We describe how we have used both these criteria to verify gene loss events reported by us.

Genome assembly and raw read based verification of *COA1* in rodents

Eurasian red squirrel (*Sciurus vulgaris*):

We used short-read Illumina data to verify genome assembly correctness and conserved synteny for the Eurasian red squirrel (*Sciurus vulgaris*). But in Eurasian red squirrel, publicly available Illumina raw read data (~15 Gb) was not enough to verify the genome assembly. We used long-read PacBio raw read data (~67 Gb) and mapped it with the red squirrel genome (Sciurus_vulgaris.mSciVul1.1). We extracted the reads > 3kb from the mapped read file and loaded it in the UCSC genome browser as a custom track (**Supplementary Figures S79-S83**). We verified that long PacBio reads span from *COA1* to the genes adjacent to *COA1*, which provides confidence about genome assembly robustness and conserved synteny (see **Supplementary Table S13**).

American beaver (*Castor canadensis):*

In the old genome assembly (Castor_canadensis.C.can_genome_v1.0) of American beaver (*Castor canadensis),* the pseudogenized *COA1* is the only gene located on the scaffold MTKA01013026.1 (**Supplementary Figures S84**). The *COA1* adjacent genes are on different scaffolds. Due to the fragmented nature of this assembly, we are unable to confirm synteny in the American beaver. We used a new genome assembly (GCA_009822645.1_ASM982264v1) of the American beaver published by Zhou et al., 2020[16] to verify the gene order. We found that the *C7orf5*, *PSMA2*, *MRPL32*, *HECW1*, and *STK17A* are adjacent to *COA1* on the scaffold RPDE01003036.1. However, other nearby genes *BLVRA* and *MRPS24*, are present on scaffold RPDE01003922.1, while *URGCP* and *UBE2D4* are on scaffold RPDE01003294.1 (**Supplementary Table S13**). We further evaluated the correctness of the genome assembly using long-read PacBio data (**Supplementary Fig. S85-S94**). However, we observed that some regions showed no spanning reads, and gaps in coverage prevent verification of the assembly. Nonetheless, the newer genome assembly helps verify the conserved gene order on the left flank. It is possible that gene order has changed on the right flank and will need to be evaluated in the future using a better quality genome assembly.

Naked mole-rat (*Heterocephalus glaber*):

A total of three genome assemblies are available for naked mole-rat (*Heterocephalus glaber*). Out of the two older assemblies, one is from a male (Heterocephalus_glaber_male.HetGla_1.0), and another is from a female (Heterocephalus_glaber_female HetGla_female_1.0). The latest assembly has been generated by Zhou et al., 2020[16] (GCA_014060925.1_Heter_glaber.v1.7) based on long-read PacBio data. As the male naked mole-rat genome is relatively fragmented, we checked the syntenic region in the female naked mole-rat genome. We found all the genes adjacent to *COA1* (i.e., *C7orf25*, *PSMA2*, *MRPL32*, *HECW1*, *STK17A*, *BLVRA*, *CYP3A9*, *GJC3*, and *AZGP1*) are on a single scaffold, JH602053.1 (see **Supplementary Table S13**). To check the correctness of

the assembly, we used the raw PacBio data of naked mole-rat and mapped it to the female genome. After visualizing the mapped reads in the UCSC genome browser, we found that the reads do not span all regions. There are several coverage gaps in the regions adjacent to *COA1* (**Supplementary Fig. S95**).

We evaluated the synteny in the latest genome assembly generated by Zhou et al., 2020[16] and found that the genes *HECW1*, *STK17A*, *COA1*, *BLVRA*, *CYP3A9*, *GJC3*, and *AZP1* are present in a collinear fashion on the scaffold RPGA01000026.1. The genes *C7orf25*, *PSMA2*, and *MRPL32* are on the scaffold RPGA01005726.1 (**Supplementary Fig. S96-S104**). As an additional step to verify genome assembly correctness, we analyzed the Hi-C (High-throughput Chromosome Conformation Capture) data from naked mole-rat using the genome assembly of a female (Heterocephalus_glaber_female HetGla_female_1.0). We did not find any apparent patterns of genome assembly errors based on the frequency of interactions between genomic regions adjacent to *COA1* (**Supplementary Fig. S105-106**). Hence, all the available data (three genome assemblies, short read Illumina, PacBio, and Hi-C data) are consistent with a conserved gene order.

Our evaluation of Hi-C interaction maps of various species did not find any such obvious patterns of gaps in the *COA1* region of the genome. Even the duplicated *COA1* regions in chicken, carnivores, primates, and rodents did not have a gap in the interaction map (**Supplementary Fig. S107-S178**). These Hi-C maps support the long-range correctness of the genome assembly and help validate the synteny relationships identified. In the case of the mouse, we found that following the CR, the *COA1* and adjacent gene *BLVRA* are translocated to two different chromosomes. The regions flanking *COA1* and *BLVRA* are on two other chromosomes. We found the Hi-C map free of gaps or other patterns indicating assembly errors in all four chromosomes (**Supplementary Fig. S141-S156**).

S4B text: *COA1* occurs in an evolutionary breakpoint region

Search for the *COA1* gene in the mammoth genome demonstrated striking heterogeneity in coverage of the four exons based on the Illumina ancient DNA sequencing datasets analyzed (see **Supplementary Fig. S180A-F**). Despite having comparable genome-wide coverage, we could see that not all exons occur in all the datasets. For instance, the re-sequencing dataset from PRJEB29510 (162 Gb) does not have reads from any of the four *COA1* exons. However, the datasets from PRJEB7929 (88.34 Gb) and PRJNA397140 (155 Gb) have reads covering three exons each despite having much lower genome-wide coverage. The third exon of *COA1* was missing or had fewer reads than the other three exons in most datasets. The dataset from PRJEB42269 had no reads from the first exon but had a few reads from exons three and four. We reasoned that this heterogeneity in the coverage of various *COA1* exons was mainly a result of the well-established sequencing bias of Illumina that results in inadequate coverage of GC-rich regions[17]. Quantification of GC content in each of the four *COA1* exons and K-mer abundance in different GC content bins in each mammoth Illumina

re-sequencing dataset explains most of the coverage heterogeneity between datasets as well as exons (see **Supplementary Fig. S180G**).

In contrast to the *COA1* gene, we did not see heterogeneity in the sequencing coverage of *TIMM21* exons despite comparable GC content for some of the exons (see **Supplementary Fig. S180G** and **Supplementary Fig. 181-187**). The heterogeneity in sequencing coverage of *COA1* exons demonstrates the challenges of detecting its presence in Illumina sequencing datasets. GC-biased gene conversion (gBGC) plays a defining role in the base composition for any particular gene or genomic region. It preferentially fixes GC in AT/GC heterozygotes and increases the GC content. The GC content of the *COA1* exons can be driven to extreme values by gBGC. The magnitude of gBGC also varies across the genome within a species as well as between species. Therefore, *COA1* orthologs from closely related species or even duplicated copies of *COA1* within the same species can have very different GC content. Such differences in GC content can result in correspondingly different coverage of the gene sequence in Illumina data and masquerade as a gene loss event[18,19].

A well-known example for high GC content impeding sequencing is the gene *PDX1*, which has striking differences in GC content between closely related rodent species and requires dedicated GC-rich DNA enrichment protocols for sequencing. We contrasted *COA1* with the *PDX1* genes of rodents by comparing the minimum and maximum (see **Supplementary Table S14**) GC contents possible given their amino acid sequence. Although *COA1* had lower GC content levels than *PDX1*, we could not rule out the possibility of gBGC affecting some exons. The values of GC* (strong and weak convergence) across more than 200 vertebrate species with intact *COA1* reading frames suggested considerable heterogeneity between taxa (see **Supplementary Fig. S188** and **Supplementary Table S15)**. In each taxonomic group, the prevalence of gBGC was separately quantified (see **Supplementary Fig. S189-S222**). Strong patterns of gBGC occur in the *COA1* sequence of several species (see **Supplementary Fig. S189-S222**: elephant (*Loxodonta africana*), kagu (*Rhynochetus jubatus*), blue-crowned manakin (*Lepidothrix coronata*), Chilean tinamou (*Nothoprocta perdicaria*), American black bear (*Ursus americanus*), North American river otter *(Lontra canadensis),* meerkat (*Suricata suricatta*), California sea lion (*Zalophus californianus*), little brown bat (*Myotis lucifugus*), large flying fox (*Pteropus vampyrus*), southern pig-tailed macaque (*Macaca nemestrina*), Brazilian guinea pig (*Cavia aperea*), sheep (*Ovis aries*), eastern brown snake (*Pseudonaja textilis*) and the Goode's thornscrub tortoise (*Gopherus evgoodei*)). However, no rodent species with intact *COA1* show any striking gBGC patterns (**Supplementary Table S16**). The GC content vs. K-mer abundance plots of PacBio, BGI-seq, and Illumina datasets spans the entire range of GC contents seen in *COA1* exons (see **Supplementary Fig. S223**). Since the GC content of individual *COA1* exons differs between species groups (see **Supplementary Fig. S224-S228** and **Supplementary Table S17**), the high GC content of certain regions might result in inadequate sequencing coverage of the *COA1* gene in some species. Hence, the lack of sequencing reads covering *COA1* cannot serve as definitive evidence of gene loss.

S4C text: *COA1* occurs in an evolutionary breakpoint region

Both *COA1* and *STK17A* are missing in the post-CR rodent genome assemblies. The search of the genome assemblies, sequencing raw read datasets, and RNA-seq datasets also failed to find evidence of an intact *COA1* or *STK17A* gene. All raw read and genome assembly hits for *STK17A* while using queries from pre-CR rodent genomes could be traced back to the *STK17B* gene that matches the *STK17A* gene at a short sequence stretch. The *STK17A* gene is lost or has sequence properties that prevent sequencing with currently available technologies. The exon-1 region of *COA1* occurs in a gene desert region between *PTPRF* and *HYI* genes in post-CR species. Using blast search of *COA1* introns, we found strong support for the existence of *COA1* intron-2 close to the exon-1 hit. Pairwise genome alignments support the presence of the *COA1* gene remains at this location (see **Supplementary File S7**). Notably, the *COA1* remnants of a truncated exon-1 and intron-2 occur in the gene desert between *PTPRF* and *HYI* genes only in post-CR species. None of the pre-CR species had any such remains. Hence, the *COA1* remnants between *PTPRF* and *HYI* genes are unlikely to have resulted from duplicated copies of *COA1*. The synteny of this region is well conserved with *KDM4A* and *PTPRF* on the left flank and *HYI* and *SZT2* on the right side and corresponds to gene order O8. Careful examination of this region in RNA-seq datasets found no evidence of transcripts.

S5 text: Implications of gene loss

Gene loss can be dealt with through compensation from another gene[20] or is associated with a biological pathway rewiring[21]. Large-scale changes in gene content are associated with major evolutionary transitions that drastically alter the fitness landscape. Prominent examples of such shifts are the origin of flight in birds[22] and the movement of mammals from land to water seen in cetaceans[23]. Recurrent gene loss events following relaxed selective constraint in various other lineages have also been documented[24–26]. Differences in the immune response of Galliformes and Anseriformes are linked to lineage-specific gene loss[27,28]. The *COA1* gene is not known to have any obvious immune functions, and its loss in galliform birds appears to be a consequence of relaxed selection on the OXPHOS pathway.

The correlation between recurrent gene loss and specific phenotypes has provided crucial insights into the evolution of traits. Stomach loss in gnathostomes co-occurs with the loss of several genes that code for digestive enzymes[29]. The loss of ketogenesis has occurred through the recurrent loss of the *HMGCS2* gene[30]. Gene losses associated with dietary composition, the patterns of feeding, and gut microbiomes have also been identified[31]. Recurrent loss of Toll-like receptors (TLRs), which play prominent roles in the innate immune system, is associated with impaired ability to detect extracellular flagellin[32]. The repeated loss of the *CORT* (cortistatin) gene is related to modifications in the circadian pathway[24]. In the *COA1* gene, we record the independent occurrence of gene disrupting changes in closely related species of galliforms and rodents. However, we cannot rule out the possibility of a common regulatory mutation that initially resulted in the loss of gene expression followed by the independent accumulation of the gene disrupting changes that we observe. Our hypothesis of relaxed selection on the OXPHOS pathway predicts gene loss following skeletal muscle fiber composition changes. The *COA1* gene does not alter the muscle fiber composition and might have experienced relaxed selective constraint due to increased fast glycolytic (FG) muscle

fibers. Hence, it is tempting to speculate that the independent gene disrupting changes reflect recurrent gene loss events. However, the mechanistic basis of changes in muscle fiber composition between species is yet to be understood. Identifying the genetic changes that determine muscle fiber composition and the sequence of events would clarify when and why the *COA1* gene loss occurred.

Species with exceptionally large body sizes or extremely long lifespans have a greater number of cell divisions. An increment in the number of cell divisions enhances cancer risk. However, paradoxically, large-bodied animals like elephants and whales do not have a higher cancer incidence[33,34]. Cancer resistance due to lineage-specific changes in gene content may explain this paradox[35–37]. While specific genetic changes in mammalian species lead to cancer resistance[38,39], the reasons for lower cancer incidence in birds compared to mammals are mostly unexplored[40].

Interestingly, the *COA1* gene is an oncogene with a role in colorectal cancer[41], and its loss could reduce cancer risk. Silencing of *COA1* by miRNAs strongly suppresses giant cell tumors of the bone[42,43]. Our discovery of *COA1* gene loss in galliforms sets a precedent for the indisputable identification of gene loss events in birds and might reveal other oncogenes which are lost. We also identify *COA1* gene loss in the beaver and naked mole-rat genomes, species that are models to study longevity[16]. High-quality near-complete vertebrate genomes with very few errors will further aid in the large-scale identification of gene loss events across the vertebrate phylogeny[44].

S6 text: Validation of *COA1* annotation

Annotation across most species endorses the existence of four coding exons that produce a ~130 to 140 amino acid (aa) protein. The *COA1* annotation in the human genome (see **Supplementary Fig. S229**) has multiple isoforms with seven exons. The additional three exons annotated in the human genome upstream from the widely conserved four exons need further investigation. Bird species such as *Nipponia nippon*, *Cuculus canorus*, *Pterocles gutturalis*, *Gavia stellata*, *Buceros rhinoceros silvestris*, *Anser cygnoides domesticus*, *Anas platyrhynchos* (corrected in XM_027451320.2), and *Fulmarus glacialis* have annotation for a fifth exon upstream from the widely conserved four exons. Annotation for multiple isoforms of the *COA1* gene also exists in *Athene cunicularia*, *Tyto alba*, *Calidris pugnax*, *Serinus canaria*, *Corvus moneduloides*, *Corvus brachyrhynchos*, *Egretta garzetta*, *Aquila chrysaetos*, *Pipra filicauda*, *Corvus cornix*, *Cygnus atratus*, and *Parus major* (see **Supplementary Table S18)**. We examined RNA-seq datasets of multiple species to evaluate the expression of the isoforms. RNA-seq data in *Colius striatus* and *Eurypyga helias* (which had partial sequences annotated) allowed the reconstruction of full-length open reading frames (ORFs). In addition to bird genomes, the *COA1* gene ortholog is annotated in lizards (*Zootoca vivipara*, *Podarcis muralis*, *Lacerta agilis*, *Anolis carolinensis*, *Gekko japonicus*, *Thamnophis sirtalis*, *Pantherophis guttatus*, *Notechis scutatus*, *Pseudonaja textilis* and *Python bivittatus*), turtles (*Trachemys scripta elegans*, *Chelonia mydas*, *Chelonoidis abingdonii*, *Chrysemys picta*, *Gopherus evgoodei* and *Pelodiscus sinensis*), alligators (*Gavialis gangeticus*, *Alligator sinensis*, *Alligator mississippiensis* and *Crocodylus porosus*), even-toed ungulates (*Bos*

taurus, *Sus scrofa*, *Odocoileus virginianus texanus*, *Bison bison bison*, *Bos indicus x Bos taurus*, *Bos mutus*, *Bubalus bubalis*, *Capra hircus*, *Ovis aries*, *Vicugna pacos*, *Camelus ferus*, *Camelus bactrianus*, *Camelus dromedarius*, *Neophocaena asiaeorientalis asiaeorientalis*, *Balaenoptera acutorostrata scammoni*, *Lipotes vexillifer*, *Lagenorhynchus obliquidens*, *Globicephala melas*, *Orcinus orca*, *Tursiops truncatus*, *Phocoena sinus*, *Monodon monoceros*, *Delphinapterus leucas*, *Physeter catodon* and *Balaenoptera musculus*), odd-toed ungulates (*Equus caballus*, *Equus asinus*, *Equus przewalskii* and *Ceratotherium simum simum*), pangolins (*Manis pentadactyla* and *Manis javanica*), *Galeopterus variegatus*, *Tupaia chinensis* and Primates (*Homo sapiens*, *Macaca mulatta*, *Pan troglodytes*, *Chlorocebus sabaeus*, *Callithrix jacchus*, *Colobus angolensis palliatus*, *Cercocebus atys*, *Macaca fascicularis*, *Macaca nemestrina*, *Papio anubis*, *Theropithecus gelada*, *Mandrillus leucophaeus*, *Trachypithecus francoisi*, *Rhinopithecus bieti*, *Rhinopithecus roxellana*, *Piliocolobus tephrosceles*, *Gorilla gorilla*, *Pan paniscus*, *Pongo abelii*, *Nomascus leucogenys*, *Hylobates moloch*, *Saimiri boliviensis*, *Sapajus apella*, *Cebus imitator*, *Aotus nancymaae*, *Carlito syrichta*, *Propithecus coquereli*, *Microcebus murinus* and *Otolemur garnettii*).

We screened the synteny pattern of the candidate *COA1* gene in Galliformes and Anseriformes using five upstream genes (*STK17A*, *HECW1*, *TNS3*, *PSMA2*, *MRPL32*) and the five downstream genes (*BLVRA*, *VOPP1*, *LANCL2*, *EGFR*, *SEC61G*) (see **Supplementary Table S6 and S19**). The chicken (*Gallus gallus*) has a chromosome level assembly, and the gene occurs on Chromosome 2, and its region is syntenic with human (*Homo sapiens*) chromosome 2 (**Supplementary Fig. S230-231**). The gene synteny is mostly conserved in these species and is present on the same scaffold/chromosome. The blast search of the genome using the query gene sequence of closely related species identified genes missing in the annotation. *Anas platyrhynchos* has chromosome level assembly with the same gene order as *Gallus gallus* (**Supplementary Fig. S232**). *Anser cygnoides* and *Anseranas semipalmata* also contain this conserved gene order. *Anas platyrhynchos*, *Numida meleagris*, *Coturnix japonica*, *Meleagris gallopavo*, *Aquila chrysaetos chrysaetos*, *Parus major*, *Strigops habroptila*, *Taeniopygia guttata*, *Felis catus*, *Canis lupus familiaris*, *Panthera leo*, *Pan troglodytes*, *Pan paniscus*, *Gorilla gorilla gorilla*, *Pongo abelii*, *Macaca mulatta*, *Ornithorhynchus anatinus*, *Balaenoptera musculus*, *Danio rerio*, *Podarcis muralis*, and *Chrysemys picta bellii* show syntenic blocks aligning with the human chromosome 7 (**Supplementary Fig. S233-S253**). Synteny-based verification was done clade-wise in birds (see **Supplementary Table S6**), rodents (**Supplementary Table S13**), Carnivores (**Supplementary Table S20**), Primates (**Supplementary Table S21**), and reptiles. Gene order and synteny relationships for representative species from each of the clades are in **Supplementary Fig. S254-S427**. For pairwise alignment of human *COA1* gene with different species, see **Supplementary Fig. S428-S457**.

Vertebrate species have a conserved *COA1* gene intron/exon organizational structure. However, two lineages (Primates and Carnivores) with evidence of intron/exon organization changes have also had *COA1* gene duplication events. To ensure that the observed differences were not a result of incorrect annotation, alignment artifacts, or duplicated copies, we

compared the *COA1* gene organization across diverse vertebrate species. Subsequently, we validated the annotations from NCBI and Ensembl using RNA-seq datasets. Sequencing read haplotypes from the functional and pseudogenized copy can be distinguished as their sequences have diverged.

Analysis of Hi-C (High-throughput Chromosome Conformation Capture) datasets

We downloaded the Hi-C data from ENA (European Nucleotide Archive) for human (*Homo sapiens*), rabbit (*Oryctolagus cuniculus*), naked mole-rat (*Heterocephalus glaber*), mouse (*Mus musculus*), chicken (*Gallus gallus*), cheetah (*Acinonyx jubatus*), rhesus macaque (*Macaca mulatta*) and dog (*Canis lupus familiaris*) (see **Supplementary Table S22)**. For human, rabbit, mouse, rhesus macaque, and dog genomes, we used the data from liver tissue. We used the data from blood tissue for the cheetah and in naked mole-rat the embryonic fibroblast cells. Embryonic tissue fibroblasts, mature erythrocytes, and immature erythrocytes tissue are used for chicken. To process Hi-C data, we used HiCUP (https://www.bioinformatics.babraham.ac.uk/projects/hicup/) with different genome digestion parameters of restriction enzymes (see **Supplementary Table S22**). We made a config file for each species which contains the details of input files and the required path of tools. HiCUP generates valid interaction pairs mapped bam files using bowtie2. The mapped files can be converted into prejuicer format using the hicup2juicer command available in HiCUP.

The prejuicer file lists valid interaction pair information for the contact map. This file was converted into a binary .hic file using juicer tools. We used the .hic format files as input to Juicebox to visualize the interaction map in the genomic region containing *COA1*. Assembly errors result in patterns of interaction between distantly located regions and gaps in the interaction map. We visualized these interaction maps using Juicebox. As a negative control, we simulated the effect of assembly gaps by removing the reads mapped to the genomic regions of different sizes (i.e., 10 kb, 50 kb, 100 kb, 500kb, and 1Mb) from the *COA1* and nearby regions using one human sample (SRR1501150). The visualization of these datasets appeared as gaps in the interaction map generated by Juicebox. When the 10kb region was subtracted, the gap in the interaction was not visible (**Supplementary Fig. S458**). However, when the 50 kb region was subtracted, a small gap is visible at the 25 kb resolution as a region lacking interaction (**Supplementary Fig. S459**). Similarly, when we increase the size of the region for which reads are subtracted to 100Kb, a more apparent gap is visible (**Supplementary Fig. S460**). For the 500kb and 1MB subtracted region, we found an apparent pattern with a large gap without any interaction (**Supplementary Fig. S461-S462**). Based on these results, we suggest that regions > 50kb with assembly errors can be identified by visual inspection of Hi-C interaction maps.

S7 text: Assessing the transcriptional status of *COA1*

No evidence for transcription of *COA1* gene in chicken exists in the RNA-seq data from 23 tissues consisting of blood, bone marrow, breast muscle, bursa, cerebellum, cerebrum, comb, eye, fascia, gallbladder, gizzard, gonad, heart, immature egg, kidney, liver, lung, mature egg,

pancreas, shank, skin, spleen, and uterus (**Supplementary Fig. S466-S534**). Among other galliform species, we found no evidence for the expression of the *COA1* gene. (The spleen and gonad of the peacock, the skin of golden pheasant, gonad, spleen, brain, muscle, liver, and heart of ring-necked pheasant, bursa, gonad spleen, blood and uterus of helmeted guineafowl, breast muscle, gonad, spleen, brain, liver, heart, and bursa of turkey, kidney, liver, muscle, lung, and heart of Japanese quail, the blood of *Colinus virginianus* and blood of *Syrmaticus Mikado*, see **Supplementary Fig. S535-603**). The only galliform species with a transcribed *COA1* gene was *Alectura lathami* in blood tissue (**Supplementary Fig. S604-S606**).

In contrast to Galliformes, the *COA1* gene is intact in the Anseriformes. However, the *COA1* gene annotation in duck (*Anas platyrhynchos platyrhynchos*) contains two isoforms. The more extended isoform codes for a 265 amino acid protein and consists of five exons. The shorter isoform (139 amino acids) is orthologous to the Galliformes ORF. Upon closer inspection of the first exon, only 24 of the 372 bases have RNA-seq read support **(Supplementary Fig. S607)**. Hence, this additional exon might be an annotation artifact or part of the untranslated region. The last four annotated exons, which correspond to the intact 139 amino acid encoding sequence, were found to be robustly expressed in the gonad, spleen, liver, brain, and skin (**Supplementary Fig. S608-S615**). A similar annotation of the fifth exon in *Anser cygnoides domesticus* appears to be an artifact. The gonad, liver, and spleen express the last four exons (see **Supplementary Fig. S616-S622**). The RNA-seq data from blood tissue for magpie goose (*Anseranas semipalmata*) and southern screamer (*Chauna torquata*) also supported the transcription of the *COA1* gene (**Supplementary Fig. S623-S626**).

Having verified the expression of the *COA1* gene in multiple Anseriformes, we screened additional bird RNA-seq datasets to evaluate the transcriptional activity of the intact ORF found in these species. Many other bird genomes have annotations for multiple isoforms of the *COA1* gene, like the duck genomes. These isoforms range in length from 136 to 265 amino acids and 4 to 7 exons. Based on careful examination of multiple RNA-seq datasets across several closely related species and sequence homology, we found that the four-exon transcript coding for a 139 amino acid protein was the only correct annotation in most cases. However, additional exons have a robust expression in rare cases and require further investigation. In the Corvidae group, annotation exists for transcripts of lengths 170 and 139 aa. The first exon of the longer transcript lacked expression.

In comparison, all four transcripts of the shorter transcript are present in the blood tissue of western jackdaw (*Corvus monedula*) as well as gonad, brain, spleen, and liver of hooded crow (*Corvus cornix*) **(Supplementary Fig. S627-S632)**. The common canary (*Serinus canaria*) has three transcripts with 177, 154, and 139 aa **(Supplementary Fig. S633-S634)**. We checked the expression using liver and skin tissue and found support for all three transcripts. However, the transcript with 139 aa was strongly expressed upon closer inspection, and the other two transcripts are potentially artifacts. The great tit (*Parus major*) has two transcripts of lengths 169 and 139 aa. While the kidney and liver express both

transcripts, the first exon has feeble expression and appears artefactual **(Supplementary Fig. S635-S636)**.

The golden eagle (*Aquila chrysaetos*) has four annotated transcripts with lengths of 219, 180, 159, and 139 aa. Transcript of 219 aa length contains six exons, transcripts of length 180 aa, and 159 aa have five exons, and 139 aa transcript contains four exons. We found that exon 1 showed negligible expression, and exons 2 to 6 have high expression levels. However, exon 1 and 2 both have an in-frame stop codon **(Supplementary Fig. S637-S641)**. Hence, we consider that the 139 aa long transcripts expressed in the liver and muscle are correct. Red-throated loon (*Gavia stellata*) has a single five exon transcript of length 155 aa annotated. We discovered a lack of expression in the first exon compared to the last four exons orthologous to the transcript of length 139 aa **(Supplementary Fig. S642-S643)**.

The ruff (*Calidris pugnax*) genome annotates three transcripts with lengths of 233, 229, and 139 aa. Transcript one and two contain seven exons each, and the third transcript contains four exons. Exons 1 and 2 lack expression in the first two transcripts, and the third exon did not have any start codon explaining the transcript. The last four exons have transcripts and are orthologous to the *COA1* gene in other species **(Supplementary Fig. 644-S648)**. In the little egret (*Egretta garzetta*), transcripts of lengths 212 and 203 are annotated and contain five exons. We found evidence of expression of *COA1* in blood tissue **(Supplementary Fig. S649-S650)**. Although the first exon has a lower expression level than the last four exons, the consistent occurrence of the fifth exon across many species suggests it might be part of the untranslated region. We annotated and verified the expression of *COA1* in *Phalacrocorax carbo, Phaethon lepturus, Opisthocomus hoazin,* and *Leptosomus discolor* (**Supplementary Fig. S651-S658)**. *Eurypyga helias* has an unverified transcript length of 121 aa. Hence, we screened the genome and RNA-seq data and found its transcript length is 139 aa **(Supplementary Fig. S659-S661)**. We verified the *COA1* gene expression using RNA-seq data in *Strigops habroptila* as it had less than 100 percent RNA-seq coverage **(Supplementary Fig. S662-S663)**. We also examined the RNA-seq data from a few other bird species to verify the *COA1* gene (see **Supplementary Fig. S664-S711**). Bird species share this conserved gene order (**Supplementary Fig. S712**). The Anolis lizard (*Anolis carolinensis*) liver also expresses the *COA1* gene (**Supplementary Fig. S713-S714**).

RNA-seq datasets from the European rabbit's (*Oryctolagus cuniculus*) heart and liver showed no evidence of transcription of *COA1* (see **Supplementary Fig. S715-S717**). In contrast to the rabbit, intact *COA1* gene is present in the Royle's pika (*Ochotona roylei*) and Daurian pika (*Ochotona dauurica*) with blood RNA-seq datasets showing robust expression (see **Supplementary Fig. S718-719**). Screening RNA-seq datasets from the root ganglion, spinal cord, ovary, liver, spleen, and testis in the naked mole-rat (*Heterocephalus glaber*) revealed no transcription of *COA1* locus (see **Supplementary Fig. S720-728**). The closely related Damaraland mole-rat (*Fukomys damarensis*) has robust *COA1* expression in the brain, liver, and testis (see **Supplementary Fig. S729-S734**). The Brazilian guinea pig (*Cavia aperea*), the guinea pig (*Cavia porcellus*), and the long-tailed chinchilla (*Chinchilla lanigera*) were all found to express the *COA1* gene robustly (see **Supplementary Fig. S735-S742**). The thirteen-lined ground squirrel (*Ictidomys tridecemlineatus*), the Arctic ground squirrel

(*Urocitellus parryii*), the groundhog (*Marmota monax*), and the Himalayan marmot (*Marmota himalayana*) do not express the *COA1* locus (see **Supplementary Fig. S743-S761**). In contrast to these species, the Eurasian red squirrel (*Sciurus vulgaris*) has an intact *COA1* expressed in the skin (see **Supplementary Fig. S762-S763**). Despite gene disrupting mutations, the North American beaver (*Castor canadensis*) *COA1* locus is expressed in the blood and spleen (see **Supplementary Fig. S764-S765**). Other tissues such as the brain, liver, stomach, ovarian follicle, skeletal muscle, and kidney do not show any expression at the *COA1* locus (see **Supplementary Fig. S766-S771**). The expressed transcript might represent a new long non-coding RNA that cannot produce a functional *COA1* protein due to the presence of premature stop codons.

Chromosomal rearrangement in rodent species has resulted in the relocation of genes flanking *COA1* to new locations. The *BLVRA* gene is transcriptionally active in the mouse's (*Mus musculus*) liver and heart even though it has translocated to an entirely different location between *AP4E1* and *NCAPH* (see **Supplementary Fig. S772**). Genes on the left flank consisting of *HECW1, PSMA2,* and *MRPL32* are now located beside *ARID4B* and are expressed in the mouse (see **Supplementary Fig. S773-S774**). The genes from the right flank (*MRPS24* and *URGCP*) are also transcriptionally active in the mouse at their new location beside *ANKRD36* (see **Supplementary Fig. S775**). Remnants of *COA1* occur between the *PTPRF* and *HYI* genes. However, no transcriptional activity is seen in the mouse in the region between *PTPRF* and *HYI* genes (see **Supplementary Fig. S776**). The new gene order and gene expression patterns are shared by rat (*Rattus norvegicus*) (see **Supplementary Fig. S777-S781**), steppe mouse (*Mus spicilegus*) (see **Supplementary Fig. S782-S786**), Gairdner's shrewmouse (*Mus pahari*) (see **Supplementary Fig. S787-S791**), Ryukyu mouse (*Mus caroli*) (see **Supplementary Fig. S792-S796**), Algerian mouse (*Mus spretus*) (see **Supplementary Fig. S797-801**), deer mouse (*Peromyscus maniculatus*) (see **Supplementary Fig. S802-S806**), prairie vole (*Microtus ochrogaster*) (see **Supplementary Fig. S807-S811**), golden hamster (*Mesocricetus auratus*) (see **Supplementary Fig. S812-S816**), Mongolian gerbil or Mongolian jird (*Meriones unguiculatus*) (see **Supplementary Fig. S817-S820**), Chinese hamster (Cricetulus griseus) (see **Supplementary Fig. S821-S825**), Northern Israeli blind subterranean mole-rat (*Nannospalax galili*) (see **Supplementary Fig. S826-S830**), white-footed mouse (*Peromyscus leucopus*) (see **Supplementary Fig. S831-S835**) and fat sand rat (*Psammomys obesus*) (see **Supplementary Fig. S836-S840**). The banner-tailed kangaroo rat (*Dipodomys spectabilis*) (see **Supplementary Fig. S841-S842**) has a different gene order and appears to represent one of the pre-CR species. However, we cannot rule out the possibility of genome assembly errors.

The genome assemblies of rodents such as the mouse and rat are well-curated and represent some of the highest-quality reference genomes[44]. To ensure that the CRs identified are correct, we evaluated the correctness of genome assemblies of the mouse (see **Supplementary Fig. S843-S849**) and white-footed mouse (*Peromyscus leucopus*) (see **Supplementary Fig. S850-S857**) using PacBio long-read sequencing datasets. The mouse genome assembly has been finished to a very high quality using artificial clones of genome fragments[45]. We further verified the mouse genome assembly by visualizing the coverage of

assembly fragments across the genomic regions of interest (see **Supplementary Fig. S858-S863**). Repeat regions occur at the boundaries of the EBRs (see the last row of the screenshots). Although repeat regions are a major contributing factor for the misassembly of genomes, the conserved gene orders across several species and concordance in the timing of the CR and support from long-read data support the presence of a genuine change in gene order.

The *COA1* gene is intact and robustly expressed in the platypus's (*Ornithorhynchus anatinus*) heart and brain (see **Supplementary Fig. S864-S866**). Gene order in the short-beaked echidna (*Tachyglossus aculeatus*) matches the platypus and other outgroup species (see **Supplementary Fig. S867)**. In contrast to the monotreme species, all marsupial genomes analyzed have a different gene order following CRs. The gray short-tailed opossum (*Monodelphis domestica*) has the gene *ACVR2B* beside the new location of the right flank genes of *COA1*. The left flank genes are beside *GPR141B*. No traces of the *COA1* gene are found either in the genome assembly or raw read datasets. The opossum's brain expresses these adjacent genes with no transcripts in the intergenic regions (see **Supplementary Fig. S868-S870**). The gene order and transcriptional activity were the same in the tammar wallaby (*Notamacropus eugenii*) (uterus: see **Supplementary Fig. S871-S872**), koala (*Phascolarctos cinereus*) (liver and PBMC (peripheral blood mononuclear cell): see **Supplementary Fig. S873-S875**), the Tasmanian devil (*Sarcophilus harrisii*) (lung and spleen: see **Supplementary Fig. S876-S878**), and the common brushtail (*Trichosurus vulpecula*) (liver: see **Supplementary Fig. S879-S881**). Long-read sequencing data in the koala supports the correctness of genome assembly (see **Supplementary Fig. S882-S884**).

The NCBI annotation documents the presence of transcripts, and the *COA1* gene is remarkably well conserved in ungulate species (see **Supplementary Table S18**). Within ungulate species, certain cervid species have remarkable sprinting abilities that allow them to escape from predators. However, in addition to sprinting ability, these species are resistant to fatigue. Hence, the prediction from our hypothesis is that gene loss would not occur in cervid species. The white-tailed deer (*Odocoileus virginianus*) liver and retropharyngeal lymph node and the red deer (*Cervus elaphus*) blood transcriptomes express *COA1* (see **Supplementary Fig. S885-S888**).

The *COA1* gene has undergone duplication within the primate lineage. We screened the genomes of 27 primate species to track down when the gene duplication event occurred. Based on the presence of the duplicate copies, the duplication event is estimated to have happened in the last 43 million years (see **Supplementary Fig. S18**). Subsequent duplications have also occurred in Nancy Ma's night monkey (*Aotus nancymaae*) and a shared duplication in the black-capped squirrel monkey (*Saimiri boliviensis*) and the Panamanian white-faced capuchin (*Cebus imitator*). Concurrent with the gene duplication, the intron-exon structure of the *COA1* gene has also changed (see **Supplementary Fig. S48**). The functional copy of the *COA1* gene is transcriptionally active in the gray mouse lemur (*Microcebus murinus*) (kidney and lung: see **Supplementary Fig. S889-S890**), the northern greater galago (*Otolemur garnettii*) (liver: see **Supplementary Fig. S891**), Coquerel's sifaka (*Propithecus coquereli*) (see **Supplementary Fig. S892**), Nancy Ma's night monkey (*Aotus*

nancymaae) (liver, heart, and kidney: see **Supplementary Fig. S893-S895**), the common marmoset (*Callithrix jacchus*) (lung, liver, and kidney: see **Supplementary Fig. S896-S897**), the Panamanian white-faced capuchin (*Cebus imitator*) (blood: see **Supplementary Fig. S898-S900**), the black-capped squirrel monkey (*Saimiri boliviensis*) (ovary and heart: see **Supplementary Fig. S901-904**), the sooty mangabey (*Cercocebus atys*) (liver: see **Supplementary Fig. S905-S906**), the olive baboon (*Papio anubis*) (kidney and heart: see **Supplementary Fig. S907-S908**), the crab-eating macaque (*Macaca fascicularis*) (blood and liver: see **Supplementary Fig. S909-S910**), the golden snub-nosed monkey (*Rhinopithecus roxellana*) (heart and blood: see **Supplementary Fig. S911-S912**), human (*Homo sapiens*) (liver : see **Supplementary Fig. S913-S918**), and the Philippine tarsier (*Carlito syrichta*) (see **Supplementary Fig. S919**).

The intron/exon structure of the *COA1* gene has undergone several changes in the carnivore lineage (see **Supplementary Fig. S48**). However, outgroup species such as the horse (*Equus caballus*) and pangolin (*Manis javanica*) lack intron/exon structure (see **Supplementary Fig. S48**). We screened the RNA-seq dataset of multiple carnivore species to validate the annotation and evaluate the intron/exon structure changes. Alternative exon usage was also carefully analyzed to quantify the transcriptional status of *COA1* in different carnivore species. The *COA1* gene is transcriptionally active in the meerkat (*Suricata suricatta*) (testis and liver: see **Supplementary Fig. S920-S922**), dog (*Canis lupus familiaris*) (spleen and skeletal muscle: see **Supplementary Fig. S923-S934**), ferret (*Mustela putorius furo*) (heart and kidney: see **Supplementary Fig. S935-S936**), the giant panda (*Ailuropoda melanoleuca*) (heart and liver: see **Supplementary Fig. S937-S938**), American black bear (*Ursus americanus*) (liver, kidney, and the brain: see **Supplementary Fig. S939-S940**), and Weddell seal (*Leptonychotes weddellii*) (lung and muscle: see **Supplementary Fig. S941-S943**). Detailed investigation of the splice junctions and actual positions of splice sites in dog transcriptome also supports the *COA1* gene annotation.

Skipping of the dog-like-exon-3 occurs in the transcriptomes of tiger (*Panthera tigris altaica*), lion (*Panthera leo persica*), cat (*Felis catus*), and puma (*Puma concolor*) (see **Supplementary Fig. S24-S42 and S944-S949**). Although annotation for the *COA1* locus exists in the cheetah (*Acinonyx jubatus*), we found no transcripts in the skin RNA-seq data (see **Supplementary Fig. S55-S57**). Closer inspection of the *COA1* locus in cheetah suggests gene loss. We further compared the splice isoforms found in canine and felid species through sashimi plots of the *COA1* locus. The sashimi plot shows the links between the splice sites and the number of reads that are splice mapped between these sites (see **Supplementary Fig. S950-S954**). Changes in the splice enhancers and splice silencer elements were also compared between cat and dog (see **Supplementary Fig. S43**).

Co-expressed genes tend to perform related functions and are lost together. Hence, to identify the loss of genes related to *COA1*, we identified the top 50 genes co-expressed with human ortholog based on the correlation values in COXPRESdb ver. 7.3[46]. The presence of orthologs of these co-expressed genes in the high-quality genomes of chicken and mouse using Ensembl BioMart suggests widespread conservation (**Supplementary Table S24**). None of these co-expressed genes appear lost in galliforms or rodents.

S8 text: Molecular evolutionary analyses

Relaxed selection signatures

Molecular signatures of relaxation in the degree of purifying selection generally accompany the loss of gene functionality and have been used as evidence of gene loss[26,47,48]. We relied upon multiple sequence alignments of Carnivores (see **Supplementary Table S1**), rodents (see **Supplementary Table S12**), and Primates (see **Supplementary Table S2**) to identify gene disrupting mutations and changes in intron-exon structure. We evaluated each taxonomic group for lineage-specific relaxed selection (see **Supplementary Table S7** and **S8**). Based on a previous report[11] of positive selection in primate species, we additionally identified positively selected sites using site models implemented in codeml (see **Supplementary Table S3**).

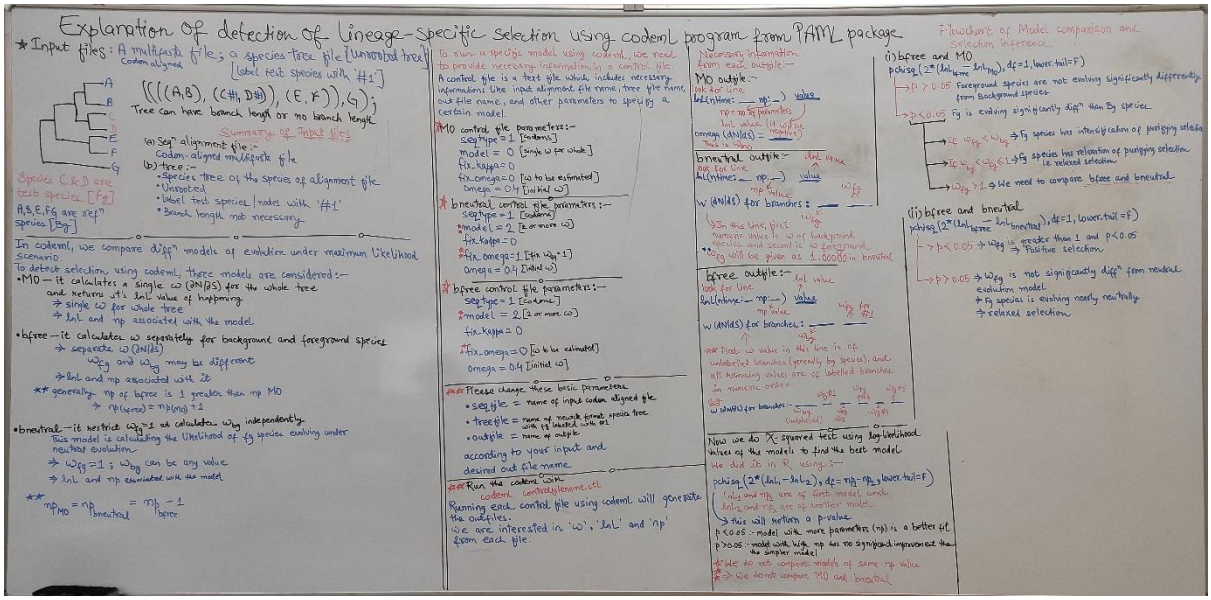
Based on the gene sequence of *COA1*, we could identify eleven galliform species with gene-disrupting mutations (see **Supplementary Table S4** and **S5**). Two other galliform species (*Chrysolophus pictus* and *Phasianus colchicus*) do not express the *COA1* gene in the RNA-seq datasets analyzed. Hence, we looked for signatures of relaxed selection in each of the terminal branches leading to each galliform species. We quantified branch-specific selection patterns using the program RELAX[49] from the HyPhy package and the codeml program from the PAML[50] package.

Identification of signatures of relaxed selection using HyPhy

We labeled the focal species as the foreground to test for relaxed selection in the terminal branches. We used the closely related species (identified based on sequence saturation test available in DAMBE[51] software) as the background species. We downloaded the phylogenetic tree with branch lengths from the TimeTree website. The codons in the ORF sequences were aligned using PRANK aligner using Guidance v2.02. We used RELAX to test for signatures of relaxed selection in each focal species using these alignments by labeling one species at a time as the foreground. The results are provided in **Supplementary Table S9**.

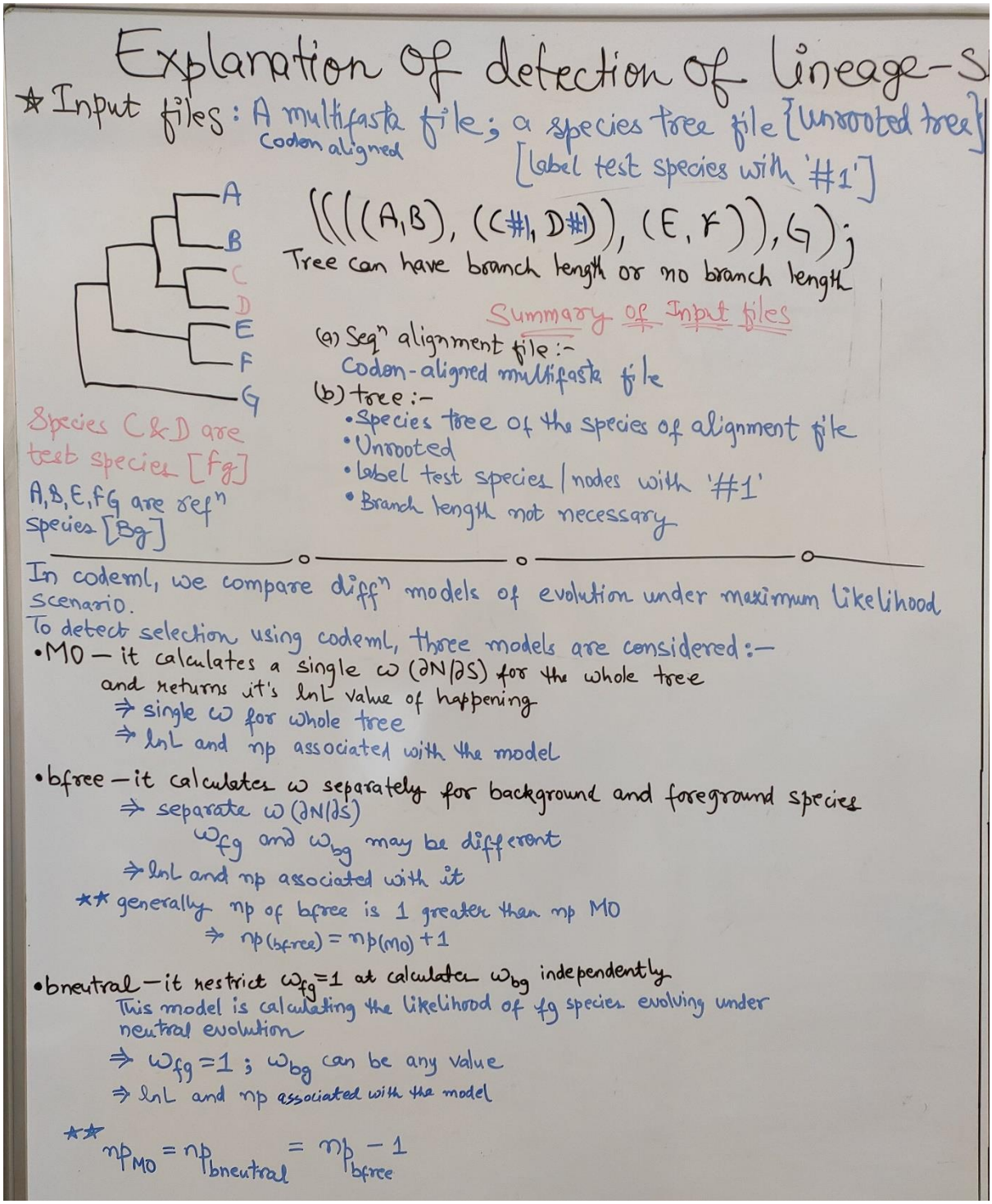
Identification of signatures of relaxed selection using codeml

The lineage-specific selection was inferred using the codeml program from the PAML package. PAML makes use of non-synonymous substitution rate/synonymous substitution rate (dN/dS or ω) to infer selection using three branch-specific models under maximum likelihood approach: M0 (null model), branch-free (alternate model), and branch-neutral (alternate model). We illustrate the various models used, config file settings and model testing procedure, and interpretation using whiteboard photographs and explanations.



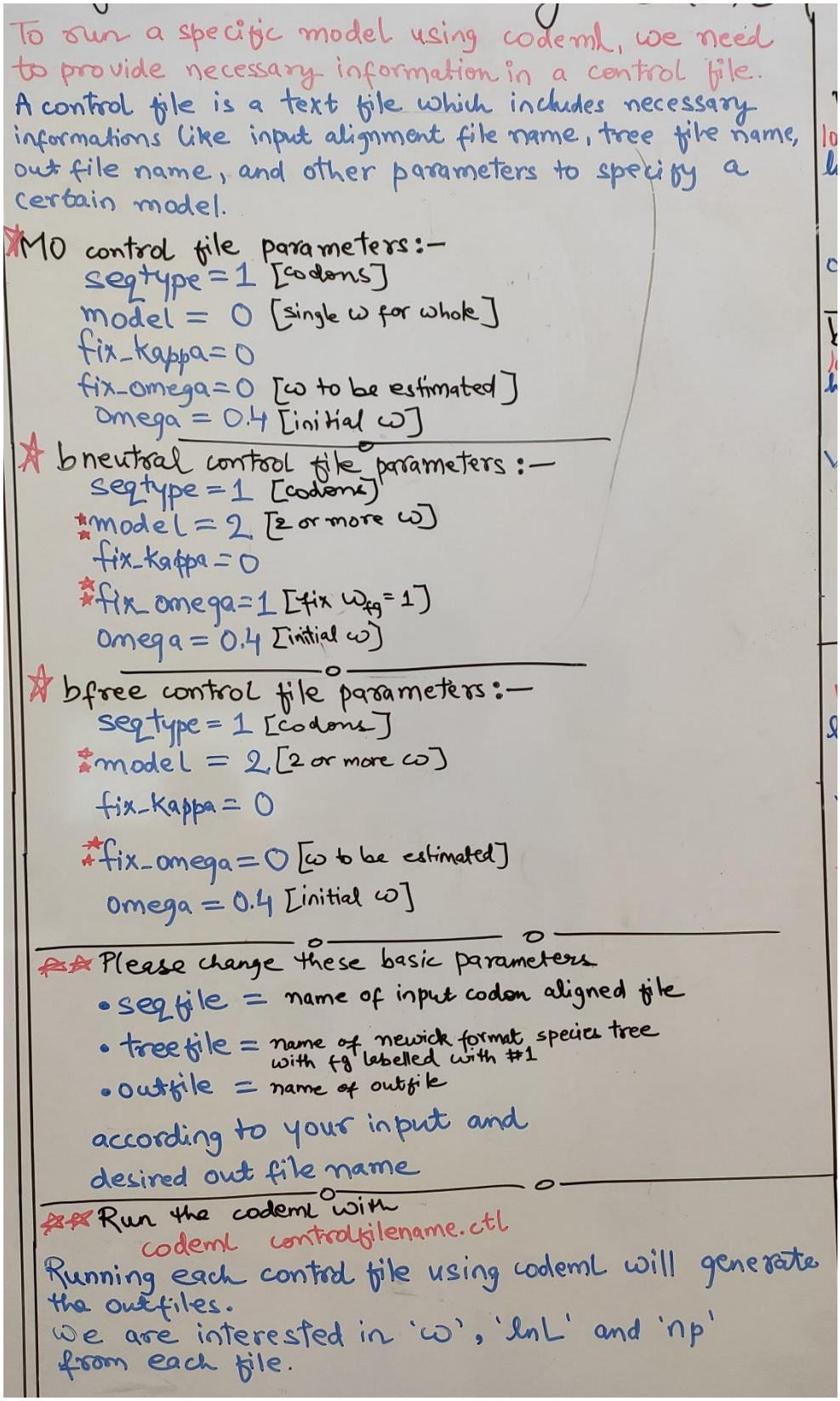
Photograph 1: Photograph of the entire whiteboard explaining the branch tests in codeml.

The M0 model assumes a single ω for all the species, while the branch-neutral model assumes the foreground species to evolve under neutral evolution ($\omega_{fg} = 1$). In contrast, the branch-free model calculates ω independently for foreground and background species. Details of each of these three models are shown in *Photograph 2*.



Photograph 2: Details of the three branch models used.

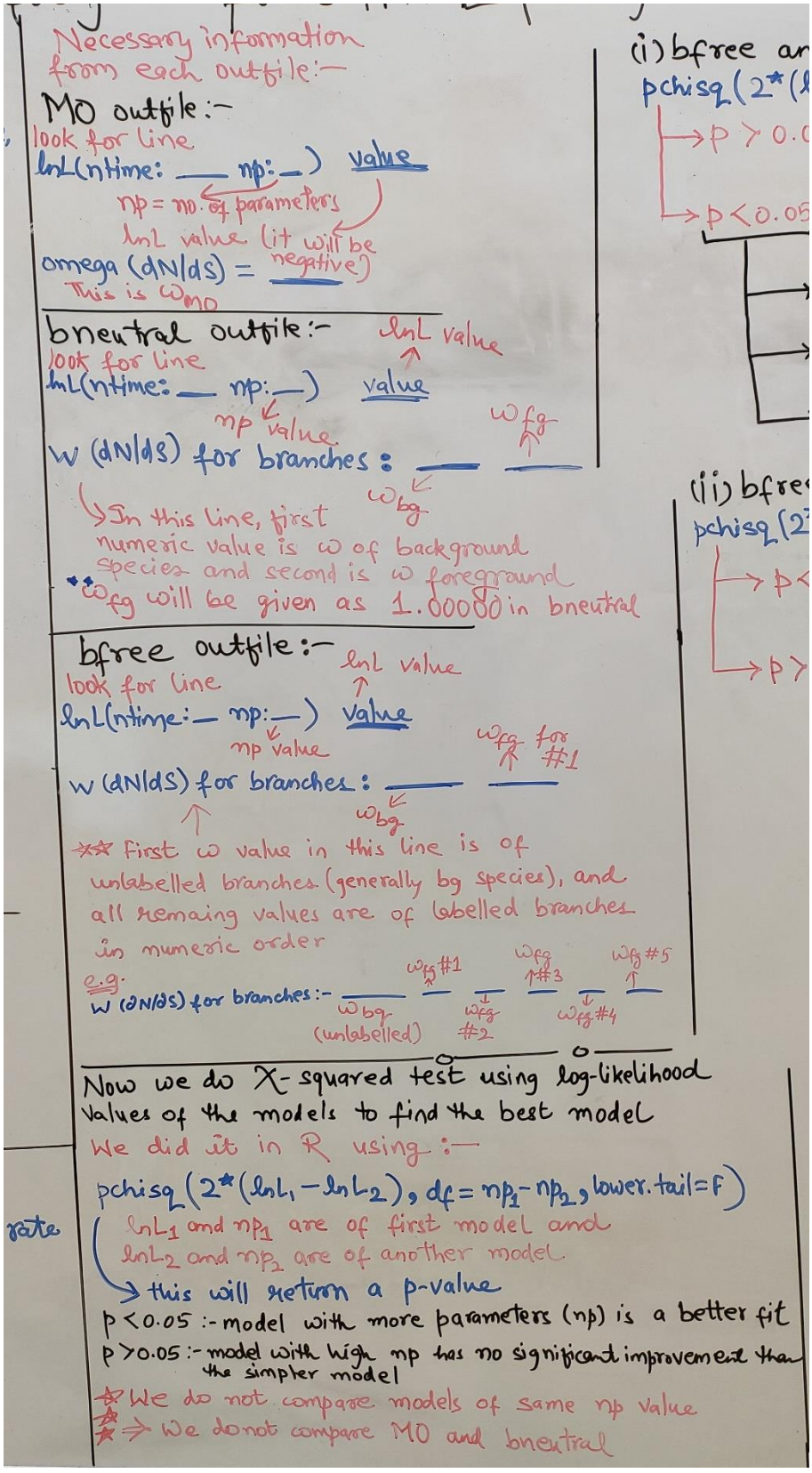
The codeml program is run using a config file with all the input files' details and parameter settings to be used for a particular run. We describe the settings to be used for each of the models in *Photograph 3*. The actual config files, input, and output files are available on the Github repository.



Photograph 3: Details of the config file settings used for each branch model.

The output files contain various details regarding the input dataset, config file settings, and output values from the run of the codeml. We require the estimates of ω and log-likelihood values from these output files for performing the likelihood ratio tests, which will determine the best fitting model. The actual lines of the output file which contain this information are

described in *Photograph 4*. The log-likelihood values are then used for calculating the p-values.

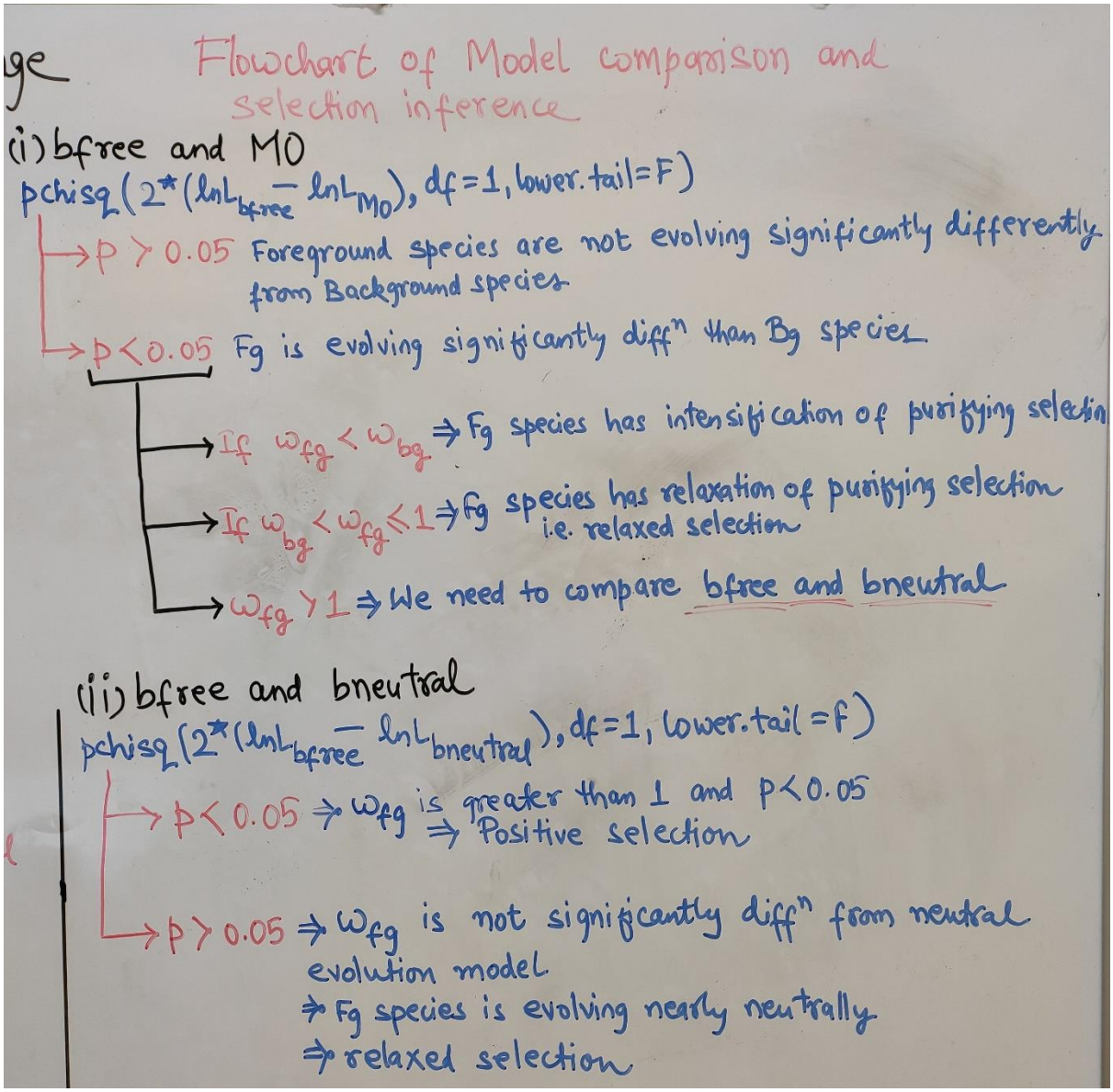


Photograph 4: Lines of the output file containing the relevant information.

The models of evolution were compared using the chi-squared test in a step-wise manner for their log-likelihood values to find the best fit model:

1. Comparison of branch-free and M0 models: A p-value less than 0.05 using the chi-squared test for their log-likelihood values implies that foreground branches are evolving significantly differently from background branches. The $\omega_{fg} < \omega_{bg}$ implies intensification of purifying selection while $\omega_{bg} < \omega_{fg} < 1$ implies relaxation of purifying selection. If $\omega_{fg} > 1$, branch-free and branch-neutral models are compared.

2. Comparison of branch-free and branch-neutral models: A p-value > 0.05 between branch-free and branch-neutral for their log-likelihood values implies that the branch-free model is not better than branch-neutral. This means the foreground species are evolving with near-neutral evolution and can be interpreted as relaxed selection. If the p-value between branch-free and branch-neutral is less than 0.05 and $\omega_{fg} > 1$, this implies that the foreground lineage is under positive selection.



Photograph 5: Steps involved in the interpretation of the likelihood ratio test results.

For our analyses, each functional species was kept as foreground, one at a time, against all other functional species as background. In case of gene duplication (both pseudogene or one functional, one pseudogene), each copy of the gene was considered alternatively against functional copies as background (**Supplementary Table S8**).

Our results for the lineage-specific selection shows that:

- In the Afrotheria group, *Trichechus manatus* is under relaxed selection, and *Orycteropus afer* is under intensification of purifying selection.
- In the Amphibian group, *Pantherophis guttatus* and *Pelodiscus sinensis* are under intensification of purifying selection.
- In Aves group, *Anser cygnoides*, *Chrysolophus pictus*, *Coturnix japonica*, *Gallus gallus*, *Lyrurus tetrix*, *Meleagris gallopavo*, *Numida meleagris*, *Odontophorus gujanensis*, *Pavo cristatus*, *Syrmaticus mikado*, and *Tympanuchus cupido* are under relaxation, and *Picoides pubescens* and *Taeniopygia guttata* are under intensification of purifying selection. The *Phasianus colchicus* (functional *COA1*) was under significant positive selection when all the 88 functional birds sequences were kept as background, but on closer inspection, we observed its dN/dS to be 999. We suspect that it is possibly an artifact.
- In the Carnivora group, *Leptonychotes weddellii* (duplicated *COA1*) is under the intensification of purifying selection.
- In the Primates group, *Chlorocebus sabaeus* is under relaxation, while *Carlito syrichta* is under intensification of purifying selection. In primates, the species with duplicate *COA1*, *Cebus imitator*, is under relaxed selection while *Chlorocebus sabaeus* and *Gorilla gorilla* are under intensification of purifying selection.
- In the rodents (using sequences that are saturated in the analysis), *Marmota marmota* and *Urocitellus parryii* are under relaxed selection.
- In the Rodents (using only unsaturated sequences in the analysis), *Fukomyses damarensis* and *Urocitellus parryii* are under relaxed selection.

We used the codeml program from the PAML package and the RELAX program from the HyPhy package to detect lineage-specific selection. The codeml program was run using the F1x4 and F3x4 codon frequency models. While results were broadly consistent between the two models, we have considered the results of the F3x4 model. Codeml can be used to infer relaxation of purifying selection, intensification of purifying selection, and positive selection. However, RELAX is used primarily to detect relaxed selection and intensified selection. RELAX program cannot delimit between positive selection and intensification of purifying selection. We compared the results generated using both programs to check the consistency and robustness of our results for lineage-specific selection.

For most of the foreground species and background species combinations, our results are consistent between both programs. However, the inference between both programs is not the same for a few species. Below is a list provided for the same:

1. *Anser cygnoides* showed relaxed selection using PAML and non-significant results using HyPhy.
2. *Buceros rhinoceros* showed non-significant results using PAML and relaxed using HyPhy.

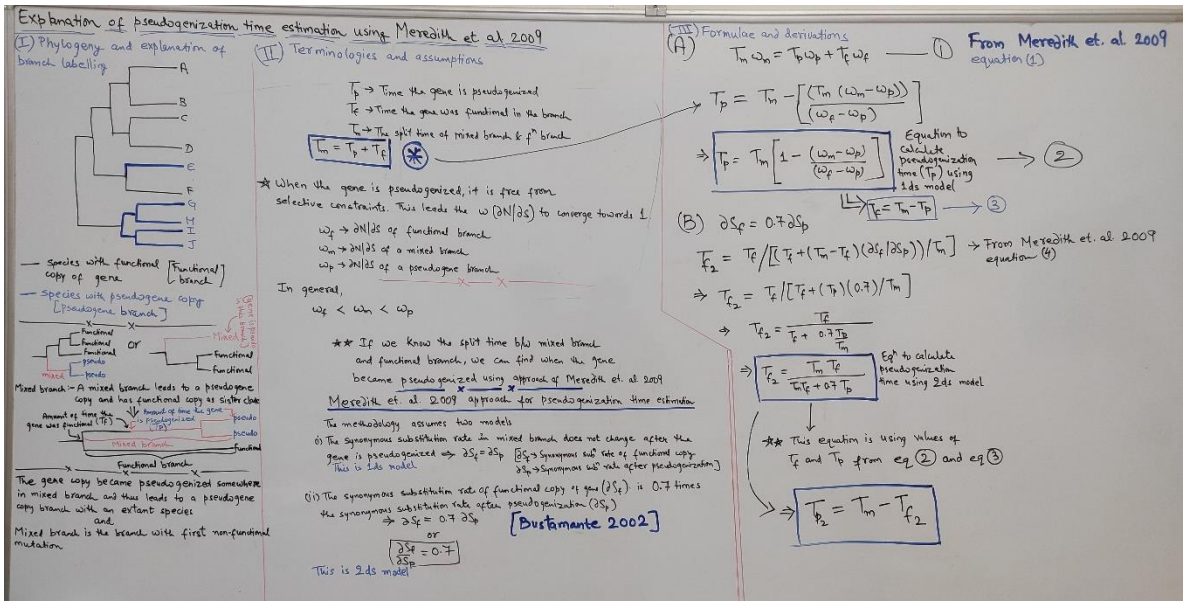
3. *Chlamydotis macqueenii* is non-significant results using PAML while intensified using HyPhy.
4. *Phalacrocorax carbo* is detected as intensified using PAML and non-significant results using HyPhy.
5. *Phasianus colchicus* is shown as positively selected using PAML, while HyPhy showed relaxation of selection.
6. *Picoides pubescens* species is detected as intensified using PAML and non-significant results using HyPhy.
7. *Stachyridopsis* ruficeps showed non-significant results using PAML and intensified using HyPhy.
8. *Taeniopygia guttata* is detected as intensified using PAML and non-significant results using HyPhy.
9. *Meleagris gallopavo* is detected as relaxed using PAML and intensified using HyPhy.
10. *Pavo cristatus* is detected as relaxed using PAML and intensified using HyPhy.
11. *Suricata suricatta* showed non-significant results using PAML and relaxed using HyPhy.
12. *Castor canadensis* showed non-significant results using PAML and relaxed using HyPhy.
13. *Ictidomys* tridecemlineatus showed non-significant results using PAML and relaxed using HyPhy.
14. *Oryctolagus cuniculus* showed non-significant results using PAML and intensified using HyPhy.
15. *Spermophilus dauricus* showed non-significant results using PAML and relaxed using HyPhy.
16. *Fukomys damarensis* showed relaxation using PAML and non-significant results using HyPhy.

These differences between the results from the two programs are not unexpected as the results are known to be influenced by the tree topology and set of background species used. Moreover, different versions of these programs also produce slightly different results due to differences in implementation. Codeml and RELAX use different models of sequence evolution and can have different results on the same dataset as well. Given the overall consistency of our results from both programs, the signatures of relaxed selection detected seem to be reasonably trustworthy.

Time of gene loss

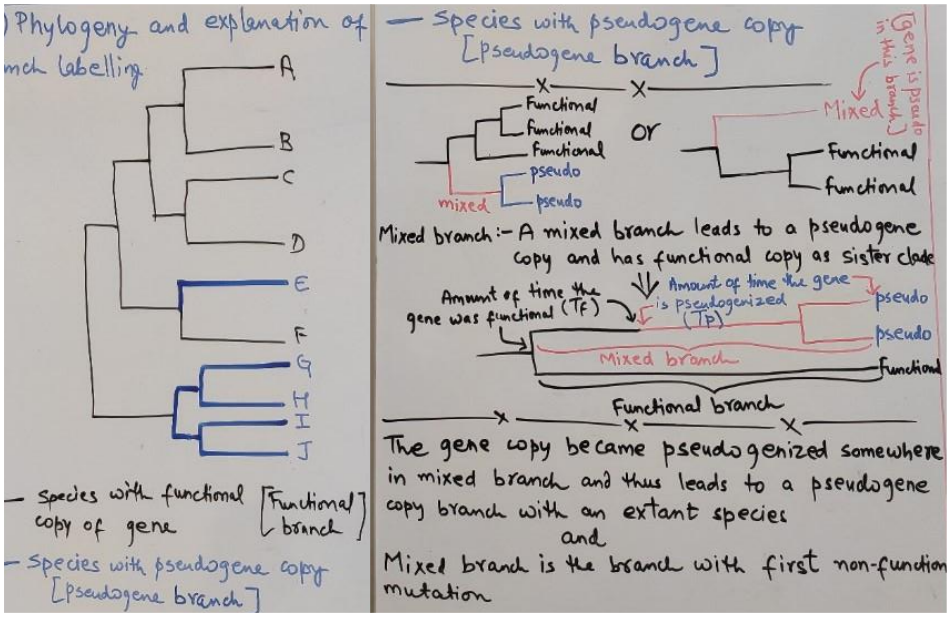
The estimates of gene loss time are obtained based on the method initially proposed by Meredith[52] and later used by several papers that have identified and timed gene loss events [22,30,53–58]. Two different models called 1ds and 2ds are used to calculate the time of gene loss. The 1ds estimates are calculated assuming a single dS (i.e., synonymous substitution rate), and 2ds estimates are calculated assuming two different values of dS (one before and another after gene loss). The 2ds approach assumes that the value of the dS before gene loss is 0.7 times the value of dS after gene loss ($dS_f=0.7*dS_p$)[59].

In *Photograph 6*, we briefly explain the approach used to calculate gene loss timing and provide detailed comments for the code used for performing the calculations.



Photograph 6: Photograph of the entire whiteboard explaining the approach of Meredith et al. 2009.

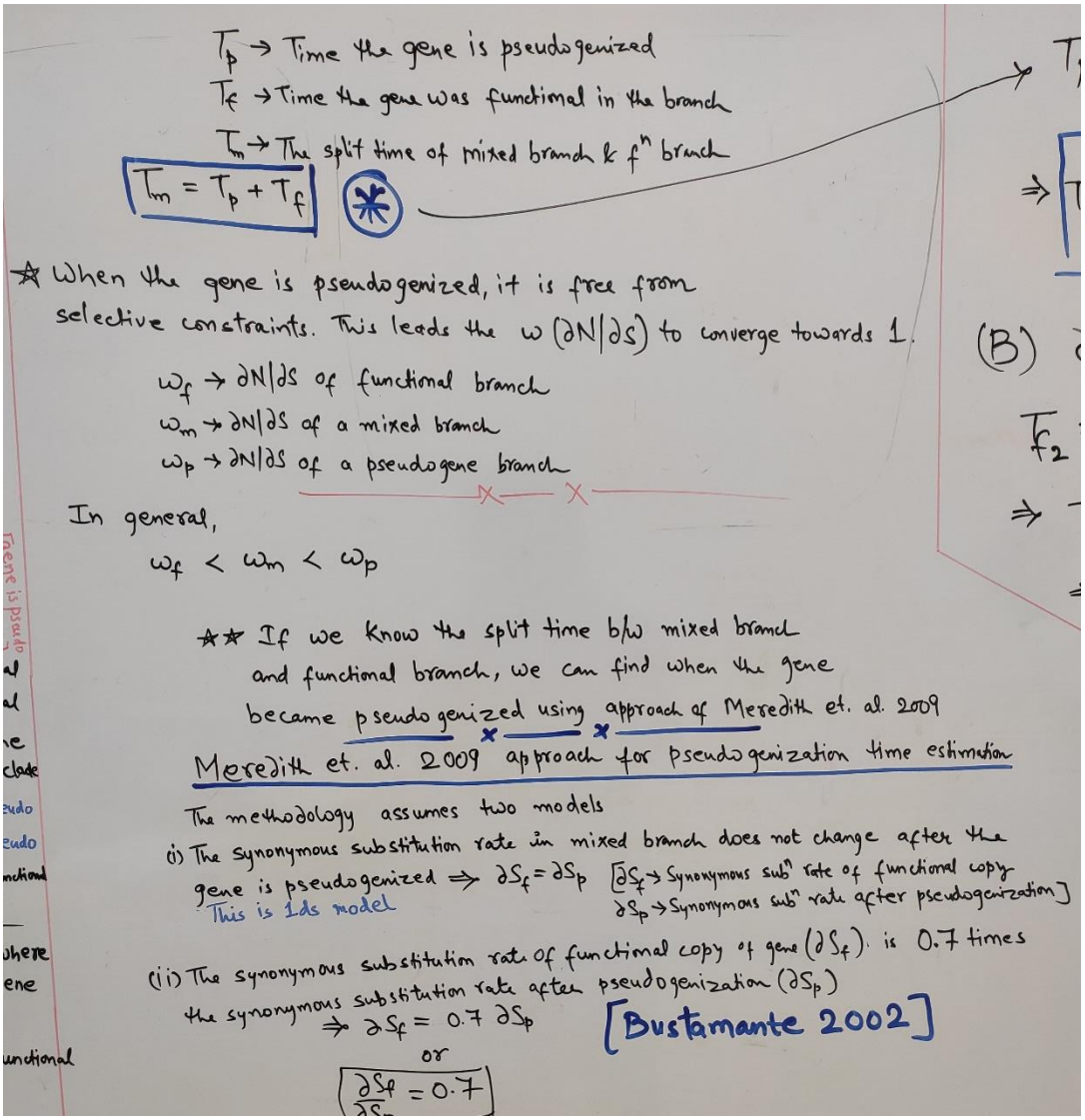
In the first part of the whiteboard, we show the phylogeny and explain branch labeling. Branches of the phylogeny are labeled as either functional, pseudogenic, or mixed. Two example scenarios of what is a mixed-branch are also shown in *Photograph 7*. Please refer to the original publication by Meredith[52] for a detailed explanation.



Photograph 7: Phylogeny and branch labeling as used in Meredith et al., 2009.

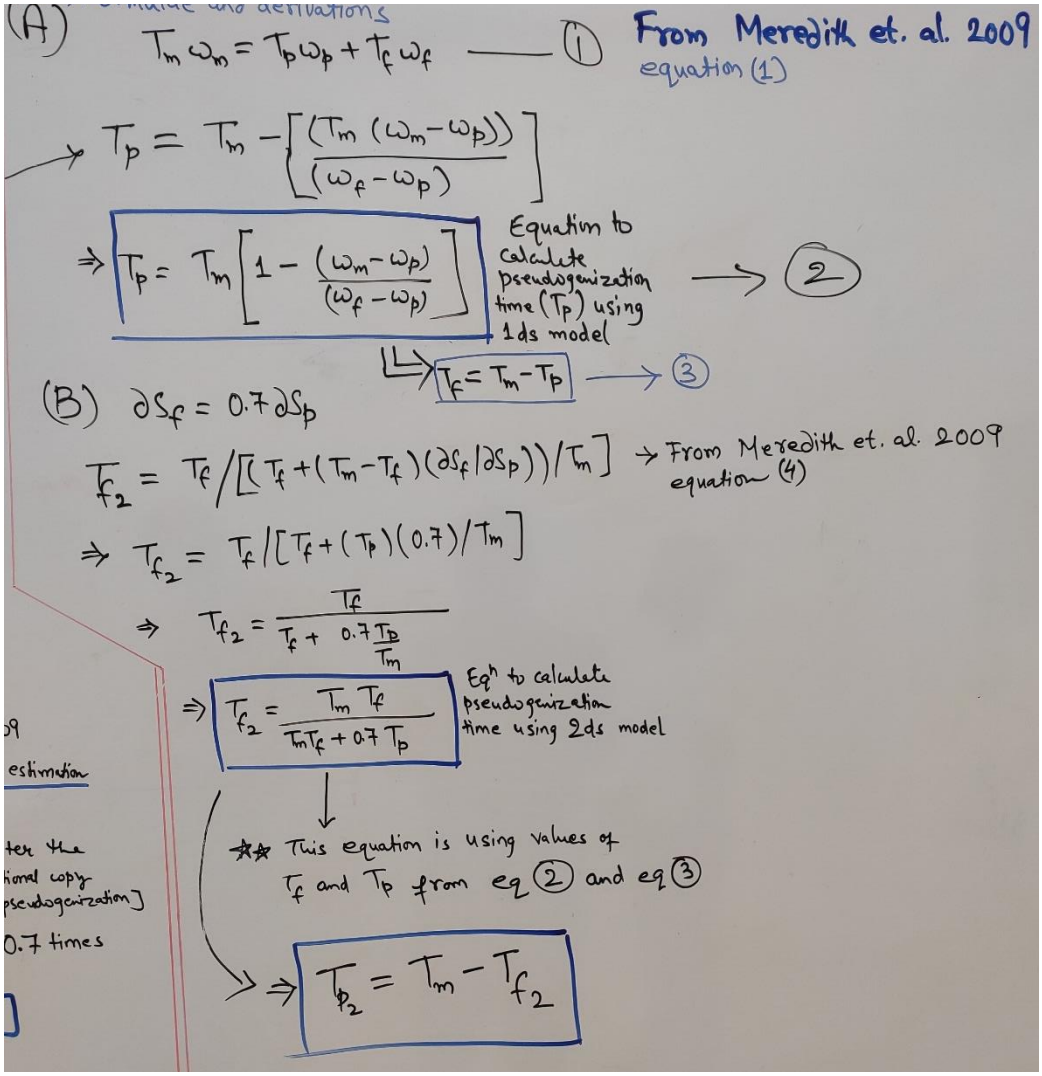
Having explained the process of labeling in *Photograph 7*, we next explain the terminologies used for estimating the time of gene loss. The estimates of ω are obtained using the branch-

free model (b_free with three estimates of ω) in the codeml program from the PAML package using unrooted trees described above. The actual config files used to run codeml, the multiple sequence alignments, tree files with labeling, and output files are provided in the Github repository. Two different models (**1ds** and **2ds**) are used to estimate the range of pseudogenization time.



Photograph 8: Terminologies and assumptions used in pseudogenization time estimation by Meredith et al., 2009.

Having explained the terminologies in *Photograph 8*, we next use the equations from Meredith et al., 2009 to calculate the pseudogenization time for each gene loss event. Actual values of each of the estimated ω obtained using codeml, T_m obtained from the TimeTree database, and all the other parameters calculated from these values are summarised in **Supplementary Table S4**. As per the standard procedure, we assume that the ω_p (i.e., ω in the pseudogene branch) equals 1. In cases where the $\omega_m > 1$, the species split time is considered as the time of gene loss.



Photograph 9:Formulae and derivations from Meredith et.al., 2009.

Part-A formula in *Photograph 9* is for the **1ds** model and Part-B formula is for the **2ds** model. The following lines of code explain the shell script (wrapper_script.sh) which extracts the ω estimates from the output file of codeml and calculates the various parameters before writing it to the summary files.

#the variable bg is assigned as the list of background species used.

bg=`cat bg.txt|tr '\n' ','|sed 's/,$/\n/g'`

tm corresponds to T_m. T_m corresponding to the split time between focal species (species with gene loss) and closest species with functional copy is obtained from the split_time.txt file. The split_time.txt file is created from screening the list of all pairwise split times of the focal species with every other species with a functional copy of the gene to obtain the lowest value. See wrapper_script.sh

tm=`grep "$sp" split_time.txt|awk '{print $2}'`

#mdl corresponds to the codon frequency model used to run codeml and is being obtained from the codeml output file.

```
mdl=`grep "Codon frequency model:" $outfile|awk -F ":" '{print $2}'|awk '{print $1}'`

# wf corresponds to ωf and is obtained from the codeml output file.

wf=`grep "w (dN/dS) for branches:" $outfile|awk '{print $5}'`

# wm corresponds to ωm and is obtained from the codeml output file.

wm=`grep "w (dN/dS) for branches:" $outfile|awk '{print $6}'`

# wp corresponds to ωp and is obtained from the codeml output file. This is set to 1 in the calculations and not used.

wp=`grep "w (dN/dS) for branches:" $outfile|awk '{print $7}'`

#tf corresponds to Tf and is being calculated using the 1ds model.

tf=`echo $tm $wm $wf | awk '{print $1*(($2 - 1)/($3 - 1))}'`

#tp corresponds to Tp and is being calculated using the 1ds model.

tp=`echo $tm $tf|awk '{print $1 - $2}'`

#tf2 corresponds to Tf2 and is being calculated using the 2ds model.

tf2=`echo $tm $tf $tp|awk '{print ($1*$2)/($2+(0.7*$3))}'`

#tp2 corresponds to Tp2 and is being calculated using the 2ds model.

tp2=`echo $tm $tf2|awk '{print $1-$2}'`
```

Pseudogenization timing estimation using Meredith et al., 2009 methodology relies upon ω estimates (dN/dS) and its shift towards 1 in the species of interest. Codeml estimates ω in a phylogenetic framework and requires a multiple sequence alignment and corresponding species tree. However, ω estimates can be affected by several factors such as sequence alignment, background species composition, number of background species, codon frequency model, and number of species labeled. Hence, it is important to check the effect of these confounding factors on pseudogenization timing calculation.

We calculated ω for each species of interest with different combinations of background species and labeling schemes under two different codon frequency models, namely, F1x4 and F3x4. Both pseudogenization timing estimation models, i.e., 1ds and 2ds, were used for each of these estimates. The variation in the estimates of pseudogenization time using these different combinations revealed considerable variation in the estimates for the same species (see **Supplementary Fig. S955**). Increasing the sampling of species closely related to the species with gene loss provides better resolution of the timing of gene loss.

The ω values for mixed (ω_m) and functional (ω_f) branches were estimated using two different codon substitution models (F1x4 and F3x4) to ensure the robustness of the estimates (see **Supplementary Table S4**). The calculation of gene loss timing relies upon estimates of T_p

(time for which the gene has been pseudogenic) using the method Meredith et al. (2009). Based on the assumptions of 1ds and 2ds, we could get a confidence interval for the estimated time of gene loss (see **Supplementary Table S4**). Gene loss timing was estimated separately in galliforms, rodents, and Carnivores (see **Supplementary Table S4**).

GC content range and K-mer abundance

The GC content range (minimum and maximum possible values of GC% for a given amino acid sequence) was calculated (see **Supplementary Table S14**) for *COA1* and *PDX1* amino acid sequences in rodent and primate species using the window-based tool CodSeqGen[60]. The ContMap function in the R package phytools extrapolates the evolution of GC content along the phylogeny for both genes (see **Supplementary Fig. S956-S957**). The program jellyfish (v2.2.8)[61] was used to get the K-mers (count command with the flags -C -m 21 -s 1000M and -t 16) and their abundance (dump command). The seqkit fx2tab (v0.10.1)[62] option calculated the abundance of K-mers at different GC content bins and the GC content of each *COA1* exon (see **Supplementary Table S17**).

Quantification of gBGC

We calculated the gBGC for *COA1* gene sequences of each species using the program mapNH (v1.3.0) implemented in the testNH package[63]. In mapNH, we used multiple sequence alignments of the *COA1* gene and species tree as input with the flag model=K80. A single gene-wide estimate of gBGC termed GC* is obtained for each species (see **Supplementary Table S15**). These estimates of GC* (GC* > 0.9 is significant) help understand the evolution of gBGC along the phylogeny when extrapolated using the ContMap function of the phytools package in R. We also calculated the gBGC for taxonomic group-wise alignments using the phastBias and phyloFit implemented in the PHAST (v1.3) package[64,65]. In the first step, we use the phyloFit program to fit phylogenetic models to multiple sequence alignments using the specified tree topology (--tree flag with species tree as argument) and substitution model (--subst-mod flag with HKY85 model as argument). Next, the phastBias program with the –bgc flag identified gBGC tracts using the ".mod" file output from phyloFit (see **Supplementary Table S16**, see **Supplementary Fig. S189-S222**). The gBGC tracts are positions along the gene with posterior probability > 0.5.

Computational prediction of RNA binding sites

The regulation of gene expression and splicing tends to be determined by the RNA binding sites present within the exons or introns of a gene[66]. A combination of such splice enhancers and splice silencer elements work in concert to facilitate the expression of different isoforms[67]. The *COA1* gene has changed the intron-exon organization and has acquired novel splice isoforms in felid species. These changes in splicing could result from changes in the RNA binding motifs present within the exons or introns of the gene. In contrast to felids, the splicing pattern in canid species matches the ancestral state. Hence, we compared the *COA1* gene sequences of canid and felid species to identify differences in the RNA binding motifs. We used the RBPmap[68] webserver to predict the RNA binding sites in each exon and intron separately (see **Supplementary Table S25-26**).

S9 text: Comparative phylogenetic logistic regression

Based on a search of the literature, we found the following research articles that have quantified the proportion of muscle fibers in birds. The nomenclature used for muscle fiber types has changed with time as identification techniques have improved. Hence, consistently distinguishing the muscle fiber types (FG, FOG, and SO) throughout the dataset is challenging. Nonetheless, white muscle fiber is consistently fast-twitch glycolytic and is used for regression analyses. The quantitative muscle fiber data used for the regression analysis are tabulated in **Supplementary Table S27.**

1. Kaiser et al., 1973[69] describe muscles mainly in two types based on morphology and biochemical properties: a red type (type 1) which are smaller in diameter, have high mitochondrial density, high myoglobin content, neutral fat, and oxidative enzymes, and a white type (type 2) characterized mainly by lack of myoglobin and fat, low density of mitochondria and oxidative enzyme activity but high phosphorylase activity. Red muscle fiber type is adapted for aerobic metabolism metabolizing fat, while white mainly relies on anaerobic metabolism and utilizes glycogen. For the classification of muscle fiber type in pectoralis muscle of birds, the authors use histochemical localization and intensity of succinate dehydrogenase (SDH) activity. Phosphorylase activity was demonstrated using glucose-1-phosphatase, while myoglobin content was measured by fixing muscle samples in phosphate-buffered 2.5% glutaraldehyde. The authors also observed an intermediate muscle fiber type that stains intermediate between white and red.
2. Wiskus et al., 1976[70] mainly describes muscle fibers as αR (Intermediate; Fast twitch), βR (Red; Tonic), and αW (White; Fast-twitch; no glycogen staining implying anaerobic). Samples were reacted for ATPase and SDH activities and stained for glycogen (PAS) and with hematoxylin-eosin. The pectoralis exhibited 90-100% αW fibers, with remaining αR. No β fibers were detected.
3. Kiessling 1976[71] describes muscle fibers as Red (slow-twitch, narrow diameter, high NADH, alkaline ATPase, FOG), White (fast-twitch, large, low NADH, high ATPase, FG), and Intermediate (twitch fiber, in between red and white). Muscle fiber identification was based on the activity of NADH-diaphorase and myosin ATPase. In this study, the author detected only red and white twitch fiber except for quail, which had an intermediate fiber in the pectoralis muscle.
4. Rosser and George 1985[72] use terminologies of FG (fast-twitch glycolytic), FOG (fast-twitch oxidative-glycolytic), and ST (slow-tonic) to describe skeletal muscle fiber types. Myofibrillar adenosine triphosphatase (mATPase) and succinate dehydrogenase (SDH) activity were used for histochemical demonstration. SDH activity revealed that ST have moderate oxidative capacity, and fast-twitch fibers are either moderately oxidative (FOG) or weakly oxidative (FG).
5. Rosser and George 1986[15] broadly describe muscle fibers as slow tonic (alkali-labile/acid-stable mATPase activity) or fast-twitch (alkali-stable/acid-labile mATPase activity). Based on SDH activity, three types of fast-twitch fibers exist: white (low SDH), intermediate (moderate SDH), and red (high SDH). The properties of these

muscles are as follow: R (narrow, myoglobin rich, fat loaded, oxidative; aerobic; FOG); W (broad, devoid of myoglobin and fat, glycolytic, FG; anaerobic); I (in between; FOG). The percentage of fiber types in the pectoralis muscle data consists only of fast-twitch muscles (FG: W; FOG: I + R).

6. Turner and Butler, in a 1988[73] study, defined muscle fiber types based on their histochemical properties as SO (low alkaline mATPase stability/high SDH staining), FOG (high alkaline mATPase stability/high SDH staining); and FG (high alkaline mATPase stability/low SDH staining). For histochemical staining, they used SDH and mATPase.
7. Kenneth Welch and Altshuler 2009[74] described twitch fiber types as SO (slow oxidative), FG (fast glycolytic), and FOG (fast oxidative glycolytic). Muscle fibers were distinguished between slow or fast-twitch by antibody-specific reaction for MHC (myosin heavy chain) isoforms. NADH-tetrazolium reductase was used to further categorize fast-twitch muscle as FG or FOG. The authors also mention that high oxidative muscles are red and low oxidative muscles are white.
8. Geldenhuys et al. 2014[75] modified the methodology of Rosser et al. 1996 and stained samples to quantify immunofluorescence and NADH activity. They identified only two types of muscle fibers: FG and FOG.
9. Schroeder et al. 2015[76] for their study used the terminologies for muscle fiber types as Slow, FG (fast glycolytic), and FOG (fast oxidative glycolytic). MHC-specific antibody tests were done to distinguish between slow and fast muscles, and NADH-TR staining was performed for glycolytic fibers.

References

1. Simakov, O. *et al.* Deeply conserved synteny resolves early events in vertebrate evolution. *Nat. Ecol. Evol. 2020 46* **4**, 820–830 (2020).

2. Mistry, J. *et al.* Pfam: The protein families database in 2021. *Nucleic Acids Res.* **49**, D412–D419 (2021).

3. Fukasawa, Y., Oda, T., Tomii, K. & Imai, K. Origin and Evolutionary Alteration of the Mitochondrial Import System in Eukaryotic Lineages. *Mol. Biol. Evol.* **34**, 1574–1586 (2017).

4. Makarova, K. S., Wolf, Y. I., Mekhedov, S. L., Mirkin, B. G. & Koonin, E. V. Ancestral paralogs and pseudoparalogs and their role in the emergence of the eukaryotic cell. *Nucleic Acids Res.* **33**, 4626 (2005).

5. Montgomerie, S. *et al.* PROTEUS2: a web server for comprehensive protein structure prediction and structure-based annotation. *Nucleic Acids Res.* **36**, W202 (2008).

6. Gautier, R., Douguet, D., Antonny, B. & Drin, G. HELIQUEST: a web server to screen sequences with specific -helical properties. *Bioinformatics* **24**, 2101–2102 (2008).

7. Webb, B. & Sali, A. Comparative protein structure modeling using MODELLER. *Curr. Protoc. Bioinforma.* **2016**, 5.6.1-5.6.37 (2016).

8. Kelley, L. A., Mezulis, S., Yates, C. M., Wass, M. N. & Sternberg, M. J. E. The Phyre2 web portal for protein modeling, prediction and analysis. *Nat. Protoc.* **10**, 845–858 (2015).

9. Waterhouse, A. *et al.* SWISS-MODEL: Homology modelling of protein structures and complexes. *Nucleic Acids Res.* **46**, W296–W303 (2018).

10. Pettersen, E. F. *et al.* UCSF ChimeraX: Structure visualization for researchers, educators, and developers. *Protein Sci.* **30**, 70–82 (2021).

11. Van Der Lee, R., Wiel, L., Van Dam, T. J. P. & Huynen, M. A. Genome-scale detection of positive selection in nine primates predicts human-virus evolutionary conflicts. *Nucleic Acids Res.* **45**, 10634–10648 (2017).

12. Xu, J. & Zhang, J. Are Human Translated Pseudogenes Functional? *Mol. Biol. Evol.* **33**, 755–760 (2016).

13. Mittal, P., Jaiswal, S. K., Vijay, N., Saxena, R. & Sharma, V. K. Comparative analysis of corrected tiger genome provides clues to its neuronal evolution. *Sci. Rep.* **9**, (2019).

14. Pan, S. *et al.* Convergent genomic signatures of flight loss in birds suggest a switch of main fuel. *Nat. Commun. 2019 101* **10**, 1–11 (2019).

15. Rosser, B. W. C. & George, J. C. The avian pectoralis: histochemical characterization and distribution of muscle fiber types. *Can. J. Zool.* **64**, 1174–1185 (1986).

16. Zhou, X. *et al.* Beaver and Naked Mole Rat Genomes Reveal Common Paths to Longevity. *Cell Rep.* **32**, (2020).

17. Chen, Y.-C., Liu, T., Yu, C.-H., Chiang, T.-Y. & Hwang, C.-C. Effects of GC Bias in Next-Generation-Sequencing Data on De Novo Genome Assembly. *PLoS One* **8**, e62856 (2013).

18. Botero-Castro, F., Figuet, E., Tilak, M. K., Nabholz, B. & Galtier, N. Avian genomes revisited: Hidden genes uncovered and the rates versus traits paradox in birds. *Mol. Biol. Evol.* **34**, 3123–3131 (2017).

19. Hargreaves, A. D. *et al.* Genome sequence of a diabetes-prone rodent reveals a mutation hotspot around the ParaHox gene cluster. *Proc. Natl. Acad. Sci. U. S. A.* **114**, 7677–7682 (2017).

20. Xiong, Y. & Lei, F. SLC2A12 of SLC2 Gene Family in Bird Provides Functional Compensation for the Loss of SLC2A4 Gene in Other Vertebrates. *Mol. Biol. Evol.* **38**, 1276–1291 (2021).

21. Vijay, N. Loss of inner kinetochore genes is associated with the transition to an unconventional point centromere in budding yeast. *PeerJ* **8**, (2020).

22. Meredith, R. W., Zhang, G., Gilbert, M. T. P., Jarvis, E. D. & Springer, M. S. Evidence for a single loss of mineralized teeth in the common avian ancestor. *Science* **346**, 1254390 (2014).

23. Huelsmann, M. *et al.* Genes lost during the transition from land to water in cetaceans highlight genomic changes associated with aquatic adaptations. *Sci. Adv.* **5**, 6671–6696

(2019).

24. Valente, R. *et al.* Convergent Cortistatin losses parallel modifications in circadian rhythmicity and energy homeostasis in Cetacea and other mammalian lineages. *Genomics* **113**, 1064–1070 (2021).

25. Schneider, K., Adams, C. E. & Elmer, K. R. Parallel selection on ecologically relevant gene functions in the transcriptomes of highly diversifying salmonids. *BMC Genomics* **20**, 1–23 (2019).

26. Sharma, V. & Hiller, M. Loss of Enzymes in the Bile Acid Synthesis Pathway Explains Differences in Bile Composition among Mammals. *Genome Biol. Evol.* **10**, 3211–3217 (2018).

27. Barber, M. R. W., Aldridge, J. R., Webster, R. G., Magor, K. E. & Magor, K. E. Association of RIG-I with innate immunity of ducks to influenza. *Proc. Natl. Acad. Sci. U. S. A.* **107**, 5913–8 (2010).

28. Sharma, S., Shinde, S. S., Teekas, L. & Vijay, N. Evidence for the loss of plasminogen receptor KT gene in chicken. *Immunogenetics* **72**, 507–515 (2020).

29. Castro, L. F. C. *et al.* Recurrent gene loss correlates with the evolution of stomach phenotypes in gnathostome history. *Proc. R. Soc. B Biol. Sci.* **281**, 20132669–20132669 (2013).

30. Jebb, D. & Hiller, M. Recurrent loss of HMGCS2 shows that ketogenesis is not essential for the evolution of large mammalian brains. *Elife* **7**, (2018).

31. Hecker, N., Sharma, V. & Hiller, M. Convergent gene losses illuminate metabolic and physiological changes in herbivores and carnivores. *Proc. Natl. Acad. Sci. U. S. A.* **116**, 3036–3041 (2019).

32. Sharma, V. *et al.* Convergent Losses of TLR5 Suggest Altered Extracellular Flagellin Detection in Four Mammalian Lineages. *Mol. Biol. Evol.* **37**, 1847–1854 (2020).

33. Peto, H., Roe, F. J. C., Lee, P. N., Levy, L. & Clack, J. Cancer and ageing in mice and men. *Br. J. Cancer* **32**, 411–426 (1975).

34. Tollis, M., Boddy, A. M. & Maley, C. C. Peto's Paradox: How has evolution solved the problem of cancer prevention? *BMC Biology* vol. 15 60 (2017).

35. Caulin, A. F. & Maley, C. C. Peto's Paradox: Evolution's prescription for cancer prevention. *Trends in Ecology and Evolution* vol. 26 175–182 (2011).

36. Caulin, A. F., Graham, T. A., Wang, L.-S. & Maley, C. C. Solutions to Peto's paradox revealed by mathematical modelling and cross-species cancer gene analysis. *Philos. Trans. R. Soc. B Biol. Sci.* **370**, 20140222 (2015).

37. DeGregori, J. Evolved tumor suppression: Why are we so good at not getting cancer? *Cancer Research* vol. 71 3739–3744 (2011).

38. Vazquez, J. M., Sulak, M., Chigurupati, S. & Lynch, V. J. A Zombie LIF Gene in Elephants Is Upregulated by TP53 to Induce Apoptosis in Response to DNA Damage. *Cell Rep.* **24**, 1765–1776 (2018).

39. Tollis, M. *et al.* Return to the Sea, Get Huge, Beat Cancer: An Analysis of Cetacean Genomes Including an Assembly for the Humpback Whale (Megaptera novaeangliae). *Mol. Biol. Evol.* **36**, 1746–1763 (2019).

40. Møller, A. P., Erritzøe, J. & Soler, J. J. Life history, immunity, Peto's paradox and tumours in birds. *J. Evol. Biol.* **30**, 960–967 (2017).

41. Xue, Y. *et al.* Cytochrome C Oxidase Assembly Factor 1 Homolog Predicts Poor Prognosis and Promotes Cell Proliferation in Colorectal Cancer by Regulating PI3K/AKT Signaling. *Onco. Targets. Ther.* **Volume 13**, 11505–11516 (2020).

42. Fellenberg, J. *et al.* Restoration of miR-127-3p and miR-376a-3p counteracts the neoplastic phenotype of giant cell tumor of bone derived stromal cells by targeting COA1, GLE1 and PDIA6. *Cancer Lett.* **371**, 134–141 (2016).

43. Herr, I. *et al.* MiR-127 and miR-376a act as tumor suppressors by in vivo targeting of COA1 and PDIA6 in giant cell tumor of bone. *Cancer Lett.* **409**, 49–55 (2017).

44. Rhie, A. *et al.* Towards complete and error-free genome assemblies of all vertebrate species. *Nature* **592**, 737–746 (2021).

45. Osoegawa, K. *et al.* Bacterial artificial chromosome libraries for mouse sequencing and functional analysis. *Genome Res.* **10**, 116–128 (2000).

46. Obayashi, T., Kagaya, Y., Aoki, Y., Tadaka, S. & Kinoshita, K. COXPRESdb v7: A gene coexpression database for 11 animal species supported by 23 coexpression platforms for technical evaluation and evolutionary inference. *Nucleic Acids Res.* **47**, D55–D62 (2019).

47. Shinde, S. S., Teekas, L., Sharma, S. & Vijay, N. Signatures of Relaxed Selection in the CYP8B1 Gene of Birds and Mammals. *J. Mol. Evol.* **87**, 209–220 (2019).

48. Hecker, N., Sharma, V. & Hiller, M. Transition to an Aquatic Habitat Permitted the Repeated Loss of the Pleiotropic KLK8 Gene in Mammals. *Genome Biol. Evol.* **9**, 3179–3188 (2017).

49. Wertheim, J. O., Murrell, B., Smith, M. D., Kosakovsky Pond, S. L. & Scheffler, K. RELAX: Detecting Relaxed Selection in a Phylogenetic Framework. *Mol. Biol. Evol.* **32**, 820–832 (2015).

50. Yang, Z. PAML 4: Phylogenetic Analysis by Maximum Likelihood. *Mol. Biol. Evol.* **24**, 1586–1591 (2007).

51. Xia, X. DAMBE5: A Comprehensive Software Package for Data Analysis in Molecular Biology and Evolution. *Mol. Biol. Evol.* **30**, 1720–1728 (2013).

52. Meredith, R. W., Gatesy, J., Murphy, W. J., Ryder, O. A. & Springer, M. S. Molecular decay of the tooth gene enamelin (ENAM) mirrors the loss of enamel in the fossil record of placental mammals. *PLoS Genet.* **5**, (2009).

53. Gaudry, M. J. *et al.* Inactivation of thermogenic UCP1 as a historical contingency in multiple placental mammal clades. *Sci. Adv.* **3**, e1602878 (2017).

54. Meredith, R. W., Gatesy, J., Cheng, J. & Springer, M. S. Pseudogenization of the tooth

gene enamelysin (MMP20) in the common ancestor of extant baleen whales. *Proc. R. Soc. B Biol. Sci.* **278**, 993–1002 (2011).

55. Jiao, H. *et al.* Trehalase Gene as a Molecular Signature of Dietary Diversification in Mammals. *Mol. Biol. Evol.* **36**, 2171–2183 (2019).

56. Meyer, W. *et al.* Ancient convergent losses of Paraoxonase 1 yield potential risks for modern marine mammals. *Science* **361**, 591–594 (2018).

57. Janiak, M., Pinto, S., Duytschaever, G., Carrigan, M. & Melin, A. Genetic evidence of widespread variation in ethanol metabolism among mammals: revisiting the 'myth' of natural intoxication. *Biol. Lett.* **16**, (2020).

58. Emerling, C. A. Genomic regression of claw keratin, taste receptor and light-associated genes provides insights into biology and evolutionary origins of snakes. *Mol. Phylogenet. Evol.* **115**, 40–49 (2017).

59. Bustamante, C. D., Nielsen, R. & Hartl, D. L. A Maximum Likelihood Method for Analyzing Pseudogene Evolution: Implications for Silent Site Evolution in Humans and Rodents. *Mol. Biol. Evol.* **19**, 110–117 (2002).

60. Al-Ssulami, A. M., Azmi, A. M. & Hussain, M. CodSeqGen: A tool for generating synonymous coding sequences with desired GC-contents. *Genomics* **112**, 237–242 (2020).

61. Marçais, G. & Kingsford, C. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics* **27**, 764–770 (2011).

62. Shen, W., Le, S., Li, Y. & Hu, F. SeqKit: A cross-platform and ultrafast toolkit for FASTA/Q file manipulation. *PLoS One* **11**, e0163962 (2016).

63. Dutheil, J. Detecting site-specific biochemical constraints through substitution mapping. *J. Mol. Evol.* **67**, 257–265 (2008).

64. Capra, J. A., Hubisz, M. J., Kostka, D., Pollard, K. S. & Siepel, A. A Model-Based Analysis of GC-Biased Gene Conversion in the Human and Chimpanzee Genomes. *PLoS Genet.* **9**, e1003684 (2013).

65. Hubisz, M. J., Pollard, K. S. & Siepel, A. Phast and Rphast: Phylogenetic analysis with space/time models. *Brief. Bioinform.* **12**, 41–51 (2011).

66. Fu, X. D. & Ares, M. Context-dependent control of alternative splicing by RNA-binding proteins. *Nature Reviews Genetics* vol. 15 689–701 (2014).

67. Dassi, E. Handshakes and fights: The regulatory interplay of RNA-binding proteins. *Frontiers in Molecular Biosciences* vol. 4 67 (2017).

68. Paz, I., Kosti, I., Ares, M., Cline, M. & Mandel-Gutfreund, Y. RBPmap: A web server for mapping binding sites of RNA-binding proteins. *Nucleic Acids Res.* **42**, W361 (2014).

69. Kaiser, C. E. & George, J. C. Interrelationship amongst the avian orders Galliformes, Columbiformes, and Anseriformes as evinced by the fiber types in the pectoralis muscle. *Can. J. Zool.* **51**, 887–892 (1973).

70. Wiskus, K. J., Addis, P. B. & Ma, R. .-I. Distribution of βR, αR and αW Fibers in Turkey Muscles. *Poult. Sci.* **55**, 562–572 (1976).

71. Kiessling, K. Muscle structure and function in the goose, quail, pheasant, guinea hen, and chicken. *Comp. Biochem. Physiol. B.* **57**, 287–292 (1977).

72. Rosser, B. W. C. & George, J. C. Histochemical Characterization and Distribution of Fiber Types in the Pectoralis Muscle of the Ostrich (Struthio camelus) and Emu (Dromaius novaehollandiae). *Acta Zool.* **66**, 191–198 (1985).

73. Turner, D. L. & Butler, P. J. The aerobic capacity of locomotory muscles in the tufted duck, Aythya fuligula. *J. Exp. Biol.* **135**, 445–460 (1988).

74. Welch, K. C. & Altshuler, D. L. Fiber type homogeneity of the flight musculature in small birds. *Comp. Biochem. Physiol. - B Biochem. Mol. Biol.* **152**, 324–331 (2009).

75. Geldenhuys, G., Hoffman, L. C. & Muller, N. HISTOLOGICAL CHARACTERIZATION OF THE FIBER TYPES IN THE M. PECTORALIS OF EGYPTIAN GEESE: A SOUTHERN AFRICAN WILDFOWL SPECIES. 17–22 (2014).

76. Schroeder, K. L., Sylvain, N. J., Kirkpatrick, L. J. & Rosser, B. W. C. Fibre types in primary 'flight' muscles of the African Penguin (Spheniscus demersus). *Acta Zool.* **96**, 510–518 (2015).