Ancient *AMY1* gene duplications primed the amylase locus for adaptive evolution upon the onset of agriculture

Feyza Yilmaz[1]*, Charikleia Karageorgiou[2]*, Kwondo Kim[1]*, Petar Pajic[2], Christine R. Beck[1,3,4], Human Genome Structural Variation Consortium, Ann-Marie Torregrossa[5,6], Charles Lee[1], Omer Gokcumen[2]****

Affiliations

1 - The Jackson Laboratory for Genomic Medicine, Farmington, CT, 06110, USA

2 - Department of Biological Sciences, University at Buffalo, 109 Cooke Hall, University at Buffalo, NY 14260

3 - University of Connecticut, Institute for Systems Genomics, Storrs, CT, 06269, USA

4 - The University of Connecticut Health Center, Farmington, CT, 06110, USA

5 - Department of Psychology, University at Buffalo, 204 Park Hall, University at Buffalo, NY 14260

6 - University at Buffalo Center for Ingestive Behavior Research, University at Buffalo, 204 Park Hall, University at Buffalo, NY 14260

** Contributed equally; ** Correspondence*

Abstract

Starch digestion is a cornerstone of human nutrition. The amylase enzyme, which digests starch, plays a key role in starch metabolism. Indeed, the copy number of the human amylase gene has been associated with metabolic diseases and adaptation to agricultural diets. Previous studies suggested that duplications of the salivary amylase gene are of recent origin. In the course of characterizing 51 distinct amylase haplotypes across 98 individuals employing long-read DNA sequencing and optical mapping methods, we detected four 31mers linked to duplication of the amylase locus. Analyses with these 31mers suggest that the first duplication of the amylase locus occurred more than 700,000 years ago before the split between modern humans and Neanderthals. After the original duplication events, amplification of the *AMY1* genes likely occurred via nonallelic homologous recombination in a manner that consistently results in an odd number of copies per chromosome. These findings suggest that amylase haplotypes may have been primed for bursts of natural-selection associated duplications that coincided with the incorporation of starch into human diets.

Main

The copy number variation at the amylase locus is frequently attributed to human health and adaptation. As such, this structurally variable locus is a prime target for research on the fundamental biology of gene duplications. Amylase gene copy number has been reported to range from 2 to 17 copies per diploid cell (Groot, Mager and Frants, 1991; Perry *et al.*, 2007; Usher *et al.*, 2015). There are two types of amylase genes, *AMY1* and *AMY2*, which are expressed in the salivary glands and pancreas, respectively (Ting *et al.*, 1992). Both genes encode the amylase enzyme, which breaks down polymeric starch into simple sugar molecules, a crucial digestive process for starch-eating species, including humans (Peyrot des Gachons and Breslin, 2016).

Mammals that consume starch-rich diets convergently underwent independent bursts of adaptive amylase gene duplication from the ancestral pancreatic *AMY2-like* gene (Pajic *et al.*, 2019). For example, a great ape-specific duplication resulted in the formation of the salivary *AMY1* gene (Meisler and Ting, 1993), which has unusually high copy numbers among human populations with increased starch consumption, especially among historically agricultural populations (Perry *et al.*, 2007). These evolutionary insights indicate that copy number variation at the amylase locus may play an adaptive role in shaping the metabolic response to starchy diets. More recent studies have further established robust connections with higher amylase gene copy number and lower body mass index, reduced obesity risk, and gastrointestinal microbiome composition, particularly the abundance of resistant starch-degrading microbes (Falchi *et al.*, 2014; Usher *et al.*, 2015; Poole *et al.*, 2019).

The expansion of amylase copy number is suggested to predate agriculture (Mathieson and Mathieson, 2018). In this study, we tested the hypothesis that the initial amylase gene duplication already existed before the introduction of starch in the human diet. However, the presence of highly similar tandem repeats within the amylase locus poses significant challenges to accurately map short-read sequences and detect variants (Sudmant *et al.*, 2015). Therefore, by integrating optical genome mapping (OGM) and long-read sequencing datasets, we resolved this locus across 98 individuals from diverse populations, which provided us with evolutionary and mechanistic insights of this locus.

Structural haplotypes at the human amylase locus

The amylase locus is a ~212.5 kbp region on chromosome 1 (GRCh38; chr1:103,554,220–103,766,732) which contains *AMY2B*, *AMY2A*, *AMY1A*, *AMY1B*, and *AMY1C* genes (**Fig. 1A**). This locus overlaps with segmental duplications (SDs), which have > 99% sequence similarity, which hindered accurate mapping of this locus using short-read sequencing (**Fig. S1**). Using the sequence similarity between SDs and the labeling pattern from the OGM data, we defined six distinct amylase segments, depicted by, and named after colored arrows (**Fig. 1B; Table S1**). To resolve structural haplotypes, we characterized the amylase locus in the genomes of 98 individuals from Africa (n = 43), America (n = 24), East Asia (n = 13), Europe (n = 9), and South Asia (n = 9) using the copy number and orientation of the aforementioned amylase segments. These datasets were made available through the Human Genome Structural Variation Consortium (n = 60) (Ebert *et al.*, 2021), the telomere-to-telomere consortium (HG002; T2T-chm13) (Nurk *et al.*, 2021), and the Human Pangenome Reference Consortium (n = 37) (Liao *et al.*, 2023) (**Table S2**). Using OGM, we constructed haplotype-resolved diploid assemblies (n = 196 chromosomes) and identified 51 distinct amylase haplotypes (**Fig. S2; Table S3**), 13 of which were previously reported (Usher *et al.*, 2015) **(Fig. S3)**. Further, we identified 30 high-confidence haplotypes (n = 117 chromosomes) that are orthogonally supported by long-read sequencing based *de novo* assemblies (**Fig. 1B**) (**Data File**). These high-confidence haplotypes represent the first nucleotide-level reconstruction of amylase haplotype variation and were used for downstream analyses.

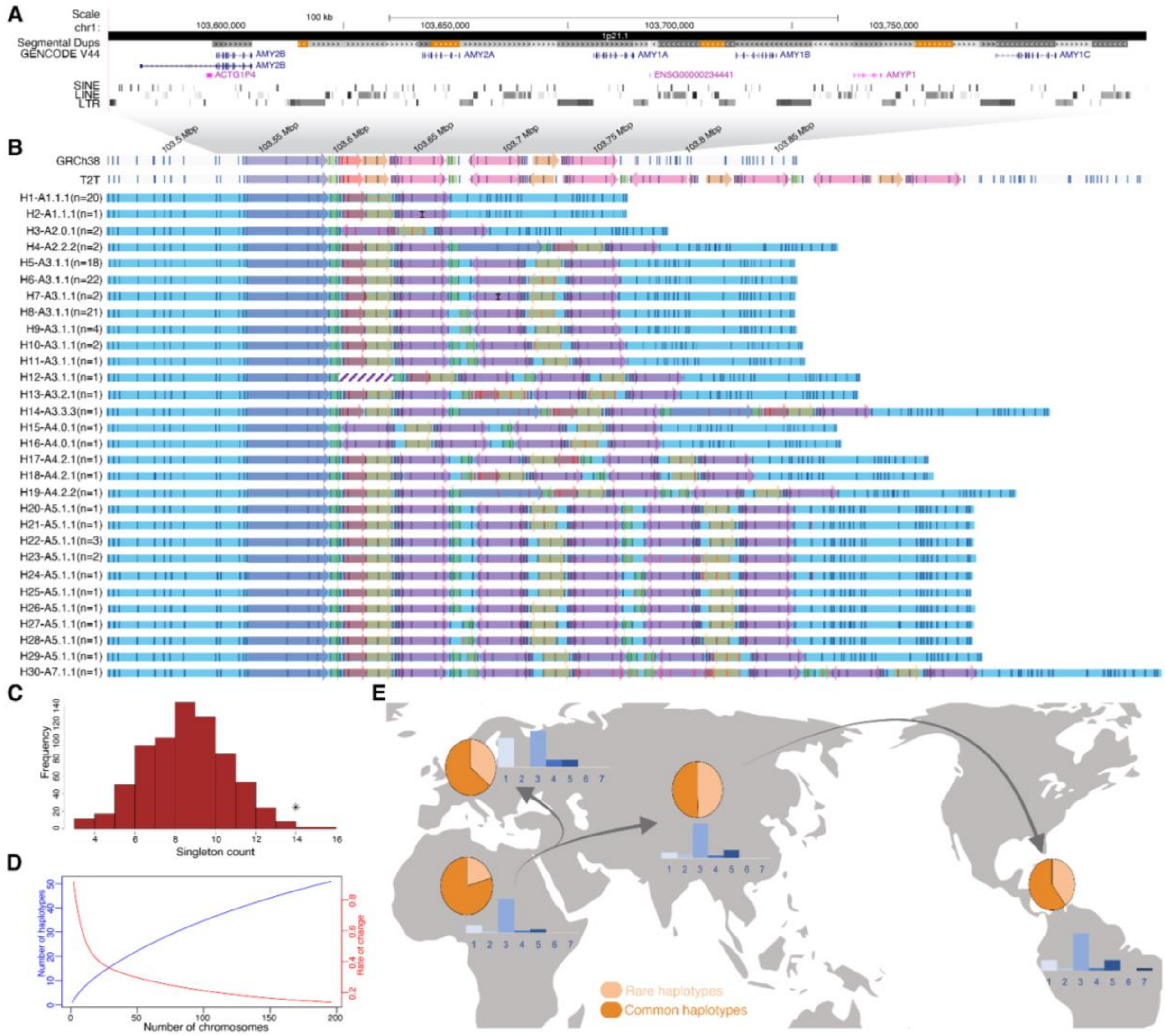

Fig. 1. Structural amylase haplotypes (n=30) identified in this study. **A.** GENCODE V44, SINEs, LINEs, LTRs, and segmental duplications are represented as tracks. **B.** Reference assemblies, GRCh38 and T2T-chm13, amylase segments (colored arrows), and *in silico* maps are represented in rectangles with white background and vertical blue lines. Size of segments are as follows: Purple: 45 kbp, green: 5 kbp, red: 12 kbp, orange: 14 kbp, maroon: 4 kbp, pink: 26 kbp. The bottom part of the panel represents the structure of amylase haplotypes identified among our samples (H1-H30). The black line in the second pink segment of H7 represents the polymorphic label present in three chromosomes. Diagonal stripes in the second purple segment of H12 indicate that the segment is a partial copy of the first purple segment. **C.** The number of singleton haplotypes detected from EnsembleTR calls generated from 1000 genomes project dataset. Star represents singletons detected from this study. **D.** Rarefaction plot displaying how the rate of change (red line) decreases when the number of unique haplotypes (blue line) starts to reach asymptote with the increase in the number of chromosomes. **E.** The distribution of amylase haplotypes across

African, American, East Asian, European, and South Asian continental populations. Orange: common haplotypes; beige: rare haplotypes (n < 5). Bar plots in blue shades represent *AMY1* copy number distribution across the globe.

The length of the amylase haplotypes ranged from 111 kbp (H1 and H2) to 402 kbp (H30) (**Fig. 1B**), including those structurally identical to the GRCh38 (H5) and T2T-chm13 (H40) (Nurk *et al.*, 2021) assemblies, respectively (**Fig. S2**). Four haplotypes, H1 (n = 20/117), H5 (n = 18/117), H6 (n = 22/117), and H8 (n = 21/117) are common (number of chromosomes > 5), have higher than 10% allele frequency among the individuals studied, and collectively constitute more than 70% of all amylase haplotypes (**Fig. 1B**). The remaining haplotypes consist of rare (5 > number of chromosomes > 1) (n = 7/30) and singleton haplotypes (number of chromosomes = 1) (n = 19/30). We found that the proportion of singleton haplotypes in the amylase locus is significantly higher than the genome-wide average of singletons observed for tandem repeat loci, suggesting a rapid evolutionary rate of the locus (**Fig. 1C,** see “Distribution of singleton numbers for genome-wide tandem repeats”).

Rarefaction analysis showed that the rate of change in the number of unique haplotypes is close to zero, suggesting that we have identified the majority of the common haplotypes in the human population (**Fig. 1D**). Despite our limited sample size, we found that all common haplotypes (H1, H5, H6, and H8) exist in all continental populations, while there is remarkable continental specificity of haplotype variation due to singletons (**Fig. 1E; Fig. S4**). Although there are many population-specific structural haplotypes, the *AMY1* copy number variation does not show geographical specificity (p-value = 0.6817, Fisher’s exact test; **Fig. 1E**). These observations suggest that common amylase haplotypes evolved before out-of-Africa migrations and copy number increase driven by agriculture may not explain the geographical distribution of *AMY1* structural variation at the continental scale.

Strong negative selection may limit functional variation among amylase gene copies

The fate of gene duplications is often pseudogenization (Innan and Kondrashov, 2010). Nevertheless, previous studies hypothesized that individual amylase gene copies code for enzymes with identical molecular functions (i.e., starch digestion) (Perry *et al.*, 2007), and that their copy number variation only regulates their expression levels. To test this hypothesis, we evaluated the extent of protein-coding sequence variation associated with our high-confidence amylase haplotypes (n = 30). We first predicted protein-coding amylase gene copies in each haplotype and identified 582 intact protein-coding amylase gene copies and 110 pseudogenes among 117 chromosomes (**Table S4**). To experimentally validate these copy number predictions, we conducted digital droplet PCR on 13 individuals (**Fig. S5; Table S4**). Subsequently, we aligned the coding sequences of intact amylase copies and reconstructed the phylogeny using the sheep amylase sequence as an outgroup (**Data Files**). Our findings revealed that all human amylase gene copies can be robustly clustered into three distinct types: *AMY2B*, *AMY2A*, and *AMY1* (**Fig. 2A**; **Fig. S6**), and the majority of SNVs within each type are singletons, 7/10, 4/12 and 12/19, respectively (**Table S5**).

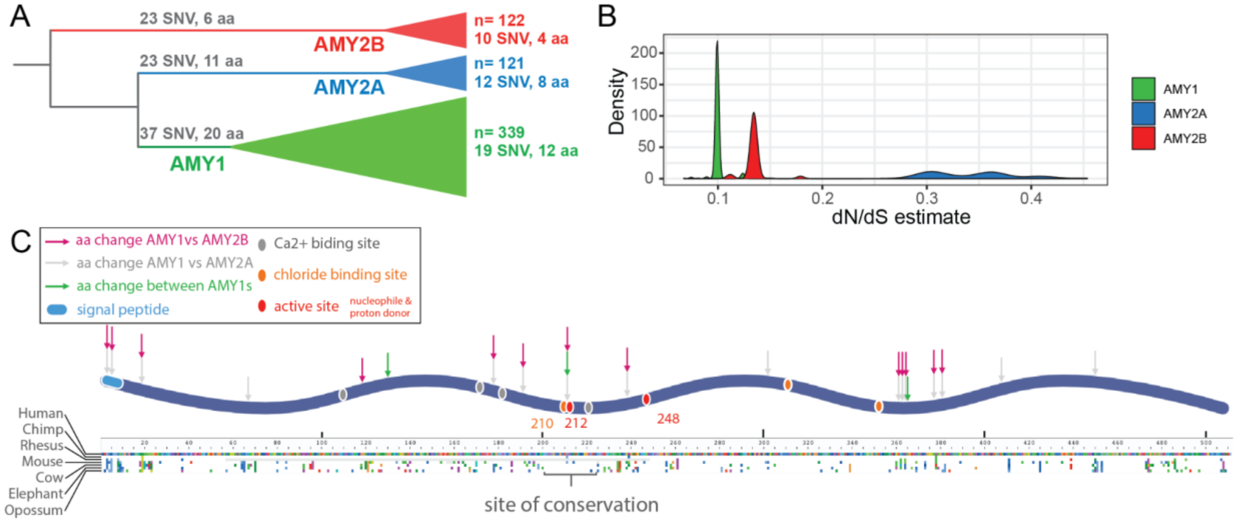

Fig. 2. Coding sequence variants and negative selection on three amylase gene types. **A.** The maximum likelihood phylogenetic tree of amylase coding sequences rooted with coding sequence from sheep genome (Oar_rambouillet_v1.0). Nucleotide and resultant amino acid changes occurring on each branch are indicated. The tree clusters the coding sequences into three branches, representing *AMY2B*, *AMY2A*, and *AMY1*. All the variants between the types are fixed except for one single nucleotide variant in the *AMY1* branch. The numbers at the tips indicate the sample size and the number of variants found within each amylase gene type. **B.** The distribution of dN/dS estimates for each amylase gene type. **C.** Locations of amino acid changes across 511 residues of amylase in the context of functional sites.

We found that the rate of synonymous mutations (dS = 0.02720+/-0.007048) was significantly higher than the rate of nonsynonymous mutations (dN = 0.003847+/-0.001540) (**Fig. 2B**). Specifically, when compared to their chimpanzee counterparts, 504 of 582 amylase gene copies had significantly lower dN/dS estimates than the null model (i.e., dN/dS = 1) (FDR adjusted p-value from χ^2 test < 0.05; **Table S6**). Additionally, we identified similarly significant signatures for negative selection within *AMY1*, *AMY2A*, and *AMY2B* types. For example, *AMY1* displayed significantly low dN/dS estimates (dN/dS = 0.09976+/-0.005797; **Fig. 2B**). These observations support the notion that negative selection (low dN/dS) has acted to retain the amino acid sequence and the original protein function of amylase gene copies, both within and between the three *AMY1* types. Some observed amino acid differences overlap with conserved, functionally relevant protein regions, including previously defined active sites (Ramasubbu, Ragunath and Mishra, 2003) (**Fig. 2C**). While these differences may not have immediate fitness consequences, they could still contribute to functional differences and have crucial biomedical relevance. Therefore, our results are consistent with the notion that amylase gene duplications encode for near-identical proteins but exhibit functional differences in terms of the regulation of their expression. The functional consequences of amylase gene duplications can be categorized as subfunctionalization (Innan and Kondrashov, 2010), resulting in variation in gene expression of the amylase genes in different tissues.

For the 110 amylase duplicates that we identified as pseudogenes (*e.g., AMYP1*) in our dataset, we hypothesize that all these amylase pseudogenes are the result of a single incomplete gene duplication that led to nonfunctionality. Indeed, consistent with this idea, we found that all amylase pseudogenes share a single phylogenetic origin from an ancestral incomplete gene duplication of *AMY2A*, as suggested previously (Carpenter *et al.*, 2015) (**Fig. S7**).

Hominin-specific AMY1 duplications that predate Human-Neanderthal divergence

Unraveling the evolutionary history and mutational mechanisms that underlie variation at the amylase locus is challenging primarily due to the presence of SDs and retrotransposon insertions. To address this issue, we systematically evaluated the haplotype variation in our data by aligning each amylase segment (**Fig. S8**) and identified an interval within the pink segment downstream of *AMY1* gene copies as the most phylogenetically informative site in the amylase locus (**Fig. 3A**). Within this interval, we observed that sequence variation can be phylogenetically structured into three distinct clusters, which we have indicated as Pink1A (n = 124), Pink1B (n =

99), and Pink1C (n = 114), (**Fig. S9**). Notably, distinct sets of retrotransposons that flank *AMY1* gene copies further contribute to variation among pink segments (**Fig. 3A**).

The phylogenetic clustering of the pink segments enabled us to delineate their evolutionary history, which is particularly important because this segment harbors the salivary human specific *AMY1* gene copies. The chimpanzee (panTro6) and gorilla (gorGor6) reference genomes have only one pink segment (**Fig. S10**), suggesting that the common ancestor of humans and chimpanzees possessed a single pink segment (Pink1C), and, thus, a single *AMY1* copy. It is worth noting that the bonobo reference genome (panPan3) has two segments similar to pink segments, both harboring *AMY1* genes. One of these bonobo pink segments aligns well with the ancestral Pink1C segment, while the other appears distinct from all other pink segments in human, chimpanzee, and gorilla (**Fig. S11**). Further, we found that amino acid position 19 in all human *AMY1* proteins carry a serine, whereas both chimpanzee and bonobo *AMY1* proteins carry a proline at the same position (**Fig. 2C**), which in the absence of convergent evolution for proline, further supports the independent duplication of *AMY1* genes in human and bonobo lineages. Collectively, rather than a scenario involving incomplete lineage sorting of haplotypes with *AMY1* duplications in humans and bonobos, an independent duplication of the *AMY1* genes in the human lineage from a single *AMY1* gene better fits the data.

Debates regarding the timing of *AMY1* duplications in the human lineage have been ongoing. To address this debate, we independently constructed alignments of pink segments from the most common haplotypes (H5 and H6) harboring all three pink segment types in our dataset. Based on these alignments, we obtained two independent Bayesian estimates. Both suggest that Pink1B originated from a duplication of Pink1C approximately 140,000 to 270,000 years ago, followed closely by the duplication of Pink1A from Pink1B approximately 120,000 to 240,000 years ago (**Fig. S12; Table S7**). Gene conversion between the GC-rich SDs is a complication for time estimation based on a molecular clock, and is a known phenomenon at the amylase locus (Groot, Mager and Frants, 1991; Vollger *et al.*, 2023). For example, we found that specific sequences exhibit a high degree of sequence similarity between Pink1A and Pink1B segments that cannot be explained easily, given that this locus is one of the rearrangement hotspots (Sharp *et al.*, 2006). This observation suggests rapid gene conversion (**Fig. S13**). Considering gene conversion, the actual duplication dates are expected to be older than our estimates, predating out-of-Africa migrations (**Fig. S12**; **Table S7**).

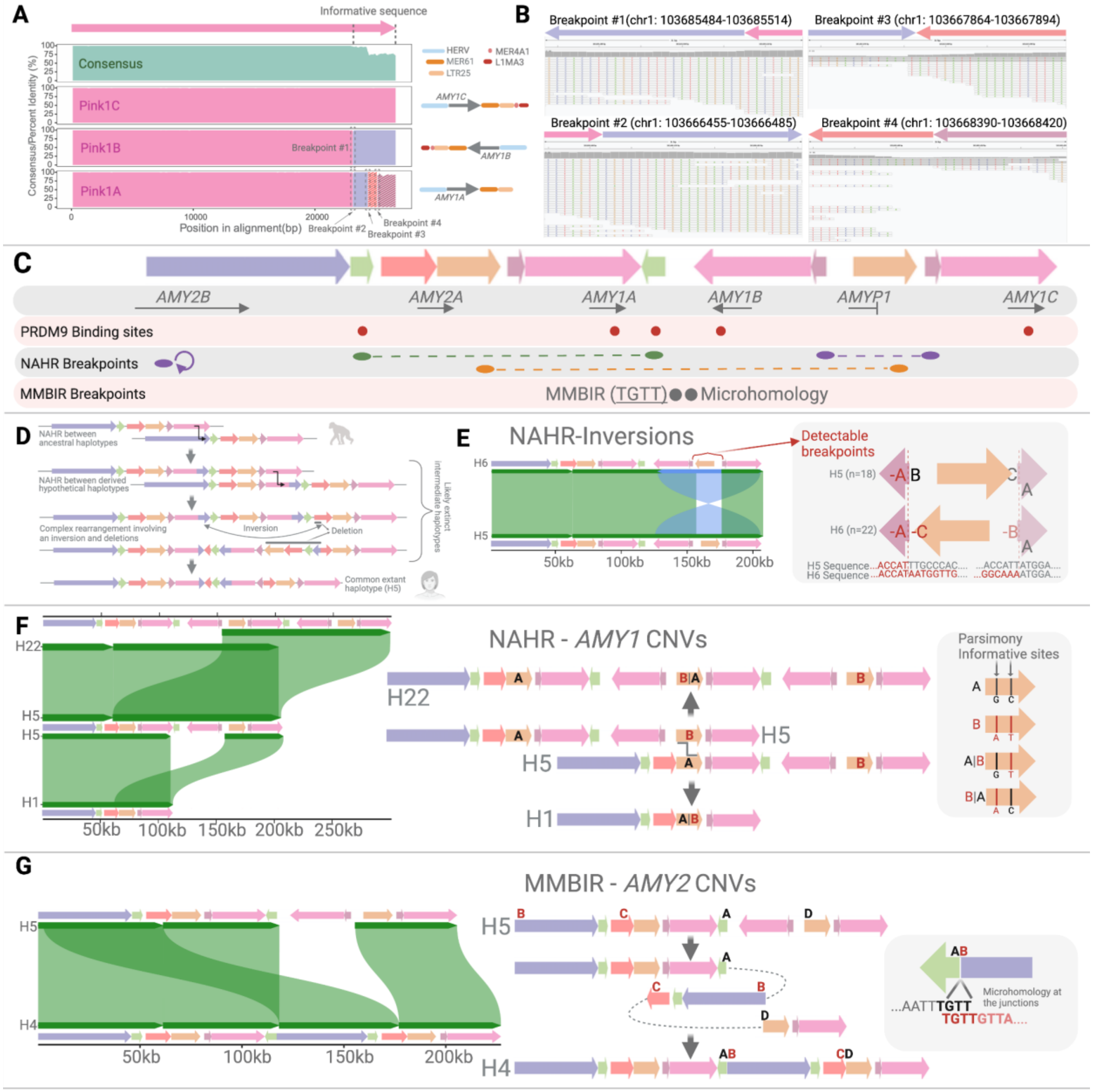
A
Informative sequence
Consensus/Percent Identity (%)
100
75
50
25
0
Consensus
Pink1C
Pink1B
Breakpoint #1
Pink1A
0
10000
20000
Position in alignment(bp)
Breakpoint #2
Breakpoint #3
Breakpoint #4
HERV
MER61
LTR25
MER4A1
L1MA3
AMY1C
AMY1B
AMY1A
B
Breakpoint #1(chr1: 103685484-103685514)
Breakpoint #3 (chr1: 103667864-103667894)
Breakpoint #2 (chr1: 103666455-103666485)
Breakpoint #4 (chr1: 103668390-103668420)
C
AMY2B
AMY2A
AMY1A
AMY1B
AMYP1
AMY1C
PRDM9 Binding sites
NAHR Breakpoints
MMBIR Breakpoints
MMBIR (TGTT) Microhomology
D
NAHR between ancestral haplotypes
NAHR between derived hypothetical haplotypes
Complex rearrangement involving an inversion and deletions
Inversion
Deletion
Likely extinct intermediate haplotypes
Common extant haplotype (H5)
E
NAHR-Inversions
H6
H5
50kb
100kb
150kb
200kb
Detectable breakpoints
H5 (n=18)
H6 (n=22)
-A B
C A
-A -C
-B A
H5 Sequence...ACCATTTGCCCAC.... ...ACCATTATGGA....
H6 Sequence...ACCATAATGGTTG.... ...GGCAAAATGGA....
F
H22
H5
H5
H1
50kb
100kb
150kb
200kb
250kb
NAHR - AMY1 CNVs
H22
H5
H5
H1
A
B|A
B
A|B
Parsimony Informative sites
A
B
A|B
B|A
G C
A T
G T
A C
G
H5
H4
50kb
100kb
150kb
200kb
MMBIR - AMY2 CNVs
H5
H4
B
C
A
D
AB
CD
...AATTTGTT
TGTTGTTA....
Microhomology at the junctions

Fig. 3. Characterization of the three amylase segments, Pink1C, Pink1B, and Pink1A, and **the evolutionary and mutational connections among common haplotypes A.** Consensus of pink segment alignment (top) and sequence composition of each pink segment cluster (bottom). The stripe pattern in Pink1A (red and maroon segments) indicates that the segment is in the opposite direction compared to the direction of the entire pink segment. The diagrams on the right represent retrotransposon compositions of each pink segment cluster. The dashed box indicates the region where the possible breakpoint for each event (e.g. inversion) is located. **B.** Read pileups of the Altai Neanderthal genome at the 31mers within the breakpoints depicted in panel A. The coordinates are based on GRCh38. The reads colored in gray indicate uniquely mapped reads. The

reads colored in white with the gray border indicate mapping quality zero reads mapped to both breakpoints #1 (Pink1B) and #2 (Pink1A) that have identical sequences. **C.** The structural variation breakpoints and recombination hotspots in the amylase locus. The colored arrow represents amylase segments. The PRDM9 binding sites are represented with red dots. The NAHR breakpoints are represented with purple, green and purple dots, and dashed lines. The MMBIR breakpoints are represented with gray dots. **D.** A plausible model for the evolution of the three distinct pink segments, starting from a chimpanzee-like ancestral haplotype and resulting in the common H5 haplotype. All haplotypes except for the chimpanzee and H5 haplotypes are hypothetical. **E**. The mutational mechanism is NAHR-based inversion. Comparison of two common haplotypes, H5 and H6 harboring the same copy number of segments, the two haplotypes are separated by a single inversion as indicated in the plot (left panel), the inversion is generated via NAHR involving the maroon inverted repeats. The detectable inversion breakpoints are depicted in dotted lines (right panel). **F.** NAHR-mediated duplication and deletion of the *AMY1A-AMY1B* cluster. The orange segment serves as the recombination substrate for the crossover, resulting in the duplication or deletion of the *AMY1A-AMY1B* cluster as illustrated in the schematic diagram in the middle of the panel. Chimeric orange segments have been identified, utilizing parsimony informative sites within the orange segment. **G**. MMBIR based copy-number gain resulting in the formation of H4. The schematic diagram shows the mutational mechanism in the middle of the panel. Four nucleotides of microhomology internal to the breakends were identified at the breakpoint junction (right panel).

An alternative approach for estimating the relative timing of gene duplications is to leverage ancient genomes. We analyzed high-coverage archaic hominin genomes to ascertain whether the duplications of Pink1B and Pink1A occurred before or after the divergence between human and Neanderthals. Specifically, we identified four 31mer sequences that harbor specific SNVs within breakpoints for each pink segment type (**Fig. 3A and B**; **Table S8**). Next, we searched for archaic hominin sequences that map to these 31mers and found support for all three types of pink segments in archaic hominin genomes (**Fig. 3B; Figures S14-S17**). These results suggest that three pink segment types, and consequently at least three *AMY1* copies, were already present in archaic hominins and thus predate Human-Neanderthal divergence. This finding contrasts the findings of previous studies that used read-depth-based estimates to conclude that Neanderthal genomes carry only one copy of *AMY1* per chromosome (Prüfer *et al.*, 2014).

Moreover, our data shows that the initial human specific *AMY1* duplications predate farming by hundreds of thousands of years. This finding is especially interesting in light of recent findings that increased starch consumption predates agriculture, where Neanderthal consumed observable amounts of starchy foods (Yates *et al.*, 2021) and Mesolithic foragers consumed starchy grains that would later become domesticated (Cristiani *et al.*, 2021). Therefore, our results raise the possibility that amylase copy number may have evolved earlier than agriculture and perhaps primed human perception and metabolism for agricultural diets.

Multiple mutational origins of diverse human amylase haplotypes

Our results suggest that the rapid structural evolution of amylase locus in the human lineage has led to functionally relevant extant variation. It is crucial to understand the mutational mechanisms that underlie this variation to better frame the putative adaptive evolution of the amylase structural haplotypes. To trace the mutational steps that underlie SVs among the amylase haplotypes, we employed dotplots and local sequence alignments to identify breakpoints of structural differences (**Supplementary Results**). In parallel, we conducted a scan for PRDM9-binding motifs within the locus to pinpoint possible recombination sites (**Fig. S18**). By integrating all these observations, we were able to infer the fewest plausible mutational steps and construct a putative evolutionary path explaining the origin of present-day amylase haplotypes commencing from common haplotypes (**Figures 3C, S18 and S19; Supplementary Results**). Our mutational models remain preliminary, they offer four major insights into the evolution of amylase locus: First, we were able to construct a putative evolutionary model that involved two NAHR events followed by an inversion-deletion event linking ancestral chimpanzee-like haplotype to extant common haplotypes (H5 and H6) harboring three *AMY1* copies (**Fig. 3D; Fig. S20-S22)**. This model gives some hints to the evolution of haplotypes carrying multiple *AMY1* copies. Specifically, given the lack of intermediate haplotypes among the extant human haplotypes, the initial formation of haplotypes harboring three *AMY1* copies might have happened only once in human evolution. It is possible that the chimpanzee-like H1 haplotype has remained polymorphic, along with other structural haplotypes in the human gene pool, since the Human-Neanderthal divergence. Alternatively, it is possible that the extant H1 haplotypes are recently and recurrently derived via deletions on extant common haplotypes. Our current data do not allow us to conclusively distinguish between these two mutually inclusive scenarios about the origins of H1 haplotypes in extant human haplotypes.

A second, related observation is that recurrent NAHR-mediated inversion events at the amylase locus, similar to those described previously (Porubsky *et al.*, 2022) underpin the mutational connections between the common H5, H6, and H8 haplotypes (**Fig. 3E**), as well as several other inversions among extant haplotypes as described in **Supplementary Results**. This result explains the observation that inversions are the predominant SVs underlying haplotypic variation in this locus and raises novel questions about the functional and adaptive relevance of inversions in the amylase locus.

Third, we found that recurrent NAHR among common haplotypes harboring three amylase copies (H5, H6, H8) with breakpoints in the orange segment would concurrently result in the H1 haplotype carrying a single *AMY1* copy, and H22, H23, and H24 haplotypes, each carrying five *AMY1* copies (*e.g.*, **Fig. 3F; Fig. S23; Supplementary Results**). Therefore, while other less likely scenarios are possible, NAHR-based deletion and duplications underlie copy number variation of the pink segments and thus *AMY1* genes. Specifically, as this proposed mechanism always adds or deletes two copies of the *AMY1* genes, it explains the puzzling observation that the human haploid genomes almost always harbor odd-numbered of *AMY1* copies (Usher *et al.*, 2015) (**Fig. S24**).

Fourth, we characterized three MMBIR events and identified the accompanying microhomologies at the breakpoint junctions (**Fig. 3G**; **Supplementary Results**). Even though these three rearrangements constitute only five chromosomes (H4, H10, and H12) (~4%), they hold substantial biological relevance since H4 and H12 harbor duplications of the *AMY2* genes. Further, NAHR events involving the resulting haplotypes with multiple *AMY2* copies may lead to NAHR events that underlie further *AMY2* gene copy number variation (**Fig. S19**). Our results implicate different mechanisms that underlie copy number variation of the salivary *AMY1* and pancreatic *AMY2* genes and explain the relatively low variation in the latter.

Conclusion

In our study, we characterized structural haplotypes in the amylase locus with unparalleled nucleotide-level precision. We detected four common, and hence likely older, amylase haplotypes that appear to have given rise to a great structural diversity across human populations, manifesting as 51 unique structural haplotypes in our study. Of these, we decoded the sequence variation for 30 high-confidence haplotypes across 117 chromosomes. Remarkably, these structural haplotypes evolve more rapidly compared to single nucleotides (Usher *et al.*, 2015) and short tandem repeat variations (**Fig. 1C**). This rapid evolution hinders imputation of structural haplotypes. Despite this variation, the amino acid sequences of amylase gene copies remain conserved due to persistent negative selection. As such, these intact gene duplicates likely contribute to the mosaic gene expression of amylase in the pancreas and the three major salivary glands (Samuelson *et al.*, 1990). Our comprehensive mapping of the amylase locus identified a small number of informative SNVs, offering a promising avenue for further investigation into duplicate-specific expression of amylase in various tissues, including the brain and blood. Additionally, our data enable the comprehensive investigation of differentially spliced isoforms that could underlie previously reported variants at the protein level (Thamadilok *et al.*, 2020).

Our evolutionary analyses revealed three distinct genomic segments containing *AMY1* genes, distinguishable by sequence variations and specific retrotransposons despite evidence for ubiquitous gene conversion. One segment predates the emergence of great apes, while the other two evolved within the hominin lineage prior to the Human-Neanderthal divergence. This suggests that modern human ancestors had at least three *AMY1* gene copies on their haploid genomes, tracing the inception of these human-specific duplications back several hundreds of thousands of years. Mechanistically, most extant haplotypes, including those with varying *AMY1* gene copies, likely arose through recurrent NAHR events exemplified in **Fig. 3**. We also found evidence for less frequent non-recurrent MMBIR-based events that underlie *AMY2* copy number variation, suggesting that the copy number variation of salivary and pancreatic amylase evolve through different evolutionary mechanisms.

Combined, our results provide an improved model for the evolution of the amylase locus and raise three intriguing hypotheses (**Fig. 4**). First, the expansion of amylase gene copy numbers in the ancestors of humans and Neanderthals raises the possibility that additional amylase gene copies may have initially become advantageous in response to a changing hominin diet (**Fig. 4A, B**). This hypothesis is supported by the recent evidence of Neanderthal starch consumption (Yates *et al.*, 2021) and perhaps further facilitated by the newfound availability of cooked starch made possible through the domestication of fire (Larbey *et al.*, 2019). Furthermore, it aligns well with the previous suggestion that the expansion of amylase copy number predates agriculture

(Mathieson and Mathieson, 2018). Our results offer a framework for scrutinizing ancient genomes to mine amylase structural variation data and to test specific hypotheses regarding the adaptive nature of amylase gene copy number variation across human evolution.

Second, different structural haplotypes exhibit distinct mutational propensities. For example, instead of stepwise duplications, our mutational model involving common 3-copy haplotypes suggests simultaneous duplications or deletions of both *AMY1A* and *AMY1B* genes, explaining the prevalence of odd-numbered *AMY1* copies in current human haplotypes (**Fig. 4A, C**). In contrast, the H1 haplotype features smaller, more divergent duplicated sequences than others, rendering it less susceptible to NAHR events. Conversely, the haplotypes with unidirectional duplicated segments are prone to recurrent NAHR leading to copy number variation, and haplotypes with inverted duplicated segments are prone to recurrent inversion events. As such, the mutation types and rates of the amylase locus and other complex loci may differ based on extant haplotype variation, especially in bottlenecked populations such as indigenous Americans (Moreno-Mayar *et al.*, 2018). It is a distinct possibility that a bottlenecked population ends up with very high allele frequencies of H1 due to drift. In this case, the absence of highly similar segments within H1 haplotype would mitigate recurrent inversion and copy number variation events, resulting in a slower accumulation of structural variation in this population. In contrast, if one of the larger amylase haplotypes were to become prevalent due to drift, the rate of structural variation would increase exponentially. Within this general context, one interesting question for future work is whether larger amylase haplotypes experience negative selection due to increased genomic instability.

Third, an attractive hypothesis is that the rapid evolution of structural variation led to exaptation, wherein selection for high copy number haplotypes in high-starch consuming agricultural populations (Perry *et al.*, 2007) acted on abundant standing variation rather than through the selection of novel variants (**Fig. 4A, D**). A more provocative argument would be that the effects of amylase gene duplications on taste preferences and starch metabolism may have facilitated the recently reported pre-agricultural consumption of agricultural grains in the Mesolithic Balkans (Cristiani *et al.*, 2021). Regardless, it was argued previously that classical sweeps that involve selection on *de novo* mutations are rare in humans (Hernandez *et al.*, 2011). Instead, multiple examples of soft sweeps that involve existing variants, including those underlying skin color, have been put forward (Crawford *et al.*, 2017; Martin *et al.*, 2017). Therefore, we suggest that comparative analyses with emphasis on soft sweeps in indigenous populations with recently diverged dietary cultures, such as indigenous Andeans (Lindo *et al.*, 2018; Jorgensen *et al.*, 2023), would provide an exciting next step for detecting signatures of selection at the human amylase locus.

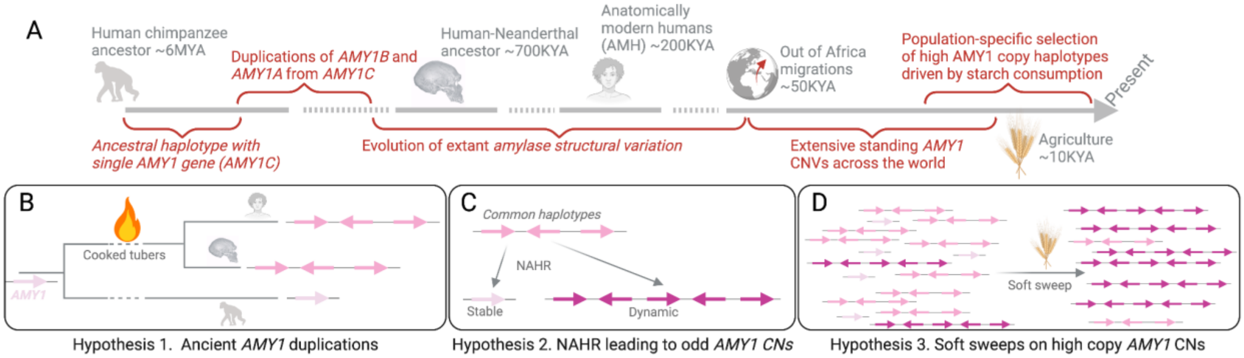
A
Human chimpanzee ancestor ~6MYA
Duplications of AMY1B and AMY1A from AMY1C
Human-Neanderthal ancestor ~700KYA
Anatomically modern humans (AMH) ~200KYA
Out of Africa migrations ~50KYA
Population-specific selection of high AMY1 copy haplotypes driven by starch consumption
Present
Ancestral haplotype with single AMY1 gene (AMY1C)
Evolution of extant amylase structural variation
Extensive standing AMY1 CNVs across the world
Agriculture ~10KYA
B
Cooked tubers
AMY1
Hypothesis 1. Ancient AMY1 duplications
C
Common haplotypes
NAHR
Stable
Dynamic
Hypothesis 2. NAHR leading to odd AMY1 CNs
D
Soft sweep
Hypothesis 3. Soft sweeps on high copy AMY1 CNs

Fig. 4. An evolutionary model of amylase gene and resulting hypotheses. **A.** A timeline of amylase locus evolution based on the results of this study. **B-D.** Specific hypotheses discussed in the Conclusion section.

Materials and Methods

Amylase Segments

In this study, we characterized a ~212.5 kbp region on chromosome 1 (GRCh38; chr1:103,554,220–103,766,732) that contains paralogous copies of SDs that have > 99% sequence similarity. The amylase segments, depicted by colored arrows, were modified from a previous study by Usher and colleagues (Usher *et al.*, 2015), and determined based on the labeling pattern from the Bionano Genomics OGM data and the sequence similarity between segments using blastn (BLASTN 2.9.0+) (Altschul *et al.*, 1990). First, using OGM labeling patterns of GRCh38 reference assembly, amylase segments were identified using pairwise alignments between smaller duplicons within SDs. Then, to obtain accurate coordinates of segments, the GRCh38 fasta sequence of the locus was aligned against itself using blastn (BLASTN 2.9.0+). The alignments with less than 99% sequence similarity and 60% coverage were filtered out and coordinates of all reference segments were finalized. We identified one copy of purple (45.7 kbp), two copies of green (5.3 kbp), one copy of red (12.4 kbp), two copies of orange (14.1 kbp), three copies of maroon (4.3 kbp), and three copies of pink (26.3 kbp) segments within the GRCh38 amylase locus (**Table S1**). Next, the GRCh38 segment sequences were used as a reference to detect amylase segments from each sample included in this study using blastn. Same sequence similarity (> 99%) and coverage thresholds (> 60%) were used to obtain the amylase segment coordinates from each sample. Amylase segments in nonhuman primates, namely: chimpanzee (Clint_PTRv2_panTro6), bonobo (Mhudiblu_PPA_v0_panPan3) and gorilla (Kamilah_GGO_v0_gorGor6), were detected by using the same approach described above with sequence similarity threshold of ³ 90% using genome reference assembly fasta files that were obtained from University of California Santa Cruz genome browser.

Sample Collection and Datasets

Samples (n = 98) from diverse populations were part of the Human Genome Structural Variation Consortium (HGSVC) (Ebert *et al.*, 2021), Human Pangenome Reference Consortium (HPRC) (Liao *et al.*, 2023) and Genome In a Bottle (GIAB) (HG002), and consists African (n = 43), American (n = 24), East Asian (n = 13), European (n = 9), and South Asian (n = 9) (**Table S2**) individuals. Bionano Genomics OGM and PacBio HiFi datasets of HGSVC samples were downloaded from http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/HGSVC2/working/ and http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/HGSVC3/working/, and HPRC and GIAB samples containing phased assemblies (PA) and OGM datasets were downloaded from https://s3-us-west-2.amazonaws.com/human-pangenomics/index.html?prefix=working/).

De novo assembly of optical genome map dataset

Optical genome maps of HGSVC and HPRC samples were assembled *de novo* and aligned to the GRCh38.p12 reference assembly using the Bionano Solve v3.5 (https://bionanogenomics.com/support/software-downloads/) assembly pipeline, with default settings as described previously (Cao *et al.*, 2014) (**Fig. S25**).

```
python2.7 Solve3.5.1_01142020/Pipeline/1.0/pipelineCL.py -T 64 -U
-j 64 -jp 64 -N 6 -f 0.25 -i 5 -w -c 3 \
-y \
-b ${bionano_bnx} \
-l ${output_dir} \
-t Solve3.5.1_01142020/RefAligner/1.0/ \
-a
Solve3.5.1_01142020/RefAligner/1.0/optArguments_haplotype_DLE1_s
aphyr_human.xml \
-r ${reference_genome}
```

A pairwise comparison of DNA molecules (min 250 kbp) was generated to produce the initial consensus genome maps. During an extension step, molecules were aligned to genome maps, and maps were extended based on the molecules aligning past the map ends. Overlapping genome maps were then merged. Extension and merge steps were repeated five times before a final refinement of the genome maps. Clusters of molecules aligned to genome maps with unaligned ends > 30 kbp in the extension step were re-assembled to identify all alleles. To identify alternate alleles with smaller size differences from the assembled allele, clusters of molecules aligned to genome maps with internal alignment gaps of size < 50 kbp were detected, and the genome maps were converted into two haplotype maps. The final genome maps were aligned to the reference genome, GRCh38.p12.

Haplotype detection and single molecule support using optical genome maps

To resolve haplotype structures at the amylase locus, *de novo* assemblies of OGMs of samples were first aligned to the GRCh38.p12 human genome reference assembly. Next, alignment results were visualized using the Bionano Access™ software to resolve amylase haplotypes. Haplotypes were identified from these visualizations using GRCh38 amylase segments and OGM labeling patterns as a guide. Next, molecule support for each haplotype was evaluated using the following steps (**Fig. S26**): contigs aligning to the region of interest (GRCh38; chr1:103,554,220–103,766,732) were detected. Any samples with greater than or less than two contigs aligning to the region of interest (ROI) were excluded. The remaining samples with two contigs aligning to ROI were included for single molecule support analysis. Next, ROIs in each contig were identified and extended 10 kbp on both proximal and distal regions. The number of molecules aligning to the ROI of the amylase locus was collected, and the average molecule length was calculated for each contig. If the average molecule length was less than the ROI, the number of molecules that were anchored in the proximal and distal unique regions were counted. Next, the length of the ROI was divided by the average molecule length, and new smaller ROIs were identified based on quotient, and the number of molecules was counted for each new ROI. Only haplotypes supported by three or more molecules were considered true haplotypes (**Data File**).

De novo assembly and orthogonal support using long-read sequencing dataset

PacBio HiFi long-read sequencing dataset from HGSVC (n = 60) were *de novo* assembled using hifiasm (Cheng *et al.*, 2021) using the following command:

```
hifiasm -o ${outputdir}${sampleID}.asm -t 32 ${reads}
```

Publicly available PA data from the HPRC were analyzed to validate the OGM-based haplotype structures using PA data, as described previously (Yilmaz *et al.*, 2023). Validation process includes the following steps: First, HPRC PA fasta files were converted to *in silico* maps using fa2cmap_multi_color.pl (Bionano Solve^{TM} v3.5.1). The resulting *in silico* maps were aligned to the GRCh38 using the refAligner tool (Bionano Solve^{TM} v3.5.1).

```
Solve3.5.1_01142020/RefAligner/1.0/RefAligner                -ref
${referenceFile} \ -maxthreads ${threads} \
-i ${inputfile} \
-o ${outputdir}EXP_REFINEFINAL1
```

Next, a visual pairwise comparison of HPRC OGM and PA assemblies for each sample was performed using Bionano Access^{TM}. Then, amylase segments in each sample with a haplotype consistent with the OGM haplotype were detected using blastn (BLASTN 2.9.0+), as described in the 'Amylase Segments' part. Finally, all orthogonally supported haplotype alignments were visualized using pygenomeviz (Shimoyama, 2022) to confirm haplotype structures.

Distribution of singleton numbers for genome-wide tandem repeats

To quantitatively compare the fraction of singletons in the amylase locus with the other fast-evolving loci, we downloaded genotypes of tandem repeats (TR) loci from (https://github.com/gymrek-lab/EnsembleTR) and selected diploid genotypes of individuals present in both the tandem repeats dataset and our dataset (n = 33). We then selected TR loci that have the same number of alleles (n = 21) as our data and have no missing genotypes. For the selected loci, the distribution of singletons was generated by counting the number of singletons for each locus.

Rarefaction analysis

Rarefunction analysis was performed to evaluate haplotype diversity, to check if adding more samples was going to increase the number of unique haplotypes, using R package called "vegan" ("Community Ecology Package [R package vegan version 2.6-4]," 2022) and the following command:

```
df <- read.table("rarefaction_allPopsCombined.txt", header = TRUE,
sep = "\t")[6,]
```

```
(raremax <- min(rowSums(df)))
Srare <- rarefy(df, raremax)
NumHap <- as.numeric(rarecurve(df, step = 1, sample = raremax, col
= "blue", cex = 1)[[1]])
Rate <- c()
for (i in 1:length(NumHap)){
  Rate[i+1] <- NumHap[i+1]-NumHap[i]
}
Plot_data <- data.frame(NumChr=1:length(NumHap), NumHap=NumHap,
Rate=Rate[-length(Rate)])
pdf("Rarefaction_AggregatedData.pdf", width=6, height=4)
twoord.plot(Plot_data$NumChr, Plot_data$NumHap, Plot_data$NumChr,
Plot_data$Rate, type="l", lcol = "blue", rcol = "red",
xlab="Number of chromosomes", ylab="Number of haplotypes",
rylab="Rate of change", lytickpos=seq(0,50,10))
dev.off()
```

Amylase gene annotation using protein sequence homology

To predict amylase gene location and copy number on each haplotype, we used homology-based gene prediction tool ("protein2genome") implemented in Exonerate v2.4.0 (Slater and Birney, 2005) with default settings except for maximum intron length 20kb (`--maxintron`

`20000`). For the homology search, we used the amylase protein sequences with experimental evidence from UNIPROT ("Evidence at protein level") (The UniProt Consortium, 2018) under the following accession numbers; *AMY2B*: P19961, *AMY2A*: P04746, *AMY1A*: P0DUB6, *AMY1B*: P0DTE7, and *AMY1C*: P0DTE8. We clustered any overlapping hits from the prediction results and selected the best hit from each cluster. The hits translated into full length amino acids (511 aa) were then kept as a final annotation. In cases where we found conflict between the homology-based annotation and the predicted copy number of *AMY* genes from amylase haplotypes, the homology-based prediction was manually curated to ensure it to be consistent with the optical mapping label. Additionally, genome annotations of *de novo* assemblies were obtained by using GRCh38 gff annotation file with liftoff (Shumate and Salzberg, 2020).

```
liftoff -db $dbfile -o $outputfile.gff -u $outputfile.unmapped -
dir $outputdir -p 8 -m $minimap2dir -sc 0.85 -copies $fastafile -
cds $refassembly
```

The liftoff results were compared with homology-based annotation above and were used when the manual curation was not able to produce annotation consistent with the optical mapping label where we found only one case (HG01175_h2_H30-7.1.1).

Validation of amylase copy number using digital droplet PCR

Custom primer and probe oligos were designed to target a conserved sequence in exon 10th across all annotated amylase gene copies (including annotated pseudogenes) in the T2T-chm13 human reference genome, with the following sequences: forward primer (TTCCGCAATGTAGTGGAT GG), reverse primer (AATGAATCCTCTGTTTCCTCTCC), and probe (TGATAATGGGAGCAACCAAGTGGCT). The ddPCR protocol, reference primer, and reagents used are as previously described in (Pajic *et al.*, 2019). Thirteen human DNA samples (Coriell Institute) were estimated for total amylase gene copies in duplicates. Restriction enzyme *HindIII* (New England Biolabs), was used in the suggested concentration to digest DNA prior to ddPCR subjugation.

Alignments of amylase segments and coding sequences

The alignments of each amylase segment and coding sequences (CDS) were constructed using MAFFT v7.310 (Katoh and Standley, 2013) and PAGAN v1.53 (Löytynoja, Vilella and Goldman, 2012) with `--codon` option, respectively. As an outgroup, chimpanzee sequences for each segment were prepared by searching a sequence that covers ³ 30% of corresponding segment of GRCh38 reference genome with ³ 97% sequence similarity in mPanTro3 genome from the Vertebrate Genome Project (https://www.genomeark.org/t2t-all/Pan_troglodytes.html). Only one of the haplotypes ("haplotype 1") from mPanTro3 was used to align segments. We also generated alignment of the 3' end region (from 22,850bp to end in the pink segment alignment, **Data File**) of pink segment from human and non-human primates; chimpanzee (Clint_PTRv2_panTro6), bonobo (Mhudiblu_PPA_v0_panPan3) and gorilla (Kamilah_GGO_v0_gorGor6) using MAFFT v7.310 (Katoh and Standley, 2013).

The raw alignments of each segment were summarized by averaging frequencies of major alleles (including a gap) across a 200 bp window with a 10 bp step (**Fig. S8**). The CDSs of *AMY2B*, *AMY2A*, and *AMY1* for chimpanzees were retrieved from the Ensembl annotation of panTro5 (release 109) (Martin *et al.*, 2023). Additionally, as an outgroup to all amylase CDSs, we incorporated amylase CDS from sheep genome, Oar_rambouillet_v1.0 (release 109) (Martin *et al.*, 2023) that is known to have only one copy of amylase gene (Pajic *et al.*, 2019).

We found that *AMY2B*, *AMY2A*, and *AMY1* genes had 23, 23, and 37 coding sequence variations, which resulted in 6, 11, and 20 gene-specific amino acid changes, respectively. In addition, we identified 10, 12, and 19 coding sequence variations within *AMY2B, AMY2A,* and *AMY1* genes, respectively (**Table S5**). However, 50 (5/10), 25 (3/12), and 26 (5/19) % of these variations were synonymous, and of the remaining nonsynonymous variations, 80% (4/5), 33% (3/9), and 64% (9/14) were found only in one gene copy.

Estimation of dN/dS for amylase genes

To estimate dN/dS value, we used codeml implemented in PAML v4.10.6 package (Yang, 2007). Orthology detection is difficult in this locus given the lineage-specific gene duplications and copy number variation in humans. Thus, we separately compared CDSs of *AMY2B, AMY2A,* and *AMY1* with those of *AMY2B, AMY2A,* and *AMY1* in the chimpanzee assembly (panTro5), assuming that chimpanzee copies are proxies to ancestral copies to subsequent human duplications. We used the pairwise comparison option (`runmode = -2`) in codeml. To obtain statistical significance, we estimated the dN/dS value for the null model where omega is fixed to one and performed a χ^2 test with the likelihood ratio between the null and alternative models. Any estimates with dS <= 0.01 and dS ³ 2, and dN/dS ³ 10 were filtered out in the result, and p-values were adjusted by Benjamini-Hochberg procedure (Benjamini and Hochberg, 1995).

Phylogenetic reconstruction of pink segments and coding sequences

We used pink segment, 3’ end region of pink segment and CDS alignments (see “*Alignments of amylase segments and coding sequences*”) to phylogenetically cluster sequences.

A maximum-likelihood tree for each alignment was reconstructed using IQ-TREE v2.2.0 (Minh *et al.*, 2020) with the following options: `-m MFP` (Kalyaanamoorthy *et al.*, 2017), `-alrt 1000`, `-B 1000` (Hoang *et al.*, 2018). The chimpanzee pink segment, and CDSs from chimpanzee and sheep amylase genes were used as outgroups for the pink segment and CDS, respectively. The tree for the 3' end region of the pink segment was rooted by midpoint. All the reconstructed trees were visualized using FigTree v.1.4.4 (http://tree.bio.ed.ac.uk/software/figtree/).

Detecting sequence origin and breakpoints in pink segments

To explore the origin of sequences within the pink segment, we first generated consensus sequences for each pink segment cluster by selecting the major allele at each site in the alignment. We then prepared a pseudo reference sequence including only a single copy amylase segment; purple, green, red, orange, maroon, and pink segment (Pink1C) from the GRCh38 reference genome. For each 200bp sliding window with 10 steps in each cluster, we searched for the best hit against the pseudo reference sequence using BLAT v. 37x1 (Kent, 2002), and determined it as the origin of the 200 bp sequence. The windows where more than or equal to 50% of the sequence is gap were excluded in this analysis. The breakpoints of possible NAHR events were manually determined around the transition of origins (**Table S9**).

To visualize homology between the consensus sequences from each pink segment cluster and the other segments in the context of haplotype, we aligned the sequences using nucmer v3.1 and generated dotplots using mummerplot v3.5 (Marçais *et al.*, 2018).

Dating AMY1 duplication events

To estimate the timing of *AMY1* duplication events, we first constructed a starting tree using the whole pink segment alignment above (see the "Alignments of segments and coding sequences" section) with IQTREE v2.2.0 (Minh *et al.*, 2020) and `--alrt 1000 -B 1000` options. We then trimmed out the 3' end of the alignment (from 22,850bp to end in the pink segment alignment, **Data File**) that is potentially poorly aligned due to the presence of rearrangement breakpoints. Exonerate v2.4.0 (Slater and Birney, 2005) with maximum intron length 20 kbp (`--maxintron 20000`) was used to determine coordinates of the coding region in the alignment and to ensure that all sequences in the alignment have the same coordinates for the coding region. Based on the annotation, we partitioned the alignment into four categories: non-coding region, 1st position of codon, 2nd position of codon, and 3rd position of codon. This partitioned alignment was used as an input for the downstream analyses. We extracted sequences and topology of H5 or H6 haplotypes and chimpanzee reference genome from the partitioned alignment and the starting tree, respectively. We restricted our analyses to H5 or H6 haplotype that is likely to be ancestral (due to their high allele frequency) to other more complex haplotypes to reduce the genetic exchange effect of inversion between different pink segments.

For estimating the timing of *AMY1* duplication events (that is, pink segment duplication events), BEAST v2.7.5 (Bouckaert *et al.*, 2014), a Bayesian framework, was used with the alignments and starting trees of H5 and H6 haplotypes separately. First, we performed bModel Test implemented in the BEAST to assess whether the four partitions are evolving differently or not using unliked site models and calibrated Yule model. Based on the results of bModel Test, we decided to use average substitution models using bModel Test in the real run to incorporate heterogeneity across partitions. Then, a path sampling approach from BEAST's

MODEL_SELECTION package was used for all model combinations to determine best-fit clock and tree models. We tested both strict and optimized relaxed clock models with constant and Bayesian skyline tree models. For the path sampling, 40 path steps with 25,000,000 iterations were used with an alpha parameter of 0.3, pre-burn-in of 75,000 iterations, and an 80% burn-in of the complete chain. While the models with a strict clock model could not produce marginal likelihood estimates, there was no significant difference among the models with an optimized relaxed clock model (bayes factor = 1.32 and 0.14 for H5 and H6; **Table S7**). Therefore, we estimated time for both models with Markov chain Monte Carlos (MCMC) with 75,000,000 iterations, sampling every 2,000 trees and using a burn-in of 10% iterations for each model. In all the time estimations above, MRCA prior to assuming the human chimpanzee divergence time to 6 million years ago (MYA) under normal distribution (a mean of 6.0 MYA and variance of 0.5) was used to calibrate time estimates. Also, we did not allow optimization of the topology of the starting tree during the estimation to avoid shuffling between pink segment clusters likely due to observed gene conversion. Tracer v1.7.2 was used to examine effective sample size (ESS) and convergence of estimates. Treeannotator v2.7.5 was used to merge and annotate the sampled trees into a single tree. Figtree v1.4.4 (http://tree.bio.ed.ac.uk/software/figtree/) was used for tree visualization.

Identification of breakpoints in archaic hominin genomes

To explore whether the initial duplications of *AMY1* genes were ancestral to archaic hominins and humans, we first selected 31mers that are unique in the consensus sequence of each pink segment cluster and kept 31mers that are exclusively found in one cluster. We then searched the remaining 31mers against GRCh38 and T2T reference genome using BLAT v. 37x1 (Kent, 2002) with the following settings to increase sensitivity; `-stepSize=5`, `-`

`repMatch=1000000, -minScore=20 -minIdentity=0`. The 31mers without any hits were kept ensuring they were unique at the genome-wide level. We prioritized 31mers by the number of SNVs in the pink segment alignment and selected 3 (Pink1C) and 2 (Pink1A) 31mers to distinguish pink segment clusters (**Table S8**). As expected, the two 31mers of Pink1A span the breakpoints that separate purple-red and red-maroon segments in Pink1A (**Fig. 3A**; Breakpoints #3 and #4), respectively. For the Pink1B, we do not expect to find unique 31mers as the breakpoint separating pink-purple segments is shared between Pink1B and Pink1A. Thus, we manually selected a 31mer spanning the breakpoint (**Table S8**). The six 31mers were mapped to the GRCh38 reference genome, and the perfectly matched regions were used to determine the presence of the breakpoints.

We collected raw reads from archaic hominin genomes (PRJEB1265: Altai Neanderthal, PRJEB21157: Vindija Neanderthal, and PRJEB3092: Denisovan), removed adaptor sequences, and merged paired-end reads in case of overlapping using leeHom v1.2.17 (Renaud, Stenzel and Kelso, 2014). To increase sensitivity, we did not trim the reads by their base quality scores. For Chagyrskaya Neanderthal genome, we used unmapped reads (http://ftp.eva.mpg.de/neandertal/Chagyrskaya/rawBAM/) that were already merged, and adapter trimmed. The clean reads were mapped to the GRCh38 reference genome, which includes decoy and HLA sequences (http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/reference/GRCh38_reference_genome/GRCh38_full_analysis_set_plus_decoy_hla.fa) using bwa aln v0.7.17 (Li and Durbin, 2009) with the following options: `-n 0.01 -o 2 -l 16500`. We investigated the 31mers spanning breakpoints to determine whether the archaic hominin genome has the same breakpoints or not.

To ensure that our approach works reasonably, we estimated the copy number of three distinct pink segments using short reads from the same individual used to reconstruct our amylase haplotype. We downloaded mapped bam files for the individuals that are available in both our diploid haplotype data and 1000 genomes project short-reads data from The International Genome Sample Resource (https://www.internationalgenome.org). The average depths for the 31mer regions were calculated and normalized by genome-wide depth using mosdepth v0.3.5 (Pedersen and Quinlan, 2018). Using the normalized depths, we estimated copy numbers of the pink segment, which significantly correlates with the copy numbers in our amylase haplotypes (**Table S10**). In addition, we profiled the DNA damage of reads mapped to the 31mer regions to examine that the reads derived from ancient sources using DamageProfiler v1.1 (Neukamm, Peltzer and Nieselt, 2021). The profile showed a typical pattern of ancient DNA with elevated C to T or G to A mismatch frequency in both read ends (**Fig. S27**).

Scanning PRDM9 binding sites across the H5 haplotype

To identify potential PRDM9 binding sites, FIMO v5.5.4 (Bailey *et al.*, 2015) implemented in the MEME package was employed using the degenerate 13-bp motif, "CCNCCNTNNCCNC" (Myers *et al.*, 2010), and the non-redundant DNA database background frequency matrix in the H5 amylase locus. For the purposes of the analysis, we generated a consensus sequence for the H5 haplotype to incorporate the standing nucleotide variation within the locus. Only hits with a FIMO score above *10* and a p-value below *0.00011* were considered (**Fig. S18)**.

Reconstruction of mutational connections among extant haplotypes

We constructed the mutational connections between extant haplotypes in a stepwise fashion starting from the common haplotypes *e*.g. H1, H5, H6, and H8, which make up for 70% of the haplotypes detected in our dataset. Common haplotypes provide a reasonable starting point as they are older and more likely to seed subsequent mutations (Albers and McVean, 2020).

Inversions among H5-H6-H8: The dot plots among H5, H6 and H8 show that a single inversion event can explain the structural differences among these haplotypes. Thus, a parsimonious explanation is that H6 is an intermediate haplotype between H5 and H8 (**Fig. S28)**. The breakpoints between H5 and H6 involve two inverted maroon segments that would serve as the substrates for NAHR (99.97%-100% identity and 4,355 bp in size) (**Fig. S29**). The breakpoints between H6 and H8 involve inverted > 15 kbp segments spanning red (partially) and orange segments (**Fig. S28**).

Emergence of H1: When compared to other common haplotypes, H1 differs from them by the absence of Pink1-Pink2 segments (**Fig. S30**). The boundaries of this deletion lie within two near identical orange segments with 99.82% similarity. Thus, we argue that an NAHR-driven deletion may explain the emergence of H1 from a haplotype that harbors these two duplications. We base this hypothesis on two observations. First, we did not find any evidence for MMBIR involving these rearrangement breakpoints among H1 and other common haplotypes. Second, we showed earlier that haplotypes that harbor three pink segments have already existed prior to Human-Neanderthal divergence (**Fig. 3C**). Thus, haplotypes that harbor three pink and two orange segments, including the common haplotypes H5, H6, and H8 are all reasonable substrates for a deletion that leads to emergence of H1.

NAHR-driven inversion and copy number variation events that involve non-common haplotypes: To incorporate the haplotypes that are not explained with a single mutational step starting from common haplotypes into the network above, we searched two non-common haplotypes that are separated by single inversion or copy number differences with breakpoints in duplicated segments: two inverted segments leading to an inversion (e.g. **Fig. 3E**), and two duplicated segments in the same direction leading to copy number variation (e.g. **Fig. 3F**). With this, we compared haplotypes to each other using pairwise dotplots to identify inversion or single copy number gain/loss events that involve two breakpoints overlapping duplicated segments. This allowed us to identify several putative mutational mechanisms (**Fig. S31**): **a.** NAHR mediated by inverted maroon segments, connecting H25 to H26; **b.** NAHR mediated by inverted maroon segments, connecting H10 to H11; **c.** NAHR mediated by purple segments between H4 and H4 leading to copy number gains resulting in H14; **d.** NAHR mediated by unidirectional purple segments between H11 and H14 leading to copy number gain resulting in H19; **e.** NAHR mediated by unidirectional orange segments between H11 and H10 leading to copy number gain resulting in H29, and subsequent NAHR involving two sets of inverted maroon segments between two H29 haplotypes leading to two inversions, resulting in H30; **f.** NAHR mediated by unidirectional green segments between H9 and H8 leading to copy number variation, resulting in H3 and H18 (See **Supplementary Results**).

By combining all the mutational connections described above, we constructed a putative mutational network (**Fig. S19**). The resulting network, despite being not definitive, provides a useful starting point for understanding main mechanisms leading to haplotypic diversity in our amylase haplotypes.

Investigating mechanisms underlying structural variation formation within haplotypes

To investigate a putative mechanism underlying SVs between two haplotypes connected in the mutational network above, we aligned the haplotypes using NUCmer v3.1 (Kurtz *et al.*, 2004). The alignments (**Data File**) were visualized as dotplots using mummerplot v3.5 and Miropeats-style plots using the R package SVbyEye (https://github.com/daewoooo/SVbyEye/tree/master) (e.g., **Fig. S29**). These sequence comparisons allowed us to identify the potential breakpoints of SVs. Once we identified the potential breakpoints, we extracted 20 kbp long sequences upstream and downstream flanking regions and then aligned them to each other, including the corresponding reference sequence using MAFFT v7.522 (Katoh and Standley, 2013). Based on the analysis of these breakpoints, we evaluated replication and DNA recombination-based processes that underlie the mutational steps leading to the extant SVs (**Fig. S32**). We inferred NAHR when the variations between two structural haplotypes included (i) breakpoints within paralogous segments and (ii) absence of sequence motifs linked to replication-based mutational mechanisms. Conversely, if we identified those sequence motifs in two non-paralogous segments, we presumed a MMBIR mechanism.

References

Albers, P. K. and McVean, G. (2020) “Dating genomic variants and shared ancestry in population-scale sequencing data,” *PLoS biology*, 18(1), p. e3000586.

Altschul, S. F. *et al.* (1990) “Basic local alignment search tool,” *Journal of molecular biology*, 215(3), pp. 403–410.

Bailey, T. L. *et al.* (2015) “The MEME Suite,” *Nucleic acids research*, 43(W1), pp. W39-49.

Benjamini, Y. and Hochberg, Y. (1995) “Controlling the false discovery rate: A practical and powerful approach to multiple testing,” *Journal of the Royal Statistical Society*. Wiley, 57(1), pp. 289–300.

Bouckaert, R. *et al.* (2014) “BEAST 2: a software platform for Bayesian evolutionary analysis,” *PLoS computational biology*, 10(4), p. e1003537.

Cao, Hongzhi *et al.* (2014) “Rapid detection of structural variation in a human genome using nanochannel-based genome mapping technology,” *GigaScience*. Oxford University Press (OUP), 3(1). doi: 10.1186/2047-217x-3-34.

Carpenter, D. *et al.* (2015) “Obesity, starch digestion and amylase: association between copy number variants at human salivary (AMY1) and pancreatic (AMY2) amylase genes,” *Human molecular genetics*, 24(12), pp. 3472–3480.

Cheng, H. *et al.* (2021) “Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm,” *Nature methods*, 18(2), pp. 170–175.

“Community Ecology Package [R package vegan version 2.6-4]” (2022). Comprehensive R Archive Network (CRAN). Available at: https://CRAN.R-project.org/package=vegan (Accessed: October 18, 2022).

Crawford, N. G. *et al.* (2017) “Loci associated with skin pigmentation identified in African populations,” *Science*, 358(6365). doi: 10.1126/science.aan8433.

Cristiani, E. *et al.* (2021) “Wild cereal grain consumption among Early Holocene foragers of the Balkans predates the arrival of agriculture,” *eLife*, 10. doi: 10.7554/eLife.72976.

Ebert, P. *et al.* (2021) “Haplotype-resolved diverse human genomes and integrated analysis of structural variation,” *Science*, 372(6537). doi: 10.1126/science.abf7117.

Falchi, M. *et al.* (2014) “Low copy number of the salivary amylase gene predisposes to obesity,” *Nature genetics*, 46(5), pp. 492–497.

Groot, P. C., Mager, W. H. and Frants, R. R. (1991) “Interpretation of polymorphic DNA patterns in the human alpha-amylase multigene family,” *Genomics*, 10(3), pp. 779–785.

Hernandez, R. D. *et al.* (2011) “Classic selective sweeps were rare in recent human evolution,” *Science*, 331(6019), pp. 920–924.

Hoang, D. T. *et al.* (2018) “UFBoot2: Improving the Ultrafast Bootstrap Approximation,” *Molecular biology and evolution*, 35(2), pp. 518–522.

Innan, H. and Kondrashov, F. (2010) “The evolution of gene duplications: classifying and distinguishing between models,” *Nature reviews. Genetics*, 11(2), pp. 97–108.

Jorgensen, K. *et al.* (2023) “Genetic adaptations to potato starch digestion in the Peruvian Andes,” *American journal of biological anthropology*. Wiley, 180(1), pp. 162–172.

Kalyaanamoorthy, S. *et al.* (2017) "ModelFinder: fast model selection for accurate phylogenetic estimates," *Nature methods*, 14(6), pp. 587–589.

Katoh, K. and Standley, D. M. (2013) "MAFFT multiple sequence alignment software version 7: improvements in performance and usability," *Molecular biology and evolution*, 30(4), pp. 772–780.

Kent, W. J. (2002) "BLAT--the BLAST-like alignment tool," *Genome research*, 12(4), pp. 656–664.

Kurtz, S. *et al.* (2004) "Versatile and open software for comparing large genomes," *Genome biology*, 5(2), p. R12.

Larbey, C. *et al.* (2019) "Cooked starchy food in hearths ca. 120 kya and 65 kya (MIS 5e and MIS 4) from Klasies River Cave, South Africa," *Journal of human evolution*, 131, pp. 210–227.

Li, H. and Durbin, R. (2009) "Fast and accurate short read alignment with Burrows-Wheeler transform," *Bioinformatics* , 25(14), pp. 1754–1760.

Liao, W.-W. *et al.* (2023) "A draft human pangenome reference," *Nature*. Springer Science and Business Media LLC, 617(7960), pp. 312–324.

Lindo, J. *et al.* (2018) "The genetic prehistory of the Andean highlands 7000 years BP though European contact," *Science advances*, 4(11), p. eaau4921.

Löytynoja, A., Vilella, A. J. and Goldman, N. (2012) "Accurate extension of multiple sequence alignments using a phylogeny-aware graph algorithm," *Bioinformatics* , 28(13), pp. 1684–1691.

Marçais, G. *et al.* (2018) "MUMmer4: A fast and versatile genome alignment system," *PLoS computational biology*, 14(1), p. e1005944.

Martin, A. R. *et al.* (2017) "An Unexpectedly Complex Architecture for Skin Pigmentation in Africans," *Cell*, 171(6), pp. 1340-1353.e14.

Martin, F. J. *et al.* (2023) "Ensembl 2023," *Nucleic acids research*, 51(D1), pp. D933–D941.

Mathieson, S. and Mathieson, I. (2018) "FADS1 and the Timing of Human Adaptation to Agriculture," *Molecular biology and evolution*, 35(12), pp. 2957–2970.

Meisler, M. H. and Ting, C. N. (1993) "The remarkable evolutionary history of the human amylase genes," *Critical reviews in oral biology and medicine: an official publication of the American Association of Oral Biologists*, 4(3–4), pp. 503–509.

Minh, B. Q. *et al.* (2020) "IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era," *Molecular biology and evolution*, 37(5), pp. 1530–1534.

Moreno-Mayar, J. V. *et al.* (2018) “Terminal Pleistocene Alaskan genome reveals first founding population of Native Americans,” *Nature*, 553(7687), pp. 203–207.

Myers, S. *et al.* (2010) “Drive against hotspot motifs in primates implicates the PRDM9 gene in meiotic recombination,” *Science*, 327(5967), pp. 876–879.

Neukamm, J., Peltzer, A. and Nieselt, K. (2021) “DamageProfiler: fast damage pattern calculation for ancient DNA,” *Bioinformatics* , 37(20), pp. 3652–3653.

Nurk, S. *et al.* (2021) “The complete sequence of a human genome,” *Science*. doi: 10.1101/2021.05.26.445798.

Pajic, P. *et al.* (2019) “Independent amylase gene copy number bursts correlate with dietary preferences in mammals,” *eLife*, 8. doi: 10.7554/eLife.44628.

Pedersen, B. S. and Quinlan, A. R. (2018) “Mosdepth: quick coverage calculation for genomes and exomes,” *Bioinformatics* , 34(5), pp. 867–868.

Perry, G. H. *et al.* (2007) “Diet and the evolution of human amylase gene copy number variation,” *Nature genetics*, 39(10), pp. 1256–1260.

Peyrot des Gachons, C. and Breslin, P. A. S. (2016) “Salivary Amylase: Digestion and Metabolic Syndrome,” *Current diabetes reports*, 16(10), p. 102.

Poole, A. C. *et al.* (2019) “Human Salivary Amylase Gene Copy Number Impacts Oral and Gut Microbiomes,” *Cell host & microbe*, 25(4), pp. 553-564.e7.

Porubsky, D. *et al.* (2022) “Recurrent inversion polymorphisms in humans associate with genetic instability and genomic disorders,” *Cell*, 185(11), pp. 1986-2005.e26.

Prüfer, K. *et al.* (2014) “The complete genome sequence of a Neanderthal from the Altai Mountains,” *Nature*, 505(7481), pp. 43–49.

Ramasubbu, N., Ragunath, C. and Mishra, P. J. (2003) “Probing the role of a mobile loop in substrate binding and enzyme activity of human salivary amylase,” *Journal of molecular biology*, 325(5), pp. 1061–1076.

Renaud, G., Stenzel, U. and Kelso, J. (2014) “leeHom: adaptor trimming and merging for Illumina sequencing reads,” *Nucleic acids research*, 42(18), p. e141.

Samuelson, L. C. *et al.* (1990) “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(6), pp. 2513–2520.

Sharp, A. J. *et al.* (2006) “Discovery of previously unidentified genomic disorders from the duplication architecture of the human genome,” *Nature genetics*, 38(9), pp. 1038–1042.

Shimoyama, Y. (2022) "pyGenomeViz: A genome visualization python package for comparative genomics." Jun.

Shumate, A. and Salzberg, S. L. (2020) "Liftoff: accurate mapping of gene annotations," *Bioinformatics* . doi: 10.1093/bioinformatics/btaa1016.

Slater, G. S. C. and Birney, E. (2005) "Automated generation of heuristics for biological sequence comparison," *BMC bioinformatics*, 6, p. 31.

Sudmant, P. H. *et al.* (2015) "An integrated map of structural variation in 2,504 human genomes," *Nature*, 526(7571), pp. 75–81.

Thamadilok, S. *et al.* (2020) "Human and Nonhuman Primate Lineage-Specific Footprints in the Salivary Proteome," *Molecular biology and evolution*, 37(2), pp. 395–405.

The UniProt Consortium (2018) "UniProt: a worldwide hub of protein knowledge," *Nucleic acids research*. Oxford Academic, 47(D1), pp. D506–D515.

Ting, C. N. *et al.* (1992) "Endogenous retroviral sequences are required for tissue-specific expression of a human salivary amylase gene," *Genes & development*, 6(8), pp. 1457–1465.

Usher, C. L. *et al.* (2015) "Structural forms of the human amylase locus and their relationships to SNPs, haplotypes and obesity," *Nature genetics*. Springer Science and Business Media LLC, 47(8), pp. 921–925.

Vollger, M. R. *et al.* (2023) "Increased mutation and gene conversion within human segmental duplications," *Nature*, 617(7960), pp. 325–334.

Yang, Z. (2007) "PAML 4: phylogenetic analysis by maximum likelihood," *Molecular biology and evolution*, 24(8), pp. 1586–1591.

Yates, J. A. F. *et al.* (2021) "The evolution and changing ecology of the African hominid oral microbiome," *Proceedings of the National Academy of Sciences*, 118(20), p. e2021655118.

Yilmaz, F. *et al.* (2023) "High level of complexity and global diversity of the 3q29 locus revealed by optical mapping and long-read sequencing," *Genome medicine*, 15(1), p. 35.

Acknowledgements

We thank Sabriya Syed, Valerie Reyes-Ortiz, Nicholas Moskwa, Pille Hallast, Stefan Ruhl, Victor Albert, Vincent Lynch, and Derek Taylor for technical help, discussions, suggestions and feedback on the manuscript; the Human Genome Structural Variation Consortium and the Human Pangenome Reference Consortium for making their data publicly available; the Scientific Services at the Jackson Laboratory including the Genome Technologies Service for their expert assistance with the work described herein and Research IT for providing computational infrastructure and support. We are grateful to the people who generously contributed samples to the 1000 Genomes Project.

Funding

C.L., and F.Y. were supported by the National Institute of Health (NIH) U24HG007497; K.K. was supported by the Jackson Laboratory Postdoctoral Scholar Award; OG. was supported by the National Science Foundation (NSF) Awards # 2049947 and #2123284.

Authors Contributions

O.G. and C.L. conceived the study. F.Y. performed the analysis and interpretation of the Human Genome Structural Variation Consortium (HGSVC), and Human Pangenome Reference Consortium (HPRC) Bionano Genomics optical mapping, PacBio HiFi sequencing and phased assemblies. F.Y. performed haplotype-resolved assemblies of HGSVC PacBio HiFi samples using hifiasm. F.Y. conducted amylase haplotype detection using HGSVC, and HPRC datasets. C.K. performed the mutational mechanisms analyses, breakpoint characterization. C.K. Contributed to evolutionary and functional analysis. K.K. performed gene annotation, selection/phylogenetic analyses of amylase coding sequences/segments, interpretation/time estimation of initial *AMY1* duplication events, and archaic hominin genome processing. P.P. performed ddPCR validation experiments and analysis of functional site differences in amylase amino acid sequences. F.Y., C.K., K.K., P.P., A-M.T., C.L. and O.G. drafted and critically revised the article. Final approval of the version to be published was given by F.Y., C.K., K.K., P.P., A-M.T, C.L. and O.G. All authors read and approved the final manuscript.

Competing Interests

C.L. is a scientific advisory board member of Nabsys and Genome Insight.

Supplementary Materials

Supplementary Results

Supplementary Materials and Methods

Figs. S1-S34

Tables S1 to S10

References