Global diversity, recurrent evolution, and recent selection on amylase structural haplotypes in humans

Davide Bolognini[1,*], Alma Halgren[2,*], Runyang Nicolas Lou[2,*], Alessandro Raveane[1,*], Joana L. Rocha[2,*], Andrea Guarracino[1,3,], Nicole Soranzo[1], Jason Chin[5], Erik Garrison[3,†], Peter H Sudmant[2,4,†]

1. Human Technopole, Milan, Italy
2. Department of Integrative Biology, University of California Berkeley, Berkeley, USA
3. Department of Genetics, Genomics, and Informatics, University of Tennessee Health Science Center, Memphis, USA
4. Center for Computational Biology, University of California Berkeley, Berkeley, USA
5. Foundation for Biological Data Science, Belmont, CA, USA

*alphabetical order - these authors contributed equally

†to whom correspondence should be addressed

Abstract

The adoption of agriculture, first documented ~12,000 years ago in the Fertile Crescent, triggered a rapid shift toward starch-rich diets in human populations. Amylase genes facilitate starch digestion and increased salivary amylase copy number has been observed in some modern human populations with high starch intake, though evidence of recent selection is lacking. Here, using 52 long-read diploid assemblies and short read data from ~5,600 contemporary and ancient humans, we resolve the diversity, evolutionary history, and selective impact of structural variation at the amylase locus. We find that both salivary and pancreatic amylase genes have higher copy numbers in populations with agricultural subsistence compared to fishing, hunting, and pastoral groups. We identify 28 distinct amylase structural architectures and demonstrate that identical structures have arisen independently multiple times throughout recent human history. Using a pangenome graph-based approach to infer structural haplotypes across thousands of humans, we identify extensively duplicated haplotypes present at higher frequencies in modern agricultural populations. Leveraging 534 ancient human genomes we find that duplication-containing haplotypes have increased in frequency more than seven-fold over the last 12,000 years providing evidence for recent selection in Eurasians at this locus comparable in magnitude to that at lactase. Together, our study highlights the strong impact of the agricultural revolution on human genomes and the importance of long-read sequencing in identifying signatures of selection at structurally complex loci.

Main

Dietary changes have played a major role in human adaptation and evolution impacting phenotypes such as lactase persistence[1,2] and polyunsaturated fatty acid metabolism[3–5]. One of the most substantial recent changes to the human diet is the shift from hunter-gatherer societies to agricultural-based subsistence. The earliest instance of crop domestication can be traced to the fertile crescent of South Western Asia ~12 kya laying the foundation for the Neolithic revolution[6]. Agriculture subsequently spread rapidly Westward into Europe by way of Anatolia by ~8.5kya and Eastward into the Indian subcontinent. Transitions to agriculture-based subsistence have happened independently several times throughout human history and today the overwhelming majority of carbohydrates consumed by humans are derived from agriculture.

Plant-based diets are rich in starches which are broken down into simple sugars by α-amylase enzymes in mammals. Human genomes contain three different amylase genes located proximally to one another at a single locus: *AMY1,* which is expressed exclusively in salivary glands, and *AMY2A* and *AMY2B,* which are expressed exclusively in the pancreas. It has long been appreciated however, that the amylase locus exhibits extensive structural variation in humans[7,8] with all three genes exhibiting copy number variation. All other great apes harbor just a single copy each of the *AMY1, AMY2A,* and *AMY2B* genes[9]. This ancestral single copy state has also been reported in Neanderthals and Denisovans[10]. *AMY1* copy number correlates with salivary amylase protein levels in humans, and an analysis of seven human populations found increased *AMY1* copy number in groups with high starch diets[11]. While it has been proposed that this gene expansion may have been an adaptive response to the transition from hunter-gatherer to agricultural societies, evidence of recent selection at this locus has been lacking[10,12]. Moreover, subsequent analyses identifying a putative association of *AMY1* copy number and BMI[13] failed to replicate[14], highlighting the challenges associated with studying structurally variable loci which are

often poorly tagged by nearby Single Nucleotide Polymorphisms (SNPs)[15]. One major challenge in characterizing selective signatures at structurally complex loci is the difficulty of phasing copy numbers onto haplotypes. Furthermore, while the human reference genome contains a single fully resolved amylase haplotype, the sequence, structure, and diversity of haplotypes on which different copy numbers have emerged is unknown.

Worldwide distribution of amylase diversity and increased copy number in traditionally agricultural societies

While extensive copy number variation has been documented at the amylase locus in humans[10,13,14,16], sampling of human diversity worldwide has been incomplete. To explore diversity at this locus we compiled 4,292 diverse high-coverage modern genomes from several sources[17–19] and used read-depth based approaches (see methods, **Fig S1**) to estimate diploid copy number in 162 different human populations (**Figs 1A-C, Extended Data Fig 1, Table S1**). Diploid *AMY1* copy number estimates ranged from 2-20 and were highest in populations from Oceanic, East Asian, and South Asian subcontinents. Nevertheless, individuals carrying high *AMY1* copy numbers were present in all continental subgroups. *AMY2A* (0-6 copies) showed the highest average copy number in African populations with deletions more prevalent in non-African populations. *AMY2B* (2-7 copies) exhibited high population stratification with duplications essentially absent from Central Asian/Siberian, East Asian, and Oceanic populations. We also assessed three high coverage Neanderthals and a single Denisovan individual, confirming all to have the ancestral copy number state (**Extended Data Fig 1**). Thus, copy number variation across all three amylase genes is human specific.

While *AMY1* copy number has been shown to exhibit a strong positive correlation with salivary protein levels[11,20], the relationship between pancreatic amylase gene expression and copy number has not been assessed. Analyzing GTEx[21] data we confirmed *AMY2A* and *AMY2B* expression was confined to the pancreas. We then genotyped diploid copy numbers in 305 samples for which expression data was available alongside high coverage genome sequencing. Both *AMY2A* (0-5 copies) and *AMY2B* (2-5 copies) copy numbers were significantly and positively correlated with gene expression levels (P=4.4e-5 and P=6.5e-4 respectively, linear model, **Fig S2**).

The strongest evidence of potential selection at the amylase locus comes from comparisons of seven modern day populations with high versus low starch intake[11]. We identified 273 individuals from 31 different populations with traditionally agricultural-, hunter-gatherer-, fishing-, or pastoralism-based diets in our dataset (**Table S2**). The copy number of all three amylase genes was significantly higher in populations with agricultural subsistence compared to those from fishing, hunting, and pastoral groups (**Fig 1D**, adjusted P=0.001, 0.035, and 0.0077 for *AMY1*, *AMY2A*, and *AMY2B* respectively, Wilcoxon-test). These results thus corroborate previous work and demonstrate that pancreatic amylase gene duplications are also more common in populations with starch-rich diets.

Pangenome-based identification of 28 distinct structural haplotypes underlying extensive amylase copy number variation

The amylase structural haplotype present in the human reference genome (GRCh38) spans ~200kb and consists of several long, nearly identical segmental duplications. While the approximate structures of several other haplotypes have been inferred through in-situ hybridization and optical mapping, these lack sequence and structural resolution[7,8,11,14]. Nevertheless, the variegated relationship between different amylase gene copy numbers (**Fig 2A**) indicates the existence of a wide range of structures.

To characterize the structural diversity of the amylase locus, we first constructed a minimizer anchored pangenome graph (MAP-graph)[22] from 94 amylase haplotypes derived from 54 long-read, haplotype resolved genome assemblies recently sequenced by the Human Pangenome Reference Consortium (HPRC)[23] alongside GRCh38 and the newly sequenced T2T-CHM13 reference[24] (**Fig 2B,** see methods). The MAP-graph captures large-scale sequence structures with vertices representing sets of homologous or paralogous sequences; thus, input haplotypes can be represented as paths through the graph. We next performed a "principal bundle decomposition" of the graph, which identifies stretches of sequence that are repeatedly traversed by individual haplotypes (colored loops in **Fig 2B**). These principal bundles represent the individual repeat units of the locus. We identified 8 principal bundles in the amylase graph corresponding to: the unique sequences on either side of the structurally complex amylase gene duplications (bundles 0 and 1), the repeat units spanning each of the three amylase genes and the *AMY2Ap* pseudogene (bundles 2, 3, and 5), as well as several other short repeat units (**Fig 2C**). For 35 individuals in which both haplotypes were incorporated into the graph, short read-based diploid genotypes were identical to the sum of the haplotype copy numbers, highlighting the concordance of both short-read genotypes and long-read haplotype assemblies (**Fig 2D**, methods).

Together we identified 28 unique structural haplotypes at the amylase locus (**Fig 2C**). The structurally variable region of the locus (hereafter SVR) spans across all of the amylase genes and ranges in size from ~95kb to ~471kb, in all cases beginning with a copy of *AMY2B* and ending with a copy of *AMY1.* To better understand the relationships between these structural haplotypes, we constructed a pangenome variation graph using PGGB (**Fig 2B**)[25]. In contrast to the MAP-graph, this graph enables base-level comparisons between haplotypes. Using this graph we computed a distance matrix between all structural haplotypes and built a neighbor-joining tree from these relationships (methods, **Fig 2C**). This tree highlights 11 different clusters of structures each defined by a unique copy number and configuration of amylase genes (**Fig 2C** right). Distinct structural haplotypes within clusters differed largely in the orientation of repeats, or only slightly in their composition. Within each cluster, we assigned one representative structural haplotype as the *"consensus"*. Several of these *consensus* structural haplotypes correspond to approximate architectures which have been previously hypothesized[14], however 3 of them are described here for the first time (H9, H3A2, and H3A3B3). Among these consensus structures, *AMY1* ranged from 1 to 9 copies with copy 6 and copy 8 states unobserved, *AMY2A* ranged from 0 to 3 copies, *AMY2Ap* ranged from 0 to 4 copies, and *AMY2B* ranged from 1 to 3 copies. Together these results reveal the wide ranging and nested-nature of diversity at the AMY locus: different haplotypes can harbor vastly different copy numbers of each of the three genes, and haplotypes with identical gene copy numbers exist in a wide array of forms.

Time-calibrated inference of haplotype evolutionary histories reveals rapid and recurrent evolution of amylase structures

To discern the evolutionary origins of the vast diversity of structures observed, we sought to explore the SNP haplotypes on which they emerged. We leveraged unique sequences (bundles 0 and 1) flanking the SVR in which SNPs can be accurately genotyped. We first quantified linkage disequilibrium (LD) around the amylase locus in 3,395 diverse human samples. To our surprise LD was extremely high between SNPs spanning the SVR (~190-370 kb apart in GRCh38, **Fig 3A, Extended Data Figs 2A-B**). Notably, LD was 7 to 20-fold higher when compared to similarly spaced pairs of SNPs across the remainder of chromosome 1 in all major continental populations (**Fig 3B**). Trio-based recombination rate estimates also indicate reduced recombination rates across the SVR (**Fig 3A** bottom panel)[26]. We hypothesize that these exceptionally high levels of LD arise from the suppression of crossover type recombination between homologs containing distinct structural architectures with vastly different lengths[27].

The high LD across the amylase locus implies that the evolutionary history of the flanking regions are a good proxy for the history of the linked complex structures of the SVR. As such, we constructed a maximum likelihood coalescent tree from these blocks using three Neanderthal haplotypes and a Denisovan haplotype as outgroups (**Figs 3C, S3**, **Extended Data Fig 3A**, methods). Time calibration of the tree was performed using an estimated 650 kyr human-Neanderthal split time. Annotating this coalescent tree with the different amylase structural architectures strikingly revealed the repeated evolution of similar and even identical structures on different haplotype backgrounds. Indeed, almost all complex amylase structures have evolved independently several times with a handful of exceptions, including the *AMY2B* gene duplications which stem from a single originating haplotype.

Our time calibrated tree further enabled us to perform an ancestral state reconstruction for each of the amylase gene copy numbers to quantify the number of times each gene has undergone duplication or deletion (**Fig 3D, Extended Data Fig 3B, S4**). We found that all amylase structural haplotypes in modern humans are descended from an $H3^r$ haplotype ~279 thousand years before present (kyr BP). This suggests that the initial duplication event, from the ancestral $H1^a$ haplotype to $H3^r$, significantly predates the out-of-Africa expansion. We identified 26 unique *AMY1* gene duplications and 24 deletions since then, corresponding to a per generation mutation rate (λ) of 2.09×10^{-4}, highlighting the exceptional turnover of this locus in recent evolution. *AMY1* gene copy number changes thus occur at a rate ~10,000-fold the genome-wide average SNP mutation rate[28]. *AMY2A* exhibited substantially fewer mutational events, undergoing 6 independent duplications and 2 deletions ($\lambda=3.07 \times 10^{-5}$) with the most recent *AMY2A* duplication occurring within the last 9.4 kyr BP (**Figs 3D, E**). While duplications of *AMY2A* have occurred several times, we identified a single origin of the complete loss of the *AMY2A* gene in our tree, which occurred 13.5-40.7 kyr BP and resulted in the H2A0 haplotype (**Figs 3D, F**). Only 2 *AMY2B* duplications were identified ($\lambda=7.36 \times 10^{-6}$), occurring sequentially on a single haplotype and thus allowing us to resolve the stepwise process of their formation (**Figs 3D, G**). We estimate the first duplication event occurred 46-107.8 kyr BP, followed by a deletion 26.9-46 kyr BP, and finally by a second duplication event 4.1-19.5 kyr BP (**Fig 3G**).

While our collection of 94 assembled haplotypes spanning the complex SVR provides the most complete picture of amylase evolution to date, it still represents just a small fraction of worldwide genetic variation. To characterize the evolution of amylase haplotypes more broadly, we performed a PCA combining the

fully assembled haplotypes with 3,395 diverse human genomes using SNPs across the unique locus used to construct the coalescent tree (**Fig 3H, Extended Data Fig 3C, S5-6**). We annotated individuals in the PCA with haploid/diploid *AMY1/2A/2B* copy numbers respectively. As expected, clusters of diploid individuals with high copy number (**Fig 3H** right panels) tended to colocalize with assembled haplotypes containing duplications (**Fig 3H** left panels). Exceptions to this indicate heterozygotes (with placements in between two haplotypes) or additional duplication/deletion events. This method identified several additional *AMY1* and *AMY2A* duplication events worldwide, as expected given their high mutation rate, and support for additional haplotypes with complete *AMY2A* deletions (**Figs 3H, S5**). However, we find no evidence of additional *AMY2B* gene duplications, supporting the single origin of these haplotypes.

Reconstruction of complex amylase structures from short read data uncovers worldwide diversity, stratification, and haplotypes associated with agriculture

Our analyses of SNP diversity at regions flanking the amylase SVR also revealed a substantial reduction in diversity compared to the chromosome-wide average (quantified by π, 2-3 fold lower, **Fig 3I**), and elevated integrated haplotype scores (iHS)[29] in some populations (**Fig S7**). Though these results are suggestive of recent positive selection at the region, our ability to detect signatures of selection is likely hampered by the repeated expansions of amylase genes and emergence of identical structures on distinct haplotype backgrounds.

Instead of relying on SNP-based methods, we developed an approach to directly identify the structural haplotype pairs present in short-read sequenced individuals. Briefly, this approach, which we term *'haplotype deconvolution'*, consists of mapping a short read-sequenced genome to the pangenome variation graph (**Fig 4A**) and quantifying read depth over each node in the graph (n=6,640 nodes in the amylase graph). This vector of read depths is then compared with a set of precomputed vectors generated by threading all pairs of 94 long-read assembled haplotypes (i.e., all possible genotypes) over the same graph. Finally, we infer the structural genotype of the short read genome to be the pair of pangenome reference haplotypes whose vector representation most closely matches to the short-read vector (**Fig 4B,** see methods). We assessed the accuracy of this approach using three orthogonal approaches. First, we compared haplotype deconvolutions in 35 individuals for which both short-read data and haplotype-resolved assemblies were available. Short read-based haplotype deconconvolutions exactly matched the long read assembly haplotypes 100% of the time (70/70 haplotypes). Second, we used 602 diverse short-read sequenced trios and estimated the accuracy of haplotype inference to be ~94% from Mendelian inheritance patterns (see methods) and 95%-97% concordant with previous inheritance-based determinations of haplotypes in 44 families[14]. Finally, we compared our previously estimated reference genome-based copy number genotypes to those predicted from *haplotype deconvolutions* across 4,292 diverse individuals. These genotypes exhibited 95-99% concordance across different amylase genes (95%, 97%, and 99% for *AMY1, AMY2A*, and *AMY2B* respectively). Cases in which the two estimates differed were generally high-copy genotypes for which representative haplotype assemblies have not yet been observed and integrated into the graph (**Fig S8**). Thus, we determine that our *haplotype deconvolution* method is robust and ~95% accurate, and limited primarily by the completeness of the reference pangenome.

We used *haplotype deconvolution* to estimate worldwide allele frequencies and continental subpopulation allele frequencies for amylase consensus structures across 7,188 haplotypes (**Figs 4B, C, Table S3**). The reference haplotype, $H3^r$, was the most common globally however several haplotypes exhibited strong population stratification. The H5 haplotype is the major allele in East Asian populations whereas the ancestral haplotype $H1^a$ was underrepresented in East Asian and Oceanic populations. The high copy H9 haplotype was largely absent from African, West Eurasian, and South Asian populations, while ranging from 1-3% in populations from the Americas, East Asia, and Central Asia and Siberia. Haplotypes with *AMY2B* duplications (i.e. H2A2B2, H3A3B3, and H4A2B2) were essentially absent from East and Central Asia, explaining our previous observation of the lack of *AMY2B* duplication genotypes in these global populations (**Fig 1C**) and consistent with their single origin.

We next compared the relative haplotype frequencies among modern human populations with traditionally agricultural-, hunter-gatherer-, fishing-, or pastoralism-based diets (**Fig 4D**). Agricultural populations differed significantly from non-Agricultural populations (p=0.00019, chi squared test) and were enriched for haplotypes with higher *AMY1* copy number, including the H5, H7, and H9 haplotypes, as well as for haplotypes with higher *AMY2A* and *AMY2B* copy number (H4A2B2, H2A2B2). In contrast, fishing, hunting, and pastoralism-based populations were enriched for the reference $H3^r$, deletion H2A0, and ancestral $H1^a$ haplotypes. These results demonstrate that haplotypes with increased amylase gene copy number are enriched in modern day populations with traditionally agricultural diets.

Ancient genomes reveal recent selection at the amylase locus in European populations

The development of agriculture ~12,000 years ago in the Fertile Crescent catalyzed a rapid shift in the diets and lifestyles of European and South Western Asian populations. To uncover how the genetic diversity of the amylase locus was shaped over this time period we collated 534 recently generated ancient European and South Western Asian genomes[30,31], which span in age from ~12,000 to ~250 BP (**Fig 5A, Table S1**). We estimated amylase gene copy numbers from these ancient individuals and compared these with modern European copy numbers (**Figs 5B-D**). *AMY1* copy number was significantly lower in ancient Hunter-Gatherer populations compared to modern day European populations, or to ancient Early Farmer or Yamnaya populations (P_{adj}= 5.1e-6 and 0.0099 for Eastern and Western Hunter Gatherer populations respectively, Wilcoxon rank sum test). By contrast, Early Farmer and Yamnaya population copy numbers were not significantly different from modern day Europeans. *AMY2A* and *AMY2B* showed similar signatures as *AMY1* with significantly lower copy number in ancient hunter gatherer populations compared to modern Europeans and Early Farmers. We next assessed how total copy numbers have changed as a function of time for each of the three amylase genes (**Figs 5E-G**). In all three cases we observed significant increases in total copy number over the last ~12,000 years (P=2.1e-6, 1.6e-6 and 0.0034 for *AMY1*, *AMY2A*, and *AMY2B* respectively, linear model). The total *AMY1* copy number increased by an average of ~2.9 copies over this time period while *AMY2A* and *AMY2B* increased by an average of 0.4 and 0.1 copies respectively. These results are suggestive of directional selection at this locus for increased copy number of each of the three amylase genes.

We next applied our haplotype deconvolution approach to these ancient genomes to infer how the frequency of amylase structural haplotypes has changed over recent time. Due to the lower coverage of these samples we were only able to confidently assign haplotypes for 328 out of 520 individuals (see

methods). Six haplotypes were found at appreciable frequencies (>1%) in either modern or ancient European populations including the H1[a] and H2A0 (*AMY2A* deletion) haplotypes, which each contain 3 total functional amylase gene copies, and the H3[r], H5, H7, and H4A2B2 haplotypes, which contain between 5 and 9 total amylase gene copies (**Figs 5H, S8-9**). Modeling the frequency trajectories of each of these haplotypes using multinomial logistic regression, we found that the ancestral H1[a] and the H2A0 haplotypes both decreased significantly in frequency over the last ~12,000 years, from a combined frequency of ~0.88 to a modern day frequency of ~0.14 (**Figs 5H, 5I inset, S10**). In contrast, duplication-containing haplotypes increased in frequency commensurately more than 7-fold (~0.12-0.86) over this time period. Using a Bayesian inference approach, we tested whether positive selection potentially might explain this substantial rise in the frequency of duplication-containing haplotypes[32]. The posterior distribution of selection coefficients strongly supported positive selection ($P<1x10^{-6}$, empirical p-value) with an average of s_{dup}=0.022 (**Fig 5I**). This selection coefficient is similar to estimates of *s* at the *MCM6/LCT* locus[30] highlighting the intensity of selection. Taken together, these results provide strong evidence for recent selection in Eurasian populations at this locus over the last 12,000 years.

Discussion

The domestication of crops and subsequent rise of farming radically reshaped human social structures, lifestyles, and diets. Though several evolutionary signatures of this transition have been identified in ancient and modern West Eurasian genomes[30,33,34], footprints of recent positive selection at the amylase locus have not been detected to date. Here, we find that haplotypes carrying duplicated copies of amylase genes have increased in frequency more than seven-fold in the last 12,000 years, consistent with strong positive selection. We also show that these complex duplicated amylase structures have arisen independently several times throughout human history on different haplotype backgrounds. Such extensive homoplasy and high mutation rate at the region are likely to obscure classical genomic signatures of a selective sweep[35,36]. Indeed, high copy number duplications with many alleles are often poorly tagged by neighboring SNPs, hampering scans that rely on this type of genetic variation to detect signatures of selection[37,38]. The recurrent emergence of increased amylase copy on distinct haplotype backgrounds may also explain conflicting results of GWAS targeting this locus[13,14]

One of the best studied examples of human adaptation to diet is the evolution of lactase persistence. Remarkably, the ability to digest milk has arisen independently in different populations[1,2]. Similarly, agriculture has been adopted independently several times throughout human history[6]. Here, in addition to showing strong signatures of positive selection in West Eurasian populations, we find that haplotypes carrying higher amylase copy numbers are found more commonly in multiple other populations with traditionally agricultural subsistence worldwide. These results suggest that selection for increased amylase copy number may have also happened several times throughout human history, though more extensive sampling of diverse ancient genomes and modern long-read assemblies are needed to further test this hypothesis.

The strong selective signature associated with duplicated amylase haplotypes illustrates the critical role SVs can play in human evolution. SVs can alter gene dosage, reconfigure the heterochromatic landscape of the genome, and reshape patterns of recombination. Yet, our understanding of human SV is still incomplete, owing largely to the inability of short reads to ascertain the sequence and configuration of long, complex duplicated architectures. Long-read sequencing approaches are revealing, for the first

time, human genetic variation that has been previously intractable. Methods that leverage pagenomes and long-read sequence assemblies in concert with short-read sequencing data in particular show great promise to detect natural selection targeting structurally complex loci and ultimately offer a more complete picture of human evolution. Here we introduce a *haplotype deconvolution* approach enabling us to genotype complex structural architectures from short read sequencing data. This method enables estimates of the allele frequency of these complex haplotypes throughout the world and to track haplotype frequencies through time.

Together, our study highlights the impact of the agricultural revolution on human genetic variation and underscores the importance of long-read sequencing and pangenomic methods to study the evolution of structurally complex regions.

Online content

Supplementary figures can be found in Supplementary Online Materials

Methods

Code: All code used can be found deposited in the following GitHub repository and https://github.com/sudmantlab/amylase_diversity_project and is archived in zenodo [ZENODO DOI FORTHCOMING].

Datasets: Short-read sequencing data were compiled from: high-coverage resequencing of 1000 genome samples[18], the Simons Genome Diversity Panel[19], and the Human Genome Diversity Panel[17]. Phased SNP calls from 1000 genomes and HGDP samples were compiled from Koenig et al[39]. Genomes from GTEx[21] samples were also assessed, but only for gene expression analyses as the ancestry of these samples was not available. Ancient genome short-read fastq samples were compiled from Allentoft et al[30] and Marchi et al[31]. Short-read modern and ancient fastq samples were mapped to the human reference genome GRCh38 with BWA (v0.7.17, `bwa mem`)[57]. While the modern genomes as well as the fifteen Marchi et al genomes are high coverage, the Allentoft et al samples are of varying read depth and therefore we removed samples that were too noisy to be accurately genotyped. To do so, we calculated the standard deviation of genome-wide copy number (after removing the top and bottom fifth percentiles of copy number to exclude outliers). We chose a standard deviation cutoff of 0.49 based on a visual inspection of the copy number data and selected 519 samples (~75% of 690) with sufficient read depth for copy number genotyping. Four archaic genomes were assessed including three high coverage Neanderthal Genomes and the high-coverage Denisova genome[40–43]. Long-read haplotype assemblies were compiled from the human pangenome reference consortium (HPRC)[23]. Year 1 genome assembly freeze data were compiled along with year 2 test assemblies. Haplotype assemblies were included in our analyses only if they spanned the amylase SVR. Furthermore, in cases where both haplotypes of an individual spanned the SVR, we checked to ensure that the diploid copy number of amylase genes matched with the read-depth based estimate of copy number. We noted that several year 1 assemblies (which were not assembled using ONT ultralong sequencing data) appeared to have been misassembled across the amylase locus as they were either discontiguous across the SVR, or had diploid assembly copy numbers that did not match with short-read predicted copy number. We thus reassembled these

genomes incorporating ONT ultralong sequence using the Verkko assembler[44] constructing improved assemblies for HG00673, HG01106, HG01361, HG01175, HG02148, HG02257. Alongside these HPRC genome assemblies, we included GRCh38 and the newly sequenced T2T-CHM13 reference[24].

Determination of subsistence by population: The diets of several populations (see **Table S2**) were determined from the literature from the following sources[11,45–53].

Read depth based copy number genotyping: Copy number genotypes were estimated using read depth as described in[15]. Briefly, read depth was quantified from BAMs in 1000bp sliding windows in 200bp steps across the genome. These depths were then normalized to a control region in which no evidence of copy number variation was observed in >4000 individuals. Depth-based "raw" estimates of copy number were then calculated by averaging these estimates over regions of interest. Copy number genotype likelihoods were estimated by fitting modified Gaussian Mixture Model (GMM) to "raw" copy estimates across all individuals with the following parameters: *k* - the number of mixture components, set to be the difference between the highest and lowest integer-value copy numbers observed; π - a k-dimensional vector of mixture weights; σ - a single variance term for mixture components; *o* - an offset term by which the means of all mixture components are shifted. The difference between mixture component means was fixed at 1 and the model was fit using expectation maximization (**Fig S1**). The copy number maximizing the likelihood function was used as the estimated copy number for each individual in subsequent analyses. Comparing these maximum likelihood copy number estimates with ddPCR yielded very high concordance with r^2 = 0.98, 0.99 and 0.96 for *AMY1*, *AMY2A*, and *AMY2B* respectively (**Fig S1**).

Analysis of gene expression: Gene expression data from the GTEx project[21] were downloaded alongside short read data (see above section). Normalized gene expression values for *AMY2A* and *AMY2B* were compared to copy number estimates using linear regression (**Fig S2**).

Minimizer Anchored Pangenome Graph Construction: Regions overlapping the amylase locus were extracted from genome assemblies in two different ways. First, we constructed a PGR-TK database from HPRC year 1 genome assemblies and used the default parameters of w=80, k=56, r=4, and min-span=64 for building the sequence database index. The GRCh38 chr1:103,655,518-103,664,551 was then used to identify corresponding AMY1/AMY2A/AMY2B regions across these individuals. Additional assemblies were subsequently added to our analysis by using minimap2[54] to extract the amylase locus from those genome assemblies. The Minimizer Anchored Pangenome Graph and the Principal Bundles were generated using revision v0.4.0 (git commit hash: ed55d6a8). The Python scripts and the parameters used for generating the principal bundle decomposition can be found in the associated GitHub Repository. The position of genes along haplotypes was determined by mapping gene modes to haplotypes using minimap2[54].

PGGB Based Graph Construction: A PGGB graph was constructed from 94 haplotypes spanning the amylase locus using PGGB v0.5.4 (commit 736c50d8e32455cc25db19d119141903f2613a63)[25] with the following parameters: `-n 94` (the number of haplotypes in the graph to be built) and `-c 2` (the number of mappings for each sequence segment). The latter parameter allowed us to build a graph that correctly represents the high copy number variation in such a locus. We used ODGI v0.8.3 (commit de70fcdacb3fc06fd1d8c8d43c057a47fac0310b)[55] to produce a Jaccard distance-based (i.e. 1-Jaccard similarity coefficient) dissimilarity matrix of paths in our variation graph (`odgi similarity -d`). These pre-

computed distances were used to construct a tree of relationships between haplotype structures using neighbor joining.

Haplotype Deconvolution Approach: We implemented a pipeline based on the workflow language Snakemake (v7.32.3) to parallelize *haplotype deconvolution* (i.e., assign to a short-read sequenced individual the haplotype pair in a pangenome that best represents its genotype at a given *locus*) in thousands of samples.
Given a region-specific PGGB graph (gfa, see **PGGB Based Graph Construction**), a list of short-read alignments (BAM/CRAM), a reference build (fasta) and a corresponding region of interest (chr:start-end; based on the alignment of the BAM/CRAM), our pipeline runs as follows:

1. extract the haplotypes from the initial pangenome using ODGI (v0.8.3, `odgi paths -f`)
2. for each short-read sample, extract all the reads spanning the region of interest using SAMTOOLS (v1.18, `samtools fasta`)[56]
3. map the extracted reads back to the haplotypes with BWA (v0.7.17, `bwa mem`)[57]. To map ancient samples, we used instead `bwa aln` with parameters suggested in Oliva A et al., 2021[58]: `bwa aln -l 1024 -n 0.01 -o 2`
4. compute a node depth matrix for all the haplotypes in the pangenome: every time a certain haplotype in the pangenome loops over a node, the path depth for that haplotype over that node increases by one. This is done using a combination of commands in ODGI (`odgi chop -c 32` and `odgi paths -H`)
5. compute a node depth vector for each short-read sample: short-read alignments are mapped to the pangenome using GAFPACK (https://github.com/ekg/gafpack, commit ad31875) and their coverage over nodes computed using GFAINJECT (https://github.com/ekg/gfainject, commit f5feb7b)
6. compare each short-read vector (see .5) with each possible pair of haplotype vectors (see .4) by means of cosine similarity using (https://github.com/davidebolo1993/cosigt, commit e247261) (which measures the similarity between two vectors as their dot product divided by the product of their lengths). The haplotype pair having the highest similarity with the short-read vector is used to describe the genotype of the sample.
7. The final genotypes were assigned as the corresponding consensus haplotypes of highest similarity pair haplotypes.

Our pipeline is publicly available on GitHub (https://github.com/raveancic/graph_genotyper) and is archived in zendo [ZENODO DOI FORTHCOMING].

We assessed the accuracy of the haplotype deconvolution approach in four different ways. First we assessed 35 individuals (70 haplotypes) for which both short-read sequencing data and long-read diploid assemblies were available. In 100% of cases (70/70 haplotypes) we accurately distinguished the correct haplotypes present in an individual from short read sequencing data. We further assessed how missing haplotypes in the pangenome graph might assess the accuracy of our approach by performing a “leave-one-out” analysis. In this approach, for each of the 35 long-read individuals individuals we rebuilt the variation graph with a single haplotype excluded and tested our ability to identify the correct consensus haplotype from the remaining haplotypes. The true positive rate was ~93% in this case. Second, we compared our haplotype deconvolutions to haplotypes determined by inheritance patterns in 44 families

in a previous study (Usher *et al* 2015, Table S3)[14]. We note that this study hypothesized the existence of an H4A4B4 haplotype without having observed it directly. In our study we also find no direct evidence of the H4A4B4 haplotype. Furthermore, we find that inheritance patterns are equally well explained by other directly observed haplotypes and thus exclude these predictions from our comparisons (2 individuals excluded). We identified the exact same pair of haplotypes in 95% of individuals(125/131 individuals) and in 97% of individuals (288/298 individuals) the haplotype pair we identify is among the potential consistent haplotype pairs identified from inheritance. Third, we compared inheritance patterns in 602 diverse short-read sequenced trios from 1000 genomes populations[18]. For each family we randomly selected one parent and assessed if either of the two offspring haplotypes were present in this randomly selected parent. Across all families, this proportion, *p*, represents an estimate of the proportion of genotype calls that are accurate in both the offspring and that parent, thus the single sample accuracy can be estimated as the square root of *p.* From these analyses we identified 533/602 parent-offspring genotype calls that are correct, corresponding to an estimated accuracy of 94%. Fourth, we compared our previously estimated reference genome read-depth-based copy number genotypes to those predicted from *haplotype deconvolutions* across 4,292 diverse individuals. These genotypes exhibited 95-99% concordance across different amylase genes (95%, 97%, and 99% for *AMY1, AMY2A*, and *AMY2B* respectively). Cases in which the two estimates differed were generally high-copy genotypes for which representative haplotype assemblies have not yet been observed and integrated into the graph (**Fig. S8**). Overall we thus estimate the *haplotype deconvolution* approach to be ~95% accurate. Haplotype deconvolution scores (**Fig. S9**) tended to be lower for low-coverage ancient genomes thus individuals below a cutoff of 0.75 were discarded.

LD estimation: To investigate pairwise linkage disequilibrium (LD) across the SVR region at a global scale, we first merged our copy number estimates with the joint SNP call set from HGDP and 1kGP[39], resulting in a variant call set of 3,395 diverse individuals with both diploid copy number genotypes and phased SNP calls. Briefly, we used bcftools-v1.9[56] to filter HGDP and 1kGP variant data for designated genomic regions in chromosome 1, including the amylase structurally variable region (SVR) and flanking regions defined as bundle 0 and bundle 1 (distal and proximal respectively) using the GRCh38 reference coordinate system (--region chr1:103456163-103863980 in GRCh38). The resulting output was saved in variant call format (vcf), keeping only bi-allelic SNPs (-m2 -M2 -v snps), and additionally filtered with vcftools-v.0.1.16[59] with –keep and –recode options for lists of individuals grouped by continental region in which we were able to estimate diploid copy numbers. Population-specific vcf files were further filtered for a minor allele frequency filter threshold of 5% (--minmaf 0.05) and used to generate a numeric genotype matrix with the physical positions of SNPs for LD calculation (R^2 statistic) and plotting with the LDheatmap[60] function in R-v4.2.2.

To further dissect the unique evolutionary history of the AMY locus, we compared regions with high R^2 across the SVR with LD estimates for pairs of SNPs across regions of similar size in chromosome 1. We specifically focused on pairs of SNPs spanning bundle 0 (chr1:103456163-103561526 in GRCh38) and the first 66-kbp of bundle 1, hereafter labeled as bundle 1a (chr1:103760698-103826698 in GRCh38), as revealed by the LD heatmap. Then we computed the R^2 values for any pair of SNPs in chromosome 1 for each superpopulation within a minimum of 190 kb distance (i.e. the equivalent distance from bundle 0 end to bundle 1a start using the GRCh38 reference coordinate system) and maximum 370 kb distance (i.e.the equivalent distance from bundle 0 start to bundle 1a end using the GRCh38 reference coordinate system). To calculate pairwise LD across the human chromosome 1 for different populations we ran plink-v1.90b6.21[61] with options -r2 –ld-window 999999 –ld-window-kb 1000 –ld-window-r2 0 –make-bed –maf

0.05, using as input population-specific vcf files for a set of biallelic SNPs of 3,395 individuals from HGDP and 1kGP. Since the resulting plink outputs only provide R^2 estimates for each pair of SNPs and respective SNP positions, we additionally calculated the physical distances between pairs of SNPs as the absolute difference between the base pair position of the second (BP_B) and first (BP_A) SNP. We then filtered out distances smaller than 190 kb and greater than 370 kb, and annotated the genomic region for each R^2 value based on whether both SNPs fall across the SVR region or elsewhere in chromosome 1. The distance between SNP pairs was also binned into intervals of 20,000 bp and each interval's midpoint was used for assessing LD decay over genomic distances. The resulting dataset was imported in R to compute summary statistics comparing LD across each major continental region, or superpopulations, and we used ggplot2 to visualize the results.

Coalescent tree, ancestral state reconstruction, and PCA:
To construct the coalescent tree, we first extracted bundle 0 and bundle 1a sequences from all 94 haplotypes (i.e. distal and proximal unique regions flanking the amylase SVR) that went through principal bundle decomposition. Based on their coordinates on the human reference genome (GRCh38), we used samtools-v1.17[62] to extract these sequences from three Neanderthal and one Denisovan genomes that are aligned to GRCh38. We used kalign-v3.3.5[63] to perform multiple sequence alignment on bundle 0 and bundle 1a sequences. We used iqtree-v2.2.2.3[64] to construct a maximum likelihood tree with Neanderthal and Denisova sequences as the outgroup, using an estimated 650 kyr human-Neanderthal split time for time calibration. We used ggtree-v3.6.2[65] in R-v4.2.1 to visualize the tree, and annotated each tip with its structural haplotype and amylase gene copy numbers. We used cafe-v5.0.0[66] to infer the ancestral copy numbers of each of the three amylase genes along the time-calibrated coalescent tree (excluding the outgroups) and to estimate their duplication/deletion rates. The timing of each duplication/deletion event was estimated based on the beginning and end of the branch along which the amylase gene copy number has changed. We used ggtree and ggplot-v3.4.2 in R to visualize these results, and used Adobe Illustrator to create illustrations for several of the most notable duplication/deletion events[67].

Next, we performed a principal component analysis (PCA) combining 94 HPRC haplotype sequences with variant calls for 3,395 individuals from HGDP and 1kGP. We first aligned all 94 bundle 0 and 94 bundle 1a haplotype sequences to the human reference genome (GRCh38) using minimap2-v2.26[54], and called SNPs from haplotypes using paftools.js. Each haplotype sequence appears as a pseudo-diploid in the resulting vcf file (i.e. when the genotype is different from the reference, it is coded as being homozygous for the non-reference allele). These haplotype-specific vcf files were merged together and filtered for biallelic SNPs (-m2 -M2 -v snps) with bcftools, resulting in a pseudo-diploid vcf file from 94 haplotype sequences for each bundle. These were then merged with the respective bundle 0 and bundle 1a vcf files from HGDP and 1kGP, also filtered for biallelic SNPs, using bcftools. Finally, we ran plink with a minor allele frequency of 5% (--maf 0.05) to obtain eigenvalues and eigenvectors for PCA and used ggplot-v3.4.2 to visualize the results. These analyses were conducted with bundle 0 and bundle 1a separately, with highly concordant results (**Fig. S5-S6**). Analyses focused on bundle 0 are reported in the main text (**Figure 3**) whereas bundle 1a results are shown as extended data (**Extended Data Figure 3**).

Signatures of recent positive selection in modern human populations: To test the hypothesis of very recent or ongoing positive selection at the amylase locus in modern humans, we looked for significant signatures of reduced genetic diversity and high integrated haplotype scores (iHS) across the

non-duplicated regions adjacent to the SVR compared to other regions of chromosome 1 in different populations worldwide. This stems from the assumption that, given low SNP density across the SVR, the high levels of LD found between pairs of SNPs spanning bundle 0 and bundle 1a indicate that SNPs in bundle 0 or bundle 1 can be used as proxies for the selective history of the linked complex structures of the SVR. We calculated nucleotide diversity (π) on sliding windows of 20,000 bp spanning GRCh38 chromosome 1 with vcftools using as input population-specific vcf files from HGDP and 1kGP filtered for a set of biallelic SNPs. Each window was annotated for the genomic region, namely bundle 0, SVR and bundle 1a. All windows comprising the SVR region were removed from the resulting output due to low SNP density. We then used ggplot2 in R to calculate and visualize summary statistics comparing nucleotide diversity for windows located windows harboring the flanking regions to amylase genes (i.e. bundle 0 and 1a) with nucleotide diversity for windows spanning the rest of chromosome 1 for each major continental region or super population.

To test for signals of very recent positive selection at the flanking regions of the SVR (i.e. in favor of variants that have not reached fixation) we computed the iHS statistic across GRCh38 chromosome 1 using the rehh package[68] in R. Vcf files from HGDP and 1kGP specific to chromosome 1 and super populations, which had been previously filtered for biallelic SNPs, were imported into R using the data2haplohh() function and filtered for minor allele frequency of 0.05. We then used scan_hh() and ihh2ihs() functions to generate iHS statistics for all the SNPs of chromosome 1. SNP positions within the SVR region were removed from the resulting output due to low SNP density. We then compared the distribution of the absolute value of iHS for SNPs located within bundle 0 and bundle 1a (labeled as 'AMY') to those from the rest of chromosome 1 for each major continental region or super population using the geom_density() function in ggplot2 (**Fig. S7**).

Inference of recent positive selection in Eurasian populations using ancient genomes: To determine if changes in the frequency of different structural haplotypes over the last 12,000 years were consistent with positive selection we used ApproxWF[32] to perform Bayesian inference of selection coefficients from allele frequency trajectories. Amylase structural haplotypes (n=11) were grouped into those with the ancestral number of amylase gene copies (three total), or with amylase gene duplications (five or more copies). Binned allele frequency trajectories were then used to run ApproxWF for 1010000 MCMC steps with parameters h=0.5 and pi=1. We assumed a generation time of 30 years to convert the age of ancient samples from years to generations. The first 10,000 steps of the MCMC process were discarded in all analyses.

References

1. Tishkoff, S. A. *et al.* Convergent adaptation of human lactase persistence in Africa and Europe. *Nat. Genet.* **39**, 31–40 (2007).
2. Enattah, N. S. *et al.* Identification of a variant associated with adult-type hypolactasia. *Nat. Genet.* **30**, 233–237 (2002).
3. Mathias, R. A. *et al.* Adaptive evolution of the FADS gene cluster within Africa. *PLoS One* **7**, e44926 (2012).
4. Ameur, A. *et al.* Genetic adaptation of fatty-acid metabolism: a human-specific haplotype increasing the biosynthesis of long-chain omega-3 and omega-6 fatty acids. *Am. J. Hum. Genet.* **90**, 809–820 (2012).
5. Fumagalli, M. *et al.* Greenlandic Inuit show genetic signatures of diet and climate adaptation.

Science **349**, 1343–1347 (2015).
6. Bellwood, P. *First Farmers: The Origins of Agricultural Societies*. (John Wiley & Sons, 2004).
7. Groot, P. C. *et al.* The human alpha-amylase multigene family consists of haplotypes with variable numbers of genes. *Genomics* **5**, 29–42 (1989).
8. Groot, P. C. *et al.* Evolution of the human alpha-amylase multigene family through unequal, homologous, and inter- and intrachromosomal crossovers. *Genomics* **8**, 97–105 (1990).
9. Pajic, P. *et al.* Independent amylase gene copy number bursts correlate with dietary preferences in mammals. *Elife* **8**, (2019).
10. Inchley, C. E. *et al.* Selective sweep on human amylase genes postdates the split with Neanderthals. *Sci. Rep.* **6**, 37198 (2016).
11. Perry, G. H. *et al.* Diet and the evolution of human amylase gene copy number variation. *Nat. Genet.* **39**, 1256–1260 (2007).
12. Mathieson, S. & Mathieson, I. FADS1 and the Timing of Human Adaptation to Agriculture. *Mol. Biol. Evol.* **35**, 2957–2970 (2018).
13. Falchi, M. *et al.* Low copy number of the salivary amylase gene predisposes to obesity. *Nat. Genet.* **46**, 492–497 (2014).
14. Usher, C. L. *et al.* Structural forms of the human amylase locus and their relationships to SNPs, haplotypes and obesity. *Nat. Genet.* **47**, 921–925 (2015).
15. Sudmant, P. H. *et al.* Global diversity, population stratification, and selection of human copy-number variation. *Science* **349**, aab3761 (2015).
16. Carpenter, D. *et al.* Obesity, starch digestion and amylase: association between copy number variants at human salivary (AMY1) and pancreatic (AMY2) amylase genes. *Hum. Mol. Genet.* **24**, 3472–3480 (2015).
17. Bergström, A. *et al.* Insights into human genetic variation and population history from 929 diverse genomes. *Science* **367**, (2020).
18. Byrska-Bishop, M. *et al.* High-coverage whole-genome sequencing of the expanded 1000 Genomes Project cohort including 602 trios. *Cell* **185**, 3426–3440.e19 (2022).
19. Mallick, S. *et al.* The Simons Genome Diversity Project: 300 genomes from 142 diverse populations. *Nature* **538**, 201–206 (2016).
20. Bank, R. A. *et al.* Variation in gene copy number and polymorphism of the human salivary amylase isoenzyme system in Caucasians. *Hum. Genet.* **89**, 213–222 (1992).
21. GTEx Consortium *et al.* Genetic effects on gene expression across human tissues. *Nature* **550**, 204–213 (2017).
22. Chin, C.-S. *et al.* Multiscale analysis of pangenomes enables improved representation of genomic diversity for repetitive and clinically relevant genes. *Nat. Methods* (2023) doi:10.1038/s41592-023-01914-y.
23. Liao, W.-W. *et al.* A draft human pangenome reference. *Nature* **617**, 312–324 (2023).
24. Nurk, S. *et al.* The complete sequence of a human genome. *Science* **376**, 44–53 (2022).
25. Garrison, E. *et al.* Building pangenome graphs. *bioRxiv* (2023) doi:10.1101/2023.04.05.535718.
26. Kong, A. *et al.* A high-resolution recombination map of the human genome. *Nat. Genet.* **31**, 241–247 (2002).
27. Ahuja, J. S., Harvey, C. S., Wheeler, D. L. & Lichten, M. Repeated strand invasion and extensive branch migration are hallmarks of meiotic recombination. *Mol. Cell* **81**, 4258–4270.e4 (2021).
28. Chintalapati, M. & Moorjani, P. Evolution of the mutation rate across primates. *Curr. Opin. Genet. Dev.* **62**, 58–64 (2020).
29. Voight, B. F., Kudaravalli, S., Wen, X. & Pritchard, J. K. A map of recent positive selection in the human genome. *PLoS Biol.* **4**, e72 (2006).
30. Allentoft, M. E. *et al.* Population genomics of Stone Age Eurasia. *bioRxiv* (2022) doi:10.1101/2022.05.04.490594.
31. Marchi, N. *et al.* The genomic origins of the world's first farmers. *Cell* **185**, 1842–1859.e18 (2022).
32. Ferrer-Admetlla, A., Leuenberger, C., Jensen, J. D. & Wegmann, D. An Approximate Markov Model for the Wright-Fisher Diffusion and Its Application to Time Series Data. *Genetics* **203**, 831–846 (2016).

33. Le, M. K. *et al.* 1,000 ancient genomes uncover 10,000 years of natural selection in Europe. *bioRxiv* (2022) doi:10.1101/2022.08.24.505188.
34. Mathieson, I. *et al.* Genome-wide patterns of selection in 230 ancient Eurasians. *Nature* **528**, 499–503 (2015).
35. Pennings, P. S. & Hermisson, J. Soft sweeps III: the signature of positive selection from recurrent mutation. *PLoS Genet.* **2**, e186 (2006).
36. Messer, P. W. & Petrov, D. A. Population genomics of rapid adaptation by soft selective sweeps. *Trends Ecol. Evol.* **28**, 659–669 (2013).
37. Handsaker, R. E. *et al.* Large multiallelic copy number variations in humans. *Nat. Genet.* **47**, 296–303 (2015).
38. Schrider, D. R. & Hahn, M. W. Lower linkage disequilibrium at CNVs is due to both recurrent mutation and transposing duplications. *Mol. Biol. Evol.* **27**, 103–111 (2010).
39. Koenig, Z. *et al.* A harmonized public resource of deeply sequenced diverse human genomes. *bioRxiv* (2023) doi:10.1101/2023.01.23.525248.
40. Prüfer, K. *et al.* A high-coverage Neandertal genome from Vindija Cave in Croatia. *Science* **358**, 655–658 (2017).
41. Meyer, M. *et al.* A high-coverage genome sequence from an archaic Denisovan individual. *Science* **338**, 222–226 (2012).
42. Mafessoni, F. *et al.* A high-coverage Neandertal genome from Chagyrskaya Cave. *Proc. Natl. Acad. Sci. U. S. A.* **117**, 15132–15136 (2020).
43. Prüfer, K. *et al.* The complete genome sequence of a Neanderthal from the Altai Mountains. *Nature* **505**, 43–49 (2014).
44. Rautiainen, M. *et al.* Telomere-to-telomere assembly of diploid chromosomes with Verkko. *Nat. Biotechnol.* **41**, 1474–1482 (2023).
45. Kirby, K. R. *et al.* D-PLACE: A Global Database of Cultural, Linguistic and Environmental Diversity. *PLoS One* **11**, e0158391 (2016).
46. Murdock, G. P. Ethnographic Atlas: A Summary. *Ethnology* **6**, 109 (1967).
47. *Encyclopedia of the world's minorities*. (Routledge, 2013). doi:10.4324/9780203935606.
48. Sukernik, R. I. *et al.* Mitochondrial genome diversity in the Tubalar, Even, and Ulchi: contribution to prehistory of native Siberians and their affinities to Native Americans. *Am. J. Phys. Anthropol.* **148**, 123–138 (2012).
49. Levin, M. G. *The Peoples of Siberia*. (1964).
50. Abryutina, L. Aboriginal peoples of Chukotka. *Etud. Inuit* **31**, 325–341 (2009).
51. Kozlov, A., Nuvano, V. & Vershubsky, G. Changes in Soviet and post-Soviet indigenous diets in Chukotka. *Etud. Inuit* **31**, 103–119 (2009).
52. Moran, E. F. Human adaptation to arctic zones. *Annu. Rev. Anthropol.* **10**, 1–25 (1981).
53. Korotayev, A., Kazankov, A., Borinskaya, S., Khaltourina, D. & Bondarenko, D. Ethnographic atlas XXX: Peoples of Siberia. *Ethnology* **43**, 83 (2004).
54. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).
55. Guarracino, A., Heumos, S., Nahnsen, S., Prins, P. & Garrison, E. ODGI: understanding pangenome graphs. *Bioinformatics* **38**, 3319–3326 (2022).
56. Danecek, P. *et al.* Twelve years of SAMtools and BCFtools. *Gigascience* **10**, (2021).
57. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. (2013) doi:10.48550/ARXIV.1303.3997.
58. Oliva, A., Tobler, R., Llamas, B. & Souilmi, Y. Additional evaluations show that specific settings still outperform for ancient DNA data alignment. *Ecol. Evol.* **11**, 18743–18748 (2021).
59. Danecek, P. *et al.* The variant call format and VCFtools. *Bioinformatics* **27**, 2156–2158 (2011).
60. Shin, J.-H., Blay, S., Graham, J. & McNeney, B. LDheatmap: AnRFunction for Graphical Display of Pairwise Linkage Disequilibria Between Single Nucleotide Polymorphisms. *J. Stat. Softw.* **16**, (2006).
61. Chang, C. C. *et al.* Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* **4**, 7 (2015).

62. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
63. Lassmann, T. Kalign 3: multiple sequence alignment of large data sets. *Bioinformatics* **36**, 1928–1929 (2019).
64. Minh, B. Q. *et al.* IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. *Mol. Biol. Evol.* **37**, 1530–1534 (2020).
65. Yu, G., Smith, D. K., Zhu, H., Guan, Y. & Lam, T. T.-Y. Ggtree: An r package for visualization and annotation of phylogenetic trees with their covariates and other associated data. *Methods Ecol. Evol.* **8**, 28–36 (2017).
66. Mendes, F. K., Vanderpool, D., Fulton, B. & Hahn, M. W. CAFE 5 models variation in evolutionary rates among gene families. *Bioinformatics* **36**, 5516–5518 (2021).
67. Wickham, H. Ggplot2. *Wiley Interdiscip. Rev. Comput. Stat.* **3**, 180–185 (2011).
68. Gautier, M. & Vitalis, R. rehh: an R package to detect footprints of selection in genome-wide SNP data from haplotype structure. *Bioinformatics* **28**, 1176–1177 (2012).

Acknowledgements

We would like to thank Morten E. Allentoft, Rasmus Nielsen, Evan K. Irving-Pease, Martin Sikora ,and Eske Willersvlev, for helpful discussion and assistance in accessing ancient datasets. Institute of General Medical Sciences [grant: R35GM142916] to PHS. Vallee Scholars Award to PHS. Ancient DNA sequencing was supported by grants from the Lundbeck Foundation (R302-2018-2155 and R155-2013-16338).

Contributions

Conceived the experimental design: PHS
Processed and analyzed the data: PHS, DB, AH, RNL, AR, JR, AG, JC, EG
Wrote and edited the manuscript: PHS, DB, AH, RNL, AR, JR, AG, JC, EG
Supervised research: PHS, EG, NS

Figures

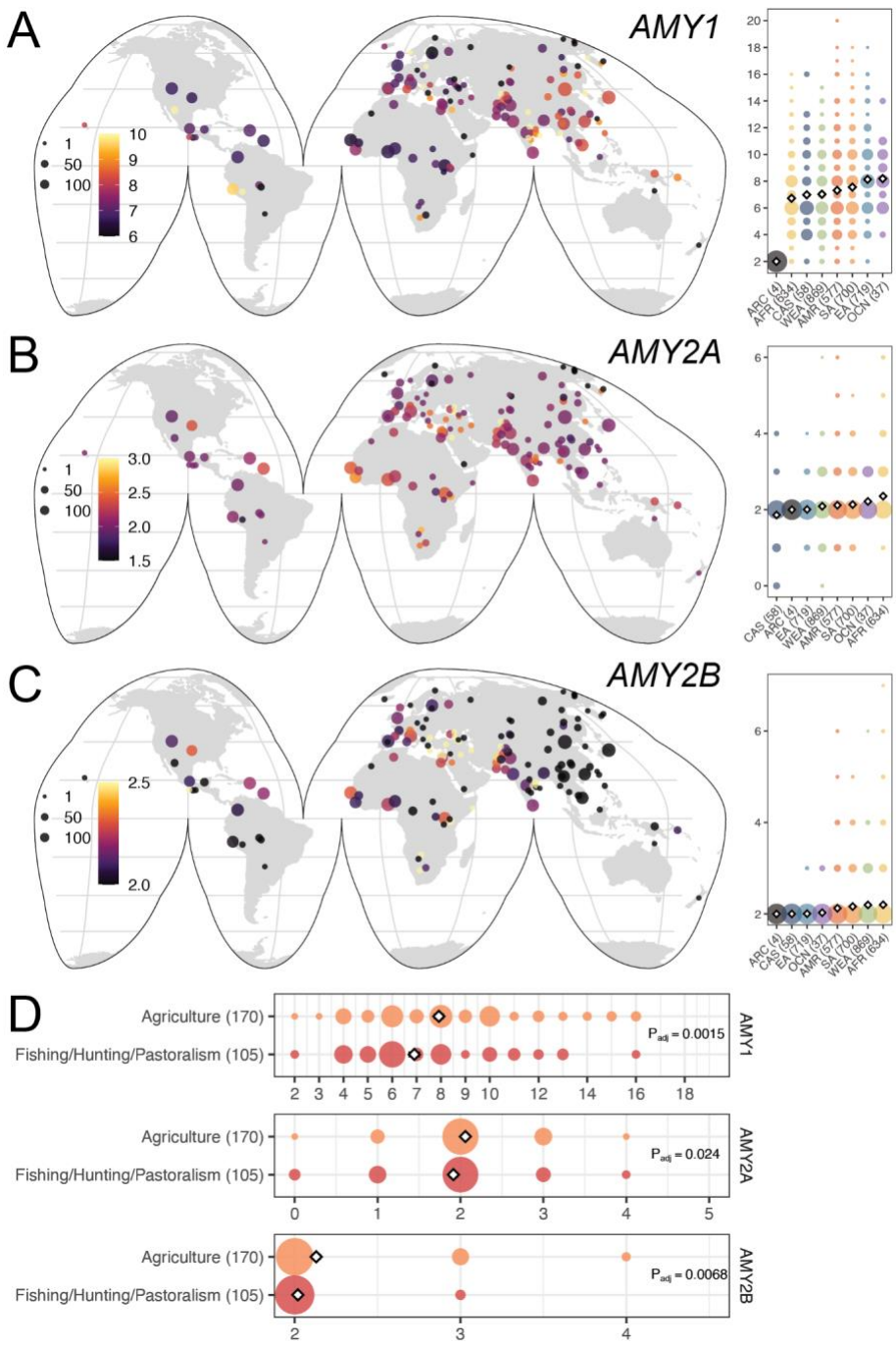


Figure 1 - Worldwide amylase copy number diversity. A-C) World maps indicating average *AMY1, AMY2A,* and *AMY2B* copy number in 162 different human populations. Point size indicates population sample sizes and color indicated mean copy number. Inset right are distribution of copy numbers (Y-axis) in continental populations (X-axis): archaic (ARC), African (AFR), Central Asia Siberia (CAS), West Eurasia (WEA), Americas (AMR), South Asia (SA), East Asia (EA), and Oceania (OCN). White diamonds indicate mean, dot sizes indicate proportion with copy number genotype. Copy number distributions across individual populations are displayed in **Extended Data Figure 1. D)** Copy number distributions of *AMY1, AMY2A,* and *AMY2B* in 31 modern human populations with traditionally agricultural subsistence compared to fishing, hunting, and pastoralism-based diets.

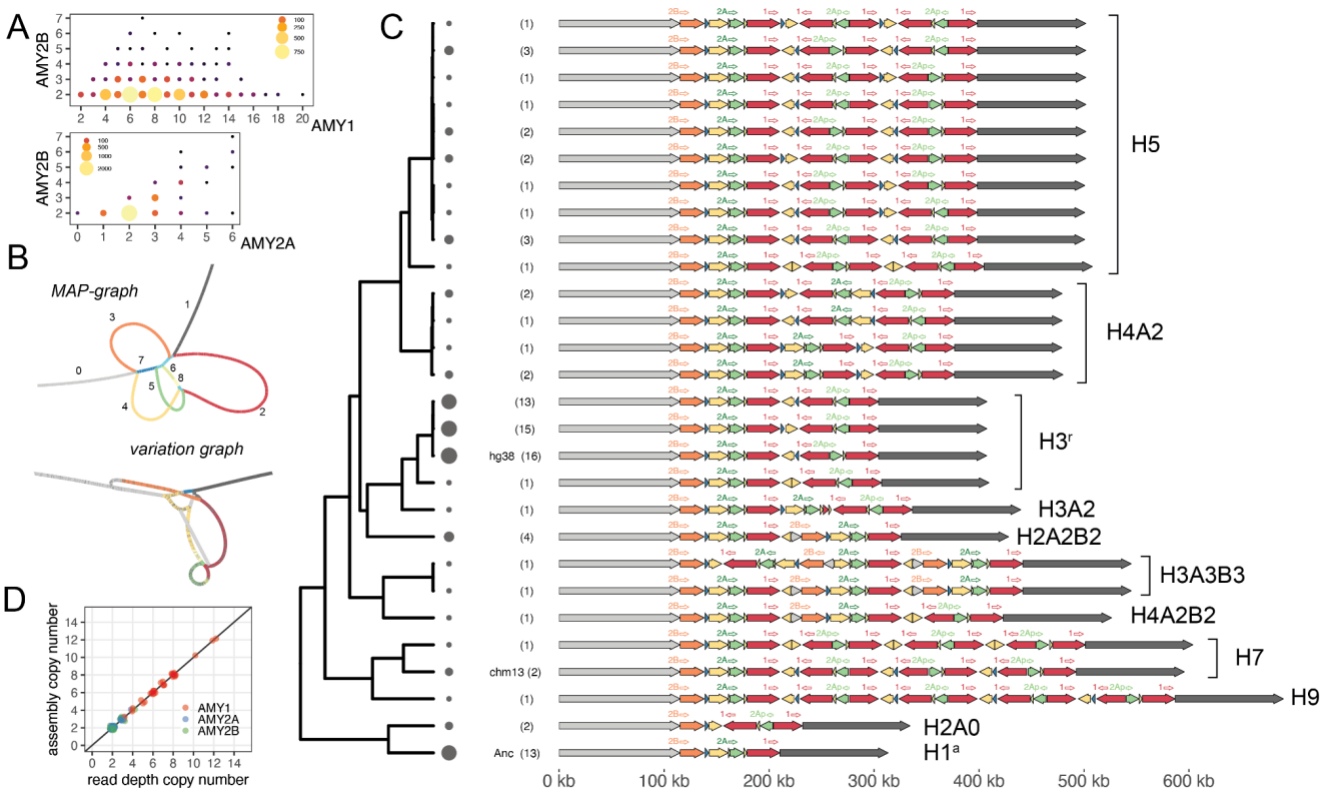


Figure 2 - Pangenome-based identification of amylase structural haplotype diversity. A) The relationship between *AMY1, AMY2A,* and *AMY2B* copy number. Size and color indicate number of individuals with copy number genotype pair. **B)** Hierarchical minimizer anchored pangenome graph (MAP-graph) and variation graph architectures. Colors and numbers in MAP graph correspond to principal bundles shown in C. **C)** 28 distinct amylase structural haplotypes identified in 94 haplotypes. Filled arrows indicate principal bundles representing paralogous and homologous relationships while labelled open arrows (above) indicate genes. Numbers in parentheses and circle sizes indicate the number of haplotypes identified with a specific structure. Haplotypes are ordered by their relationship in tree (left) which is generated from the jaccard distance between haplotypes from the variation graph. *Consensus structures,* referring to clusters of similar structures, are indicated to the right. **D)** The relationship between read-depth based copy number and assembly-based copy numbers for amylase genes for 35 individuals (70 haplotypes) in which both haplotypes were assembled across the amylase region.

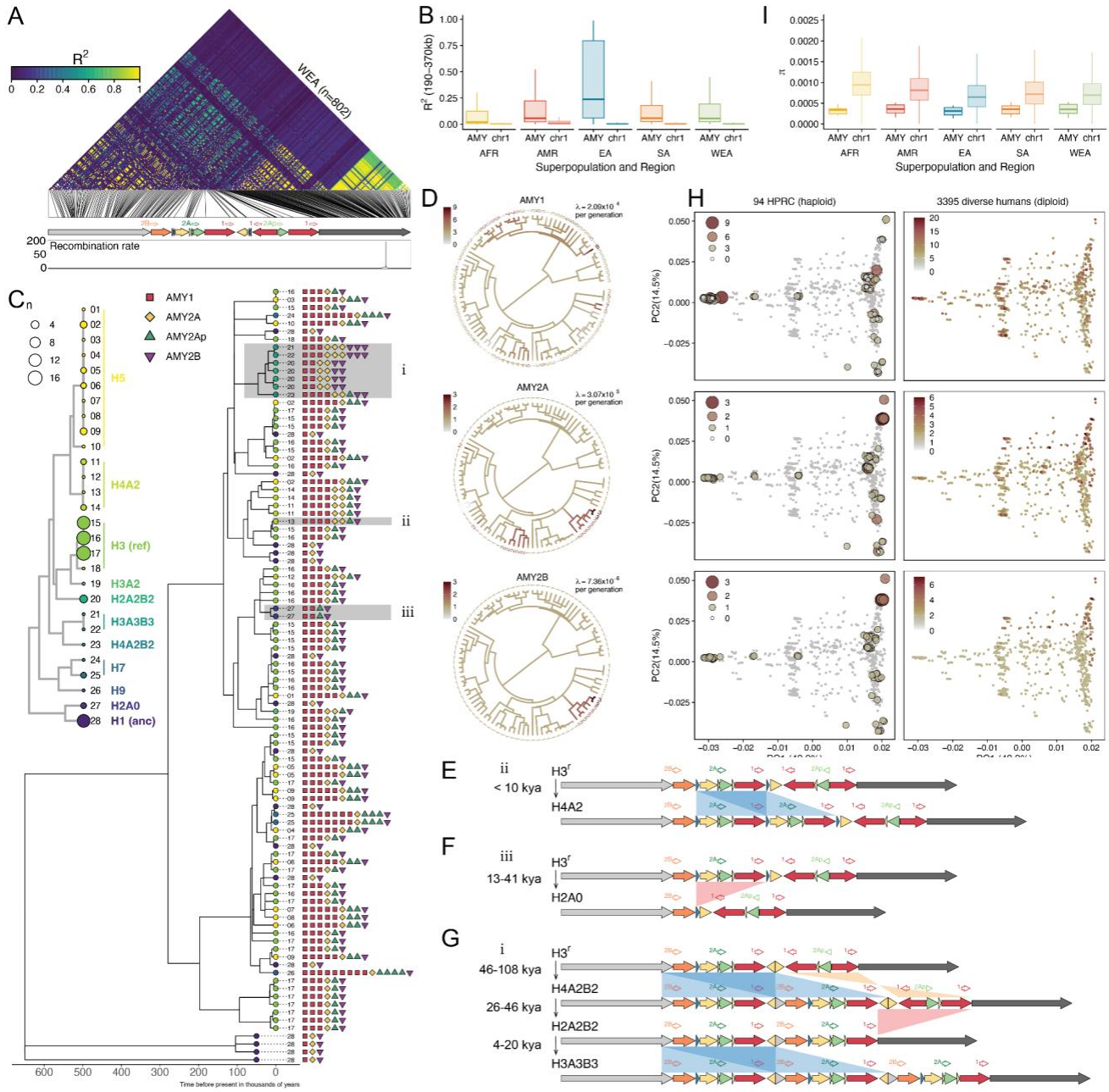


Figure 3 - Evolutionary history of amylase structural haplotypes. A) Heat map of linkage disequilibrium (LD) for SNPs across a ~406 kb region spanning unique sequences on either side of the structurally variable region of amylase (SVR) for 802 Western Eurasians (WEA) (see **Extended Data Fig. 2A** for all populations). Schematic of GRCh38 structure and recombination rate are shown below. **B)** Boxplots comparing R^2 between pairs of SNPs on either side of the SVR (i.e. 190 kb - 370 kb apart) to identically spaced SNPs across chromosome 1 for major human populations with more than 100 samples (see **Extended Data Fig. 2B** for LD decay over genomic distances). **C)** A time-calibrated coalescent tree from the distal non-duplicated region flanking the SVR (leftmost gray arrow in A) across 94 assembled haplotypes (tree from the proximal region in **Extended Data Fig. 3)**. The number next to each tip corresponds to the structural haplotype that the sequence is physically linked to and the color of the circle at each tip corresponds to its consensus haplotype structure (see inset structure tree). The copy numbers of each amylase gene and pseudogene are also shown next to the tips of the tree. **D)**

Ancestral state reconstruction and mutation rate estimates for amylase gene copy number (archaic outgroups excluded). Branch color corresponds to copy number. **E-G)** Illustrations of the most recent AMY2A gene duplication, the complete loss of AMY2A gene, and the sequential and joint duplication of AMY2A and AMY2B genes (shaded in gray in C). **H)** A PCA from 94 haplotype assemblies and 3,395 diverse diploid human genomes from the distal non-duplicated region flanking the SVR (PCA from the proximal region in **Extended Data Fig. 3**). In the left column diploid genomes are shown in gray while assembled haplotypes are colored and sized by their haploid amylase copy number. In the right column assembled haplotypes are hidden and diploid genomes are colored by their diploid copy number. **I)** Boxplots comparing π calculated in 20 kbp sliding windows across the distal non-duplicated region adjacent to the SVR for major continental human populations with more than 100 individuals.

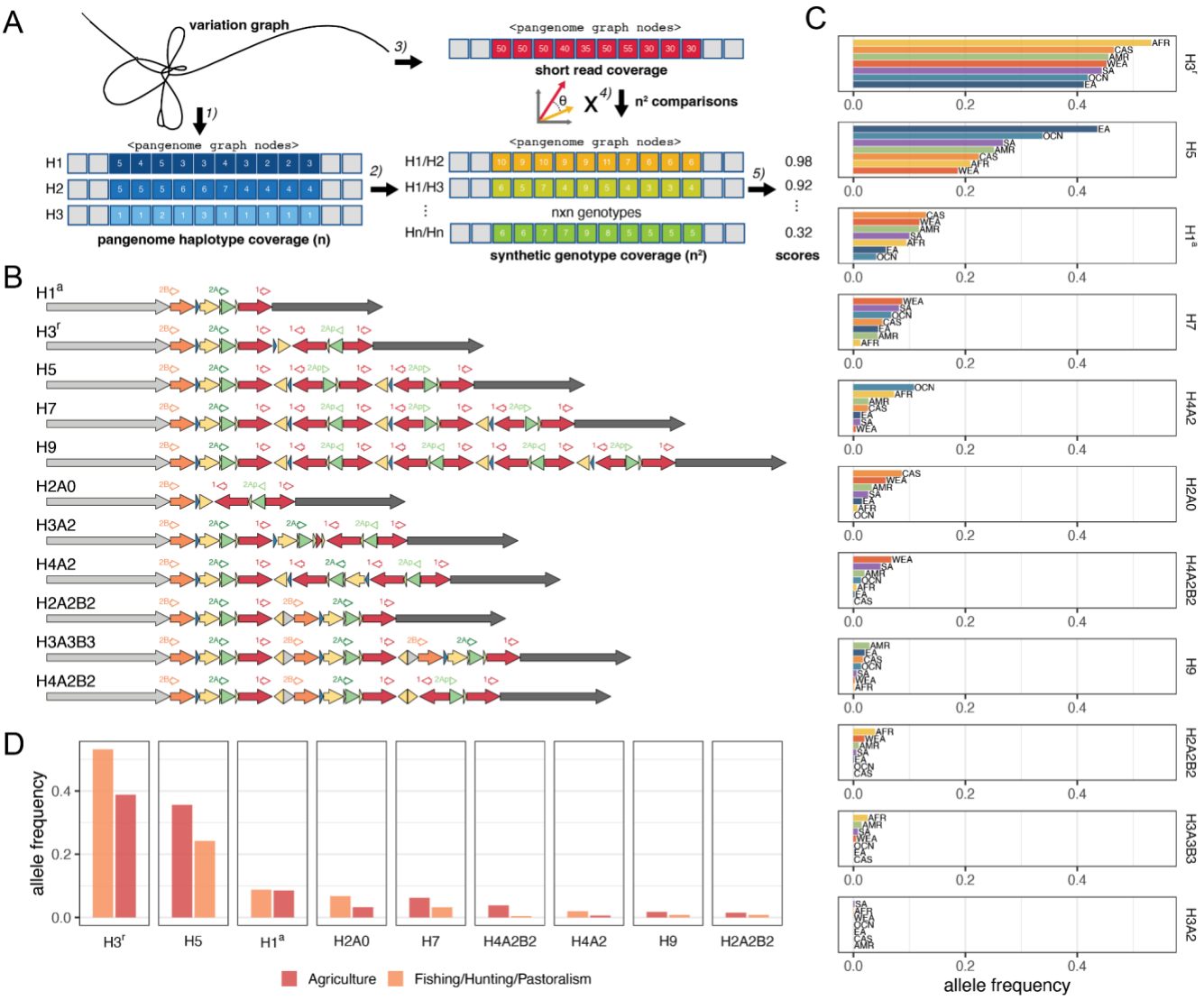


Figure 4 - Inference of complex structural haplotypes from short-read data. A) A schematic of the *haplotype deconvolution* approach to infer the pair of structural haplotypes present in a short-read sequenced individual. 1) A set of assembled haplotypes are mapped to a variation graph and coverage vectors are quantified over all nodes of the graph. 2) Synthetic genotype vectors are constructed from summing all pairs of haplotype vectors. 3) A short-read genome is mapped to the variation graph and

read depth is quantified over all nodes in the graph. 4) The short-read coverage vector is compared to all synthetic genotype vectors and scored (5) to identify the most likely haplotype pair present in the short-read sequenced individual. **B)** Consensus haplotype structures. **C)** Structural haplotype frequencies across continental populations in 3394 diverse humans (7188 haplotypes). **D)** Haplotype frequencies in 273 individuals from 31 modern populations with traditionally agricultural subsistence compared to fishing, hunting, and pastoralism based diets.

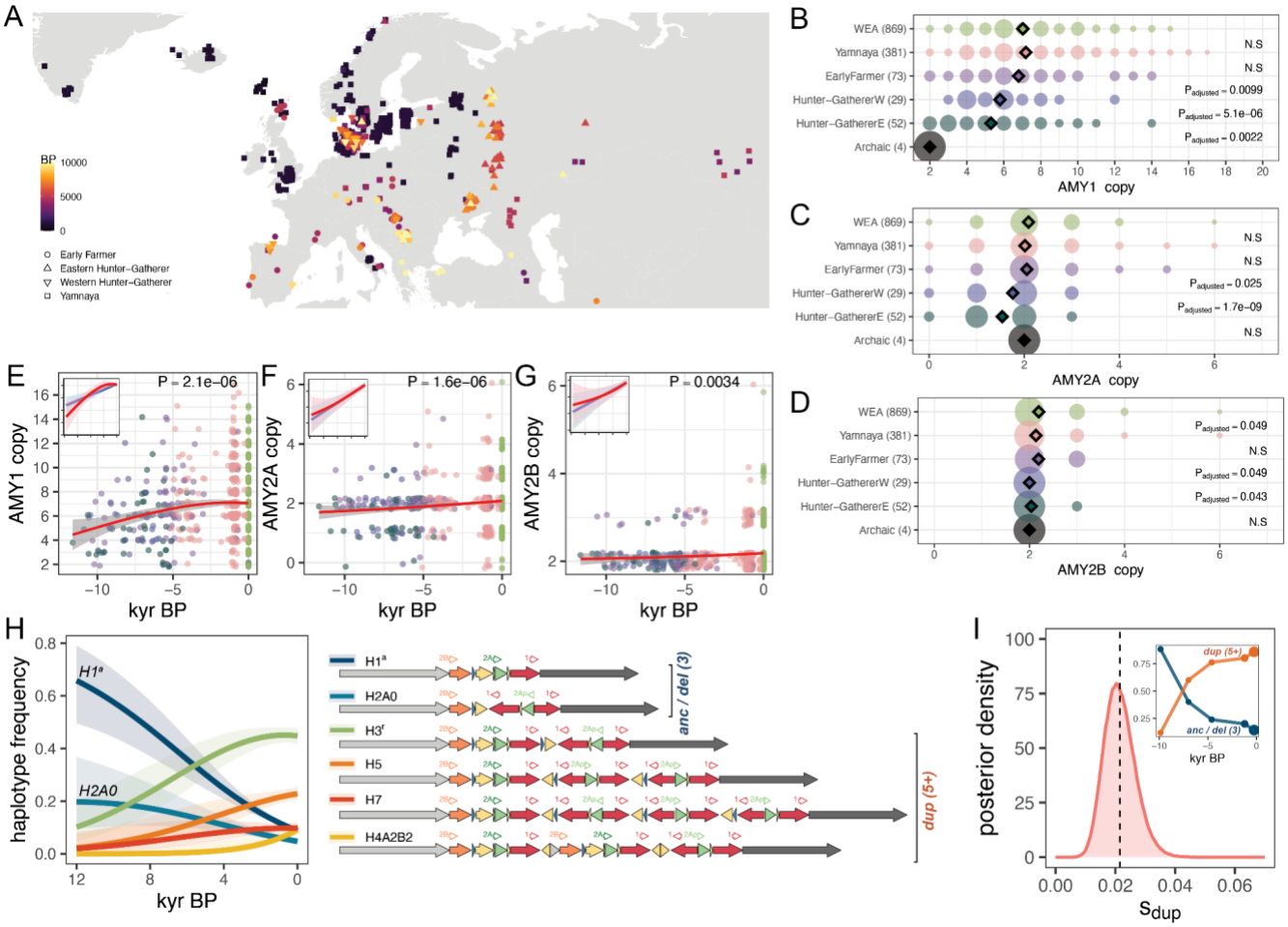
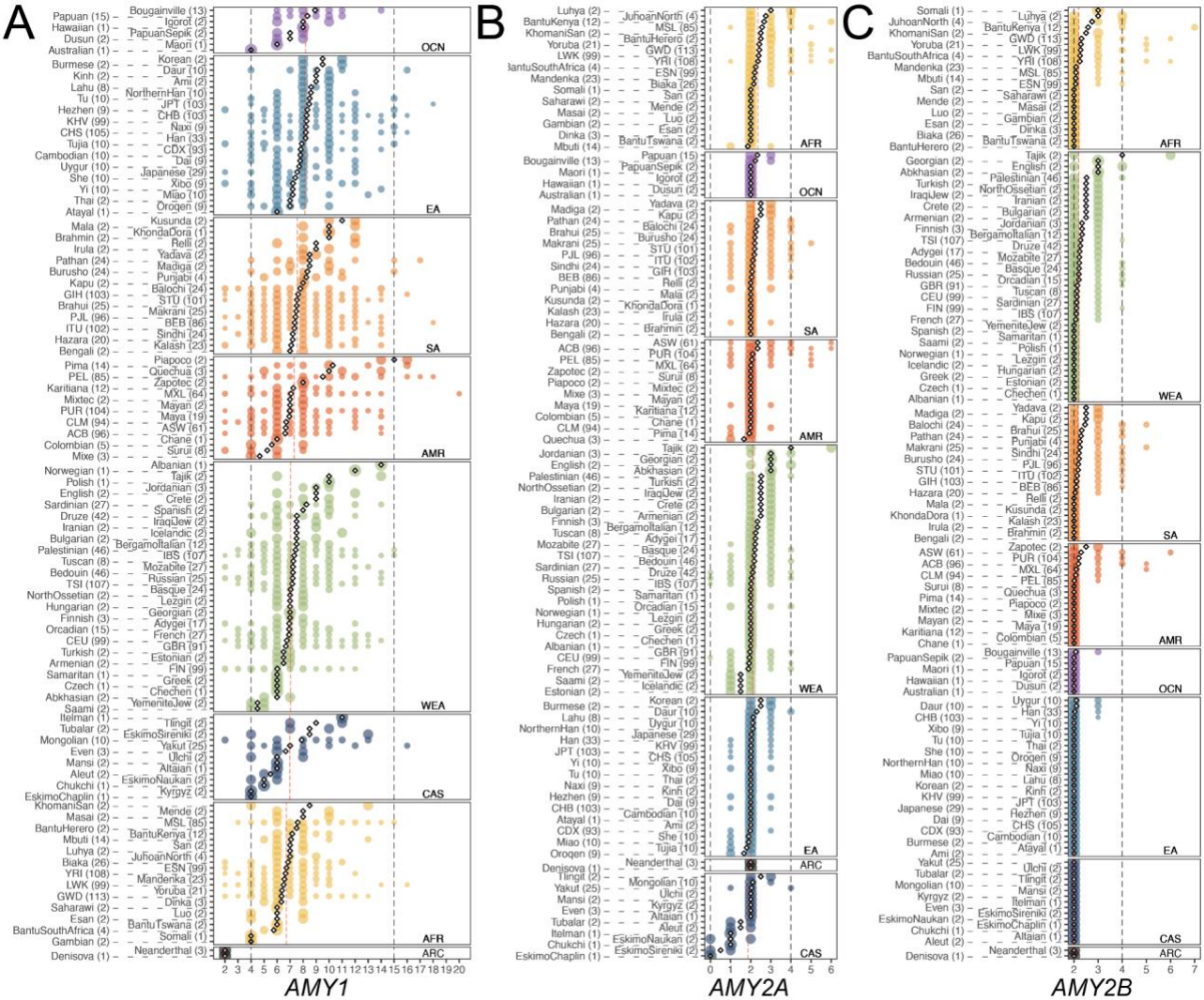
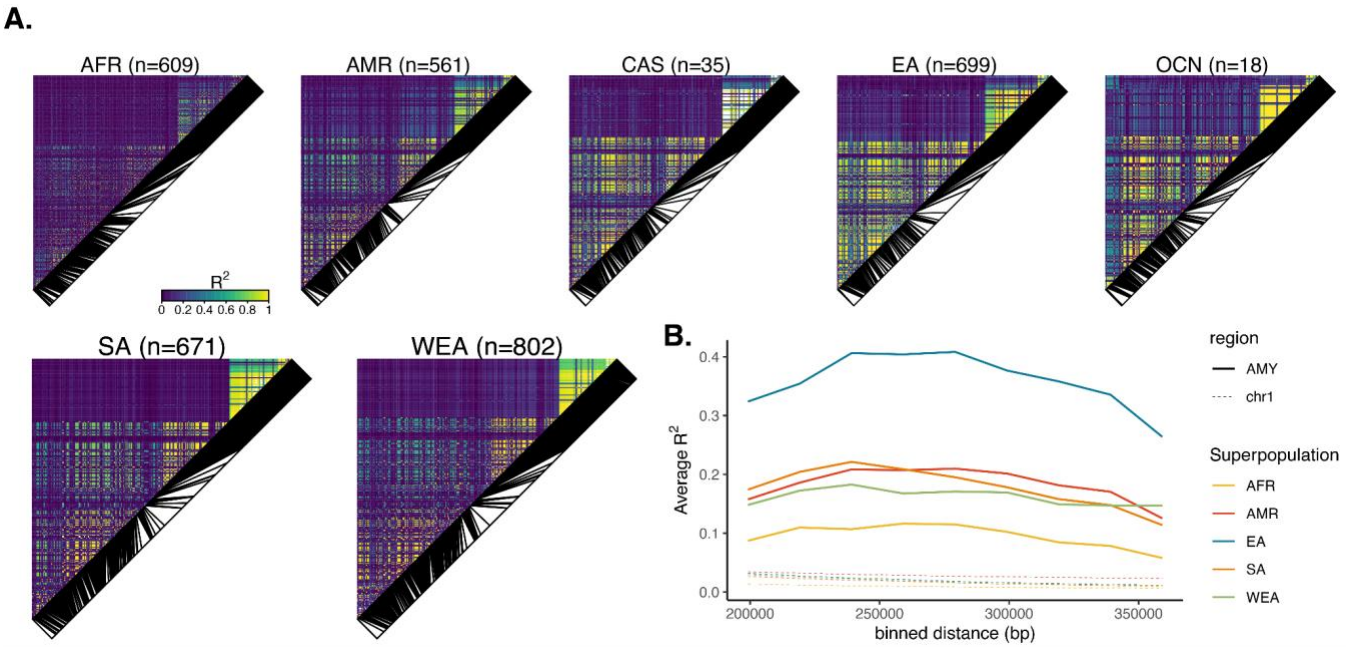


Figure 5 - Recent selection at the amylase locus in Europeans. A) Locations and ages of 534 West Eurasian ancient genomes from which amylase copy numbers were estimated. **B-D)** The distribution of *AMY1, AMY2A,* and *AMY2B* copy numbers in archaic hominids, ancient human Hunter Gatherer, Early Farmer, and Yamnaya individuals, and modern West Eurasians. **E-G)** Copy number genotypes plotted as a function of age overlaid with a smooth generalized additive model fit. Inset shows isolated linear model and generalized additive model fit to data. **H)** Haplotype trajectories fit by multinomial logistic regression for 6 haplotypes (right) present at >1% frequency in ancient and modern West Eurasians. Structures with the ancestral 3 total amylase copies (anc / del) are distinguished from duplicated haplotypes with ≥5 amylase genes. **I)** Posterior density of the estimated selection coefficient for duplicated haplotypes over the last 12,000 years (mean 0.022, indicated by dotted line, no estimates ≤ 0

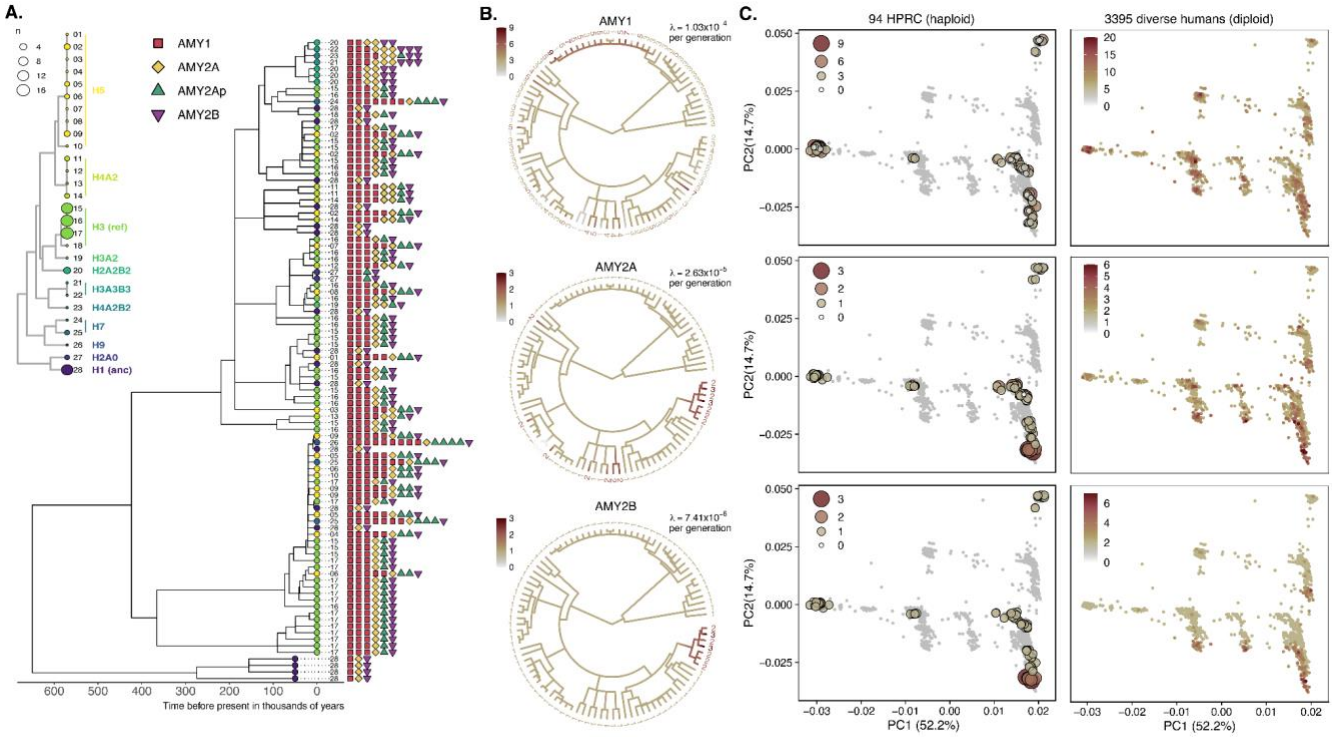
were observed in 1,000,000 MCMC iterations). Inset are binned observations of duplicated versus non-duplicated haplotype frequency trajectories.



Extended Data Figure 1 - Worldwide amylase subpopulation copy number diversity. A-C) Copy number distributions of *AMY1, AMY2A,* and *AMY2B* in 162 modern human populations and four archaic hominids. The size of each point is proportional to the proportion of individuals in the population with that genotype. Diamonds indicate the subpopulation mean, red dashed lines indicate the continental population mean, grey dashed lines indicate minimum and maximum subpopulation means.



Extended Data Figure 2 - LD in different populations worldwide . A) Heat maps of linkage disequilibrium (LD) for SNPs across a ~406 kb region spanning unique sequences on either side of the structurally variable region of amylase (SVR) in different populations from seven continental regions (Africa - AFR, America - AMR, Central Asia - CAS, East Asia - EA, Oceania - OCN, South Asia - SA and Western Eurasia - WEA). **B)** LD decay over genomic distances for groups with more than 100 samples, measured as the average R2 between SNP pairs on either side of the SVR (i.e. 190 kb - 370 kb apart) binned into intervals of 20,000 bp, compared to identically spaced SNPs in chromosome 1.



Extended Data Figure 3 - Reconstruction of the evolutionary history of amylase structural haplotypes using proximal unique sequence. A) A time-calibrated coalescent tree from the proximal non-duplicated region flanking the SVR (rightmost gray arrow in A until the recombination hotspot) across 94 assembled haplotypes (tree from the distal region in **Fig. 3**). The number next to each tip corresponds to the structural haplotype that the sequence is physically linked to and the color of the circle at each tip corresponds to its consensus haplotype structure (see inset structure tree). The copy numbers of each amylase gene and pseudogene are also shown next to the tips of the tree. **B)** Ancestral state reconstruction and mutation rate estimates for amylase gene copy number (archaic outgroups excluded). Branch color corresponds to copy number. **C)** A PCA from 94 haplotype assemblies and 3,395 diverse diploid human genomes from the proximal non-duplicated region flanking the SVR (PCA from the distal region in **Fig. 3**).

865