Convergent evolution through independent rearrangements in the primate amylase locus

Authors
Charikleia Karageorgiou[1], Stefan Ruhl[2], Omer Gokcumen[1]*

Affiliations
[1]*Department of Biological Sciences, University at Buffalo, Buffalo, NY, USA.*
[2]*Department of Oral Biology, School of Dental Medicine, University at Buffalo, Buffalo, NY, USA.*
**Corresponding*

ABSTRACT

Structurally complex regions of the genome are increasingly recognized as engines of evolutionary convergence due to their propensity to generate recurrent gene duplications that give rise to similar gene expression patterns and traits across lineages. However the mutational mechanisms driving these duplications and the regulatory changes enabling novel expression patterns remain poorly understood. The primate amylase locus, marked by independent gene duplications, provides an ideal model to investigate these dynamics. Leveraging high-quality genome assemblies from 53 primates and multi-tissue transcriptomes from Old World monkeys, we reconstructed the evolutionary history of the recurrent gene duplications across the primate phylogeny. Our data suggest that lineage-specific LTR retrotransposon insertions are associated with initial structural instability, while subsequent duplications are primarily driven by non-allelic homologous recombination. Recurrent independent duplications in rhesus macaques, olive baboons, and great apes gave rise to distinct amylase gene copies with convergent expression in the pancreas and salivary glands. We found that these independent gene duplications are accompanied by episodic diversifying selection on lineage-specific copies, likely driving the emergence of functional divergence. Our comparative analyses in primates indicate that the gene ancestral to great ape *AMY1* and *AMY2A* was expressed in both salivary glands and pancreas in the Catarrhini ancestor. The great ape–specific duplication of this ancestral gene likely facilitated subfunctionalization into salivary gland- and pancreas-specific expression, respectively. Comparative analysis of primate amylase promoter regions reveals regulatory rewiring, driven by motif turnover mediated by structural rearrangements, and partially explaining evolutionary shifts in expression. Together, our findings highlight how structural and regulatory modularity in complex genomic regions drives evolutionary innovation and molecular convergence, and we provide a genomic framework for dissecting these processes across diverse lineages.

TEASER

Structural plasticity drives convergence in primate amylase

GLOSSARY (Envisioned as a box within the manuscript)

Segmental Duplications: Segmental duplications (SDs) (also called low-copy repeats, LCRs) are stretches of DNA present two or more times per haploid genome, with each copy exhibiting a high degree of sequence identity (commonly >95%) and typically spanning at least ~1 kilobase in length. These duplicated segments may occur in tandem (*i.e.* adjacent to one another) or may be dispersed to different locations on the same chromosome or on different chromosomes.

Recurrent Duplications: Recurrent duplications refer to duplication events that arise multiple times, independently, at the same genomic locus across different individuals or lineages. These events are typically mediated by underlying genomic features, such as the presence of flanking segmental duplications or low-copy repeats or transposons or mutational hotspots, which predispose these loci to recurrent rearrangement through mechanisms like non-allelic homologous recombination (NAHR).

Independent Duplications: Independent duplications denote duplication events that have arisen separately, (not from a shared ancestral event) as separate mutational events, either at the same or different loci. In this context, "independent" emphasizes that each duplication occurred as a unique mutational event, not inherited from a common ancestor.

Recurrent Independent Duplication: Recurrent independent duplications are structural rearrangements in which the same genomic locus is duplicated multiple times, but each duplication arises as a distinct, independent mutational event in different lineages or individuals, rather than being inherited from a common ancestor.

- Recurrent: The duplication happens repeatedly at a specific locus, often due to the region's inherent genomic instability or predisposition to rearrangement.
- Independent: Each observed duplication event is phylogenetically separate i.e. arising de novo in different evolutionary lineages or individuals, rather than being a single ancestral duplication that is inherited.

Lineage-specific events: Lineage-specific events are genetic or evolutionary changes, such as deletions, duplications, regulatory shifts etc, that occur uniquely within a particular evolutionary lineage after it has diverged from its closest relatives. In phylogenetic terms, these are changes that take place along the branch (edge) connecting two nodes in a species tree. Such events distinguish that lineage and the species it contains from all others due to novel mutational or biological processes that arose within that lineage alone.

Convergent Evolution: The independent emergence of similar traits or features in related lineages that had different ancestral starting points. In convergence, similar phenotypes arise via different genetic changes or developmental routes, but are shaped by similar ecological or selective pressures.

AMY: Alpha-amylase gene family: A family of genes encoding alpha-amylase enzymes, which hydrolyze starch into sugars. Members of this family often arise via gene duplication and can differ in tissue specificity, expression level and function.

***AMY1/AMY1A|B|C*:** Great ape salivary gland amylase genes: Salivary gland-specific amylase genes in great apes, derived from the ancestral *AMY1′* gene, primarily involved in starch digestion in the oral cavity.

***AMY2A*:** Great ape pancreatic amylase gene: A pancreatic-specific amylase paralog in great apes, evolved from the same the ancestral *AMY1′* gene that gave rise to *AMY1*, and specialized for starch digestion in the small intestine.

***AMY2B*:** The presumed ancestral amylase gene in primates, with conserved expression in the pancreas across multiple lineages.

***AMY1′*:** Ancestral Catarrhini amylase duplicate: The first duplication of *AMY2B* in the Catarrhini ancestor, predating the split between Old World monkeys and apes; progenitor of human *AMY1* and *AMY2A*.

***AMYm*:** Macaque-specific amylase duplicate: A lineage-specific duplication of *AMY2B* in macaques (fascicularis and sinica groups), arising after divergence from the silenus group.

***AMYp1*:** Baboon-specific first amylase duplication: The first baboon-lineage duplication generated via non-allelic homologous recombination of the *AMY2B* and *AMY1′* paralogs, shared with the Guinea baboon.

***AMYp2*:** Olive baboon-specific second amylase duplication: A second, more recent duplication in the olive baboon lineage, derived from *AMY2B* and *AMYp1* via non-allelic homologous recombination.

Abbreviations:

OWM: Old World monkey

NWM: New World monkey

NAHR: Non-allelic homologous recombination

TFBS: Transcription factor binding sites

MAIN TEXT

INTRODUCTION

Convergent evolution, the independent emergence of similar traits in distinct lineages, has long intrigued evolutionary biologists seeking to understand how similar phenotypes arise from different genetic starting points (*1*, *2*). Recent advances in long-read sequencing have allowed us to characterize structurally complex genomic regions accurately at the nucleotide resolution, revealing that these regions frequently undergo recurrent structural variation (*3–6*), including gene duplications, that can drive major shifts in biological function (*7–11*). Integrating these insights, an emerging model posits that structurally complex loci, through repeated gene duplications and regulatory rewiring, can serve as substrates for molecular convergence. In particular, such loci may produce similar spatial expression patterns of gene families in phylogenetically distant lineages. Yet, the specific mutational mechanisms that give rise to these duplications, the evolutionary forces shaping nucleotide variation among paralogs, and the processes by which regulatory elements are reshuffled during structural rearrangements remain largely unresolved. These questions are at the heart of understanding how structural and regulatory complexity contributes to evolutionary innovation and the repeated emergence of similar traits across lineages.

The amylase locus is one of the most intriguing structurally complex regions in mammalian genomes, notable for its exceptionally rapid structural evolution. It is one of the fastest-evolving loci in the human genome, despite being essential in starch metabolism (Roberts & Whelan, 1960). Amylase in mammals is primarily expressed in the pancreas. However, in some lineages, the gene has undergone a regulatory shift to include expression in the salivary glands (*12–14*). The evolution of amylase expression across species is a hallmark of dietary adaptation to the consumption of starch. In humans, for instance, copy number variation of the human salivary amylase gene (*AMY1*) correlates with starch-rich diets (*4*, *15*, *16*). Similar patterns of amylase gene duplications and concordant expression shifts of the duplicated copies have evolved multiple times, independently in different mammalian lineages, thus suggesting convergent mechanisms in response to dietary shifts (*14*, *16–18*). However, the mutational mechanisms underlying the independent gene duplications in the amylase locus and proximate regulatory sequences remain unexplored. Investigating these processes, using this locus as a model, could provide key insights into fundamental aspects of genomic evolution, including neofunctionalization, subfunctionalization, and gene expression dosage regulation.

In humans, saliva amylase is the most abundantly secreted enzyme in the oral cavity (*19*), with expression levels *6-8*-fold higher than in other great apes (*15*, *20*, *21*). This heightened expression was considered a

human-specific trait (*12*, *14*, *15*). However, across the primate phylogeny, other species also exhibit high salivary amylase expression, including Old World monkeys (*22*) and capuchins (*14*). The prevailing model suggests that the ancestral amylase gene was expressed initially in the pancreas, and was duplicated independently in different primate lineages, where some duplications acquired expression in the salivary glands. In apes, where this process has been best studied, the shift to salivary expression of one *AMY1* duplicate has been attributed to the insertion of an endogenous retroviral element (ERV) upstream of *AMY1* (*14*, *23–25*). In humans, additional *AMY1* duplications led to increased saliva expression levels (*26*). In other nonhuman primates, the genetic and regulatory mechanisms underlying salivary-gland-specific amylase expression remain unknown.

Therefore, we compared the evolutionary history of the amylase locus in great apes and Old World monkeys, offering an ideal framework for investigating how rapidly evolving, structurally complex regions can give rise to convergent gene expression patterns. Specifically, we asked whether the recurrent gain of salivary gland expression of *AMY* in these lineages is driven by shared mutational mechanisms and regulatory shifts, or by distinct molecular events that lead to similar outcomes. To address this, we leverage 244 high-quality primate genome assemblies and transcriptome data from pancreas, liver, and salivary glands of rhesus macaques (*Macaca mulatta*) and olive baboons (*Papio anubis*). Using this data, we identified features of the amylase locus associated with salivary gland expression, characterize the mutational processes underlying lineage-specific gene duplications, and examine how structural and regulatory variation is linked to *AMY* function and expression. In doing so, our study not only advances understanding of amylase locus evolution and regulation in primates but also provides a broader model for investigating how gene duplications in complex genomic regions can drive convergent evolution of tissue-specific expression.

RESULTS AND DISCUSSION

Ancestral and recurrent independent duplications shape primate amylase genetic structural diversity

Previous studies have documented extensive structural variation within the human amylase locus, identifying multiple independent copy number changes and inversions (*3*, *4*). To determine if similar structural complexity extends across non-human primates, we analyzed 244 primate genomes, successfully curating continuous contigs spanning the amylase locus in 53 species (total 69 genomes, **Table S1;** see **Methods**). For 16 species, multiple high-quality genome assemblies (ranging from 2 to 4 genomes per species) exist, allowing us to assess within-species variation (**Table S2**). In the remaining

174 genomes, the amylase locus could not be completely resolved within a single contig, underscoring the challenges of assembling this complex region, even with high-quality long-read data. Nevertheless, the 69 curated genomes provided a robust framework for reconstructing *AMY* structural variation across primates, documenting the expansions and contractions of the copy number variation of *AMY* genes, and predicting the ancestral structural states (**Figure 1A, Table S3,** see **Methods**).

Our results confirmed previous work (*14*, *23*) that identified an ancestral duplication of the amylase locus in the *Catarrhini* lineage, lineage-specific duplications in New World monkeys, and a loss of a single *AMY* gene in leaf-eating monkeys (*14*). The recently available high-quality genome assemblies analyzed in this study provide sequence-level resolution of lineage-specific copy number variations and led to the discovery of novel duplication events. These include a burst of duplications in the orangutan lineage, a previously noted (*15*) but uncharacterized copy number increase in bonobos relative to chimpanzees (**Figure 1B**), and recurrent independent duplications within Old World monkey genera, as described below. In addition, we found copy number variation among lesser ape species (i.e., in gibbons) (**Figure 1A**), yet it remains unclear whether this variation reflects an ancestral loss followed by lineage-specific gains, independent lineage-specific losses, or incomplete lineage sorting (**Figure S1**), posing an intriguing question for future research

Lemurs provide an ideal outgroup for studying the rest of the primate phylogeny. We found that all (n=11) but one of the analyzed lemur species harbor a single *AMY* gene per haploid cell, a structure shared with several non-lemur primates, suggesting that this configuration likely represents the ancestral amylase haplotype in primates. One exception is *Microcebus murinus*, which carries an additional *AMY* copy (**Figure 1C**). Taken together, our findings illustrate a complex evolutionary history of the amylase locus characterized by recurrent independent duplications and losses, thus contributing a remarkable structural diversity in primates. These patterns are consistent with the amylase region behaving as a mutational hotspot across primates as defined by (*27–29*). More genomes from additional species and more comprehensive dietary data would be required to robustly test the relationship between diet and *AMY* copy number in primates. Our study sets the stage to more comprehensively analyze the evolution of functional variation within a hotspot of genomic variation within closely related primate species with distinct dietary habits.

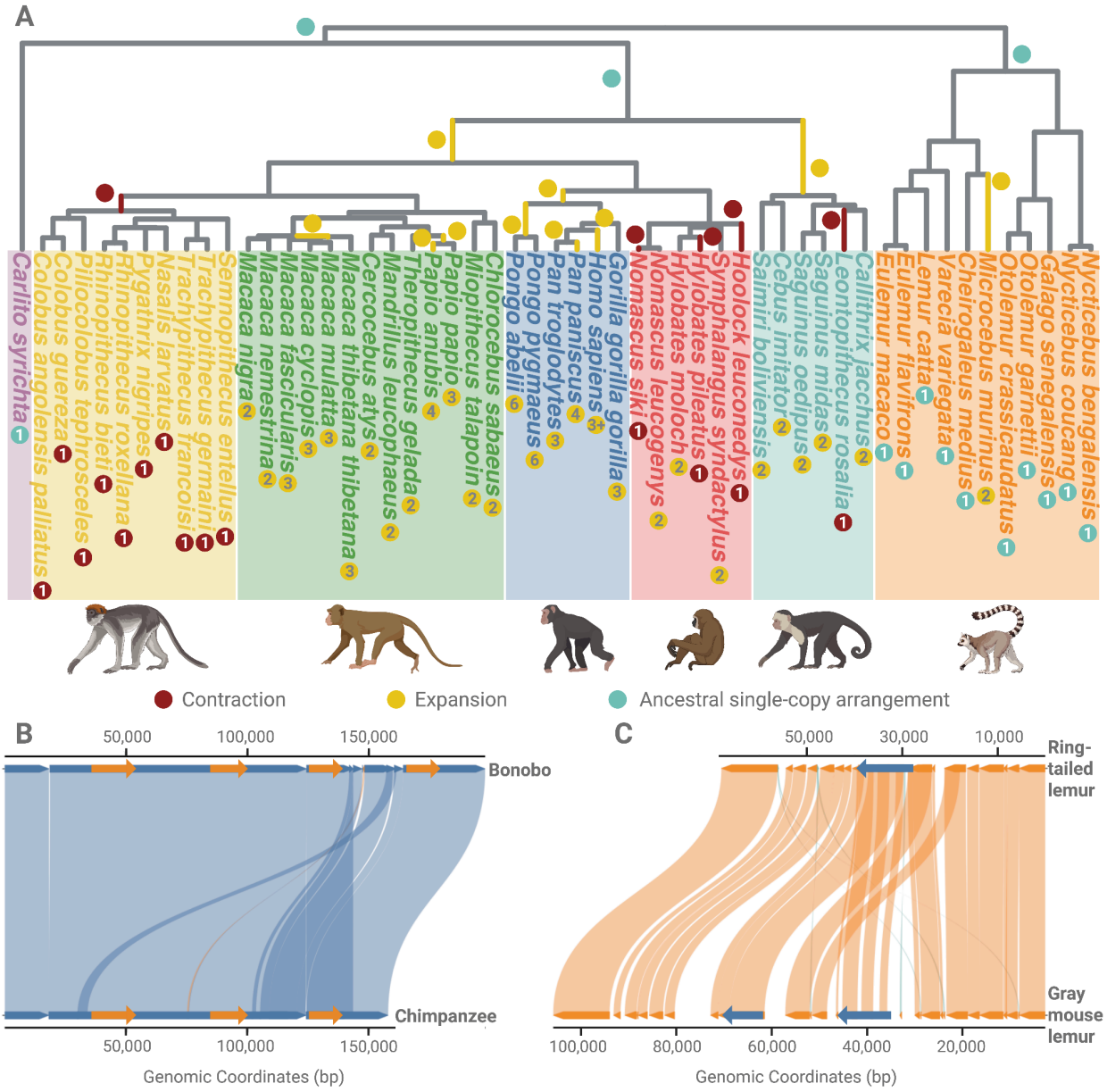

Figure 1. **Structural evolutionary history of the amylase locus across primates.** (A) Contractions and expansions in the amylase locus reconstructed from 69 high-quality genomes representing 53 primate species **(Table S3 and S4)**. Lineages are color-coded by clade: purple for tarsiers (outgroup), yellow for leaf-eating monkeys, green for Old World monkeys, dark blue for great apes, red for lesser apes (gibbons), cyan for New World monkeys, and orange for lemurs. Red dots indicate independent contractions in the amylase locus. Yellow dots indicate expansions. Cyan dots mark lineages retaining the ancestral single-copy configuration, inferred as the ancestral primate state. Independent contraction events are observed in the gibbon, leaf-eating monkey, and New World monkey genera. Numbers inside the dots give the total number of AMY gene copies detected in each species (per haploid genome). (B) Synteny comparison between bonobo and chimpanzee at the amylase locus. The copy number increase in bonobo, previously reported (*15*), is shown here to involve a chimeric duplication. The 5′ flanking region of the duplicated segment resembles the downstream region of *AMY1*. The internal genic region corresponds to a full duplication of the *AMY1* coding sequence. The 3′ flanking region aligns with the upstream flanking region of *AMY2A*. This chimeric architecture is consistent with a nonallelic homologous recombination mechanism. (C) Comparison of the amylase locus in ring-tailed lemur

and gray mouse lemur. The gray mouse lemur harbors a tandem duplication of the amylase gene, including both upstream and downstream flanking regions. This is the only duplication identified among the lemur species analyzed.

Reconstruction of the amylase duplications in catarrhini

To further elucidate the mutational basis of *AMY* structural variation in primates, we closely examined the rhesus macaque and olive baboon genomes, contrasting their evolutionary histories with recent findings from human haplotypes (*4*). Previous studies leveraged Old World monkeys as an outgroup to explore *AMY* copy number expansions in the human lineage (*25*). These analyses successfully identified an ancestral Catarrhini duplication event, followed by a great ape-specific duplication featuring a lineage-specific 5' ERV retrotransposon associated with salivary expression (*23*) (**Figure 2**). In addition, recent assemblies available for these species suggested lineage-specific duplications within rhesus macaque and olive baboon genomes that were undetected in earlier work.

Our findings provide phylogenetic evidence indicating that the primate ancestor possessed a single amylase gene, orthologous to human *AMY2B* (**Figure 1A**). In the Catarrhini ancestor, after the divergence of new world monkeys, an insertion of the 3' untranslated region of a γ-actin pseudogene insertion occurred 5' upstream of the ancestral amylase gene, followed by the duplication of the γ-actin-*AMY2B* segment, thereby generating a new amylase copy (*AMY1')* (**Figure 2**). Later in the great ape lineage, an endogenous retrovirus (ERV) inserted into the γ-actin region flanking *AMY1'*. This combined γ-actin–ERV–*AMY1'* segment duplicated into *AMY1* and the precursor of *AMY2A* in great apes. Eventually, the progenitor of *AMY2A* underwent an ectopic deletion of a portion of the ERV element, leading to its current structure as described further along with mutational mechanisms (**Figure 2**). Our observations within the great ape lineage builds upon previous findings reported by Meisler & Ting (*25*) but provides a more granular picture. The observation that *AMY1'* represents the ancestral copy leading to great ape *AMY2A* and *AMY1* is remarkable because these genes have distinct functions with specialized expression in pancreas and salivary glands, respectively.

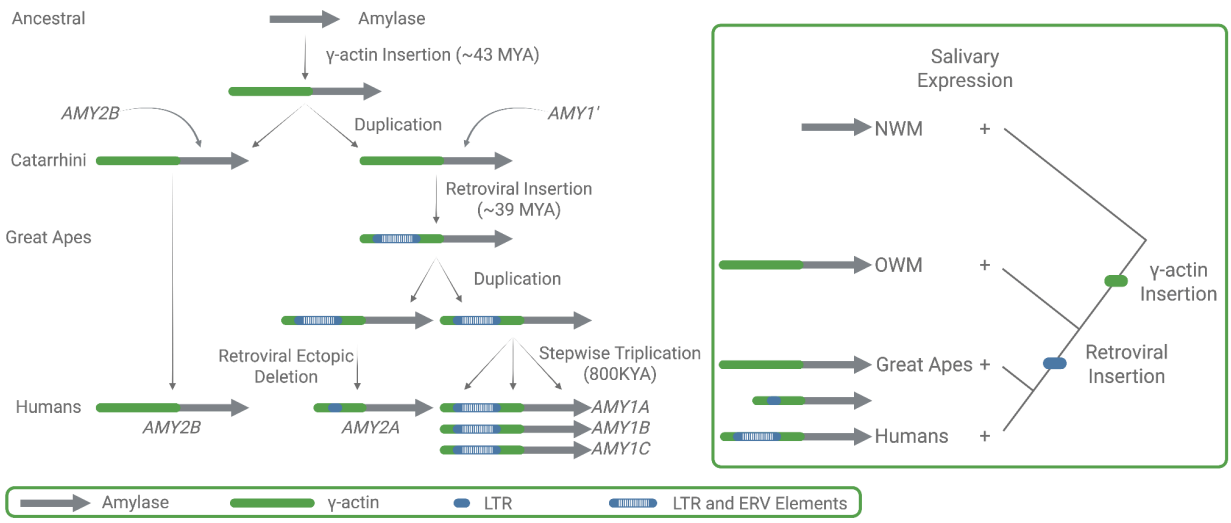

Figure 2. **Evolutionary origins of the amylase locus in primates.** Model of amylase locus evolution across primates adjusted from Samuelson et al. 1990 (*23*). The ancestral primate genome contained a single-copy amylase gene (*AMY2B*). In the Catarrhini lineage, a γ-actin insertion occurred upstream of *AMY2B* (~43 MYA), followed by duplication of the γ-actin-*AMY2B* segment, generating a second gene (*AMY1'*). In great apes, a retroviral (ERV) insertion occurred (~39 MYA), followed by retroviral ectopic deletion in the human lineage, giving rise to *AMY2A*. In humans, a stepwise triplication of the salivary-expressed amylase gene (*AMY1*) generated *AMY1A*, *AMY1B* and *AMY1C*. The right panel summarizes the inferred regulatory and structural events, showing independent gains of salivary expression in New World monkeys, Old World monkeys and great apes, with subfunctionalization of *AMY1* and *AMY2A* in humans.

<u>Lineage-specific duplications in rhesus macaques:</u>

In Old World monkeys, we found that baboons and rhesus macaques retain both the ancestral *AMY2B* gene as well as the derived *AMY1'* gene, which originated in the Catarrhini ancestor (**Figure 2A**). Additionally, we identified one lineage-specific duplicate in macaques and two additional, species-specific duplicates in baboons (**Figure 1A**). To better understand these events, we assessed the gene orthologies of these duplicates and resolved the duplication breakpoints relative to the ancestral Catarrhini and great ape haplotypes (**Figure S2**).

The novel lineage-specific duplication in rhesus macaques, which we termed *AMYm*, is located between *AMY2B* and *AMY1'* (**Figure 3A**). Sequence comparisons indicated that *AMYm* exhibits the highest similarity to *AMY2B*, suggesting its origin from the ancestral *AMY2B* gene. *AMY1'* in rhesus macaques on the other hand, shares orthology with great apes' *AMY2A* and *AMY1*, although without clear one-to-one orthology with either gene, suggesting that *AMY1'* represents the ancestral state from which *AMY1* and *AMY2A* evolved.

By comparing the ancestral Catarrhini haplotype (*i.e.*, two-copy haplotype in **Figure 3A**) (see **Methods**) with the rhesus macaque (*Macaca mulatta*) haplotype, we resolved the breakpoints of the duplicated

sequences (**Figure 3B**). We identified non-allelic homologous recombination (NAHR) as the primary mechanism driving the segmental duplications. NAHR, in this case, is characterized by two key signatures: first, *AMYm* is flanked by segmental duplications, *AMY2B* at the 5' end and *AMY1'* at the 3' end; and second, the duplicated segment containing *AMYm* exhibits mosaic characteristics derived from both flanking regions (**Figure S2**). Additionally, the duplication breakpoints overlap with the γ-actin insertion element, suggesting a possible role of this element in having facilitated the duplication events via NAHR. Among macaques, we were further able to determine that the duplication is specific to *sinica* and *fascicularis* groups, and must have occurred after the divergence of the *silenus* group (**Figure 3B**), dating this duplication to the relatively tight window of 4.5 to 5 million years ago based on the previously published phylogenetic dating of this clade (*30*). Based on these findings, we constructed what we believe to be the most plausible model that explains the mutational mechanism underlying this macaque-specific duplication event (**Figure S3**).

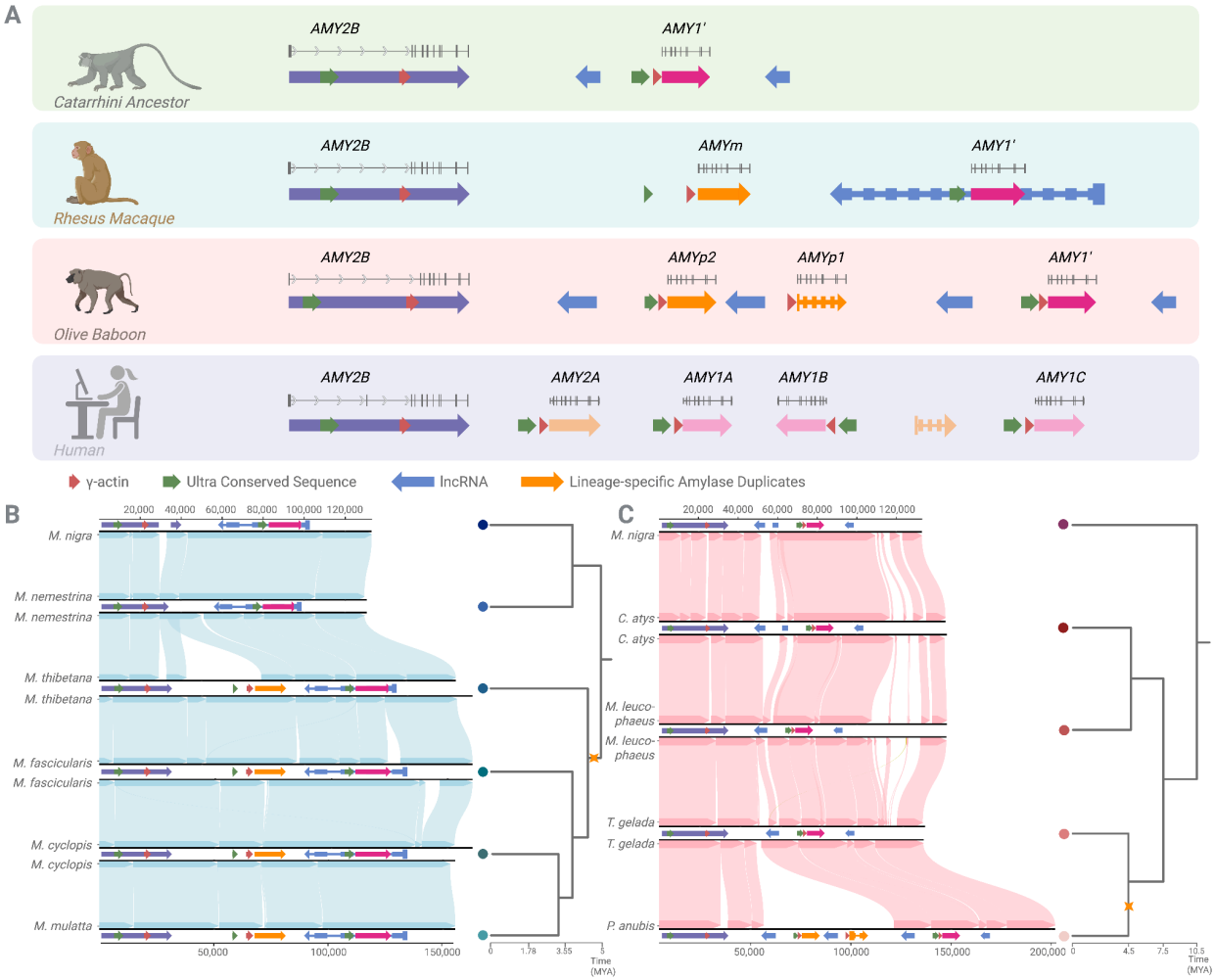

Figure 3. **Independent duplication events and structural evolution of the amylase locus in Old World monkeys.** (A) Schematic reconstruction of the amylase locus in the Catarrhini ancestor, rhesus macaque, olive baboon, and human reference genome (which is also shown to be the ancestral human arrangement (*4*) (hg38). The Catarrhini ancestor contains two genes: *AMY2B* and *AMY1'*, the latter derived from a duplication of *AMY2B*. In rhesus macaques, the locus includes *AMY2B*, *AMYm* , and *AMY1'* . In olive baboons, the locus consists of *AMY2B* , *AMYp2* and *AMYp1* , and *AMY1'*. *AMYp1* is annotated as a pseudogene in *NCBI*, and is indicated with a dashed outline. *AMYp1* and *AMYp2* arose from independent duplication events with distinct breakpoints from those of *AMYm*. In the human reference genome, the locus contains *AMY2B* , *AMY2A* and three *AMY1* paralogs, *AMY1A*, *AMY1B*, and *AMY1C*. The *AMY2B* genes are one-to-one orthologs across species. *AMY1'* in macaques, baboons and the Catarrhini ancestor is orthologous among Old World monkeys and represents the ancestral precursor to human *AMY2A* and *AMY1* genes. (B) Synteny analysis and phylogenetic context of the rhesus macaque-specific duplication. *AMYm* (orange) is located between *AMY2B* (purple) and *AMY1'* (pink), and is present only in the *fascicularis* and *sinica* macaque groups, but absent in the *silenus* group. This distribution dates the duplication to approximately 4.5-5 million years ago. The gene structure of *AMYm* is identical for coding sequence to *AMY2B*, but differs on the 5' untranslated region. Breakpoint analysis indicates a nonallelic homologous recombination (NAHR) mechanism. (C) Synteny comparison across *Papionini* species reveals two independent NAHR-mediated duplication events in olive baboons. The first duplication, giving rise to *AMYp1* (orange), is shared with Guinea baboon (*Papio papio)* and dates to approximately 2-4.5 million years ago. The second duplication, resulting in *AMYp2* (orange), is specific to olive baboons and dates to ~1.85 million years ago. Because of incomplete assembly in Guinea baboon (*Papio papio)*, we carried out the synteny analysis using closely related *Theropithecus gelada* and *Mandrillus leucophaeus* which are members of the Papionini group, supporting this reconstruction.

<u>Species-specific duplications in olive baboons:</u>

We conducted a similar comparative analysis of the olive baboon amylase locus to identify shared and lineage-specific amylase gene duplications and the mechanisms through which they arise. Specifically, we investigated the structural configuration of the amylase locus by comparing the olive baboon haplotype carrying four gene copies to the ancestral Catarrhini haplotype. We identified two lineage-specific amylase copies in *Papio anubis* flanked by *AMY2B* on the 5' and *AMY1'* on the 3' (**Figure 3C**), which we termed *AMYp1* and *AMYp2*.

These two novel genes were absent from other Old World monkey genomes, consistent with independent duplication events specific to the Papio lineage. To infer the sequential order of these duplications, we included the genome of Guinea baboon (*Papio papio)*, which is a sister species to olive baboons. We found the *AMYp1* gene (third in genomic order) to be identical to the derived amylase copies in Guinea baboon, while the *AMYp2* (second in order) could not be found in Guinea baboon (**Figure S4**), suggesting that *AMYp1* was duplicated before *AMYp2.* The sequential order of duplications was further assessed using sequence similarity metrics and competitive mapping among the four *olive baboon AMY* paralogs (see **Methods**). Combined, these analyses revealed that the two lineage-specific gene copies represent successive duplication events: *AMYp1* likely emerged in the ancestral baboon population, while *AMYp2* represents an olive baboon specific expansion highlighted in orange in **Figure 2C**.

Based on the data sources available to us and our findings outlined above, the most parsimonious mutational model for these olive baboon-specific duplications to be as follows: The ancestral haplotype, containing *AMY2B* and *AMY1'*, underwent an initial non-allelic crossover between these two genes, giving

rise to the novel *AMYp1* copy. Subsequently, haplotypes carrying *AMY2B*, *AMYp1*, and *AMY1'* experienced a second recombination event, this time between *AMY2B* and *AMYp1*, resulting in the emergence of the second novel copy, *AMYp2*, which is now positioned between *AMY2B* and *AMYp1* in the extant olive baboon haplotype (**Figure 3C**). We further infer, based on previously published phylogenetic estimates (*31*, *32*), that the first duplication shared between olive and guinea baboons occurred prior to their divergence from other baboon lineages, approximately 2 to 4.5 million years ago, while the second, olive baboon-specific duplication likely arose after their complete split from guinea baboons around 1.85 million years ago.

Taken together, our findings highlight lineage-specific duplications in rhesus macaques and olive baboons that occurred independently of one another through NAHR and within approximately 10 million years of evolutionary divergence. Two major questions remain, which we address in the next two sections. First, what was the initial mutational driver of these duplications? Second, what could be the adaptive and functional impact of these lineage-specific duplications?

The abundance of long terminal repeats (LTRs) correlates with gene copy number gains in the amylase locus

The amylase locus is exceptional in that it has evolved rapidly through independent structural rearrangements across the tree of life, including in species such as fruit flies, mice, rats, and dogs (*14*, *18*, *33*, *34*). Several studies have linked this extensive variation to dietary adaptations (*14*, *15*, *18*). As described above, the mutational mechanisms driving additional duplications, following the initial structural changes in the Catarrhini ancestors, have been characterized as non-allelic homologous recombination (NAHR). However, the origins of primary duplications, which arose from a single-copy ancestral haplotype (likely representing the ancestral state in primates), remain poorly understood.

Our previous work (*14*) identified lineage-specific insertions of transposable elements coinciding with *AMY* gene duplications across various mammalian taxa. The association between transposable elements (TEs) and structural variation has been explored in multiple contexts, and recent studies have shown that TE-mediated rearrangements can arise through diverse molecular mechanisms and induce structural instability (*35*, *36*). Building on these observations, we hypothesized that TEs elements might contribute to the formation of primary duplications in primates, thereby predisposing the amylase locus to structural instability. To test this hypothesis, we used a standardized primate repeat dataset to annotate transposable elements across 53 primate genomes, thereby minimizing genome-specific biases (**Table S5**; see **Methods** for details).

Our analyses revealed a wide range of transposable element content in the amylase locus across species, from 25.23% in the northern greater galago (*Otolemur garnettii*) to 59.52% in the Bornean orangutan (*Pongo pygmaeus*) (**Figure 4A**). In contrast to our expectations, we found that the amylase locus is generally depleted in transposable elements compared to genome-wide averages; nevertheless, we observe an enrichment in LTRs (**Figure 4B & 4C**). The general depletion is primarily driven by reduced representation of common, active, short retrotransposons such as Alu sequences (**Figure 4B**), suggesting that the locus is not broadly permissive to transposable element retention.

While other common transposable elements were underrepresented within the amylase locus in most primate species, we observed an enrichment of LTR transposons (**Figure 4,** panels **C** and **D**). The proportion of LTR elements in the locus, even after accounting for phylogenetic relatedness, correlates with the number of amylase genes, ($p < 10^{-5}$, $R^2 = 0.62$; **Figure S5**). This correlation is most significant within Catarrhini, where the majority of amylase gene gains and losses were observed, supporting the hypothesis that LTRs contribute to structural instability in the amylase locus. If indeed LTRs are driving the structural changes, we would expect to find LTRs proximal to the breakpoint junctions, as exemplified by the *AMYp2* duplication in olive baboons (**Figure S6**). We also detected a strong negative correlation between amylase gene copy number and the presence of DNA transposons, particularly *Charlie* and *Tigger* elements ($p = 1.4e^{-6}$ and $p = 0.0005$ respectively; **Figure S7**), further raising the possibility that non-LTR transposons might have limited retention in the locus. Together, these findings point to a transposable element-specific pattern of enrichment and depletion within the amylase locus, with these dynamics showing strong correlations with amylase gene copy number variation across primates.

Although we cannot conclusively demonstrate LTRs as the direct cause of structural rearrangements, we hypothesize that their insertion may contribute not only to structural instability but may also facilitate functional changes in regulatory regions as suggested earlier (*25*). Due to their role in regulatory rewiring, their retention may generate long stretches of homologous sequence, which in turn may cause subsequent NAHR events. Such dynamics in primates may explain both the recurrence of amylase gene duplications , as well as their functional changes in tissue expression.

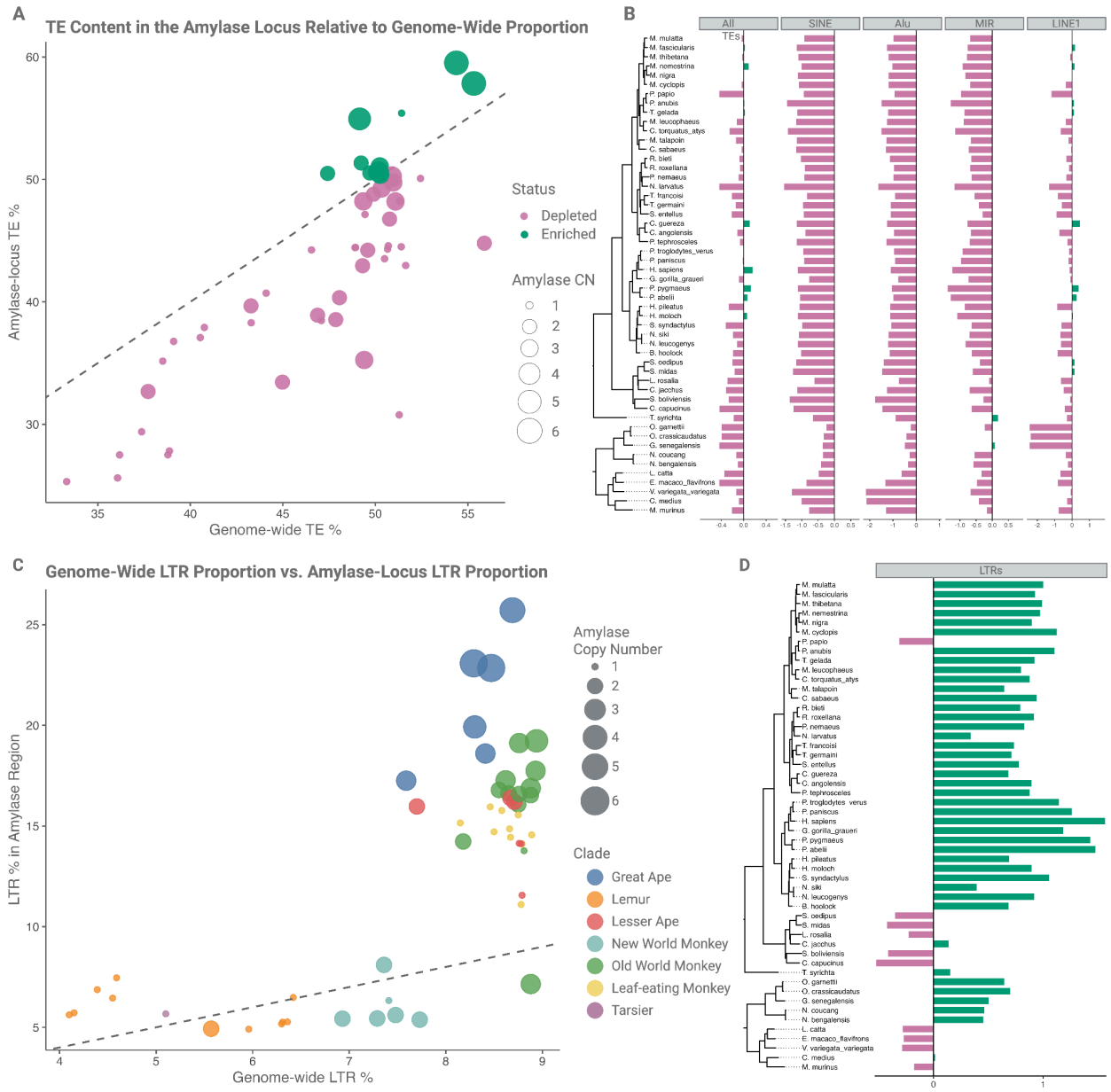

***Figure 4*: The transposable element landscape in the primate amylase locus.** (A) Transposable element (TE) content (given as the total TE in bp/the total length of the locus in bp, see **Methods** for more) in the amylase locus relative to genome-wide TE proportion across 53 primate species. Each point represents a species, with the size of the circle scaled by the respective amylase copy number and the color indicating enrichment status. The dashed line marks the 1:1 expectation. The majority of species show depletion of TEs in the amylase locus compared to genome-wide levels. (B) TE family-specific enrichment ($\log_2$ transformed) in the amylase locus across primates for SINEs, Alus, MIRs and LINE1s, alongside total TE content ("All TEs") (**Table S5**). Bars represent enrichment (green) or depletion (pink) relative to genome-wide TE representation. Short retrotransposons (e.g. Alus & SINEs) are consistently depleted from the amylase locus, suggesting that this region might have limited retention of active mobile elements across lineages. (C) Relationship between genome-wide and amylase locus-specific LTR content across primate species. Each point represents a species, with the circle size indicating *AMY* gene copy number and the color representing phylogenetic clade. The dashed line indicates a 1:1 ratio. Most species show an enrichment of LTRs at the amylase locus relative to genome-wide levels. (D) Log_2-transformed enrichment and depletion of LTR content at the amylase locus across 53 primate species, organized by phylogeny. Bars indicate whether LTR representation in the locus is higher (green) or lower (pink) than genome-wide LTR proportions. Several species with reported amylase gene duplications, such as olive baboon, exhibit pronounced LTR enrichment.

Signals of selection suggest potential functional variation among primate amylase paralogs

While human amylase paralogs in general appear to evolve under negative selection without significant amino acid divergence from each other (*4*), recent studies highlighted potentially functional variations among human paralogs (*37*). To identify such functional differences and signals of potential positive selection, we aligned the coding sequences of all identified paralogs from olive baboons and rhesus macaques with paralogs from ape genomes.

Our analysis yielded three major insights (**Figure 5A & 5B)**. First, we identified a significant positive selection signal on the internal branch leading to baboon and rhesus macaque paralogs compared to great ape paralogs (aBSREL, $p = 0.038$). Second, we detected a premature stop codon mutation in the ancestral *AMY2B* gene copy within baboons, previously unannotated in the NCBI database (**Figure 5A**). Third, we detected strong evidence for positive selection specifically on *AMYp2* in olive baboons, supported by both aBSREL ($p < 10^{-6}$) and RELAX analyses ($p < 0.0001$). Additionally, we identified six codons exhibiting episodic positive selection that likely contribute to this overall selection signal (**Methods; Figure S8**). The concurrent presence of a premature stop codon in *AMY2B* and positive selection in *AMYp2* is remarkable. It has been shown previously that following gene duplications, one of the copies may become pseudogenized and the other one retains the original function (*38*, *39*). Thus, it is plausible here that the newly derived gene (*AMYp2*) functionally compensates for the observed loss of function in the ancestral gene (*AMY2B*).

Next, we analyzed the functional annotations of primate amylase amino acid sequences, including known glycosylation sites, streptococci-binding motifs, catalytic and active sites, and calcium-binding domains (**Figure 5C**, see **Methods**). To evaluate whether the observed amino acid substitutions might affect protein structure or function, we generated AlphaFold2 models for the amylase gene products of each paralog, with a particular focus on the baboon-specific *AMYp2* sequence, which showed the strongest signal of positive selection (**Figure S9**). The functional predictions suggest that most of the positively selected substitutions do not overtly disrupt the protein fold, glycosylation motifs, or key catalytic residues. However, subtle impacts on substrate affinity or protein stability cannot be ruled out (**see Table S6** for predicted pathogenic and neutral substitutions). Notably, we identified one strongly selected site ($p = 0.01$, MEME) in *AMYp2*, involving a threonine-to-serine substitution at the position 178 predicted as an active site by NLM's conserved domain database (*40*), which may reflect fine-tuned functional divergence in the olive baboon lineage.

Collectively, our findings suggest that, unlike reported for humans (*4*), Old World monkey amylase gene duplications involve significant amino acid diversification, indicative of potential neofunctionalization.

Reconstructing the evolution of salivary gland-specific expression

To investigate the expression patterns of lineage-specific amylase gene paralogs, we generated transcriptome data from parotid, sublingual, and submandibular salivary gland, as well as liver and pancreas tissues from six rhesus macaques and five olive baboons (**Table S7**). We also leveraged our previously published transcriptomic data from the corresponding human salivary gland tissues (*41*), and added data for pancreas and liver tissues from the GTEx database. Taking advantage of sequence differences among paralogs within each species, we used Kallisto (*42*) to quantify transcript abundance of each gene, as it offers high sensitivity and accuracy for distinguishing among closely related gene copies (*42*) (see **Methods** for more detailed discussion). These findings led us to form several hypotheses to explain the regulation of amylase duplicates.

The first clear pattern we observed in both baboons and rhesus macaques is that, similar to humans, the last gene (at the 3' end) in the amylase cluster consistently shows elevated expression (parotid vs. pancreas $\log_2$FC: rhesus *AMY1'*=6.08, baboon *AMY1'*=1.91) in salivary tissues relative to the other paralogs (**Figure 5D**). However, in contrast to humans, where *AMY1* is exclusively expressed in saliva, this gene in baboons and rhesus macaques retains expression (pancreas vs. liver $\log_2$FC: rhesus *AMY1'*=0.40, baboon *AMY1'*=3.48) in the pancreas and liver, indicating broad expression patterns (**Figure 5E)**. Additionally, the relative contribution of each amylase paralog to total salivary expression differs among species (**Figure 5D**). In humans, *AMY1* accounts for nearly all salivary gland expression (parotid vs. pancreas $\log_2$FC: *AMY1A*=14.71, *AMY1B*=9.74, *AMY1C*=14.35). The lineage-specific duplications *AMYm* and *AMYp2* also contribute to the overall salivary expression in rhesus macaques and olive baboons, respectively.

Given that *AMY1'* in Old World monkeys represents the ancestral gene from which the great ape *AMY1* and *AMY2A* genes evolved (**Figure 2A**), it offers a valuable framework for investigating how tissue-specific expression arose in these newly duplicated genes. In humans, *AMY1* is expressed exclusively in salivary glands, while *AMY2A* is expressed only in the pancreas (**Figure 5E**). Our transcriptomic analyses in rhesus macaques and baboons shows that *AMY1'* is expressed in both pancreas and salivary glands, likely reflecting the ancestral state for Catarrhini (**Figure 5D & 5E**). Based on these observations, the most parsimonious explanation is that the ancestral *AMY1'* gene had already acquired expression in both the pancreas and salivary glands, particularly the parotid gland, prior to the divergence

of great apes. Following duplication in the great ape lineage, subfunctionalization occurred: *AMY1* retained salivary gland expression, while *AMY2A* lost salivary expression and became restricted to the pancreas. This shift may have been facilitated by an ERV insertion, as previously proposed (*23*, *25*). Together, these findings support a subfunctionalization model for the great ape *AMY1* and *AMY2A* following their divergence from *AMY1'*.

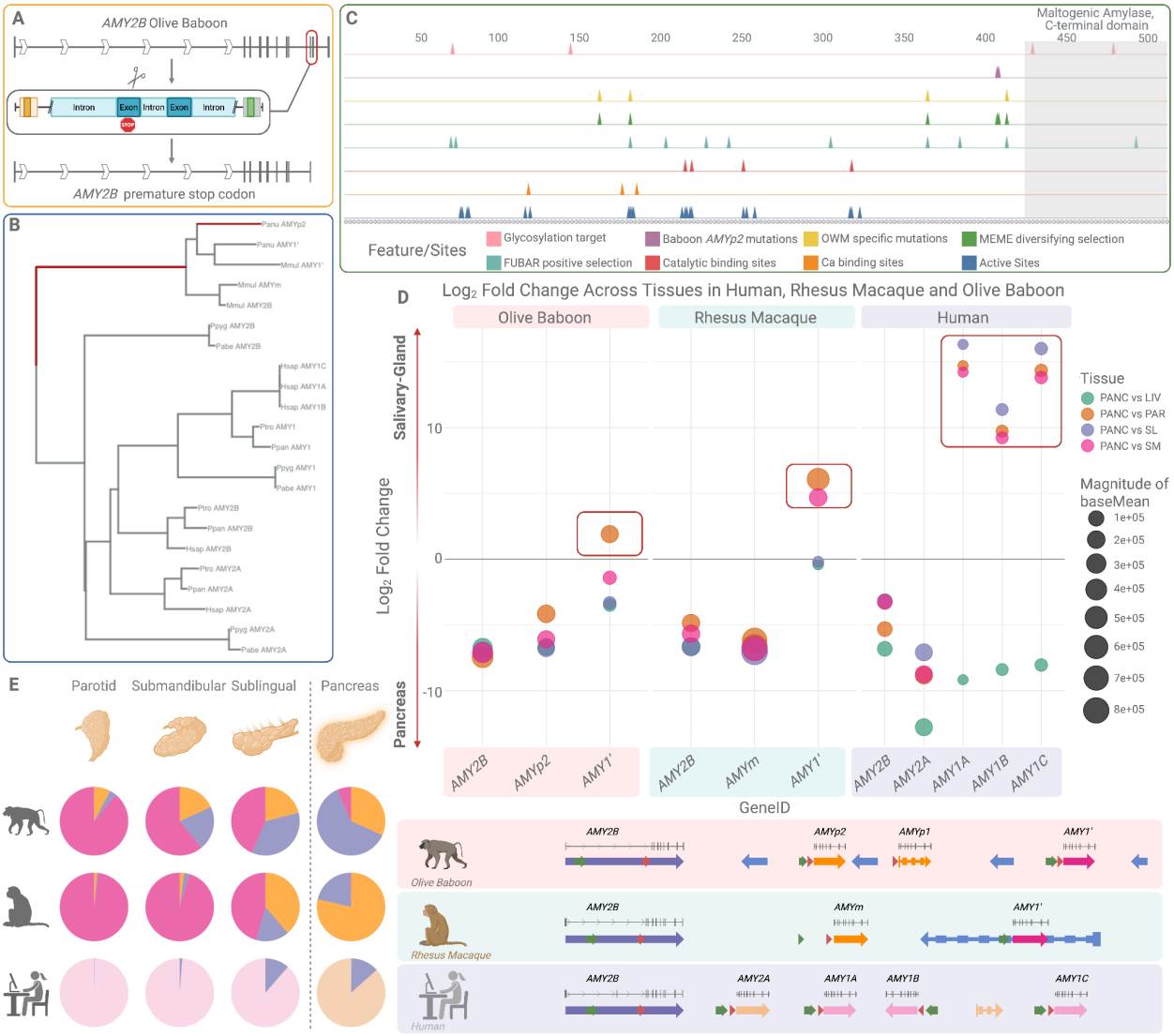

Figure 5: **Functional divergence and tissue-specific expression of amylase paralogs in olive baboons, rhesus macaques and humans.** (A) Identification of a premature stop codon in *AMY2B* of olive baboon. A schematic of the disrupted gene structure (top) shows the location of the nonsense mutation (red box) within the ninth exon. (B) A maximum likelihood phylogeny of amylase coding sequences from olive baboon, rhesus macaque and great apes, highlighted branches under significant episodic diversifying selection (dark red) affecting the lineage leading to Old World monkey paralogs (aBSREL, p = 0.038). (C) Old world monkey and baboon-specific variants and their overlap with predicted functional sites across the amylase protein. Tracks show the position of catalytic and calcium-binding residues, glycosylation sites, positively selected codons (FUBAR, MEME), Old World monkey-specific mutations and baboon-specific mutations in *AMYp2*. A MEME- and FUBAR-identified site (codon position 178) involves a threonine-to-serine substitution overlaps an active site. (D) Differential expression of amylase paralogs across tissues in olive baboons, rhesus macaques and humans. The scatterplot shows $\log_2$-fold change for each paralog across

tissue comparisons (pancreas vs parotid, sublingual or submandibular glands and pancreas vs. liver). The size of the circle corresponds to the base mean expression level. In baboons and macaques, lineage-specific genes (*AMYp2*, *AMYm*) contribute to salivary expression, unlike in humans where *AMY1* is the dominant/primary salivary paralog. Arrows represent the structure and orientation of amylase paralogs across species, colored by paralog identity: *AMY2B* (purple), *AMY2A* (light orange), *AMY1* (light pink), *AMY1'* (pink), *AMYm*, *AMYp1* and *AMYp2* (orange). (E) Pie charts showing proportional expression of each amylase paralog in salivary glands (parotid, submandibular, sublingual) and pancreas. Expression levels of *AMY1'* in both salivary glands and pancreas in macaques and baboons supports subfunctionalization as the likely evolutionary mechanism for the tissue-specific roles of *AMY1* and *AMY2A* in humans.

Multifactorial regulation of tissue-specific expression of amylase paralogs

To investigate how regulatory elements contributed to the evolution of amylase gene expression in primates, we examined transcription factor binding sites across paralogs and species. Using *in silico* predictions of transcription factor binding sites and promoter regions for amylase paralogs in rhesus macaques, olive baboons, and humans, we identified 262 distinct binding sites (consisting of 108 unique TFBS) across the promoters of the nine annotated amylase gene paralogs analyzed (**Table S8**).

A simplistic model of tissue-specific expression assumes that the presence of a specific transcription factor binding site determines tissue specificity. If this were the case, we would expect the transcription factor binding motifs among primate amylase paralogs to cluster according to their expression in either salivary glands or pancreas. However, we found no such pattern: paralogs with known tissue-specific expression did not show consistent enrichment of salivary- or pancreas-biased motifs (**Figure 6**). Instead, our results suggest a partial conservation in promoter binding site composition among primate paralogs, where all three rhesus macaque paralogs (*AMY2B*, *AMYm*, and *AMY1'*) share a similar binding site profile with olive baboon *AMY1'* and *AMYp2*. In contrast, the promoter regions of the pseudogenized olive baboon *AMY2B* and the human paralogs (*AMY2B*, *AMY2A*, and *AMY1s*) exhibit a distinct transcription factor binding sequence composition, consistent with the lineage-specific evolution of these gene copies (**Figure 2A**).

FOXC1 is a key transcription factor involved in salivary gland development and expression (*43*). Notably, we found that all primate *AMY* paralogs, regardless of whether they are salivary gland- or pancreas-biased, contain FOXC1 binding sites, with the exception of human *AMY2B* and the pseudogenized olive baboon *AMY2B*. This observation suggests that the presence of a salivary-biased regulatory motif alone is insufficient to dictate tissue-specific expression. Instead, expression patterns of *AMY* genes may be further shaped by context-dependent factors such as chromatin accessibility, competitive binding, or the presence of co-regulators. For example, despite containing a FOXC1 binding site, human *AMY2A* is primarily expressed in the pancreas, highlighting the complexity of regulatory control at this locus.

These findings suggest that the presence or absence of salivary gland or pancreas-biased transcription factor binding sites alone does not fully account for the observed tissue-specific expression patterns, highlighting the potentially combinatorial nature of transcriptional regulation, an observation consistent with previous findings from other tissue-specific systems (*41*).

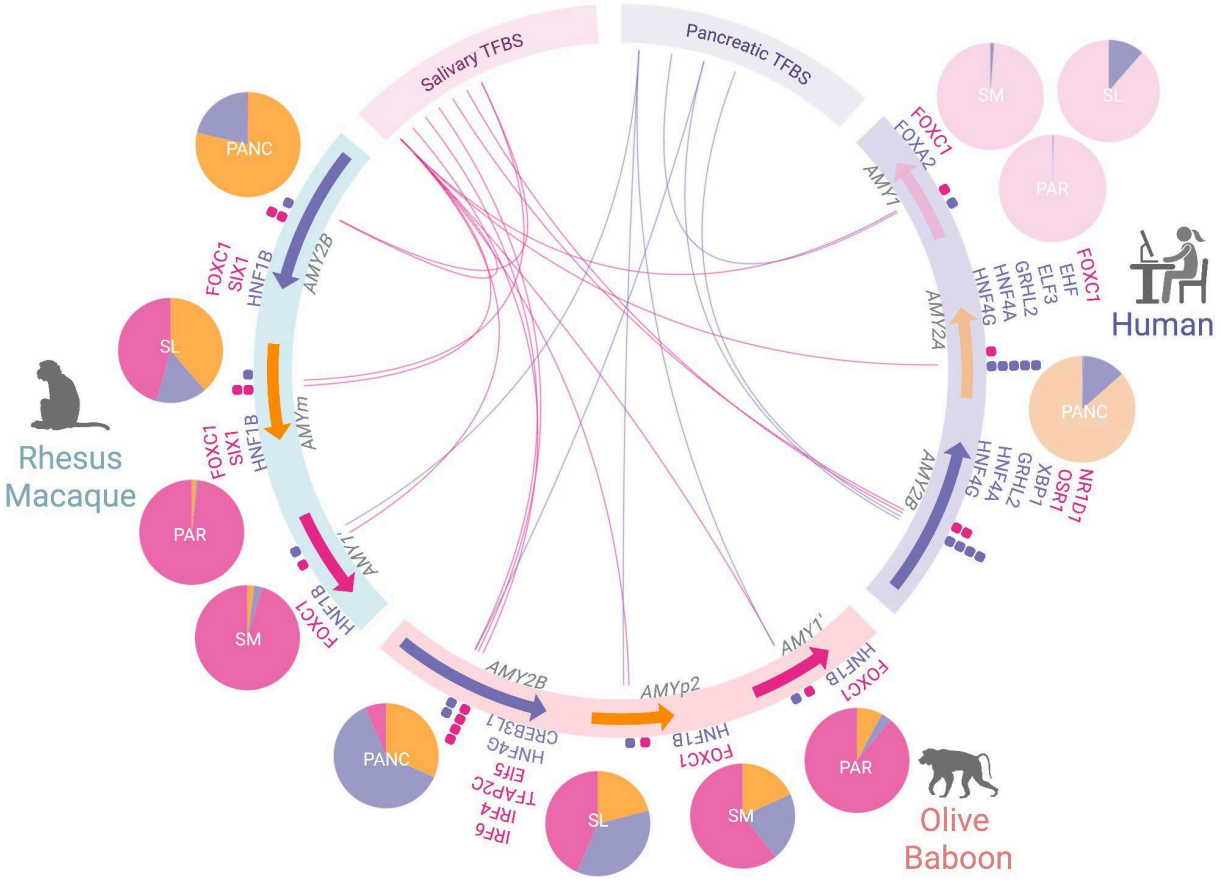

Figure 6. **Predicted transcription factor binding sites (TFBS) associated with salivary-gland and pancreatic expression across amylase gene paralogs.** Circos plot showing *in silico* predictions of salivary (pink) and pancreatic (purple) TFBS across promoter regions (~170 bp upstream of the start codon and annotated promoter) of amylase paralogs in rhesus macaques, olive baboons and humans. Colored arrows represent gene orientation. Dots adjacent to each gene denote the presence of TFBS (pink for salivary, purple for pancreatic). Ribbon links connect paralogs to TFBS categories. Pie charts summarizing relative gene expression across tissues: PAR (parotid), SM (submandibular), SL (sublingual), and PANC (pancreas). The color of each slice corresponds to the proportion of total expression of the given gene in each tissue, following the same color scheme used for the gene annotation arrows. All three rhesus macaque paralogs (*AMY2B*, *AMYm*, *AMY1'*) share similar TFBS profiles with olive baboons *AMY1'* and *AMYp2*. In contrast, *AMY2B* in olive baboon and human exhibits distinct TFBS composition. The salivary gland-biased transcription factor *FOXC1* is present in most paralogs, but absent from human and olive baboon *AMY2B*, both of which lack salivary expression.

CONCLUSION

In this study, we used the amylase gene locus as a model to investigate within the primate lineage how structurally complex loci contribute to molecular convergence. Convergent evolution, where similar traits arise independently across lineages, is a hallmark of adaptive evolution and highlights non-random patterns shaped by natural selection. Structurally complex regions have recently emerged as key players in this process, and it is thought that independent gene duplications might play an important role in this context. Focusing on the amylase locus, which is one of the most structurally dynamic regions in the human genome (*4*), we expanded prior work in humans to include the broader primate phylogeny. We identified multiple lineage-specific amylase gene duplications, including previously uncharacterized expansions in bonobos, orangutans, lemurs, and New World monkeys. This comprehensive analysis of amylase evolution enabled us to dissect the interplay between mutational mechanisms, positive selection, and regulatory divergence underlying functional convergence.

Our findings revealed that NAHR-mediated rearrangements in baboons, macaques, and humans occurred through independent breakpoints, underscoring the recurrent nature of duplication events. Given that NAHR relies on existing homologous sequences, we asked how primary duplications arise in the first place. We observed a strong correlation between amylase copy number and LTR abundance across species, suggesting that lineage-specific transposable element insertions may seed the homology required for initiating structural instability. This provides preliminary evidence for a broader hypothesis, that transposable elements may play a pivotal role in priming structurally complex loci for recurrent evolution, a compelling direction for future studies across mammalian genomes.

Beyond structural variation, we detected signatures of positive selection at specific codons, suggesting functional divergence among amylase paralogs in a species-specific manner. Such changes may be of functional relevance because they can alter the conformation of the protein and change or create motifs governing posttranslational modifications. Evidence supports that human AMY paralogs differ in glycosylation potential and may influence oral microbiota composition (*37*). The interplay of nucleotide-level and structural variation lays the groundwork for future population-level studies in closely related species with differing diets or pathogen pressures, offering a powerful framework to investigate the adaptive relevance of *AMY* variation.

By integrating genomic variation, gene expression, and transcription factor motif analyses, we show that *AMY* paralogs have independently evolved tissue-specific expression, particularly in salivary glands. Rather than relying on distinct, tissue-specific transcription factors, expression biases appear to result

from varying combinations of binding motifs among paralogs and across species. This regulatory rewiring is linked to structural rearrangements at the locus, including duplications, inversions, and deletions, that reshape the genomic context of regulatory elements. Our findings, thus, highlight the likely importance of distal regulatory elements beyond core promoters in mediating tissue-specific expression. Future functional assays such as ATAC-seq, ChIP-seq, or Fiber-seq will be essential to fully resolve the evolving regulatory landscape of the amylase locus.

A vivid example of the co-evolution of gene duplication and regulatory rewiring is seen in the great ape *AMY2A* and *AMY1* genes. Our results show that these paralogs arose from an ancestral gene (*AMY1'*) with dual expression in the pancreas and salivary glands. Following duplication, *AMY2A* and *AMY1* subfunctionalized into largely pancreas- and salivary gland-specific expression, respectively. This instance of regulatory and functional co-divergence exemplifies the distinct evolutionary innovation at the primate amylase locus, ranging from fixed duplications (as in macaques and baboons), to extensive intraspecific variation (as in humans), and even gene loss (in leaf-eating monkeys).

Notably, we found no species lacking the amylase gene entirely, suggesting that while the locus tolerates considerable structural and regulatory flexibility, it remains under consistent functional constraint. This balance, between mutational plasticity and adaptive necessity, places the amylase locus in what Ponting (*44*) has termed the "evolutionary twilight zone." It is within this zone that convergence has repeatedly emerged during primate evolution, driven by distinct molecular mechanisms acting on a structurally complex genomic landscape.

We propose that structurally complex loci across mammalian genomes are hotspots of molecular convergence, harboring exceptional evolutionary potential. Their structural complexity not only promotes gene copy number expansion and contraction but also enables regulatory innovation, fueling expression divergence and functional diversification of duplicated genes.

MATERIAL & METHODS

Primate Genome Assemblies and Locus Extraction

To capture the phylogenetic diversity of the amylase locus across primates, we analyzed 244 high-quality genome assemblies representing 222 species (**Table S1**). To assess copy number variation and structural organization, we first delineated the locus using two flanking, non copy-number-variable anchor

sequences: one upstream (5′) of the *AMY2B* gene corresponding to the *RNPC3* genic region, and one downstream (3′) of the *AMY1C* gene. Both sequences were derived from the human GRCh38 reference genome and used as queries in blastn (v. 2.14.1+) (*45*) searches against each assembly.

We automated this analysis using a custom SLURM-based Bash pipeline. The script identified assemblies where both anchors mapped to a single contig or scaffold, extracted the intervening locus using samtools faidx (v. 1.22) (*46*) and generated reverse complements with seqtk (v. 1.5-r133; https://github.com/lh3/seqtk) for loci located on the reverse strand. Only assemblies in which the full amylase locus was contained within a single contig were retained for downstream analyses; loci fragmented across multiple scaffolds were excluded but recorded. The queried sequences and the SLURM-based Bash pipeline have been deposited in Zenodo; see Data Availability.

Synteny Analyses and Gene Annotation

To characterize the structural composition and assess synteny of the amylase locus across primates, we analyzed the 70 assemblies in which the complete locus was successfully extracted as a single contiguous sequence. Each locus was aligned to the human H1^{a}.1 haplotype, which carries a single *AMY1* gene and represents the ancestral configuration in great apes, using NUCmer (v. 3.1) (*47*) and LAST (*48*). Dotplots were then generated using mummerplot (v. 3.5) (*47*) and visually inspected to identify structural rearrangements relative to the human reference.

To further investigate patterns of structural evolution both within and between genera, we performed additional pairwise alignments among species of the same genus, as well as representative comparisons across genera. These alignments were again generated with NUCmer (v3.1) and visualized using dotplots and miropeats-style plots. For the latter, we used custom scripts to convert NUCmer coordinate files into PAF format and visualized the alignments using the R package SVbyEye (v. 0.99.0) (*49*).

To annotate amylase gene copies across primates, we employed a multi-step approach combining reference annotations, CDS mapping, and manual curation. For each genome assembly with a contiguous extracted amylase locus, we first retrieved available NCBI gene annotations for that species and used exonerate (v. 2.4.0) (*50*) to map the corresponding coding sequences (CDS) to the extracted regions. In assemblies lacking gene annotations, we utilized CDS from closely related sister species and mapped them to the respective genomes using the same exonerate-based pipeline, which is well-suited for aligning sequences in the presence of sequence divergence and detecting partial exons.

To identify putative orthologs and resolve paralogous relationships, we implemented a recursive reciprocal BLAST approach. Putative amylase gene models from each species were queried against one another using blastn (v. 2.14.1+) and reciprocal searches were used to establish one-to-one or one-to-many orthology/paralogy relationships. In a one-to-one orthology relationship, reciprocal BLAST searches yield the same gene pair as the top hit in both directions e.g. gene X from species 1 identifies gene Y from species 2, and gene Y reciprocally returns gene X. By contrast, a one-to-many pattern arises when duplication has occurred in only one lineage, and this can appear in two complementary ways. If species 2 experienced the duplication, a single gene in species 1 matches several genes in species 2, and each of those paralogs reciprocally lists the original query gene as its best hit. Conversely, if the duplication occurred in species 1, multiple paralogs in that species all identify the same ortholog in species 2 as their top hit, while the lone gene in species 2 reciprocally aligns most strongly to just one, commonly the most conserved, of the copies in species 1. Either configuration reveals a one-to-many orthology generated by a lineage-specific duplication event. This approach allowed us to confidently distinguish ancestral copies (e.g. *AMY2B*) from lineage-specific duplications.

Lastly, in olive baboons (*P. anubis*), we detected a premature stop codon within the *AMY2B* gene, suggesting likely pseudogenization. This mutation is absent from the NCBI annotation and was not present in the CDS of *Papio papio*, but was consistently identified in both haploid assemblies of *P. anubis*.

Transposable Element Annotation and Quantification

All 53 primate assemblies that yielded a contiguous amylase locus were analyzed using RepeatMasker (v. 4.1.5; http:// repeatmasker.org) with the -species primates flag and default parameters. This approach applies a uniform primate-specific repeat library to every genome, avoiding the inconsistencies that would arise from mixing lineage-specific libraries. However, this choice can systematically underestimate species-restricted transposable element (TE) families absent from the reference library, an acknowledged limitation of our approach. RepeatMasker output files (.tbl and .out) were parsed to calculate genome-wide TE content for each assembly. Assembly quality introduces additional bias to these estimates, as chromosome-level assemblies often recover TE-rich centromeric and telomeric regions that incomplete or nearly-complete assemblies commonly miss, thereby inflating total TE proportions.

To analyze TE content within the amylase locus, we applied RepeatMasker separately to the extracted amylase sequences from each of the 53 species. These sequences were bounded by non copy-number-variant RNPC3 5′ and AMY1C 3′ flanks (detailed in the previous section) and restricted to

loci assembled as single, gap-free contigs/scaffolds, as a prerequisite to avoid potential segmental duplication collapse. Masking isolated locus sequences, rather than extracting annotations from whole-genome GFF3 files, prevented coordinate errors and ensured that every species was compared across identical boundaries.

TE abundance was quantified as the proportion of masked bases (bp masked / region length) for both genome-wide and locus-specific analyses. This normalization accounts for length variation introduced by gene duplications. All downstream statistical analyses and visualizations were conducted in R (v. 4.3.3). For each species, we compared the percentage of masked sequence in the amylase locus to its genome-wide percentage. For each TE family, we calculated log2-enrichment, defined as log2(% in locus)/(% genome-wide). This metric was then merged with amylase copy-number estimates and clade assignments. Enrichment or depletion was assessed with a two-sided Wilcoxon signed-rank test across the 53 species in which the locus was contiguous; family-specific analyses (e.g. Alu, LTR, DNA transposons) were performed analogously. P-values were adjusted for multiple testing using the Benjamini-Hochberg procedure, considering FDR<0.05 significant.

We assessed the potential confounding effect of assembly contiguity (scaffold N50/Assembly N50) on TE content estimates. Assembly N50 showed a strong positive correlation with genome-wide TE content (Spearman $\rho=0.554$, $P<0.001$; $n=53$). However, this relationship was primarily driven by assembly length. After controlling for total assembly length, the association between N50 and TE content remained significant but was substantially reduced (partial correlation $\rho=0.327$, $P=0.018$), indicating that assembly quality has a modest independent effect on TE detection beyond simple genome size effects.

To test whether TE abundance predicts amylase gene copy number while accounting for shared ancestry, we fit phylogenetic generalized least-squares (PGLS) models in R (nlme 3.1 and caper 1.0.1). A ultrametric primate tree pruned to the 53 focal species was obtained from TimeTree (accessed January 2025). The response variable was amylase gene copy number while the predictor was locus-specific TE proportion (total or by family). Pagel's λ was estimated by maximum likelihood and retained if significantly different from zero. Model fit was evaluated by R^2 and residual diagnostics. All scripts for TE analysis, including those used to reproduce Figure 4, are available in a Zenodo repository (see Data Availability)

Structural Variant Inference and Duplication Mechanism Analysis

To investigate the mechanisms underlying the duplication in the amylase locus, we extended our annotation of the region beyond the existing annotations available on NCBI. We utilized BISER (v. 1.4)

(*51*) to identify segmental duplications in the olive baboon amylase locus, analyzing the masked and unmasked genome to capture all duplicated units, irrespective of functional annotation. This approach led to the identification of 43 distinct duplicated segments (**Table S9**). Additionally, we identified four lncRNA sequences within the locus, positioned downstream of each gene with reverse orientations relative to the genes. Using self-alignments, dotplots, and BLAST searches, we partitioned the amylase locus into four distinct segments, each corresponding to a single duplicon (**Figure S10A**). Each segment includes the coding sequence of an amylase gene and extends to the terminal sequence of an adjacent lncRNA. This partition enables a systematic examination of duplicated regions across the locus.

We then compared the four distinct segments to one another using pairwise NUCmer alignments and visualized these comparisons with dotplots. Query coverage and sequence similarity metrics allowed us to infer the order of duplication events. A closer examination of the nucleotide sequences further informed us of the duplication mechanisms involved. Namely, the third segment (containing the amylase gene *AMYp1*) appears to be a composite of the fourth segment (encompassing *AMY1'*) and the first segment (encompassing *AMY2B*) (**Figure S10B**). Meanwhile, the second segment is nearly identical to the third, with the only difference being a ~10 kb deletion within the third segment (**Figure S10C**). Because this interval is bounded by long stretches of Ns in the olive baboon assembly, we cannot determine whether it represents a true biological deletion or a local misassembly/scaffolding artifact, and we therefore do not interpret this feature further. Taken together, the sequence-similarity patterns indicate that two non-allelic homologous recombination (NAHR) events duplicated the ancestral *AMY2B* and *AMY1'* blocks, yielding the present-day amylase locus structure in olive baboons (**Figure S10**).

To delineate precisely the novel duplications in the olive baboon amylase locus, we utilized closely related *Papionini* and Old World monkey species that still retain the ancestral Catarrhini two-copy configuration (*AMY2B* and *AMY1'*). Continuous contigs spanning the locus were available for gelada (*Theropithecus gelada*), drill (*Mandrillus leucophaeus*), sooty mangabey (*Cercocebus atys*), and crested macaque (*Macaca nigra*). We extracted the orthologous intervals (RNPC3 5′ flank to AMY1C 3′ flank; see above) and aligned each of them against the olive baboon amylase locus with NUCmer (v 3.1; default settings), visualising the results with mummerplot and extracting the corresponding coordinates. Between these species, uninterrupted collinearity was observed across the junctions that define the "first segment" and "fourth segment" as defined above. The "second segment" and "third segment" were then mapped separately against the ancestral two-copy configuration, allowing us to delineate the exact breakpoints introduced by the first and second non-allelic homologous recombination events. The sequence that is contiguous in olive baboons but split in the two-copy genomes corresponds to the novel duplications.

These alignments confirmed that no additional rearrangements in the olive baboon amylase locus and that the derived *AMYp1* and *AMYp2* blocks are absent from the ancestral Catarrhini two-copy configuration, validating the BISER-based segmentation. We additionally aligned the olive baboon locus to that of its sister species Guinea baboon (*Papio papio*). Although the Guinea baboon assembly contains six tandem amylase copies, in addition to the ancestral *AMY2B* and *AMY1'*, all of these extra copies are identical in nucleotide sequence, strongly suggesting an assembly artifact caused by assembly over-expansion rather than true structural variation in the locus. Importantly, every Guinea baboon expanded amylase block matched the olive baboon *AMYp1* segment, confirming that the *AMYp1* represents the first and oldest duplication shared by both baboon species. This cross-species comparison therefore corroborates the step-wise duplication model inferred above. Finally, these comparisons allow us to anchor the two NAHR events to the 2-4.5 MYA and ~1.85 MYA windows inferred from *Papionini* phylogeny.

To investigate the macaque-specific duplication, we applied the same segmentation and alignment workflow to the rhesus macaque locus. Self‑alignments (dotplots) and pairwise NUCmer comparisons partitioned the region into three paralogue segments, each corresponding to a single duplicon: segment 1 harbouring *AMY2B*, segment 2 carrying *AMYm* (the novel copy), and segment 3 harbouring *AMY1′*. Pairwise comparisons showed that *AMYm* shares extensive 5′ homology with *AMY1′* and 3′ homology with *AMY2B*, with crossover points falling within the intergenic sequence upstream the novel gene. The preserved orientation and high identity of the flanks, together with these shared blocks, implicate NAHR as the duplication mechanism (**Figure S2**). In parallel, BISER (v. 1.4) identified 26 unique duplicons within the macaque locus (**Table S10**), which we treat as the smallest repeat units; by contrast, the three segments defined above represent the largest repeated modules spanning each paralogue block across the locus.

We then aligned orthologous contiguous amylase locus from multiple *Macaca* species representing the three major clades (fascicularis, sinica, silenus). *AMYm* is present in the *fascicularis* and *sinica* groups but absent from the *silenus* group, placing the duplication after the silenus split and before diversification of the fascicularis and sinica clade, i.e. approximately ~4.5-5 MYA. All *AMYm*-harbouring loci were recovered on single, gap-free contigs with identical breakpoints across species and no additional rearrangements, supporting a single NAHR event that generated the three-segment architecture in macaques (**Figure S3**).

Transcriptomic Data Generation and Processing, and Differential Expression Analysis

Biopsies (parotid, submandibular, sublingual, pancreas and liver) from five olive baboons (*Papio anubis*) and six rhesus macaques (*Macaca mulatta*) were flash-frozen in liquid nitrogen immediately after

collection and stored at -80°C. These samples were collected at Texas Biomedical Institute by veterinarians from 7-18 year old animals right after planned euthanization for health reasons. Samples are kept at -20°C. Library preparation and standard Illumina HiSeq RNA sequencing (2x150bp) experiments were carried out following standard operating procedures by GENEWIZ/Azenta. Publicly available adult human salivary gland dataset (*41*) was downloaded as FASTQ files and processed similarly to the olive baboon and rhesus macaque RNAseq samples from the quality-control step onwards. The pancreas and liver expression datasets were obtained from GTEx v8 and normalized as detailed below.

Quality assessment of raw reads was performed with FastQC (v. 0.12.1) (*52*) and summarised with MultiQC (v. 1.25) (*53*). Adapter sequences, low quality bases at the 3' and 5' end of the read, and reads shorter than 36 bp were removed using Cutadapt (v. 3.5) (*54*). To investigate tissue-specific expression, trimmed reads were aligned to the olive baboon (*Papio anubis* Annotation Release 104) and rhesus macaque (*Macaca mulatta* Annotation Release 103) transcriptomes, and transcript abundances were quantified with Kallisto (v. 0.51.1) (*42*). Transcript-level transcripts per million (TPMs) were summarized to the gene level. The gene-level quantifications were further normalized using the "lengthScaledTPM" method, for consistency in downstream analyses. These data were then filtered to include only high-confidence gene expression estimates (minimum of ten reads across all samples). The normalized gene expression profiles were used for differential expression analysis and visualization. For differential expression analysis, we utilized DESeq2 (v. 1.46.0) (*55*) to compare gene expression profiles between tissues within species. We applied a Wald test to detect significant expression differences between groups. Genes with an adjusted p-value (FDR < 0.05) were considered significantly differentially expressed.

Human salivary gland reads (*41*) were quantified with Kallisto against the NCBI *Homo sapiens* Annotation Release 110 transcriptome. Pancreas and liver read-count tables were obtained from GTEx v8. Before combining these data with the Kallisto-derived human salivary gland quantifications, we removed Ensembl version suffixes from every gene ID, retained only genes present in all samples and rounded Kallisto's fractional estimates to integers so that every entry represented a raw count compatible with DESeq2. A DESeq2 object was then created using tissue as the design factor, and size factors were estimated with the default median-of-ratios procedure to generate library-normalised counts. Variance-stabilising transformation was applied for PCA-based quality control (GTEx pancreas, GTEx liver or each salivary gland tissue) to confirm that batch effects did not dominate the variance structure. The resulting size-factor-normalised count matrix was used for all subsequent differential-expression analyses.

In parallel, we developed a polymorphism-aware pipeline to achieve higher resolution quantification of gene expression across tissues. This pipeline leverages nucleotide polymorphisms within RNA-seq reads to partition expression into transcript-specific contributions. We first aligned all human, rhesus macaque, and olive baboon transcripts, accounting for alternative splicing isoforms, to identify coding-sequence polymorphisms. For each species, we generated a single consensus FASTA sequence that captured the shared polymorphisms across the different transcripts within species. We then extracted RNA-seq reads previously mapped to each genome using HISAT2 (v. 2.2.1) (*56*), retaining only those overlapping the amylase locus, and realigned them to the corresponding species-specific consensus FASTA with BWA (v. 0.7.19-r1273) (*57*). NCBI annotations were curated for known polymorphic regions in the olive baboon and rhesus macaque transcripts, and the data were integrated with species-specific annotations to refine the analysis.

Using Old World Monkeys transcripts as outgroups, we assigned the detected polymorphisms as ancestral, derived, or lineage-specific. Using custom Python scripts we extracted and quantified reads at these polymorphic sites, enabling paralog-specific expression estimates (scripts have been deposited to Zenodo:https://doi.org/10.5281/zenodo.16809248). By focusing on transcript- and paralog-level variation, our approach aimed to reveal tissue-specific expression patterns that would otherwise remain obscured in conventional bulk RNA-seq data. To ensure robust detection of gene-level differences, we specifically utilized polymorphisms that distinguish one paralogous gene from another. In the rhesus macaque genome, for instance, three paralogous genes, *AMY2B*, *AMYm* and *AMY1'*, are annotated in that order. For each paralogous gene, we selected four polymorphisms: two located toward the proximal (5') end and two toward the distal (3') end of the coding region. At positions 93 and 111, *AMY2B* carries the polymorphisms G, T, *AMYm* carries A, G, and *AMY1'* carries A, T (**Figure S11**). At positions 737 and 753, *AMY2B* carries A, T, *AMYm* carries T, T, and *AMY1'* carries T, C respectively. The polymorphism-aware pipeline allowed us to unambiguously assign reads to individual paralogs and confirmed the accuracy of Kallisto's expression estimates. It was used only as an internal validation step and did not contribute to the differential expression analyses or any other downstream analyses reported here.

Finally, we queried all annotated long-non-coding RNAs located within, or immediately flanking, the amylase locus in the rhesus macaque and olive baboon genome builds. The Kallisto-derived TPMs for every such lncRNA were below ten reads in every tissue, indicating negligible expression; they therefore cannot account for the tissue-specific patterns described in Results.

Positive Selection Analyses

Coding sequences (CDS) from the two Old World monkey species analysed here (olive baboons and rhesus macaques) together with orthologous CDS from all extant Great Ape species (human, chimpanzee, bonobo, Sumatran and Bornean orangutan) were aligned at the amino-acid level with MAFFT (v. 7.515) (*58*), back-translated to codons with PAL2NAL (v. 14) (*59*), and manually trimmed to preserve reading frame. We assessed the sequences for internal stop codons and removed any truncated CDS. A maximum-likelihood gene tree was inferred with IQ-TREE (v 2.4.0) (*60*) using the MG+F3X4+R2 model and was input to all HyPhy analyses. All selection analysis results are provided in Supplementary **Tables S11-S14**.

Branch-level tests (aBSREL)

To identify entire lineages that experienced bursts of adaptive change we ran the adaptive Branch-Site Random-Effects Likelihood (aBSREL) model in HyPhy (v. 2.5.48) (*61*) with default settings. aBSREL compares for every branch, a null model in which all sites evolve neutrally or under purifying selection ($\omega \leq 1$) to an alternative model that allows a proportion of sites on that branch to have $\omega > 1$. Likelihood-ratio tests are corrected for multiple comparisons with the built-in Holm procedure. Foreground branches were not pre-specified; instead HyPhy evaluates each branch in turn. The aBSREL p-values were corrected for multiple testing with HyPhy's built-in Holm-Bonferroni procedure (**Table S11**).

Site-level episodic diversifying tests (MEME)

Codon-specific episodic diversifying selection was evaluated with the Mixed-Effects Model of Evolution (MEME) in HyPhy (*62*). MEME allows the selective regime (ω) at a site to vary among branches, thus detecting positive selection that operates only on a subset of lineages. We used the default significance cut-off of $p \leq 0.10$ after applying the False Discovery Rate (FDR) correction using the Benjamini-Hochberg procedure (**Table S12**).

Site-level pervasive tests (FUBAR)

Pervasive, site-wide selection was assessed with the Fast, Unconstrained Bayesian AppRoximation (FUBAR) implemented in HyPhy (*63*), which estimates posterior probabilities for $\omega > 1$ under a Bayesian framework that assumes a constant selection regime across the tree. Analyses were run for 5 million MCMC iterations (burn-in=1 million); codons with posterior probability larger than 0.90 were considered positively selected (**Table S13**).

Relaxation or intensification of selection (RELAX)

To test variation in selective pressures across specific lineages, we applied RELAX (*64*). The baboon *AMYp2* branch was assigned as foreground and all other branches as background. RELAX fits two models differing by a scaling parameter K that inflates (K>1) or deflates (K<1) the background ω distribution; significance is assessed by a likelihood-ratio test (**Table S14**).

Structural Modeling and Functional Domain Prediction

The coding sequence for *AMYp2* was modelled *in silico* using AlphaFold2 (v. 2.3.1) (*65*) via the ColabFold implementation with the "monomer_ptm" preset and default recycling.The top-ranked model by pLDDT was retained; predicted-aligned-error (PAE) matrices were inspected to verify global fold confidence. Annotated PDB files were generated in PyMOL (v. 2.5) and used for all subsequent structure-based alignments.

Active, catalytic and calcium-binding sites were identified using the NCBI Conserved Domain Database (CDD, accessed March 2025). Calcium and chloride binding sites were further cross-referenced with the identified binding sites by Ramasubbu et al. (*66*, *67*). Glycosylation candidates were predicted with NetNGlyc 1.0 (*68*) and cross-referenced to the proposed glycosylation sites reported by Kamitaki et al. (*37*).

Regulatory Motif Analysis

Promoter sequences for the amylase paralogs in humans and rhesus macaques were defined as the 170-bp window spanning 100 bp upstream to 70 bp downstream of the experimentally supported TSS recorded in Eukaryotic Promoter Database (EPD) (*69*). When an EPD entry was unavailable, we took the RefSeq transcription-start site (TSS) from the corresponding annotation (NCBI *Papio anubis* Release 104, *Macaca mulatta* Release 103, and *Homo sapiens* Release 110) and defined the promoter as the region 100 bp upstream to 70 bp downstream of that TSS. We retrieved these windows for every amylase paralog in humans, rhesus macaques, and olive baboons. In parallel, we analyzed the 50 most highly expressed salivary gland genes in each species (ranked by TPM) and annotated their promoter sequences. Each promoter was scanned with MEME (MEME-suite v. 5.5.7) (*70*) for *de novo* motif discovery to identify novel, enriched sequence motifs.

The identified motifs were then annotated with Tomtom (MEME-suite v. 5.5.8) (*71*) against the JASPAR 2024 CORE non-redundant vertebrate library (*72*), retaining hits with $P<10^{-4}$ to identify potential transcription factors (TFs). To assess specificity, we conducted parallel analyses using promoter

sequences from 50 randomly selected genes per species, which did not show significant motif enrichment for the motifs identified with the salivary gland dataset, supporting the specificity of our results. The resulting TF list was cross-referenced with salivary- and pancreas-specific transcription factors with enriched expression from the FANTOM5 database ($P<0.05$ and $\log_{10}$(relative expression over median)>1.3) (*73*) and the salivary gland TF catalogue from Michael et al. (2019) (*43*). TFs present in either salivary gland set and absent from the pancreatic set were labelled salivary-gland-biased, while the converse defined pancreatic-biased TFs, with the remainder classed as core. The amylase-paralog promoter windows were then scanned with FIMO (*74*) against the JASPAR 2024 CORE library to identify TFBS within each promoter; only hits with $P<10^{-4}$ were retained. These TFBS were subsequently grouped into core, pancreatic-biased, and salivary-gland-biased categories based on the above TF assignments.

Data Visualization

All plots and summary graphics were created in R (v. 4.3). We used BioRender to combine individual panels into complete figures and to add all highlights and annotations. The R scripts for the figure panels are deposited in Zenodo (see **Data Availability**). The underlying source data for each figure are provided in the associated Supplementary Tables, for full reproduction of all visualisations.

Data and code availability

Bulk RNA-seq expression data (TPM, GTEx v8) is available from the GTEx open-access portal (https://gtexportal.org/home/downloads/adult-gtex/bulk_tissue_expression). All analysis scripts and the input files for downstream analyses and visualization have been deposited in Zenodo (DOI: https://doi.org/10.5281/zenodo.16809248). RNA-seq datasets for olive baboon and rhesus macaque have been deposited in GEO under accession numbers GSE305241 and GSE305255, respectively.

References

1. D. L. Stern, The genetic causes of convergent evolution. *Nat. Rev. Genet.* **14**, 751–764 (2013).

2. G. Ferraretti, A. Rill, P. Abondio, K. Smith, C. Ojeda-Granados, S. De Fanti, M. Alberti, M. Izzi, P. T. Sherpa, P. Cocco, M. Tiriticco, M. Di Marcello, A. Dezi, G. A. Gnecchi-Ruscone, L. Natali, A. Corcelli, G. Marinelli, P. Garagnani, D. Peluzzi, D. Luiselli, D. Pettener, S. Sarno, M. Sazzini, Convergent evolution of complex adaptive traits modulates angiogenesis in high-altitude Andean and Himalayan human populations. *Commun. Biol.* **8**, 377 (2025).

3. D. Bolognini, A. Halgren, R. N. Lou, A. Raveane, J. L. Rocha, A. Guarracino, N. Soranzo, C.-S. Chin, E. Garrison, P. H. Sudmant, Recurrent evolution and selection shape structural diversity at the amylase locus. *Nature* **634**, 617–625 (2024).

4. F. Yilmaz, C. Karageorgiou, K. Kim, P. Pajic, K. Scheer, Human Genome Structural Variation Consortium, C. R. Beck, A.-M. Torregrossa, C. Lee, O. Gokcumen, P. A. Audano, O. Austine-Orimoloye, C. R. Beck, E. E. Eichler, P. Hallast, W. T. Harvey, A. R. Hastie, K. Hoekzema, S. Hunt, J. O. Korbel, J. Kordosky, C. Lee, A. P. Lewis, T. Marschall, K. M. Munson, A. Pang, F. Yilmaz, Reconstruction of the human amylase locus reveals ancient duplications seeding modern-day variation. *Science* **386**, eadn0609 (2024).

5. C. Karageorgiou, O. Gokcumen, M. Y. Dennis, Deciphering the role of structural variation in human evolution: a functional perspective. *Curr. Opin. Genet. Dev.* **88**, 102240 (2024).

6. S. Schloissnig, S. Pani, J. Ebler, C. Hain, V. Tsapalou, A. Söylev, P. Hüther, H. Ashraf, T. Prodanov, M. Asparuhova, H. Magalhães, W. Höps, J. E. Sotelo-Fonseca, T. Fitzgerald, W. Santana-Garcia, R. Moreira-Pinhal, S. Hunt, F. J. Pérez-Llanos, T. E. Wollenweber, S. Sivalingam, D. Wieczorek, M. Cáceres, C. Gilissen, E. Birney, Z. Ding, J. N. Jensen, N. Podduturi, J. Stutzki, B. Rodriguez-Martin, T. Rausch, T. Marschall, J. O. Korbel, Structural variation in 1,019 diverse humans based on long-read sequencing. *Nature*, 1–11 (2025).

7. S. Ohno, *Evolution by Gene Duplication* (Springer, Berlin, Germany, 2014).

8. R. C. Moore, M. D. Purugganan, The early stages of duplicate gene evolution. *Proc. Natl. Acad. Sci. U. S. A.* **100**, 15682–15687 (2003).

9. X. Lan, J. K. Pritchard, Coregulation of tandem duplicate genes slows evolution of subfunctionalization in mammals. *Science* **352**, 1009–1013 (2016).

10. A. I. Teufel, M. M. Johnson, J. M. Laurent, A. H. Kachroo, E. M. Marcotte, C. O. Wilke, The many nuanced evolutionary consequences of duplicated genes. *Mol. Biol. Evol.* **36**, 304–314 (2019).

11. D. C. Soto, J. M. Uribe-Salazar, G. Kaya, R. Valdarrago, A. Sekar, N. K. Haghani, K. Hino, G. La, N. A. F. Mariano, C. Ingamells, A. Baraban, Z. Jamal, T. N. Turner, E. D. Green, S. Simó, G. Quon, A. M. Andrés, M. Y. Dennis, Human-specific gene expansions contribute to brain evolution. *Cell*, doi: 10.1016/j.cell.2025.06.037 (2025).

12. C. Boehlke, O. Zierau, C. Hannig, Salivary amylase - The enzyme of unspecialized euryphagous animals. *Arch. Oral Biol.* **60**, 1162–1176 (2015).

13. M. C. Janiak, Digestive enzymes of human and nonhuman primates: Digestive enzymes of human and nonhuman primates. *Evol. Anthropol.* **25**, 253–266 (2016).

14. P. Pajic, P. Pavlidis, K. Dean, L. Neznanova, R. Romano, D. Garneau, E. K. Daugherity, A. Globig, S. Ruhl, O. Gokcumen, Independent amylase gene copy number bursts correlate with dietary preferences in mammals. *Elife* **8**, e44628 (2019).

15. G. H. Perry, N. J. Dominy, K. G. Claw, A. S. Lee, H. Fiegler, R. Redon, J. Werner, F. A. Villanea, J. L. Mountain, R. Misra, N. P. Carter, C. Lee, A. C. Stone, Diet and the evolution of human amylase gene copy number variation. *Nat. Genet.* **39**, 1256–1260 (2007).

16. K. Scheer, L. J. B. Landau, K. Jorgensen, C. Karageorgiou, L. Siao, C. Alkan, A. M. Morales-Rivera, C. Osbourne, O. Garcia, L. Pearson, M. Kiyamu, F. Leon-Velarde, M. Rivera-Chira, F. Lee, T. Brutsaert, A. Bigham, O. Gokcumen, Adaptive increase of amylase gene copy number in Peruvians driven by potato-rich diets, *bioRxiv* (2025)p. 2025.03.25.644684.

17. U. Schibler, A. C. Pittet, R. A. Young, O. Hagenbüchle, M. Tosi, S. Gellman, P. K. Wellauer, The mouse alpha-amylase multigene family. Sequence organization of members expressed in the pancreas, salivary gland and liver. *J. Mol. Biol.* **155**, 247–266 (1982).

18. E. Axelsson, A. Ratnakumar, M. Arendt, K. Maqbool, M. Webster, M. Perloski, O. Liberg, J. Arnemo, Å. Hedhammar, K. Lindblad-Toh, The genomic signature of dog domestication reveals adaptation to a starch-rich diet. *Nature* **495**, 360–364 (2013).

19. V. Behringer, C. Borchers, T. Deschner, E. Möstl, D. Selzer, G. Hohmann, Measurements of salivary alpha amylase and salivary cortisol in hominoid primates reveal within-species consistency and between-species differences. *PLoS One* **8**, e60773 (2013).

20. R. L. McGeachin, J. R. Akin, Amylase levels in the tissues and body fluids of several primate species. *Comp. Biochem. Physiol. A Comp. Physiol.* **72**, 267–269 (1982).

21. S. Thamadilok, K.-S. Choi, L. Ruhl, F. Schulte, A. L. Kazim, M. Hardt, O. Gokcumen, S. Ruhl, Human and nonhuman primate lineage-specific footprints in the salivary proteome. *Mol. Biol. Evol.* **37**, 395–405 (2020).

22. L. C. Samuelson, R. S. Phillips, L. J. Swanberg, Amylase gene structures in primates: retroposon insertions and promoter evolution. *Mol. Biol. Evol.* **13**, 767–779 (1996).

23. L. C. Samuelson, K. Wiebauer, C. M. Snow, M. H. Meisler, Retroviral and Pseudogene Insertion Sites Reveal the Lineage of Human Salivary and Pancreatic Amylase Genes from a Single Gene during Primate Evolution. *Molecular and Cellular Biology* **10**, 2513–2520 (1990).

24. C. Ting, M. Rosenberg, C. Snow, L. Samuelson, M. Meisler, Endogenous retroviral sequences are required for tissue-specific expression of a human salivary amylase gene. *Genes Dev.* **6**, 1457–1465 (1992).

25. M. H. Meisler, C.-N. Ting, The Remarkable Evolutionary History of the Human Amylase Genes. *Critical Reviews in Oral Biology & Medicine* **4**, 503–509 (1993).

26. A. L. Mandel, C. Peyrot des Gachons, K. L. Plank, S. Alarcon, P. A. S. Breslin, Individual differences in AMY1 gene copy number, salivary α-amylase levels, and the perception of oral starch. *PLoS One* **5**, e13352 (2010).

27. S. J. Lindsay, M. Khajavi, J. R. Lupski, M. E. Hurles, A chromosomal rearrangement hotspot can be identified from population genetic variation and is coincident with a hotspot for allelic recombination. *Am. J. Hum. Genet.* **79**, 890–902 (2006).

28. G. H. Perry, J. Tchinda, S. D. McGrath, J. Zhang, S. R. Picker, A. M. Cáceres, A. J. Iafrate, C. Tyler-Smith, S. W. Scherer, E. E. Eichler, A. C. Stone, C. Lee, Hotspots for copy number variation in chimpanzees and humans. *Proc. Natl. Acad. Sci. U. S. A.* **103**, 8006–8011 (2006).

29. Y.-L. Lin, O. Gokcumen, Fine-scale characterization of genomic structural variation in the human genome reveals adaptive and biomedically relevant hotspots. *Genome Biol. Evol.* **11**, 1136–1151 (2019).

30. E. B. Cooper, L. Brent, N. Snyder‐Mackler, M. Singh, A. Sengupta, S. Khatiwada, S. Malaivijitnond, Z. Q. Hai, J. Higham, The natural history of model organisms: the rhesus macaque as a success story of the Anthropocene. *Elife* **11** (2022).

31. J. Rogers, M. Raveendran, R. A. Harris, T. Mailund, K. Leppälä, G. Athanasiadis, M. H. Schierup, J. Cheng, K. Munch, J. A. Walker, M. K. Konkel, V. Jordan, C. J. Steely, T. O. Beckstrom, C. Bergey, A. Burrell, D. Schrempf, A. Noll, M. Kothe, G. H. Kopp, Y. Liu, S. Murali, K. Billis, F. J. Martin, M. Muffato, L. Cox, J. Else, T. Disotell, D. M. Muzny, J. Phillips-Conroy, B. Aken, E. E. Eichler, T. Marques-Bonet, C. Kosiol, M. A. Batzer, M. W. Hahn, J. Tung, D. Zinner, C. Roos, C. J. Jolly, R. A. Gibbs, K. C. Worley, Baboon Genome Analysis Consortium, The comparative genomics and complex population history of Papio baboons. *Sci. Adv.* **5**, eaau6947 (2019).

32. C. Roos, S. Knauf, I. S. Chuma, A. Maille, C. Callou, R. Sabin, R. Portela Miguez, D. Zinner, New mitogenomic lineages in Papio baboons and their phylogeographic implications. *Am. J. Phys. Anthropol.* **174**, 407–417 (2021).

33. C. F. Aquadro, A. L. Weaver, S. W. Schaeffer, W. W. Anderson, Molecular evolution of inversions in Drosophila pseudoobscura: the amylase gene region. *Proc. Natl. Acad. Sci. U. S. A.* **88**, 305–309 (1991).

34. F. Staubach, A. Lorenc, P. W. Messer, K. Tang, D. A. Petrov, D. Tautz, Genome patterns of selection and introgression of haplotypes in natural populations of the house mouse (Mus musculus). *PLoS Genet.* **8**, e1002891 (2012).

35. M. E. Johnson, National Institute of Health Intramural Sequencing Center Comparative Sequencing Program, Z. Cheng, V. A. Morrison, S. Scherer, M. Ventura, R. A. Gibbs, E. D. Green, E. E. Eichler, Recurrent duplication-driven transposition of DNA during hominoid evolution. *Proc. Natl. Acad. Sci. U. S. A.* **103**, 17626–17631 (2006).

36. P. Balachandran, I. A. Walawalkar, J. I. Flores, J. N. Dayton, P. A. Audano, C. R. Beck, Transposable element-mediated rearrangements are prevalent in human genomes. *Nat. Commun.* **13**, 7115 (2022).

37. N. Kamitaki, R. E. Handsaker, M. L. A. Hujoel, R. E. Mukamel, C. L. Usher, S. A. McCarroll, P.-R. Loh, Human and bacterial genetic variation shape oral microbiomes and health, *medRxiv* (2025)p. 2025.03.31.25324952.

38. S. Takuno, T. Nishio, Y. Satta, H. Innan, Preservation of a pseudogene by gene conversion and diversifying selection. *Genetics* **180**, 517–531 (2008).

39. H. Velová, M. W. Gutowska-Ding, D. W. Burt, M. Vinkler, Toll-like receptor evolution in birds: Gene duplication, pseudogenization, and diversifying selection. *Mol. Biol. Evol.* **35**, 2170–2184 (2018).

40. J. Wang, F. Chitsaz, M. K. Derbyshire, N. R. Gonzales, M. Gwadz, S. Lu, G. H. Marchler, J. S. Song, N. Thanki, R. A. Yamashita, M. Yang, D. Zhang, C. Zheng, C. J. Lanczycki, A. Marchler-Bauer, The conserved domain database in 2023. *Nucleic Acids Res.* **51**, D384–D388 (2023).

41. M. Saitou, E. A. Gaylord, E. Xu, A. J. May, L. Neznanova, S. Nathan, A. Grawe, J. Chang, W. Ryan, S. Ruhl, S. M. Knox, O. Gokcumen, Functional specialization of human salivary glands and origins of proteins intrinsic to human saliva. *Cell Rep.* **33**, 108402 (2020).

42. N. L. Bray, H. Pimentel, P. Melsted, L. Pachter, Near-optimal probabilistic RNA-seq quantification. *Nat. Biotechnol.* **34**, 525–527 (2016).

43. D. G. Michael, T. J. F. Pranzatelli, B. M. Warner, H. Yin, J. A. Chiorini, Integrated epigenetic mapping of human and mouse salivary gene regulation. *J. Dent. Res.* **98**, 209–217 (2019).

44. C. P. Ponting, Biological function in the twilight zone of sequence conservation. *BMC Biol.* **15**, 71 (2017).

45. S. F. Altschul, W. Gish, W. Miller, E. W. Myers, D. J. Lipman, Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).

46. P. Danecek, J. K. Bonfield, J. Liddle, J. Marshall, V. Ohan, M. O. Pollard, A. Whitwham, T. Keane, S. A. McCarthy, R. M. Davies, H. Li, Twelve years of SAMtools and BCFtools. *Gigascience* **10**, giab008 (2021).

47. S. Kurtz, A. Phillippy, A. L. Delcher, M. Smoot, M. Shumway, C. Antonescu, S. L. Salzberg, Versatile and open software for comparing large genomes. *Genome Biol.* **5**, R12 (2004).

48. S. M. Kiełbasa, R. Wan, K. Sato, P. Horton, M. C. Frith, Adaptive seeds tame genomic sequence comparison. *Genome Res.* **21**, 487–493 (2011).

49. D. Porubsky, X. Guitart, D. Yoo, P. C. Dishuck, W. T. Harvey, E. E. Eichler, SVbyEye: a visual tool to characterize structural variation among whole-genome assemblies. *Bioinformatics* **41**, btaf332 (2025).

50. G. S. C. Slater, E. Birney, Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics* **6**, 31 (2005).

51. H. Išerić, C. Alkan, F. Hach, I. Numanagić, Fast characterization of segmental duplication structure in multiple genome assemblies. *Algorithms Mol. Biol.* **17**, 4 (2022).

52. S. W. Wingett, S. Andrews, FastQ Screen: A tool for multi-genome mapping and quality control. *F1000Res.* **7**, 1338 (2018).

53. P. Ewels, M. Magnusson, S. Lundin, M. Käller, MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics* **32**, 3047–3048 (2016).

54. M. Martin, Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet J.* **17**, 10 (2011).

55. M. I. Love, W. Huber, S. Anders, Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).

56. D. Kim, J. M. Paggi, C. Park, C. Bennett, S. L. Salzberg, Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat. Biotechnol.* **37**, 907–915 (2019).

57. H. Li, R. Durbin, Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).

58. K. Katoh, D. M. Standley, MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–780 (2013).

59. M. Suyama, D. Torrents, P. Bork, PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic Acids Res.* **34**, W609–12 (2006).

60. L.-T. Nguyen, H. A. Schmidt, A. von Haeseler, B. Q. Minh, IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.* **32**, 268–274 (2015).

61. M. D. Smith, J. O. Wertheim, S. Weaver, B. Murrell, K. Scheffler, S. L. Kosakovsky Pond, Less is more: an adaptive branch-site random effects model for efficient detection of episodic diversifying selection. *Mol. Biol. Evol.* **32**, 1342–1353 (2015).

62. B. Murrell, J. O. Wertheim, S. Moola, T. Weighill, K. Scheffler, S. L. Kosakovsky Pond, Detecting individual sites subject to episodic diversifying selection. *PLoS Genet.* **8**, e1002764 (2012).

63. B. Murrell, S. Moola, A. Mabona, T. Weighill, D. Sheward, S. L. Kosakovsky Pond, K. Scheffler, FUBAR: a fast, unconstrained bayesian approximation for inferring selection. *Mol. Biol. Evol.* **30**, 1196–1205 (2013).

64. J. O. Wertheim, B. Murrell, M. D. Smith, S. L. Kosakovsky Pond, K. Scheffler, RELAX: detecting relaxed selection in a phylogenetic framework. *Mol. Biol. Evol.* **32**, 820–832 (2015).

65. J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko, A. Bridgland, C. Meyer, S. A. A. Kohl, A. J. Ballard, A. Cowie, B. Romera-Paredes, S. Nikolov, R. Jain, J. Adler, T. Back, S. Petersen, D. Reiman, E. Clancy, M. Zielinski, M. Steinegger, M. Pacholska, T. Berghammer, S. Bodenstein, D. Silver, O. Vinyals, A. W. Senior, K. Kavukcuoglu, P. Kohli, D. Hassabis, Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).

66. N. Ramasubbu, V. Paloth, Y. Luo, G. D. Brayer, M. J. Levine, Structure of human salivary α-amylase at 1.6 Å resolution: Implications for its role in the oral cavity. *Acta Crystallogr. D Biol. Crystallogr.* **52**, 435–446 (1996).

67. N. Ramasubbu, C. Ragunath, P. J. Mishra, Probing the role of a mobile loop in substrate binding and enzyme activity of human salivary amylase. *J. Mol. Biol.* **325**, 1061–1076 (2003).

68. R. Gupta, S. Brunak, Prediction of glycosylation across the human proteome and the correlation to protein function. *Pac. Symp. Biocomput.*, 310–322 (2002).

69. R. Dreos, G. Ambrosini, R. C. Périer, P. Bucher, The Eukaryotic Promoter Database: expansion of EPDnew and new promoter analysis tools. *Nucleic Acids Res.* **43**, D92–6 (2015).

70. T. L. Bailey, C. Elkan, Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* **2**, 28–36 (1994).

71. S. Gupta, J. A. Stamatoyannopoulos, T. L. Bailey, W. S. Noble, Quantifying similarity between motifs. *Genome Biol.* **8**, R24 (2007).

72. I. Rauluseviciute, R. Riudavets-Puig, R. Blanc-Mathieu, J. A. Castro-Mondragon, K. Ferenc, V. Kumar, R. B. Lemma, J. Lucas, J. Chèneby, D. Baranasic, A. Khan, O. Fornes, S. Gundersen, M. Johansen, E. Hovig, B. Lenhard, A. Sandelin, W. W. Wasserman, F. Parcy, A. Mathelier, JASPAR 2024: 20th anniversary of the open-access database of transcription factor binding profiles. *Nucleic Acids Res.* **52**, D174–D182 (2024).

73. S. Noguchi, T. Arakawa, S. Fukuda, M. Furuno, A. Hasegawa, F. Hori, S. Ishikawa-Kato, K. Kaida, A. Kaiho, M. Kanamori-Katayama, T. Kawashima, M. Kojima, A. Kubosaki, R.-I. Manabe, M. Murata, S. Nagao-Sato, K. Nakazato, N. Ninomiya, H. Nishiyori-Sueki, S. Noma, E. Saijyo, A. Saka, M. Sakai, C. Simon, N. Suzuki, M. Tagami, S. Watanabe, S. Yoshida, P. Arner, R. A. Axton, M. Babina, J. K. Baillie, T. C. Barnett, A. G. Beckhouse, A. Blumenthal, B. Bodega, A. Bonetti, J. Briggs, F. Brombacher, A. J. Carlisle, H. C. Clevers, C. A. Davis, M. Detmar, T. Dohi, A. S. B. Edge,

M. Edinger, A. Ehrlund, K. Ekwall, M. Endoh, H. Enomoto, A. Eslami, M. Fagiolini, L. Fairbairn, M. C. Farach-Carson, G. J. Faulkner, C. Ferrai, M. E. Fisher, L. M. Forrester, R. Fujita, J.-I. Furusawa, T. B. Geijtenbeek, T. Gingeras, D. Goldowitz, S. Guhl, R. Guler, S. Gustincich, T. J. Ha, M. Hamaguchi, M. Hara, Y. Hasegawa, M. Herlyn, P. Heutink, K. J. Hitchens, D. A. Hume, T. Ikawa, Y. Ishizu, C. Kai, H. Kawamoto, Y. I. Kawamura, J. S. Kempfle, T. J. Kenna, J. Kere, L. M. Khachigian, T. Kitamura, S. Klein, S. P. Klinken, A. J. Knox, S. Kojima, H. Koseki, S. Koyasu, W. Lee, A. Lennartsson, A. Mackay-Sim, N. Mejhert, Y. Mizuno, H. Morikawa, M. Morimoto, K. Moro, K. J. Morris, H. Motohashi, C. L. Mummery, Y. Nakachi, F. Nakahara, T. Nakamura, Y. Nakamura, T. Nozaki, S. Ogishima, N. Ohkura, H. Ohno, M. Ohshima, M. Okada-Hatakeyama, Y. Okazaki, V. Orlando, D. A. Ovchinnikov, R. Passier, M. Patrikakis, A. Pombo, S. Pradhan-Bhatt, X.-Y. Qin, M. Rehli, P. Rizzu, S. Roy, A. Sajantila, S. Sakaguchi, H. Sato, H. Satoh, S. Savvi, A. Saxena, C. Schmidl, C. Schneider, G. G. Schulze-Tanzil, A. Schwegmann, G. Sheng, J. W. Shin, D. Sugiyama, T. Sugiyama, K. M. Summers, N. Takahashi, J. Takai, H. Tanaka, H. Tatsukawa, A. Tomoiu, H. Toyoda, M. van de Wetering, L. M. van den Berg, R. Verardo, D. Vijayan, C. A. Wells, L. N. Winteringham, E. Wolvetang, Y. Yamaguchi, M. Yamamoto, C. Yanagi-Mizuochi, M. Yoneda, Y. Yonekura, P. G. Zhang, S. Zucchelli, I. Abugessaisa, E. Arner, J. Harshbarger, A. Kondo, T. Lassmann, M. Lizio, S. Sahin, T. Sengstag, J. Severin, H. Shimoji, M. Suzuki, H. Suzuki, J. Kawai, N. Kondo, M. Itoh, C. O. Daub, T. Kasukawa, H. Kawaji, P. Carninci, A. R. R. Forrest, Y. Hayashizaki, FANTOM5 CAGE profiles of human and mouse samples. *Sci. Data* **4**, 170112 (2017).

74. C. E. Grant, T. L. Bailey, W. S. Noble, FIMO: scanning for occurrences of a given motif. *Bioinformatics* **27**, 1017–1018 (2011).

Acknowledgments

We thank Leo Speidel, Petar Pajic and Luane Landau for carefully reading this manuscript.

Funding:

O.G. was supported by the National Science Foundation (NSF) Awards 2049947 and 2123284, and National Institute of General Medical Sciences (NIH) Award R35-GM156519. S.R. was supported by the National Science Foundation (NSF) Award 2049947.

Competing interests:

All other authors declare they have no competing interests.

Data and materials availability:

All data supporting this study are available in the main text and Supplementary Materials. RNA-seq datasets for olive baboon (*Papio anubis*) and rhesus macaque (*Macaca mulatta*) are available in GEO under accessions GSE305241 and GSE305255, respectively. Analysis scripts and

input files for downstream analyses and visualization are deposited on Zenodo (DOI: https://doi.org/10.5281/zenodo.16809248).

Figures and Tables

Fig. 1. Structural evolutionary history of the amylase locus across primates. (A) Contractions and expansions in the amylase locus reconstructed from 69 high-quality genomes representing 53 primate species **(Table S3 and S4)**. Lineages are color-coded by clade: purple for tarsiers (outgroup), yellow for leaf-eating monkeys, green for Old World monkeys, dark blue for great apes, red for lesser apes (gibbons), cyan for New World monkeys, and orange for lemurs. Red dots indicate independent contractions in the amylase locus. Yellow dots indicate expansions. Cyan dots mark lineages retaining the ancestral single-copy configuration, inferred as the ancestral primate state. Independent contraction events are observed in the gibbon, leaf-eating monkey, and New World monkey genera. Numbers inside the dots give the total number of AMY gene copies detected in each species (per haploid genome). (B) Synteny comparison between bonobo and chimpanzee at the amylase locus. The copy number increase in bonobo, previously reported (*15*), is shown here to involve a chimeric duplication. The 5′ flanking region of the duplicated segment resembles the downstream region of *AMY1*. The internal genic region corresponds to a full duplication of the *AMY1* coding sequence. The 3′ flanking region aligns with the upstream flanking region of *AMY2A*. This chimeric architecture is consistent with a nonallelic homologous recombination mechanism. (C) Comparison of the amylase locus in ring-tailed lemur and gray mouse lemur. The gray mouse lemur harbors a tandem duplication of the amylase gene, including both upstream and downstream flanking regions. This is the only duplication identified among the lemur species analyzed.

Fig. 2. Evolutionary origins of the amylase locus in primates. Model of amylase locus evolution across primates adjusted from Samuelson et al. 1990 (*23*). The ancestral primate genome contained a single-copy amylase gene (*AMY2B*). In the Catarrhini lineage, a γ-actin insertion occurred upstream of *AMY2B* (~43 MYA), followed by duplication of the γ-actin-*AMY2B* segment, generating a second gene (*AMY1′*). In great apes, a retroviral (ERV) insertion occurred (~39 MYA), followed by retroviral ectopic deletion in the human lineage, giving rise to *AMY2A*. In humans, a stepwise triplication of the salivary-expressed amylase gene (*AMY1*) generated *AMY1A*, *AMY1B* and *AMY1C*. The right panel summarizes the inferred regulatory and structural events, showing independent gains of salivary expression in New World monkeys, Old World monkeys and great apes, with subfunctionalization of *AMY1* and *AMY2A* in humans.

Fig. 3. Independent duplication events and structural evolution of the amylase locus in Old World monkeys. (A) Schematic reconstruction of the amylase locus in the Catarrhini ancestor, rhesus macaque,

olive baboon, and human reference genome (which is also shown to be the ancestral human arrangement (*4*) (hg38). The Catarrhini ancestor contains two genes: *AMY2B* and *AMY1'*, the latter derived from a duplication of *AMY2B*. In rhesus macaques, the locus includes *AMY2B*, *AMYm* , and *AMY1'* . In olive baboons, the locus consists of *AMY2B* , *AMYp2* and *AMYp1* , and *AMY1'*. *AMYp1* is annotated as a pseudogene in *NCBI*, and is indicated with a dashed outline. *AMYp1* and *AMYp2* arose from independent duplication events with distinct breakpoints from those of *AMYm*. In the human reference genome, the locus contains *AMY2B* , *AMY2A* and three *AMY1* paralogs, *AMY1A*, *AMY1B*, and *AMY1C*. The *AMY2B* genes are one-to-one orthologs across species. *AMY1'* in macaques, baboons and the Catarrhini ancestor is orthologous among Old World monkeys and represents the ancestral precursor to human *AMY2A* and *AMY1* genes. (B) Synteny analysis and phylogenetic context of the rhesus macaque-specific duplication. *AMYm* (orange) is located between *AMY2B* (purple) and *AMY1'* (pink), and is present only in the *fascicularis* and *sinica* macaque groups, but absent in the *silenus* group. This distribution dates the duplication to approximately 4.5-5 million years ago. The gene structure of *AMYm* is identical for coding sequence to *AMY2B*, but differs on the 5' untranslated region. Breakpoint analysis indicates a nonallelic homologous recombination (NAHR) mechanism. (C) Synteny comparison across *Papionini* species reveals two independent NAHR-mediated duplication events in olive baboons. The first duplication, giving rise to *AMYp1* (orange), is shared with Guinea baboon (*Papio papio)* and dates to approximately 2-4.5 million years ago. The second duplication, resulting in *AMYp2* (orange), is specific to olive baboons and dates to ~1.85 million years ago. Because of incomplete assembly in Guinea baboon (*Papio papio)*, we carried out the synteny analysis using closely related *Theropithecus gelada* and *Mandrillus leucophaeus* which are members of the Papionini group*,* supporting this reconstruction.

Fig. 4. The transposable element landscape in the primate amylase locus. (A) Transposable element (TE) content (given as the total TE in bp/the total length of the locus in bp, see **Methods** for more) in the amylase locus relative to genome-wide TE proportion across 53 primate species. Each point represents a species, with the size of the circle scaled by the respective amylase copy number and the color indicating enrichment status. The dashed line marks the 1:1 expectation. The majority of species show depletion of TEs in the amylase locus compared to genome-wide levels. (B) TE family-specific enrichment ($\log_2$ transformed) in the amylase locus across primates for SINEs, Alus, MIRs and LINE1s, alongside total TE content ("All TEs") (**Table S5**). Bars represent enrichment (green) or depletion (pink) relative to genome-wide TE representation. Short retrotransposons (e.g. Alus & SINEs) are consistently depleted from the amylase locus, suggesting that this region might have limited retention of active mobile elements across lineages. (C) Relationship between genome-wide and amylase locus-specific LTR content across primate species. Each point represents a species, with the circle size indicating *AMY* gene copy number

and the color representing phylogenetic clade. The dashed line indicates a 1:1 ratio. Most species show an enrichment of LTRs at the amylase locus relative to genome-wide levels. (D) Log_2-transformed enrichment and depletion of LTR content at the amylase locus across 53 primate species, organized by phylogeny. Bars indicate whether LTR representation in the locus is higher (green) or lower (pink) than genome-wide LTR proportions. Several species with reported amylase gene duplications, such as olive baboon, exhibit pronounced LTR enrichment.

Fig 5. Functional divergence and tissue-specific expression of amylase paralogs in olive baboons, rhesus macaques and humans. (A) Identification of a premature stop codon in *AMY2B* of olive baboon. A schematic of the disrupted gene structure (top) shows the location of the nonsense mutation (red box) within the ninth exon. (B) A maximum likelihood phylogeny of amylase coding sequences from olive baboon, rhesus macaque and great apes, highlighted branches under significant episodic diversifying selection (dark red) affecting the lineage leading to Old World monkey paralogs (aBSREL, $p = 0.038$). (C) Old world monkey and baboon-specific variants and their overlap with predicted functional sites across the amylase protein. Tracks show the position of catalytic and calcium-binding residues, glycosylation sites, positively selected codons (FUBAR, MEME), Old World monkey-specific mutations and baboon-specific mutations in *AMYp2*. A MEME- and FUBAR-identified site (codon position 178) involves a threonine-to-serine substitution overlaps an active site. (D) Differential expression of amylase paralogs across tissues in olive baboons, rhesus macaques and humans. The scatterplot shows log_2-fold change for each paralog across tissue comparisons (pancreas vs parotid, sublingual or submandibular glands and pancreas vs. liver). The size of the circle corresponds to the base mean expression level. In baboons and macaques, lineage-specific genes (*AMYp2*, *AMYm*) contribute to salivary expression, unlike in humans where *AMY1* is the dominant/primary salivary paralog. Arrows represent the structure and orientation of amylase paralogs across species, colored by paralog identity: *AMY2B* (purple), *AMY2A* (light orange), *AMY1* (light pink), *AMY1'* (pink), *AMYm*, *AMYp1* and *AMYp2* (orange). (E) Pie charts showing proportional expression of each amylase paralog in salivary glands (parotid, submandibular, sublingual) and pancreas. Expression levels of *AMY1'* in both salivary glands and pancreas in macaques and baboons supports subfunctionalization as the likely evolutionary mechanism for the tissue-specific roles of *AMY1* and *AMY2A* in humans.

Fig 6. Predicted transcription factor binding sites (TFBS) associated with salivary-gland and pancreatic expression across amylase gene paralogs. Circos plot showing *in silico* predictions of salivary (pink) and pancreatic (purple) TFBS across promoter regions (~170 bp upstream of the start codon and annotated promoter) of amylase paralogs in rhesus macaques, olive baboons and humans. Colored arrows represent gene orientation. Dots adjacent to each gene denote the presence of TFBS (pink for salivary, purple for

pancreatic). Ribbon links connect paralogs to TFBS categories. Pie charts summarizing relative gene expression across tissues: PAR (parotid), SM (submandibular), SL (sublingual), and PANC (pancreas). The color of each slice corresponds to the proportion of total expression of the given gene in each tissue, following the same color scheme used for the gene annotation arrows. All three rhesus macaque paralogs (*AMY2B*, *AMYm*, *AMY1'*) share similar TFBS profiles with olive baboons *AMY1'* and *AMYp2*. In contrast, *AMY2B* in olive baboon and human exhibits distinct TFBS composition. The salivary gland-biased transcription factor *FOXC1* is present in most paralogs, but absent from human and olive baboon *AMY2B*, both of which lack salivary expression.

Supplementary Materials

Figures S1 to S11

Tables S1 to S14

Supplementary Materials for

Convergent evolution through independent rearrangements in the primate amylase locus

Charikleia Karageorgiou *et al.*

*Corresponding author. Email: Omer Gokcumen, omergokc@buffalo.edu

This PDF file includes:

Figs. S1 to S11
Tables S1 to S14 (Table titles only)

Other Supplementary Materials for this manuscript include the following:
Tables S1 to S14 (full data) provided as a separate Excel file

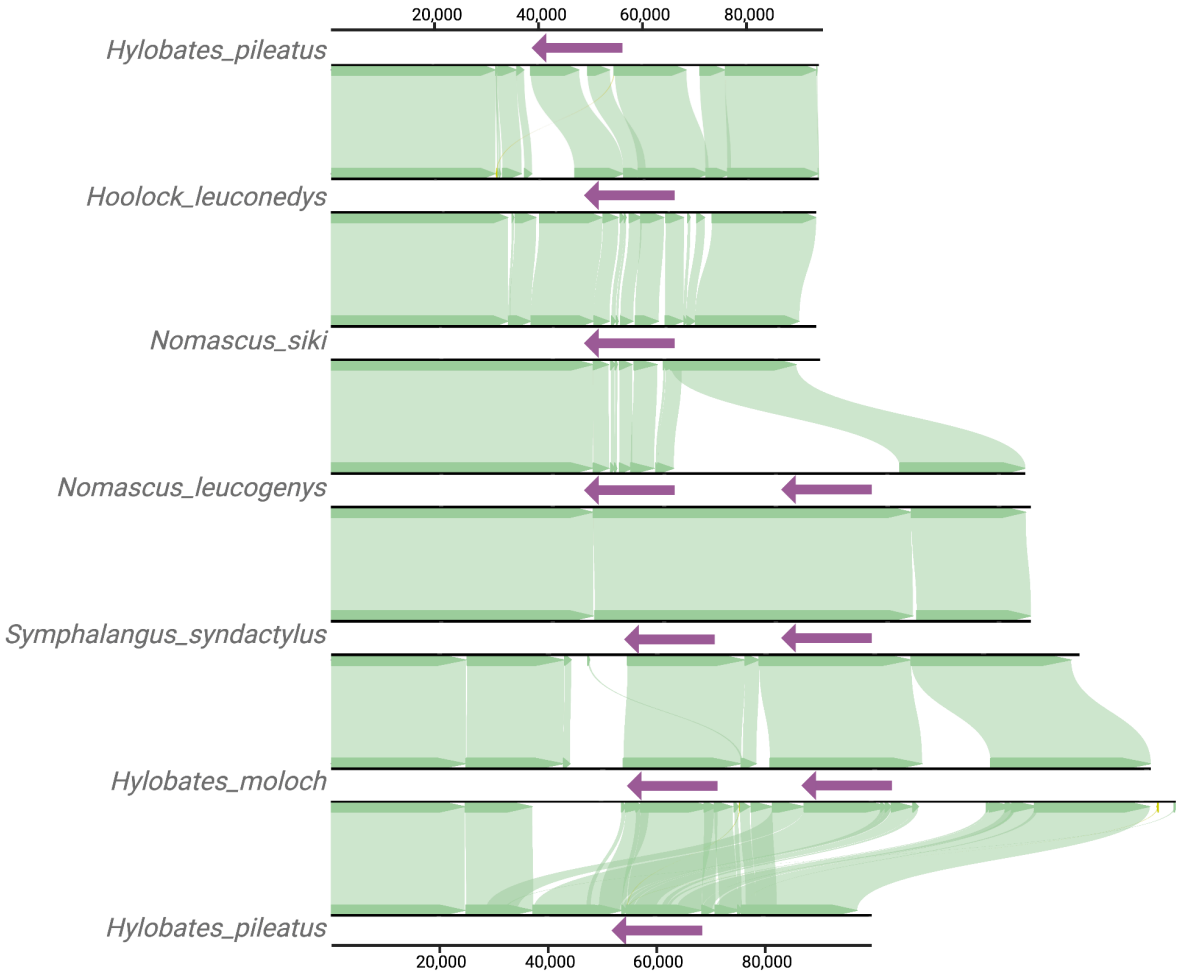

Fig. S1.

Copy-number and structural diversity of the amylase locus in gibbons (lesser apes). Miropeat style plots across representatives of *Hoolock leucodenys*, *Symphalangus syndactylus*, *Hylobates pileatus, Hylobates moloch* and *Nomascus leucogenys* and *Nomascus siki* illustrate lineage-specific differences consistent with either independent losses or incomplete lineage sorting (see main text). Axes denote locus coordinates.

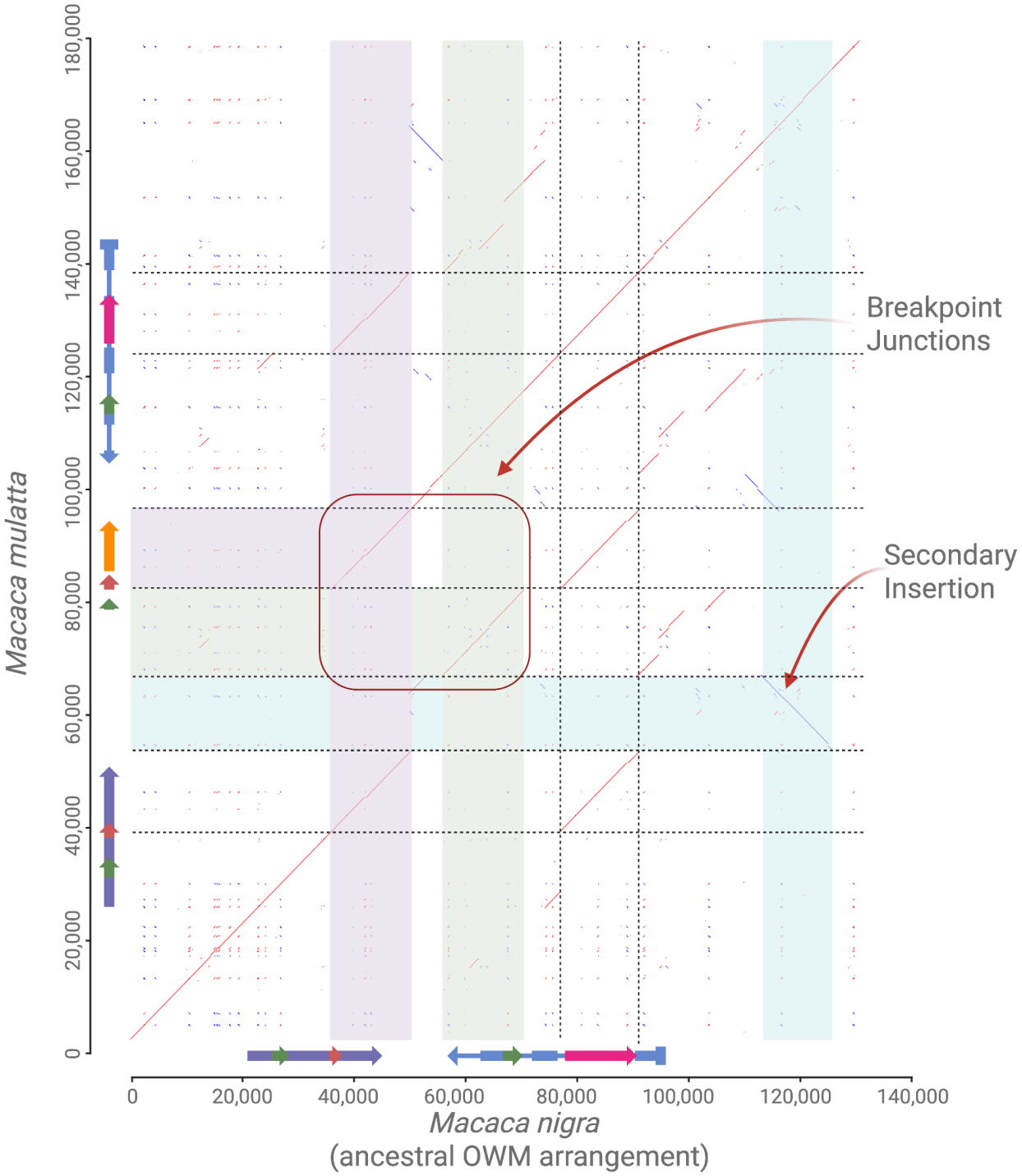

Fig. S2.

Synteny of the macaque amylase locus across *Macaca* and mapping of the *AMYm* duplication. Comparative dotplots use *Macaca nigra* as the two-copy ancestral Old World monkey configuration and highlight the novel *AMYm* block in the fascicularis and sinica clades (represented by *M. mulatta*), with breakpoints marked at the NAHR junction and a small secondary insertion within the novel duplicon.

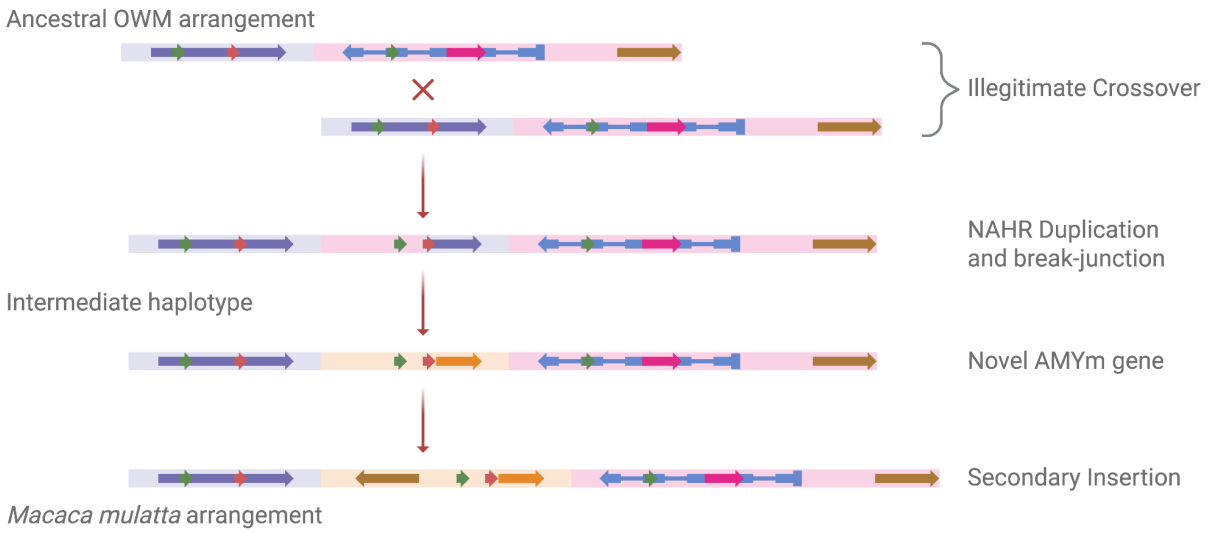

Fig. S3.

Mechanistic model for the macaque *AMYm* origin. Cartoon summarizing an initial illegitimate crossover (NAHR) between *AMY1′* (5′ homology) and *AMY2B* (3′ homology), generating the chimeric *AMYm* (intermediate haplotype). Following the NAHR duplication the locus experienced a small secondary insertion upstream the *AMYm*. Panels show the ancestral OWM configuration, *Macaca mulatta* arrangement, and the inferred intermediate arrangement prior to the secondary insertion.

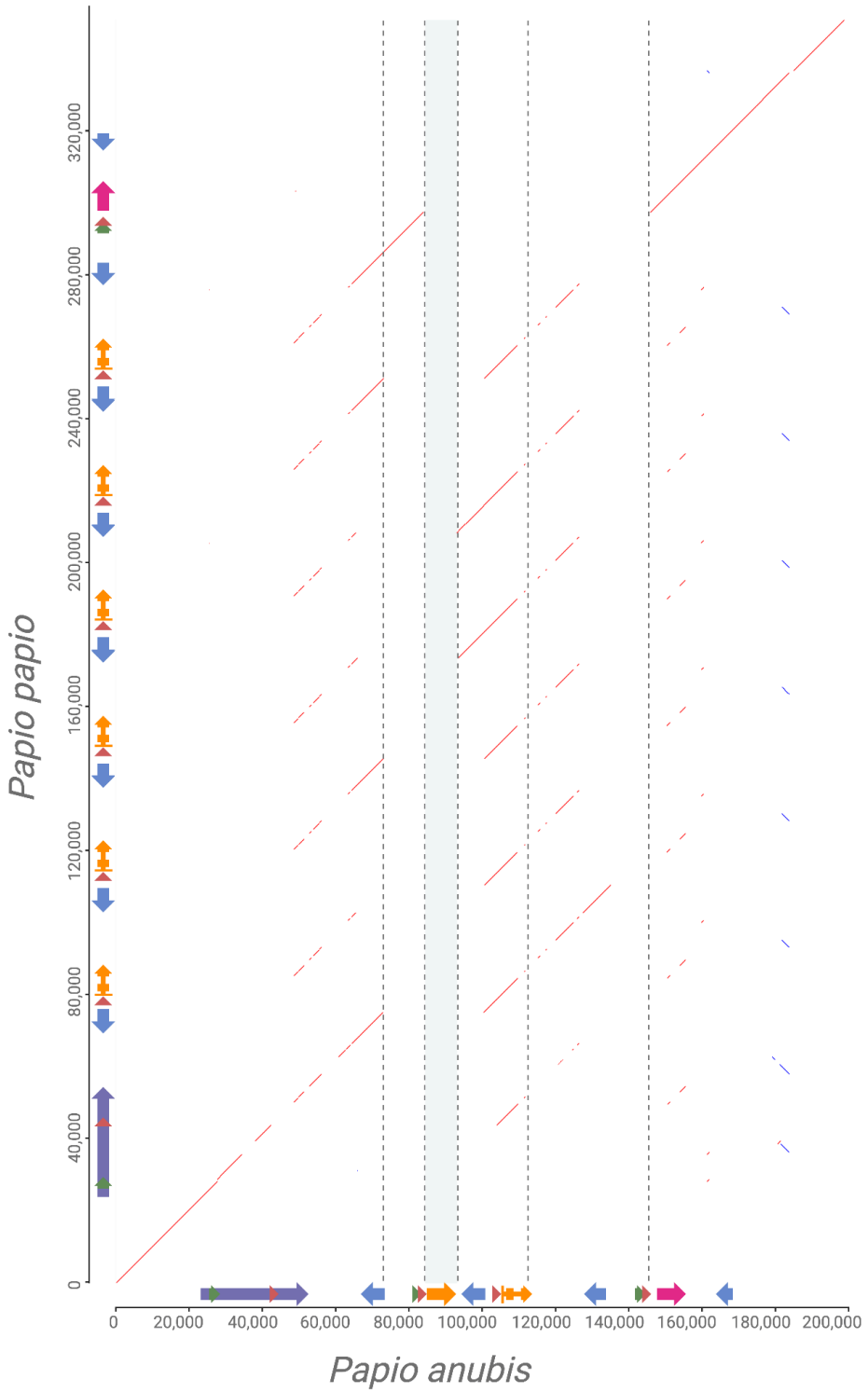

Fig. S4.

Dotplot of the amylase locus between *Papio anubis* (olive baboon; x-axis) and *Papio papio* (Guinea baboon; y-axis). *P. papio* lacks the *AMYp2* segment present in *P. anubis*, confirming *AMYp1* as the first and older duplication shared by both species. In *P. papio*, multiple tandem

tracts align identically to *AMYp1* (6 copies), a pattern consistent with assembly over-expansion rather than a true biological increase (see Methods). Alignments were generated with LAST (using lastal) and filtered with last-split to retain only primary, best-scoring placements and the dotplot was rendered from these primary chains. Breakpoint-proximal, single-copy flanks were used as anchors, increasing confidence that the observed absence of *AMYp2* in *P. papio* reflects genuine synteny rather than paralogous cross-mapping.

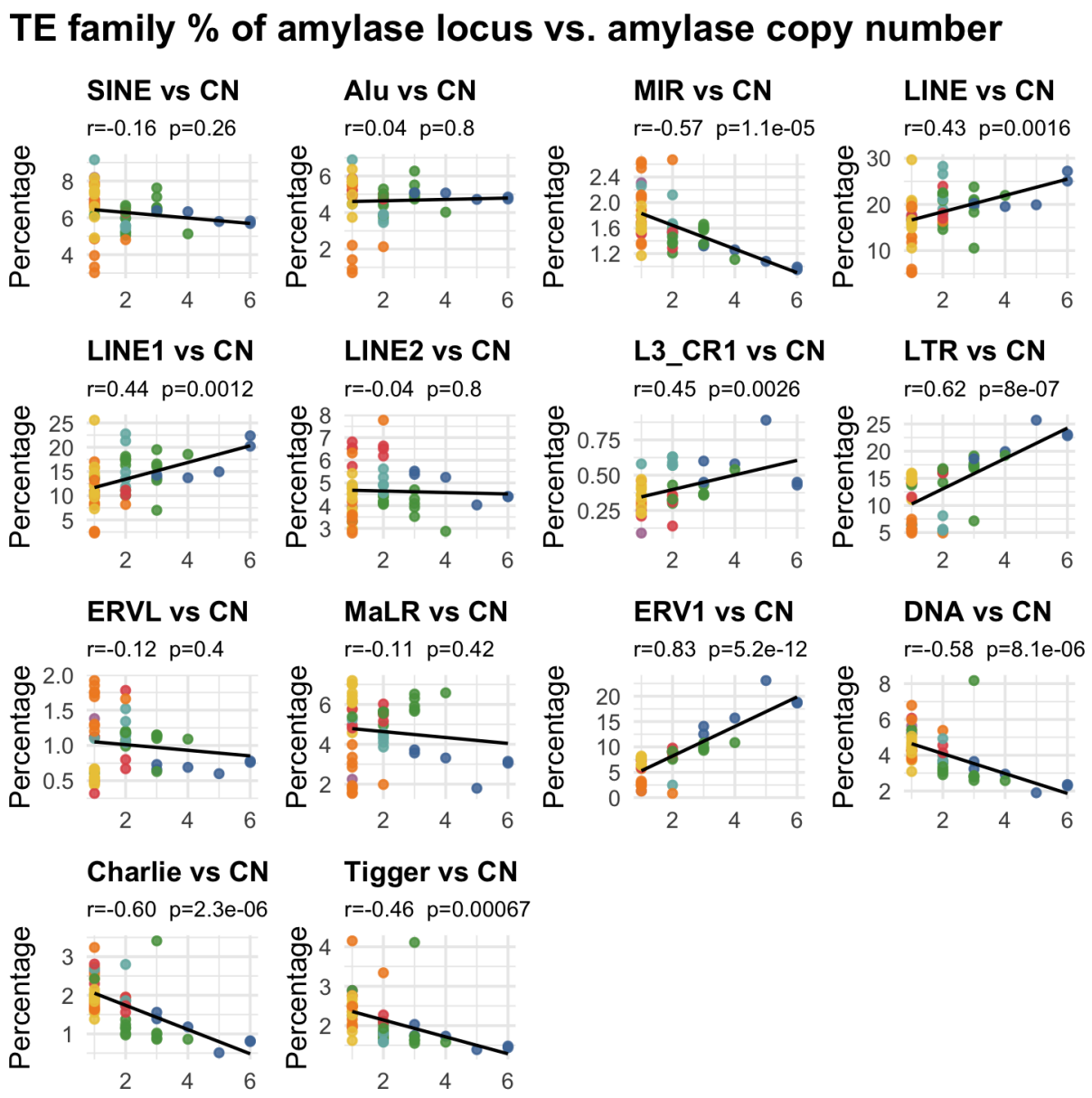

Fig. S5.

TE family abundance in the amylase locus versus AMY copy number across primates. Scatterplots show, for 53 species, the percentage of each TE family within the RNPC3-AMY1C amylase interval (y-axis) against the total AMY copy number per haploid genome (x-axis). Points are species (colored by clade as in Fig. 1); lines are least-squares fits. Panels report Pearson's r and two-sided P test. LTRs, LINEs (LINE1, L3/CR1) and ERV1 show positive correlations with AMY copy number, whereas MIRs and DNA transposons (Charlie, Tigger) are negatively correlated. TE content was called with RepeatMasker using a uniform primate library on locus-bounded sequences.

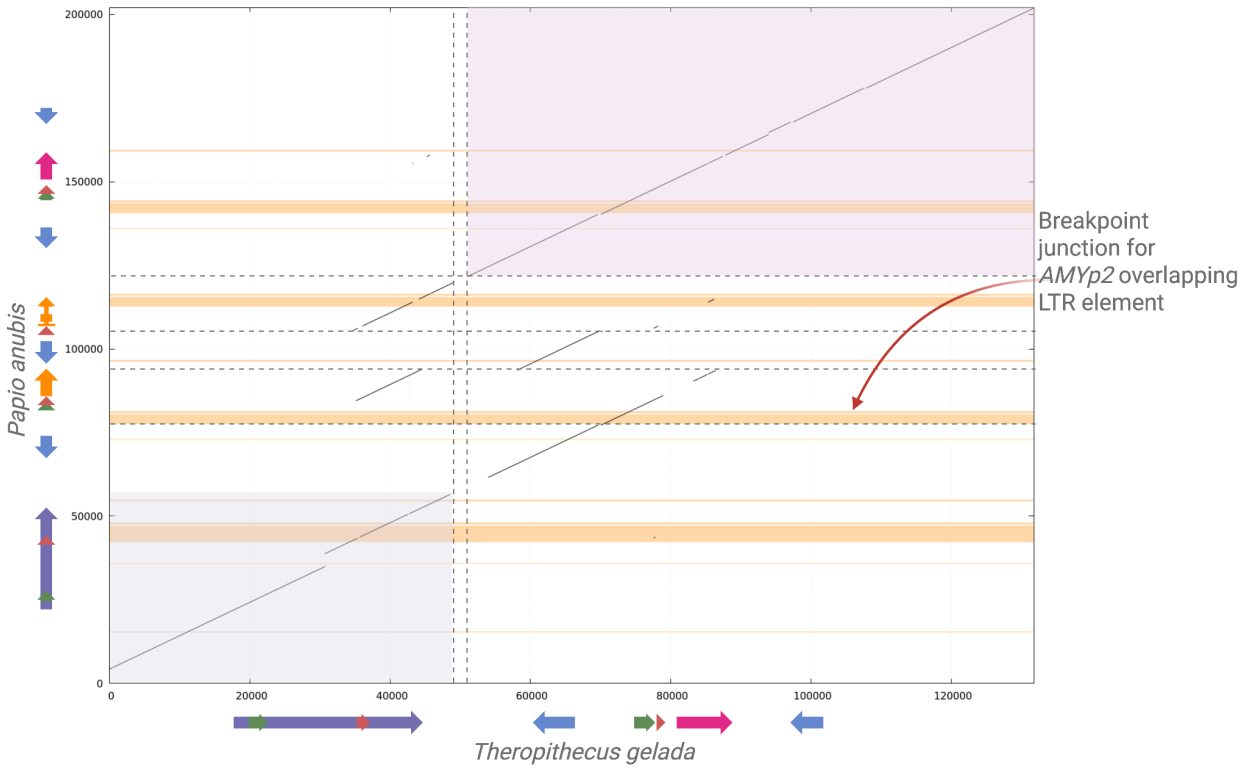

Fig. S6.

Dotplot of the olive baboon (*Papio anubis*) amylase locus (y-axis) versus the inferred ancestral Old World monkey two-copy locus from *Theropithecus gelada* (x-axis), aligned with LAST (primary alignments only). Light purple background blocks mark the two ancestral OWM segments. Horizontal orange bands are LTR annotations in olive baboons. The dashed horizontal lines mark the inferred crossover positions. The labeled *AMYp2* breakpoint junction overlaps an LTR tract (LTR25-int), consistent with LTR-seeded homology at the recombination site.

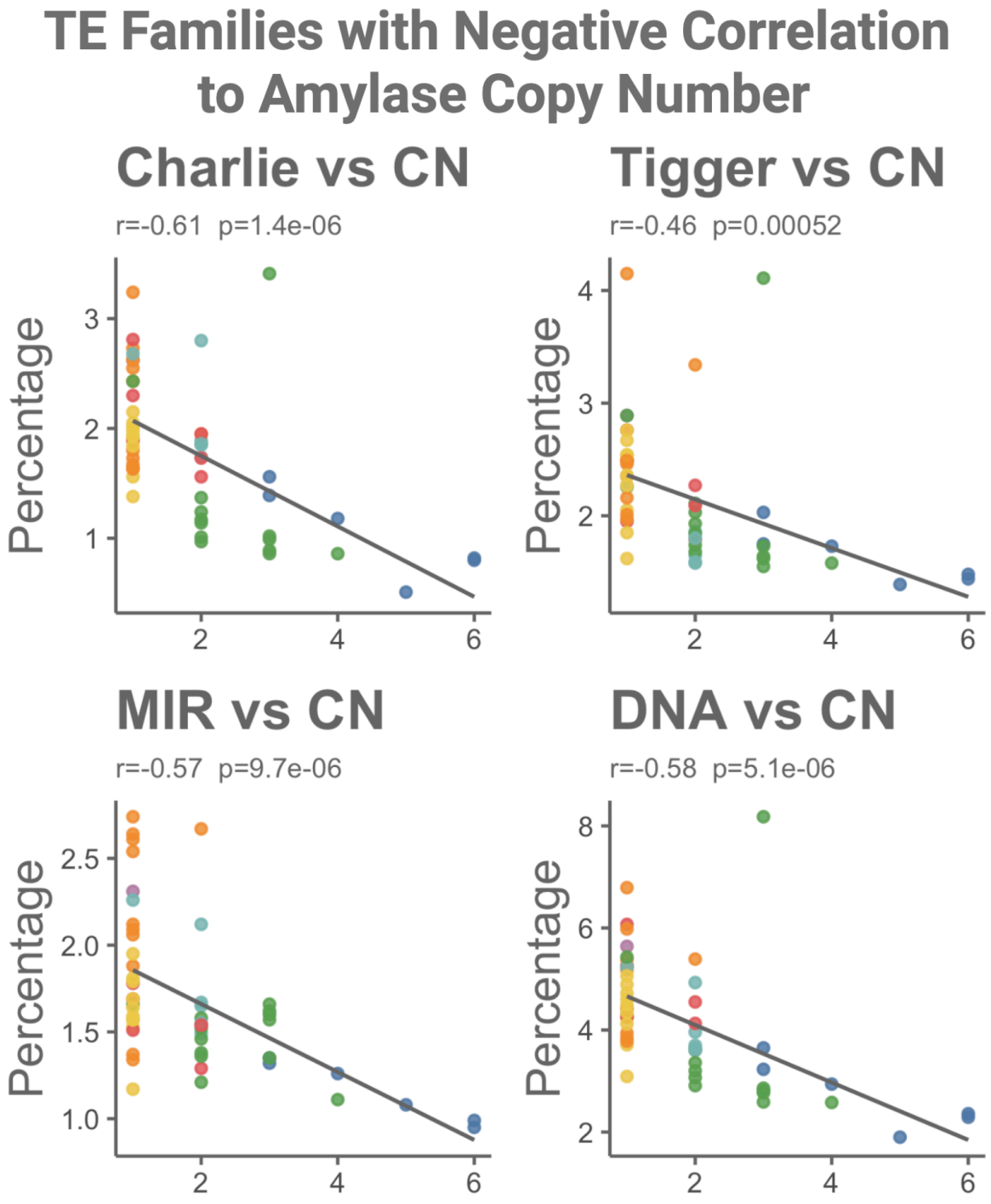

Fig. S7.

TE families negatively correlated with AMY copy number. Zoomed scatterplots for the four families with significant negative associations in Fig. S5 (Charlie, Tigger, MIR, and total DNA transposons). Axes, point colors, fit lines, and statistics are as in Fig. S5.

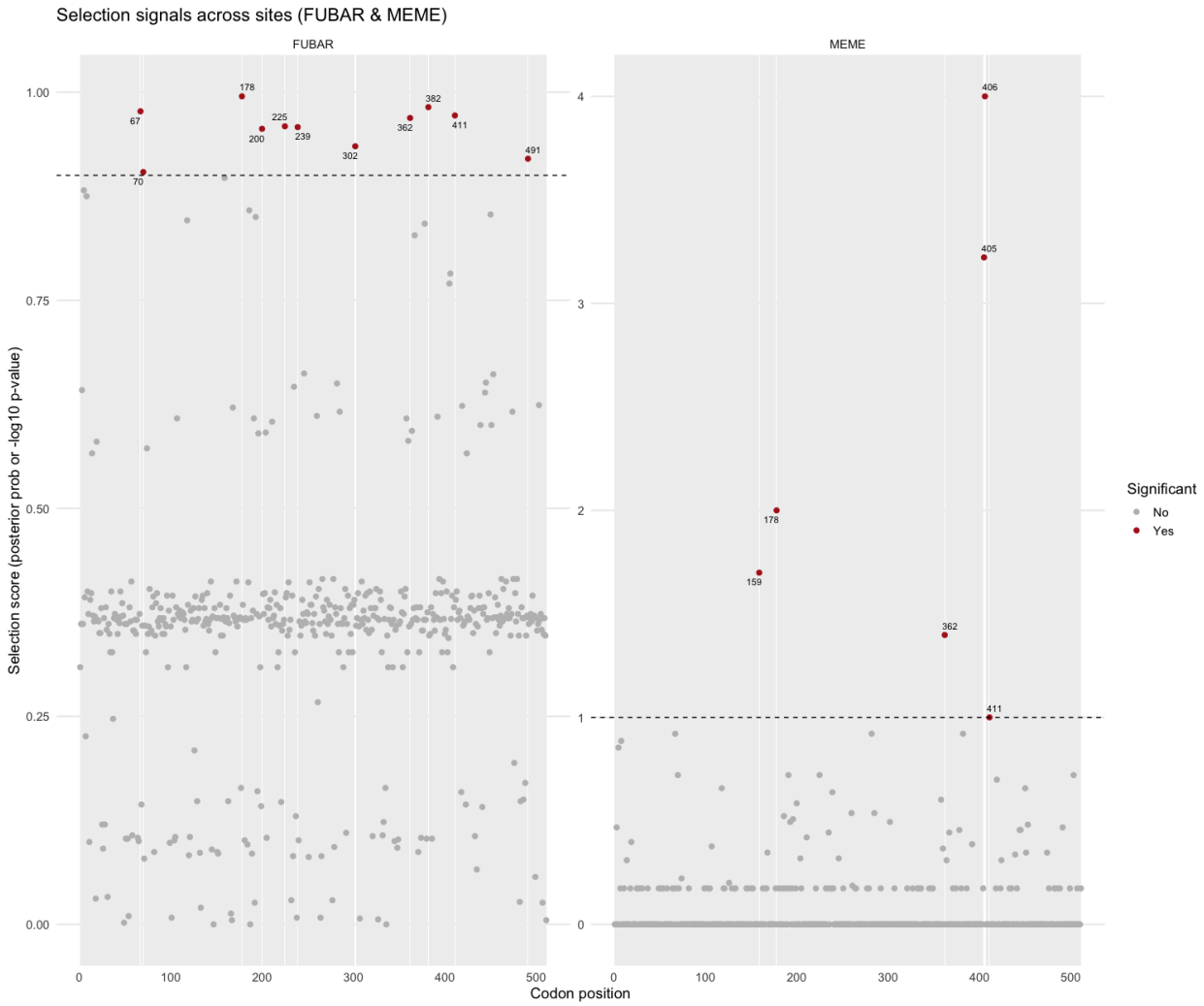

Fig. S8.

Site-level signals of selection across amylase coding sequences. Left: FUBAR posterior probability that ω>1 at each codon; the dashed line marks the 0.90 significance threshold. Right: MEME $-\log_{10}(P)$ for episodic selection; the dashed line marks P=0.10. Red dots denote sites called significant by each method (e.g. MEME: 159, 178, 362, 405, 406, 411; FUBAR: multiple sites ≥0.90). Analyses were run on the codon alignment of Old World monkey and great ape *AMY* paralogs.

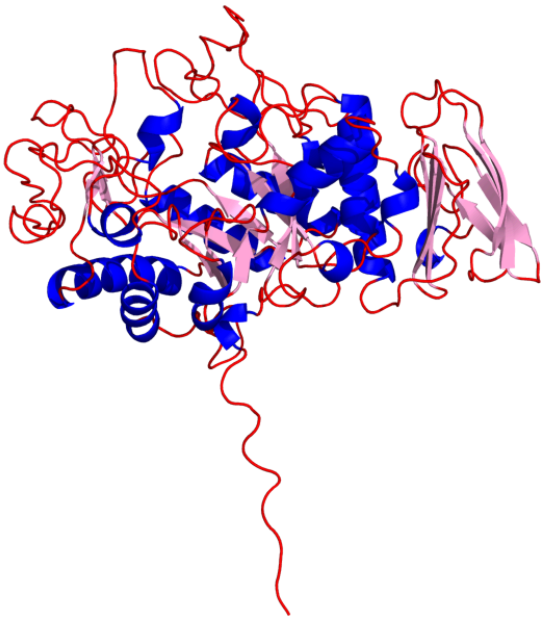

Fig. S9.
Predicted 3D structure of olive baboon *AMYp2*. AlphaFold2 model rendered as a PyMOL cartoon with α-helices in blue, β-strands in light pink, and loops/coil in red.

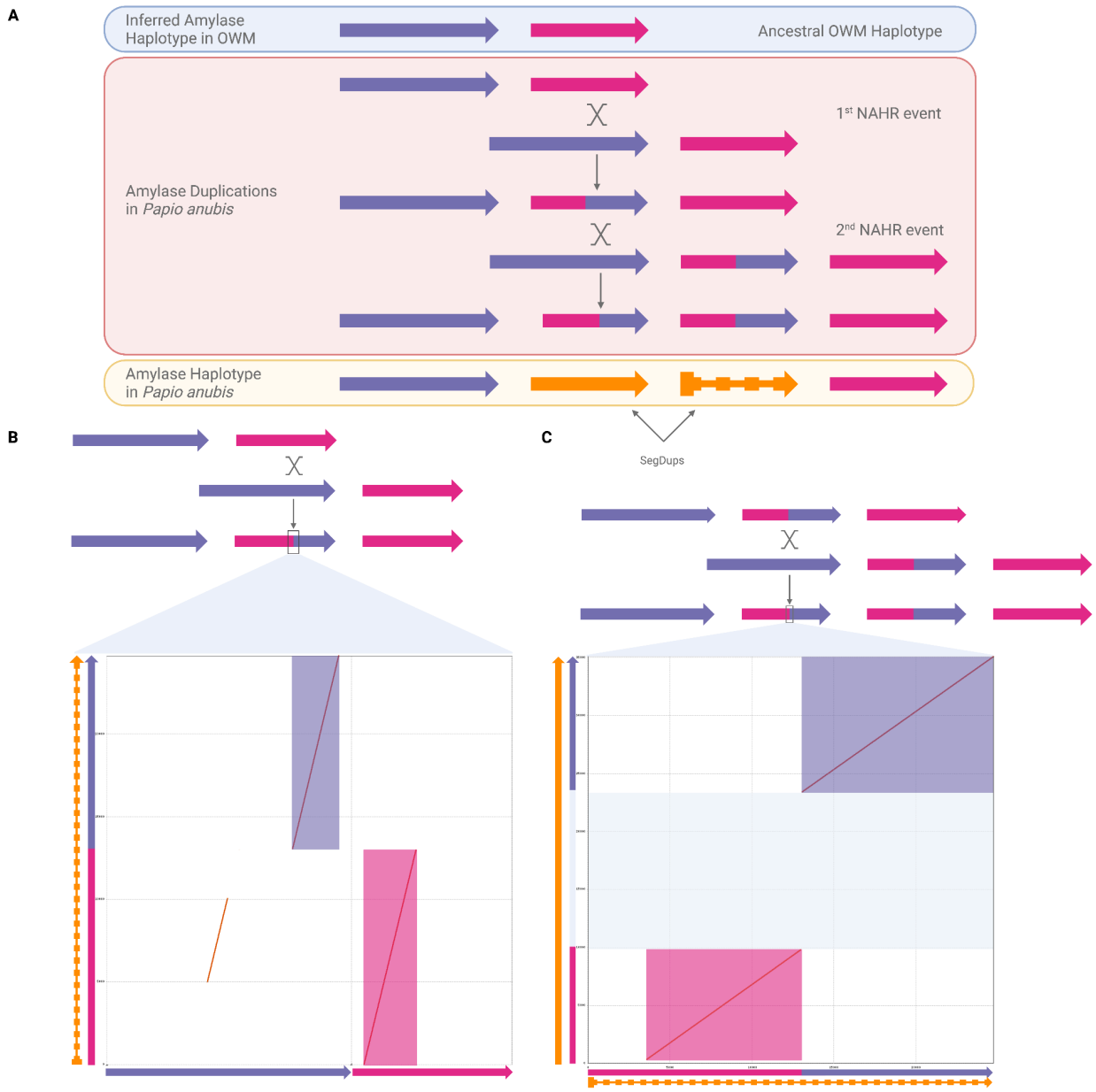

Fig. S10.

Figure S10. Stepwise NAHR duplications shaping the olive baboon amylase locus. The locus is partitioned into four duplicated segments, each drawn as a colored arrow to indicate block identity and orientation; colored arrows denote segments, not individual genes. Segment 1 (purple arrow) carries *AMY2B*, segment 4 (pink arrow) carries *AMY1'*, segment 3 (orange-dashed arrow) carries *AMYp1*, and segment 2 (orange arrow) harbors *AMYp2*. (A) Pairwise NUCmer alignments and dotplots show segment 3 is a mosaic of segment 1 and segment 4, consistent with a first NAHR event between segments 1 and 4 that created the *AMYp1*-containing block. (B) Segment 2 (which hardbours *AMYp2*) is near-identical to segment 3, except for an ~10 kb DNA stretch absent from segment 3. Because this interval is flanked by long N stretches in the olive baboon assembly, we cannot determine whether it reflects a true deletion or a scaffolding artifact.

These dotplots and cartoon schematic representation support a two-step NAHR model: NAHR between segments 1 and 4 produced segment 3 (harboring *AMYp1*), followed by NAHR between segment 1 and segment 3 produced segment 2 (harboring *AMYp2*), yielding the present-day order *AMY2B AMYp2*, *AMYp1*, *AMY1'* in olive baboon. Segments were defined from self-alignments/BISER and compared with NUCmer dotplots as described in Methods.

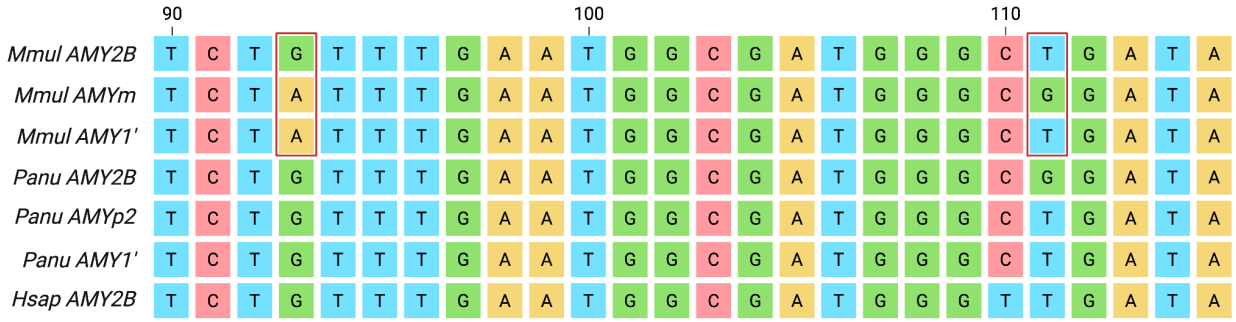

Fig. S11.

Example of the diagnostic polymorphisms used for paralog-specific expression in macaque and baboon. Alignment snapshot (codons ~90-10) across AMY paralogs shows fixed nucleotide differences used to assign RNA-seq reads to specific copies. Red outline boxes highlight parsimony informative sites in macaque (e.g. positions 93 and 111 distinguishing *AMY2B*, *AMYm*, and *AMY1'*). These markers underlie the SNP-aware quantification described in Methods.

Table S1.

Primate genome assemblies surveyed.
Columns: Assembly Accession (NCBI GCA/GCF), Assembly Name (submitter label), Organism Name (NCBI taxon), Annotation Name (if/when available), Assembly Stats Total Sequence Length (bp), Assembly Level (e.g. Scaffold/Chromosome), Assembly Release Date, WGS Project Accession, Assembly BioProject Accession. This table lists all assemblies screened; inclusion does not imply that the amylase locus was recovered on a single contiguous contig.

T**able S2.**

Species represented by multiple genome assemblies (n=16 species; 2-4 assemblies per species).
For each assembly of these species, we report: Assembly Accession, Assembly Name, Organism Name, Annotation Name (if available), Total Sequence Length, Assembly Level, Assembly Release Date, WGS Project Accession and BioProject Accession. This table is descriptive and does not include locus contiguity or copy-number estimations.

Table S3.

Genome assemblies with a contiguous amylase locus (n=69).
Assemblies listed here are the subset in which the amylase interval (RNPC3 5′ flank to AMY1C 3′ flank) was recovered on a single, gap-free contig/scaffold and used for locus-level analyses (synteny, dotplots, and copy-number reconstruction in Fig. 1A). For each assembly we report Assembly Accession, Assembly Name, Organism Name, Annotation Name (if available), Total Sequence Length, Assembly Level, Assembly Release Date, WGS Project Accession, BioProject Accession, and N50. The N50 value is the assembly-wide contig/scaffold N50 (bp) as reported by NCBI at the time of download.

Table S4.

Representative assemblies used for downstream analyses (one per species; 53 species).
From the 69 genomes with a contiguous RNPC3-AMY1C interval (Table S3), we selected one assembly per species for all downstream analyses (e.g. synteny analysis, dotplots, breakpoint mapping, TE quantification within the locus and species-level summaries).

Table S5.

Transposable-element (TE) composition at the amylase locus versus genome-wide across 53 primate assemblies.
For each species, genome-wide TE content is compared to TE content within the curated amylase locus (RNPC3 5′ flank to AMY1C 3′ flank; gap-free contigs only). Annotations were generated with RepeatMasker (v4.1.5; -species primates). Percentages are the fraction of bases masked ×100; ratios are unitless enrichments calculated as (% masked in amylase locus) / (% masked genome-wide), where values >1 indicate enrichment and <1 indicate depletion at the locus. The copy number column provides the number of AMY paralogs in that assembly's locus.
Columns: Species - scientific name; General Class - phylogenetic clade (e.g., OWM, NWM, tarsier, ape); copy number - AMY paralog count; % masked (Genome wide / Amylase locus / Ratio) - total TE burden; then, for each TE group/family, triplets of genome-wide %, amylase-locus %, and Ratio are reported: SINEs (overall), Alu, MIR; LINEs (overall), LINE1, LINE2, L3/CR1; LTR elements (overall), ERVL, ERVL-MaLRs, ERV class I, ERV class II; DNA elements (overall), hAT-Charlie, TcMar-Tigger.

Table S6.
Predicted functional impact of amino-acid substitutions in primate amylase paralogs.
SNVs are listed with GRCh38 coordinates (CHROM, POS, REF, ALT), the mapped UniProt protein (uniprot_id), and the study paralog (transcript_id). protein_variant shows the amino-acid change (RefAA-position-AltAA; e.g. K2Q). am_pathogenicity is a 0-1 predicted impact score (higher: larger effect), and am_class is its category (e.g. likely_benign, ambiguous, likely_pathogenic etc.) Only substitutions unique relative to the UniProt sequence are included.

Table S7.
RNA-seq sample metadata for olive baboon and rhesus macaque tissues.
Each row lists the specimen and tissue used for transcriptomics: SampleID, Species (olive baboon or rhesus macaque), Tissue (parotid, submandibular, sublingual, pancreas, liver), preservation Method (e.g. FlashFrozen), body side (Left/Right) when paired tissues were sampled, Age (years) and Sex. Samples from five olive baboons and six rhesus macaques underpin all expression analyses reported in the manuscript.

Table S8.
Predicted transcription factor binding sites (TFBS) in amylase promoters.
TFBS identified by FIMO (MEME Suite v5.5.8) using the JASPAR 2024 CORE non-redundant vertebrate library. Promoters were scanned as 170-bp windows (100 bp upstream to 70 bp downstream of the TSS). Only matches with $P < 1\times10^{-4}$ are reported (q-values shown for FDR control). Columns: motif_id (JASPAR matrix ID), motif_alt_id (TF symbol), sequence_name (promoter window: gene and species), start/stop (1-based coordinates within the promoter window), strand (motif orientation), score (FIMO match score), p-value (match probability), q-value (Benjamini-Hochberg adjusted p-value), matched_sequence (genomic k-mer). The full table contains 262 sites (108 unique motifs) across promoters of nine amylase paralogs in human, rhesus macaque and olive baboon.

Table S9.
Segmental-duplication calls in the olive baboon amylase locus (BISER v1.4).

Table S10.
Segmental-duplication calls in the rhesus macaque amylase locus (BISER v1.4).

Table S11.
aBSREL branch‑level tests of positive selection.
HyPhy v2.5.48 aBSREL results on AMY paralog codon alignments (olive baboon, rhesus macaque, great apes). P values are Holm-Bonferroni-corrected LRTs. Columns: branch (Species_Gene; NodeXX=internal), tested, p-value (evidence for $\omega>1$), sites (positively selected classes), rates (ω classes), ω distribution (class ω with proportions), ω plot (mean ω, CoV). Branches with $p<0.05$ indicate episodic diversifying selection, notably the internal OWM branch and *Papio anubis AMYp2* (Fig. 5B).

Table S12.
MEME (HyPhy) sites under episodic diversifying selection.
All tested sites are listed. Columns: Site (1-based codon), Partition (alignment partition), α

(synonymous rate), β−/p− (background dN and its weight), β+/p+ (episodic dN>1 and its weight), LRT (likelihood-ratio statistic), p-value (BH-FDR), # branches under selection (branches inferred with ω>1 at this site), Total branch length (from HyPhy), MEME LogL and FEL LogL (site log-likelihoods), Variation p (within-site rate-variation test). Sites with $p < 0.10$ indicate episodic selection.

Table S13.
FUBAR site-wise tests for pervasive selection.
Columns: Site (1-based codon), Partition (alignment block), α (dS), β (dN), β−α (effect size), Prob[α>β] (purifying selection), Prob[α<β] (positive selection), BayesFactor[α<β] (support for positive selection). Posterior summaries are from the FUBAR run described in Methods. Sites with Prob[α<β] ≥ 0.90 are called as positively selected; Prob[α>β] ≥ 0.90 indicates purifying selection.

Table S14. RELAX branch-wise relaxed selection test.
Columns: Branch name (terminal/internal label), Branch partition (Reference=background; Test=foreground), Branch length (expected substitutions/site), k (selection-intensity parameter; k>1 = intensified selection, k<1 = relaxed, k≈1 = no change). The designated foreground (Test) branch is *Papio anubis AMYp2*; model fitting and LRT significance are described in Methods. Test for selection intensification (K = 2.63) was significant (p = 0.000, LR = 15.27).