Adaptive Increase of Amylase Gene Copy Number in Peruvians Driven by Potato-rich Diets

Authors: Kendra Scheer[1]*, Luane J. B. Landau[1]*, Kelsey Jorgensen[2,3], Charikleia Karageorgiou[1], Lindsey Siao[1], Can Alkan[9], Angelis M. Morales-Rivera[10], Christopher Osborne[1], Obed Garcia[3], Laurel Pearson[4], Melisa Kiyamu[5], María Rivera-Chira[6,7], Fabiola Leon-Velarde[5], Frank Lee[11], Tom Brutsaert[8], Abigail W. Bigham[2]**, Omer Gokcumen[1]**

** Contributed equally; ** Correspondence*

Affiliations:

[1]Department of Biological Sciences, University at Buffalo, Buffalo, NY, USA

[2]Department of Anthropology, University of California, Los Angeles, CA, USA

[3]Department of Anthropology, University of Kansas, Lawrence, KS

[4]Department of Anthropology, The Pennsylvania State University, State College, PA

[5]Laboratorio de Fisiología del Transporte de Oxígeno y Adaptación a la Altura, Laboratorios de Investigación y Desarrollo, Facultad de Ciencias e Ingeniería, Universidad Peruana Cayetano Heredia, Lima, Peru

[6]Departamento de Ciencias Biológicas y Fisiológicas, Laboratorios de Investigación y Desarrollo, Facultad de Ciencias e Ingeniería, Universidad Peruana Cayetano Heredia, Lima, Perú

[7]Centro de Investigación para el Desarrollo Integral y Sostenible, Universidad Peruana Cayetano Heredia, Lima, Peru

[8]Department of Exercise Science, Syracuse University, Syracuse, NY, 13244

[9]Department of Computer Engineering, Bilkent University, Ankara, Turkey

[10]Departamento de Biología dentro de las Ciencias Naturales, Universidad de Puerto Rico en Cayey, Cayey, Puerto Rico

[11]Department of Pathology and Laboratory Medicine, University of Pennsylvania Perelman School of Medicine, Philadelphia, PA, USA

Summary/Abstract:

The salivary amylase gene (*AMY1*) exhibits remarkable copy number variation linked to dietary shifts in human evolution. While global studies highlight its structural complexity and association with starch-rich diets, localized selection patterns remain under explored. Here, we analyzed *AMY1* copy number in 3,723 individuals from 84 populations, revealing that Indigenous Peruvian Andean populations possess the highest *AMY1* copy number globally. A genome-wide analysis showed significantly higher amylase copy numbers in Peruvian Andean genomes compared to closely related populations. Further, we identified positive selection at the nucleotide level on a haplotype harboring at least five haploid *AMY1* copies, with a Peruvian Andean-specific expansion dated to/shortly after potato domestication (~6–10 kya). Using ultra-long-read sequencing, we demonstrated that recombination-based mutational mechanisms drive the formation of high-copy *AMY1* haplotypes. Our study provides a framework for investigating structurally complex loci and their role in human dietary adaptation.

Introduction:

The amylase enzyme plays a central role in the modern human diet. Its main function is to digest starch, which is a cornerstone of most contemporary diets, into simple sugars (Roberts & Whelan, 1960). Among humans, salivary amylase (*AMY1*) copy number variation demonstrates exceptional diversity, with individuals carrying between 2 and 20 copies per diploid genome, potentially affecting salivary amylase dosage and downstream starch hydrolysis (Usher et al., 2015). Early duplications of *AMY1* in hominin evolution facilitated further copy number expansions in humans (Yılmaz et al., 2024), which some studies have linked to agricultural diets (Bolognini et al., 2024; Mathieson & Mathieson, 2018; Perry et al., 2007). A similar trend has been observed in other species consuming broad range diets, such as dogs, which recently began cohabiting with humans and have also experienced an increase in amylase copy number in regions associated with the spread of agriculture (Arendt et al., 2016; Axelsson et al., 2013; Pajic et al., 2019). Further, this variation has been associated with metabolic traits, including the composition of the gastrointestinal microbiome, which may shape metabolic health (Poole et al., 2019). Despite these exciting associations, the structural complexity of the *AMY1* locus, characterized by recurrent duplications, deletions, and inversions, has complicated efforts to connect gene dosage to local dietary evolutionary pressures in extant populations.

The Andean Altiplano, where potato domestication began approximately 6,000–10,000 years ago (Haas Jr & Llave, 2015; Spooner et al., 2005), offers a compelling model for investigating the adaptive significance of *AMY1* copy number variation in extant human populations. Potatoes, a dietary staple in this region, are among the highest glycemic index foods, meaning that their consumption can cause spikes in blood glucose levels far more easily than other foods (Atkinson et al., 2021). For thousands of years, Andeans have relied heavily on the nutrition provided by this staple, which still constitutes over 50% of daily caloric intake in the modern Andean diet (Berti et al., 2010; Chen et al., 2024; Rose et al., 2009). Additionally, prior work has already demonstrated selection on the digestive enzyme, *MGAM*, in relation to the potato starch digestion process in Andeans (Jorgensen et al., 2023; Lindo et al., 2018). Thus, potatoes may have played a significant role in shaping the evolution of genes involved in carbohydrate metabolism among Andeans.

Here, we accurately genotyped *AMY1* copy numbers in a cohort of Peruvian Andean individuals, combining population genomic, haplotypic, and structural analyses. By integrating ultra-long-read sequencing with genome-wide selection tests, we reveal a signature of local adaptation to a starch-rich

diet in Andean populations. These findings contribute to our understanding of local dietary adaptations and the evolutionary significance of copy number variation in complex genomic loci.

Results:

Peruvians have the highest *AMY1* copy number in the world

We calculated gene copy number variation from 3,723 individuals across 84 global populations using several read-depth-based genotyping approaches (**Table S1,** see **Methods)**. We verified previous observations that Peruvians in Lima, Peru (PEL) possess strikingly elevated *AMY1* copy numbers compared to other global populations (**Figure 1A-D)** (Bolognini et al., 2024). In fact, PEL display the highest median copy number (PEL median = 10.07) among all 84 populations (worldwide median = 7.25) tested in this study.

Prior work has demonstrated shared continent-level trends regarding *AMY1* copy number (Bolognini et al., 2024; Usher et al., 2015). To assess the *AMY1* copy number pattern of variation among populations within each continent, we performed an ANOVA. This analysis revealed significant population structure in the Americas ($p < 2.2e-16$), while other continents showed relatively homogeneous *AMY1* distributions (**Figure 1B & 1C**). This high level of population structure is driven in part by PEL possessing the highest median copy number for all 84 populations tested, while other American populations exhibit some of the lowest median copy numbers **(Figure 1A)**. Notably, the Pima population (preferably known as Akimel O'odham) also shows a high median *AMY1* copy number (10.24), but has a much smaller sample size (Akimel O'odham/Pima n = 14 versus PEL n = 85).

In an independent cohort of Quechua speakers from Lima and Cerro de Pasco, Peru (n = 81), we replicated our findings of exceptionally high *AMY1* copy numbers. As a control, we included an additional cohort of Mexican Maya speakers (Tzeltal, Tzotzil, and Ch'ol) from Palenque, Chiapas, Mexico (n = 48) with shared Indigenous American ancestry with Andeans but distinct historical and cultural backgrounds (see **Table S1**, **S2**, and **Methods** for admixture analysis). We found that the Quechua population showed *AMY1* copy numbers comparable to the PEL population (mean = 9.9 and 9.74, respectively; range = 3.91-20.26 and 2.00-18.69), while the Maya population showed *AMY1* copy numbers (mean = 6.62, range = 2.4-13.25) that did not differ significantly from other Indigenous American populations (**Figure 1A & 1D**).

The results of previous studies suggest that the phenotypic effects of *AMY1* copy number may depend on surpassing a specific threshold, rather than increasing proportionally (Mejía-Benítez et al., 2015; Poole et al., 2019; Venkatapoorna et al., 2019). To explore this possibility, we conducted a one-sided binomial test comparing the proportion of individuals with higher *AMY1* copy numbers at different thresholds from 8-13 across all 84 study populations (**Table S3**). The results revealed that Andean populations most significantly deviated from the null expectation at ≥ 9 *AMY1* copies (Quechua and PEL Bonferroni-corrected p-values = 1.87×10^{-12} and 5.82×10^{-08}, respectively), indicating a remarkable increase in the frequency of "high" *AMY1* copy number individuals in these Peruvian groups as compared to other populations at this threshold (**Figure 1E**). These findings align with (Poole et al., 2019), which reported a positive correlation between *AMY1* copy number (low: ≤ 5, medium: 6-8; high: ≥ 9) and salivary amylase expression. It is of note that of all the Quechua and PEL samples, only a single PEL individual (HG01893) with large amounts of non-American ancestry has an *AMY1* copy number below four copies (**Figure 1D, Table S2**). Overall, Peruvian populations represent a compelling case for investigating the selection of high *AMY1* copy number haplotypes, making them the primary focus of our further analyses.

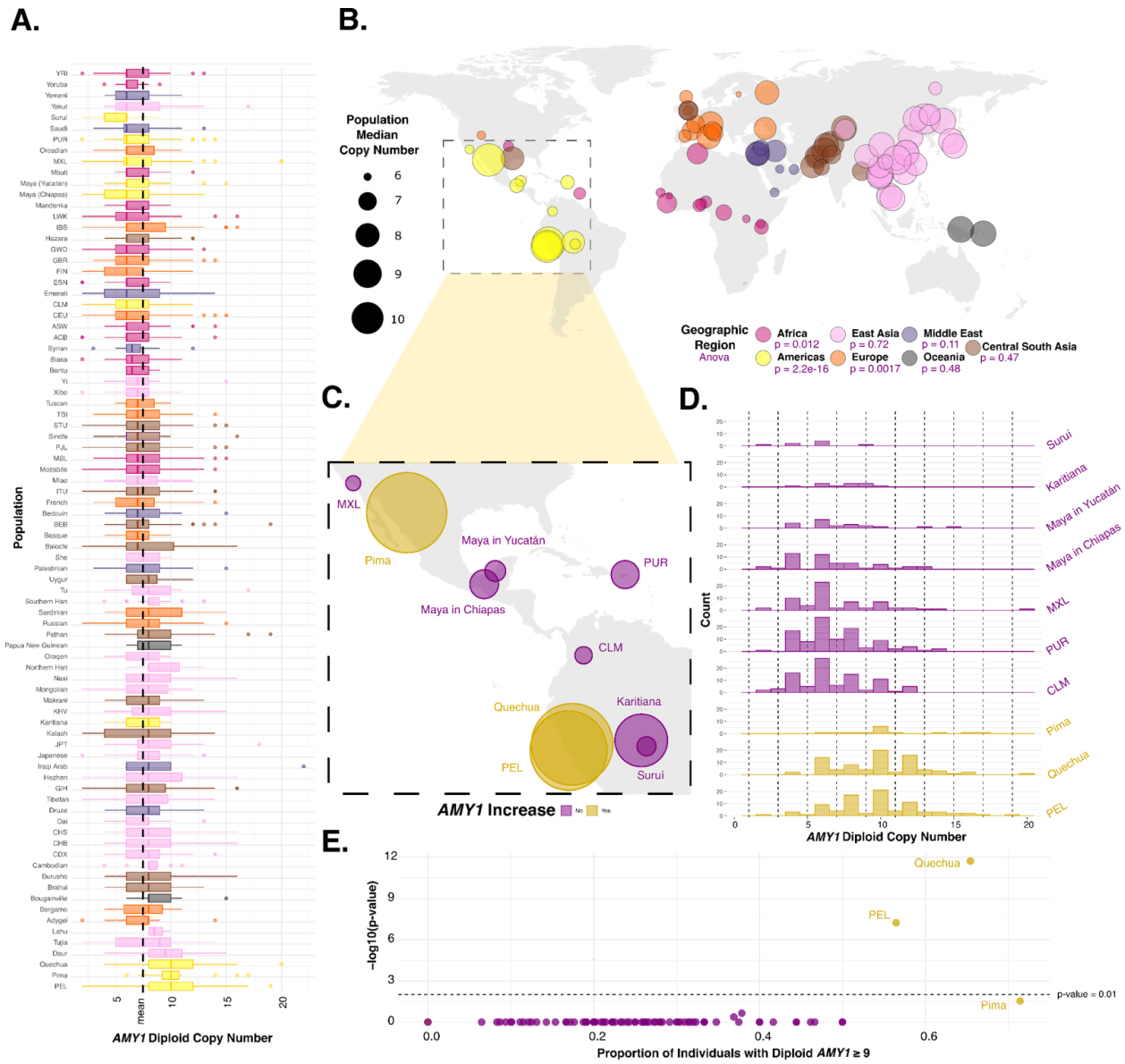

Figure 1. (**A**) Boxplots of *AMY1* diploid copy numbers across all populations, ranked by median. Indigenous American populations are in yellow. A black dotted line indicates the overall mean. (**B**) Geographic distribution of all populations with circle size indicating median *AMY1* copy number (per diploid cell) and colors representing regions. Within the geographic region, one-way ANOVA values are listed. (**C**) Zoom in on American populations from panel B with populations demonstrating a higher proportion in *AMY1* genes labeled in gold. (**D**) The histogram of the *AMY1* copy numbers of the American populations from panel C, the three previously highlighted indigenous populations show an increase in the *AMY1* copy number. (**E**) Dot plot that illustrates populations with significant proportions of individuals with *AMY1* diploid copy numbers ≥ 9. The x-axis represents the ratio of individuals with copy numbers ≥ 9 in each population, while the y-axis shows the Bonferroni-corrected -log(p-values) of a binomial significance test for *AMY1* ≥ 9.

AMY1 copy number is a genome-wide outlier in Peruvians, indicating a selective sweep

We hypothesized that selection in response to a starch-rich potato heavy diet drove the high *AMY1* copy number in Peruvian populations. To test this hypothesis, we conducted a genome-wide V_{ST} analysis, an F_{ST} analog designed for copy number variation (Redon et al., 2006). We prioritized the comparison between Quechua and Maya individuals due to their relatively higher Indigenous American ancestry compared to 1000 Genomes populations (**Figure 2D**). Specifically, we compared gene copy number variants from genes called using the program mrCaNaVaR between Peruvian Quechua individuals from Cerro de Pasco (n = 46) and Maya individuals from Palenque (n = 48) (**Figure 2A, Table S4**). After starting with ~18,000 protein coding genes, we filtered for genes that had a median difference or VST value of zero to we a dataset of 1,346 genes for our PEL and MXL comparison and a dataset of 7,163 genes for our Quechua and Maya comparison (**Table S5**). Our analysis revealed that the amylase genes (*AMY1* and *AMY2*) displayed one of the highest levels of population differentiation in copy number among all protein-coding genes, ranking in the top five genes (top ~0.02% of ~18,000). As this analysis cannot distinguish genes in the same gene families, we used an additional copy-number caller to demonstrate that this increase is only attributed to *AMY1* as the *AMY2* genes show no significant variation across American populations (**Figure S3 & S4**). To verify this observation, we performed 10,000 permutations where we generated V_{ST} values and median gene copy number differences from permutated samples composed of randomly assigned Quechua and Maya individuals. The observed values for *AMY1* copy number falls well outside of the permutated distributions (> 99th percentile, **Figure 2B** and **2C**).

To further validate our findings, we calculated population differentiation using 1000 Genomes data, comparing PEL and Mexicans from Los Angeles (MXL) (**Figure 2D**). This analysis corroborated our results, with *AMY1* copy number variation ranking among the highest median differences observed for common copy number variable genes (in the top 18 genes, top ~0.1%, **Figure S5**). Overall, population differentiation tests in two independent cohorts consistently indicated significantly higher *AMY1* copy numbers in Peruvian populations compared to closely related groups, supporting the hypothesis of positive selection acting on this variation.

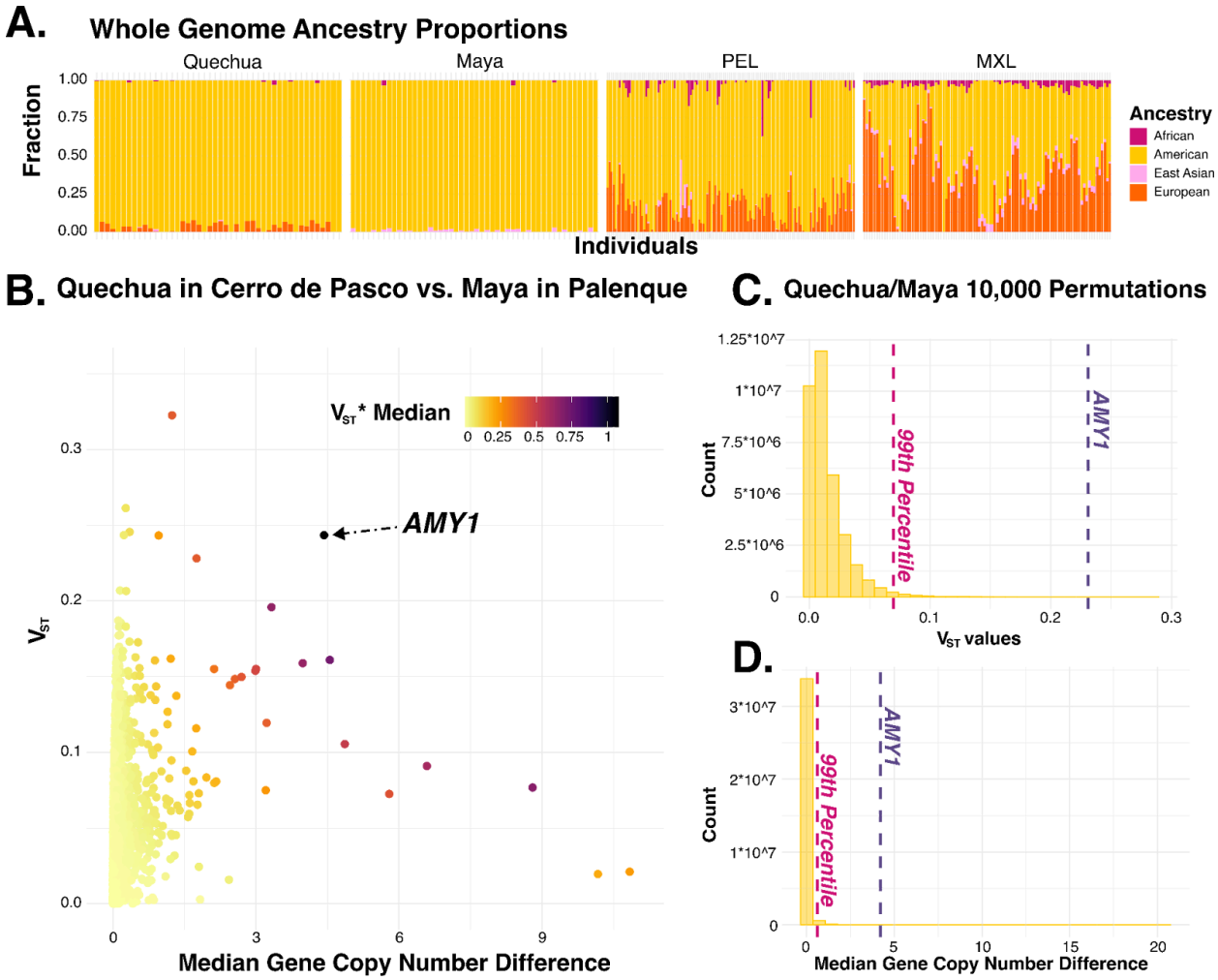

Figure 2: (**A**) Bar plots displaying the fraction of different ancestry proportions for populations tested in this study. The Maya from Chiapas population contains the least amount of non-Amerindian ancestry while MXL contains the most. (**B**) Scatterplot showing the relationship between V_{st} and median gene copy number difference in all protein-coding genes in Quechua from Cerro de Pasco and Maya from Palenque populations. Genes that have zero difference between the populations (both V_{st} and the Median difference = 0) are not plotted. (**C**) Histogram showing all positive Vst values across 10,000 random permutations of Quechua and Maya labels for all protein-coding genes (see **Methods**). The real *AMY1* value is labeled in purple and the 95th percentile is labeled in pink. (**D**) Histogram showing all positive median gene copy number differences across 10,000 random permutations of Quechua and Maya labels for all protein-coding genes.

SNV *AMY1* Haplotypes Provide Additional Evidence for Local Adaptation

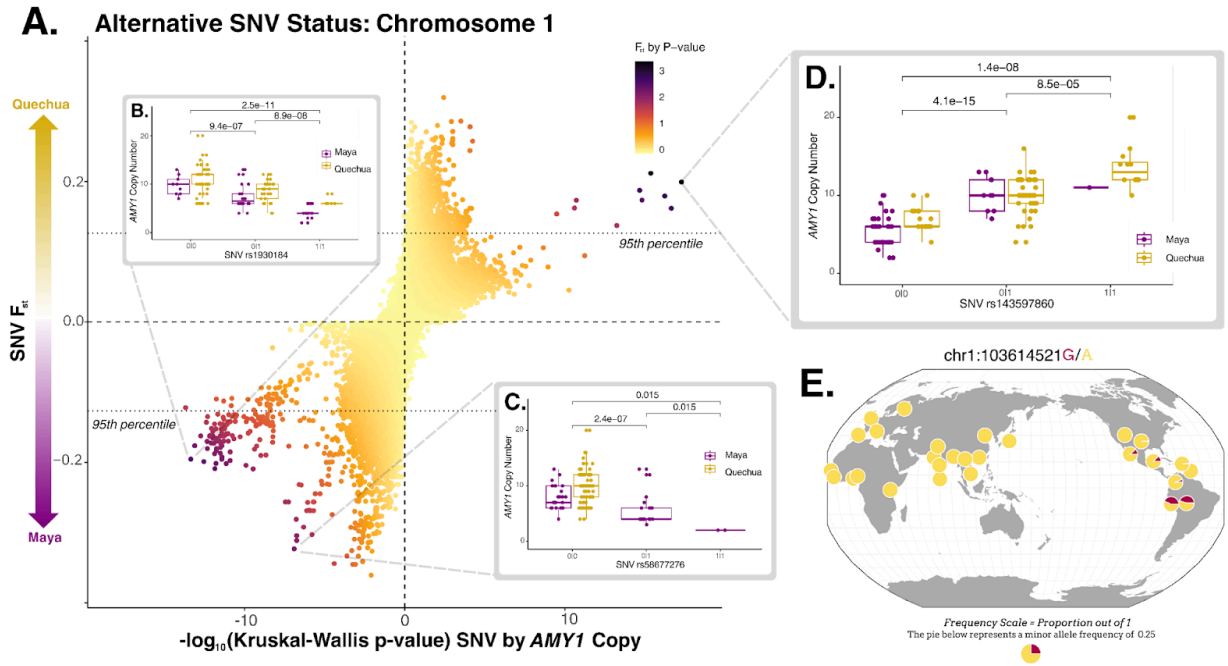

Figure 3: (**A**) Chromosome 1 SNVs in Quechua from Peru (top) and Maya from Chiapas (bottom) with Indigenous American ancestry in the amylase region (**Table S6**) in the alternative state. The y-axis shows directional F_{ST} with SNVs that have a higher frequency in the Quechua population represented with positive values and SNVs with a higher frequency in the Maya population represented with negative values. The x-axis depicts SNV association with diploid *AMY1* copy number. Colors indicate the relationship between these statistics. (**B**) The SNV (rs1930184) with the highest negative correlation to *AMY1* copy number in the alternative state is significantly more frequent in Maya. A zoom-in shows a stepwise association with *AMY1* across populations. The position is based on the HG38 reference genome. (**C**) A SNV (rs58677276) found only in the alternative state in Maya, which is negatively associated with the *AMY1* copy number. (**D**) The SNV (rs143597860) with the strongest association with *AMY1* copy number across populations, with a stepwise association as in (C). (**E**) rs143597860 frequency across the 1000 Genomes project, Quechua, and Maya populations.

To understand the haplotype structure associated with high *AMY1* copy numbers, we analyzed linkage disequilibrium (LD) at the amylase locus and flanking regions (HG38 Chr1: 103,348,464-103,830,994) (Kruskal & Wallis, 1952). We found that flanking single nucleotide variants (SNVs) in the Peruvian Quechua cohort are significantly associated with the *AMY1* gene copy number (**Figure 3A-D**). This finding is particularly exciting because it contrasts with the global trend, where LD between flanking SNVs and *AMY1* copy number is typically low due to high recombination rates, recurrent copy number variation, and gene conversion (Yılmaz et al., 2024). The presence of SNVs in LD with diploid *AMY1* copy number in Peruvian Quechua suggests that a recent selective sweep may have acted on haplotypes harboring high numbers of *AMY1* copies, leading to long range locus LD, unlike in

other populations. Indeed, we observed a noticeable increase in LD across the amylase locus in both the PEL and Quechua cohorts compared to other populations (**Figure S6**).

We hypothesized that if the increased LD at the amylase locus is a result of a recent selective sweep, then the same SNVs that showed LD with *AMY1* copy number should exhibit increased F_{ST} between Peruvian Quechua and Maya populations, similar to the trends observed in the V_{ST} analysis. Concordant with these expectations, we found that the SNVs tagging the *AMY1* copy numbers show significantly high F_{ST} values (for example F_{ST} for rs1930184 = 0.195, rs58677276 = 0.311, rs143597860 = 0.199, all above 95th percentile) compared to the rest of chromosome 1 (**Figures 3A-D**).

These results support the notion that haplotypes harboring high *AMY1* copies are under selection in Peruvian Andean populations. For example, the SNV rs143597860, which shows the strongest association with high *AMY1* copy numbers in Quechua and Maya individuals, is found at low frequencies across non-Andean populations in the Americas excluding Akimel O'odham/Pima (50%) in the Americas (average of 6.1% in 1000 Genomes populations and 16.6% in HGDP populations) (**Figure 3E**). In contrast, the alternative allele (G) frequency rises sharply to 44% in PEL and 46% in Quechua. Using a 25-year generation time, a joint clock estimate based on 1000 Genomes and SGDP (Simons Genome Diversity Project) data places the allele's age at ~17,945 years, coinciding with early migrations to the Americas (**Figure S7**) (Albers & McVean, 2020; Willerslev & Meltzer, 2021). Thus, this SNV likely existed among the ancestral population to both Maya and Quechua increasing in frequency among Andeans in a manner concordant with selection on standing variation due to the hypothesized selective sweep on the high amylase copy numbers.

To test whether demographic bottlenecks alone could explain the observed allele frequency differentiation, we performed neutral simulations using Ms-prime (see **Methods**). Under published divergence times and effective population sizes, as well as our own estimates via *Relate* (**Figure S8**), we could not replicate the observed allele frequency shift for rs143597860 (**Table S7**). Only when assuming unrealistically low effective population sizes (e.g., $N_e = 250$) in either Quechua or Maya could we recover comparable differentiation in greater than 5.2% of simulated SNVs (**Figure S9, Table S7**). Collectively, these findings indicate that selection on pre-existing haplotypes carrying high AMY1 copies, rather than neutral processes, drove the allele's increased frequency in Peruvian Andeans.

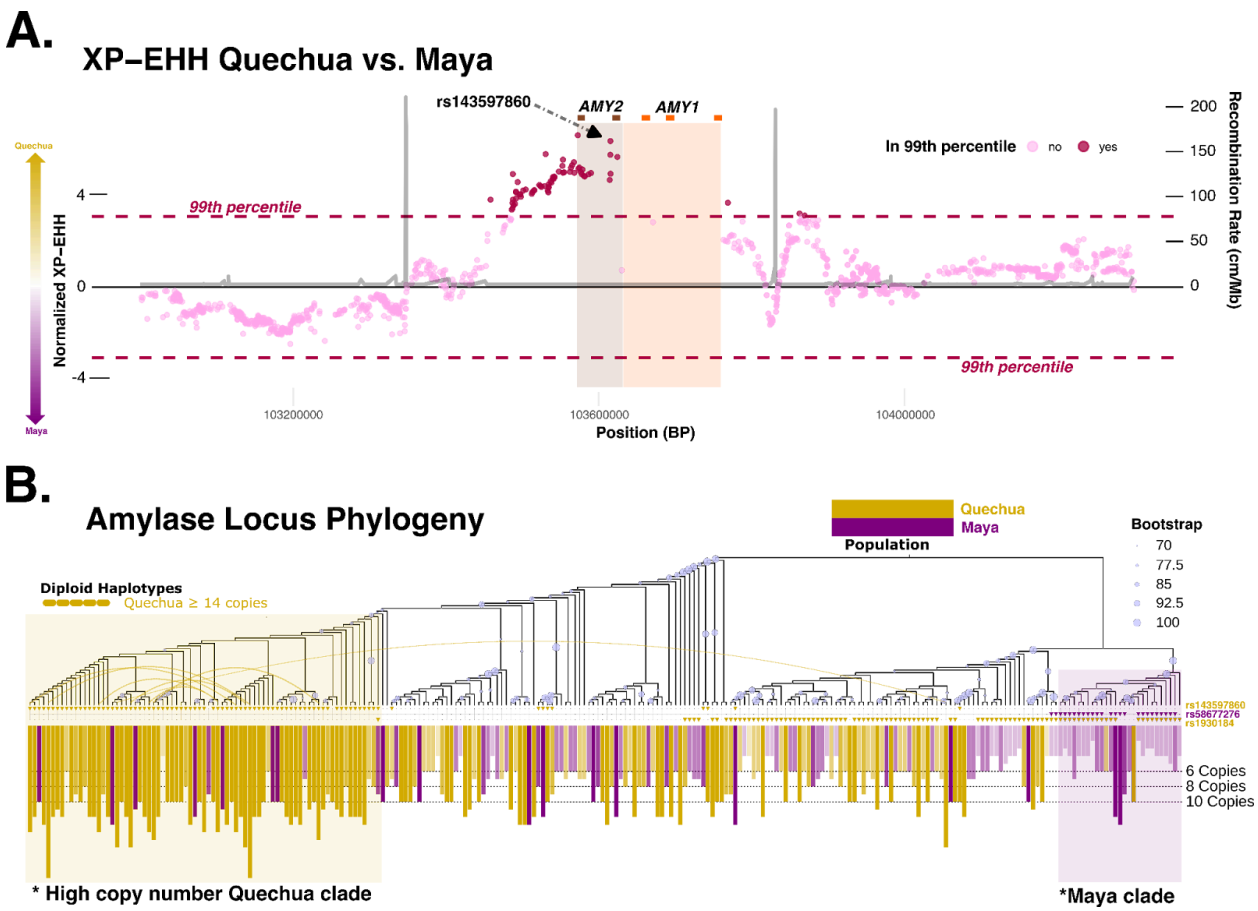

Figure 4: (**A**) XP-EHH between Quechua and Maya samples in the amylase locus and flanking regions along with regional recombination rates. Brown squares are *AMY2B* and *AMY2A* while orange squares are *AMY1A*, *AMY1B*, and *AMY1C*. Chromosome-wide significant values are in red. Positive XP-EHH values have longer haplotypes in Quechua; negative XP-EHH values have longer haplotypes in Maya. (**B**) A phylogeny of the haploid amylase and flanking regions. Bars are colored by population and shaded by *AMY1* copy number, with SNVs from **Figure 3B**, **C**, and **D** labeled. All Quechua individuals with ≥ 14 *AMY1* copies (exceeding the max Maya copy numbers) have linked haplotypes.

Our observation that SNVs that are in unusually high LD with *AMY1* copy numbers exhibit significant population differentiation in Quechua suggests that LD patterns for these SNVs should differ markedly between Andean Peruvians and the closely related Maya population. To test this, we calculated cross-population extended haplotype homozygosity (XP-EHH) between Maya and Quechua genomes (Sabeti et al., 2007). Multiple SNVs flanking the *AMY1* locus displayed chromosome-wide significant XP-EHH values (> 95th percentile **Figure 4A, Figure S10**), including SNV rs143597860. This result is conservative, as most genetic maps are based on SNP data and exclude recombination events within the *AMY1* locus due to segmental duplications that evolve more rapidly than SNPs. Moreover, the significant XP-EHH block flanks the *AMY1* locus and contains multiple linked SNVs associated with *AMY1* copy number, providing strong evidence that the haplotype is under recent positive selection in Andean Peruvians.

Next, to visualize the amylase haplotypic variation, we constructed a phylogenetic tree using phased SNVs from Quechua and Maya individuals encompassing the amylase locus and flanking regions (Hg38: Chr1: 103,348,464-103,830,994) (**Figure 4B**). On this phylogeny, we superimposed the diploid *AMY1* copy number and the presence or absence of the three high F_{ST} SNVs described above. We found that rs58677276 appears mainly in a clade harboring diverse haplotypes that is predominantly composed of Maya individuals (25/27 haploid genomes) with low *AMY1* copy numbers (median = 4). This phenomenon may have been caused by a deletion event triggering the loss of *AMY1A* and *1B*, the core mutational unit, and resulting in a static single *AMY1* copy number, which may be present in this clade's haploid genomes.

More relevant to the proposed selective sweep, we identified a clade primarily containing the reference allele for rs1930184, within which a subclade predominantly carries the alternative allele for rs143597860, associated with the selected haplotype. This subclade is largely Quechua (67/77 haploid genomes) and exhibits high *AMY1* copy numbers (median = 12), accounting for over half (68 of 126) of all haploid genomes in our dataset who have diploid *AMY1* copy numbers ≥ 9. This remains the case when recreating the phylogeny to include Luhya in Webuye, Kenya (LWK) as an outgroup (**Figure S11**). Since this phylogeny is constructed from haploid phased SNV data, it represents only one haploid genome per individual. Therefore, we hypothesize that individuals showing high *AMY1* copy numbers in branches outside the high copy number clade likely carry their additional copies through the haplotype found in this clade. This discrepancy arises because while the phylogeny is haploid, the gene copy number data is diploid. As a result, some individuals with high total copy numbers may possess one high-copy haplotype from the identified clade and another low-copy haplotype present elsewhere in the phylogeny. This interpretation aligns with the data, as 13 out of 14 Quechua SNV haplotypes associated with a diploid *AMY1* copy number of 14—an unusually high count requiring an average of seven copies per haplotype—fall within this subclade. We used *Relate* to date the expansion of the haplotypes associated with this clade (defined by the presence of rs143597860) and found that the timing of the primarily Quechua-specific haplotypes carrying this SNV immediately postdates the domestication of potatoes in this region (**Figure S12**) (Haas Jr & Llave, 2015; Spooner et al., 2005).

Structural haplotypes of Peruvian populations

The structural haplotypic variation of the amylase locus in Peruvians is not well-characterized. Previous studies showed that *AMY1* copy number variation involves non-allelic homologous recombination (NAHR) of a mutational unit comprising *AMY1A* and inverted *AMY1B* (**Figure S2**) (Yılmaz et al., 2024). To address whether the unusually high *AMY1* copy number haplotypes in Peruvian

Andeans were due to the same mutational mechanism, we analyzed data from 40 existing haplotype assemblies, including 10 Peruvian samples (**Table S8**) (Bolognini et al., 2024; Yılmaz et al., 2024). The existing assemblies allowed us to construct a pangenome of the amylase locus representing 33 distinct structural haplotypes (**Table S8**, **see Github**). Among these, 14 haplotypes were present in Peruvians, yet all of them also occurred in other global populations (**Figure 5B**).

To further investigate, we generated long-read sequencing data from six PEL individuals representing a range of copy number alleles, including those carrying the highest *AMY1* copy numbers **Figure 5A**). These six PEL samples are children in independent trios with previously sequenced genomes (**Supplementary Methods)**. Specifically, we generated Oxford Nanopore long-read sequencing data ranging from 14–45× coverage (**Supplementary Table S9**), achieving an N50 of ~20 kb. For two samples, we additionally obtained ultra-long reads with an N50 of ~50 kb. Using this dataset, we constructed phased contigs through local assembly tools (**Figure S13, Supplementary Methods, Supplementary Table S9**). We analyzed the pangenome and individual contigs for sequence signatures indicative of previously characterized microhomology-mediated break-induced replication (MMBIR) events, which have been implicated in rare occurrences of amylase duplications (Yılmaz et al., 2024). No such signatures were detected in the Peruvian samples, either within the pangenome or the contigs (**Supplementary Methods**), suggesting that observed structural variation at the *AMY1* locus was not driven by MMBIR mechanisms in this population. This conclusion is further supported by the absence of significant variation in *AMY2A* and *AMY2B* copy numbers in Peruvian Andeans (**Figure S3 & S4**), which are shown to be typically affected by MMBIR.

Next, we examined the above-mentioned 40 assembled haplotypes to confirm that the observed *AMY1* copy number variation in the Peruvian samples could be explained by the recently described NAHR mechanisms involving both *AMY1A* and *AMY1B* (**Figure 5B**). We competitively mapped the newly sequenced phased long reads from six PEL individuals (12 phased samples) onto the 40 assembled haplotypes. For each phased sample, we calculated the mean read depth and standard deviation of read depth across all assembled haplotypes to identify the best fit (e.g., **Figure 5C, Supplementary Methods**). We assumed that the best-matching haplotype would exhibit the most uniform coverage across assemblies (i.e., lowest standard deviation) and an appropriate mean read depth (see **Supplementary Methods**). These assignments were further validated by copy number estimates derived from GeneToCN and Droplet Digital PCR analyses (**Table S1, S10)**. Overall, we successfully assigned 8 out of 12 newly sequenced phased samples to known reference haplotypes (**Table S9, S11**).

Two phased samples (HG01975 and HG02106), with an expected 14 diploid copies each, did not match the expected copy numbers by the read-depth approach (**Supplementary Methods**). We manually inspected the mapping to known haplotypes and found that the alignments to other known 3-, 5-, and 9-copy haplotypes were not uniform. Furthermore, despite being the best match, we observed apparent gaps in the mapping even for the 9-copy previously resolved Peruvian haplotype (**Figure 5D**). This led us to hypothesize that, despite structural similarities, the Peruvian haplotypes may differ in local sequence variation. To test this, we generated an estimated 7-copy haplotype by deleting one mutational unit from the 9-copy Peruvian haplotype (**Figure S2**). This adjustment significantly improved the mapping (**Table S11**), providing the best fit for these phased samples (**Figure 5C**), and allowed us to conclude that this individual may carry two 7-copy haplotypes, previously unreported and observed exclusively in PEL to dates (**Figure 5D**).

Together, these results suggest that the majority of high *AMY1* copy number haplotypes observed in Peruvians already existed in global populations and have evolved through the same mutational mechanisms reported earlier (Yılmaz et al., 2024). Most Peruvian haplotypes have been documented previously, except a seven-copy haplotype, which is within the expected realm of mutational possibilities but may harbor Peruvian-specific variants at the local scale.

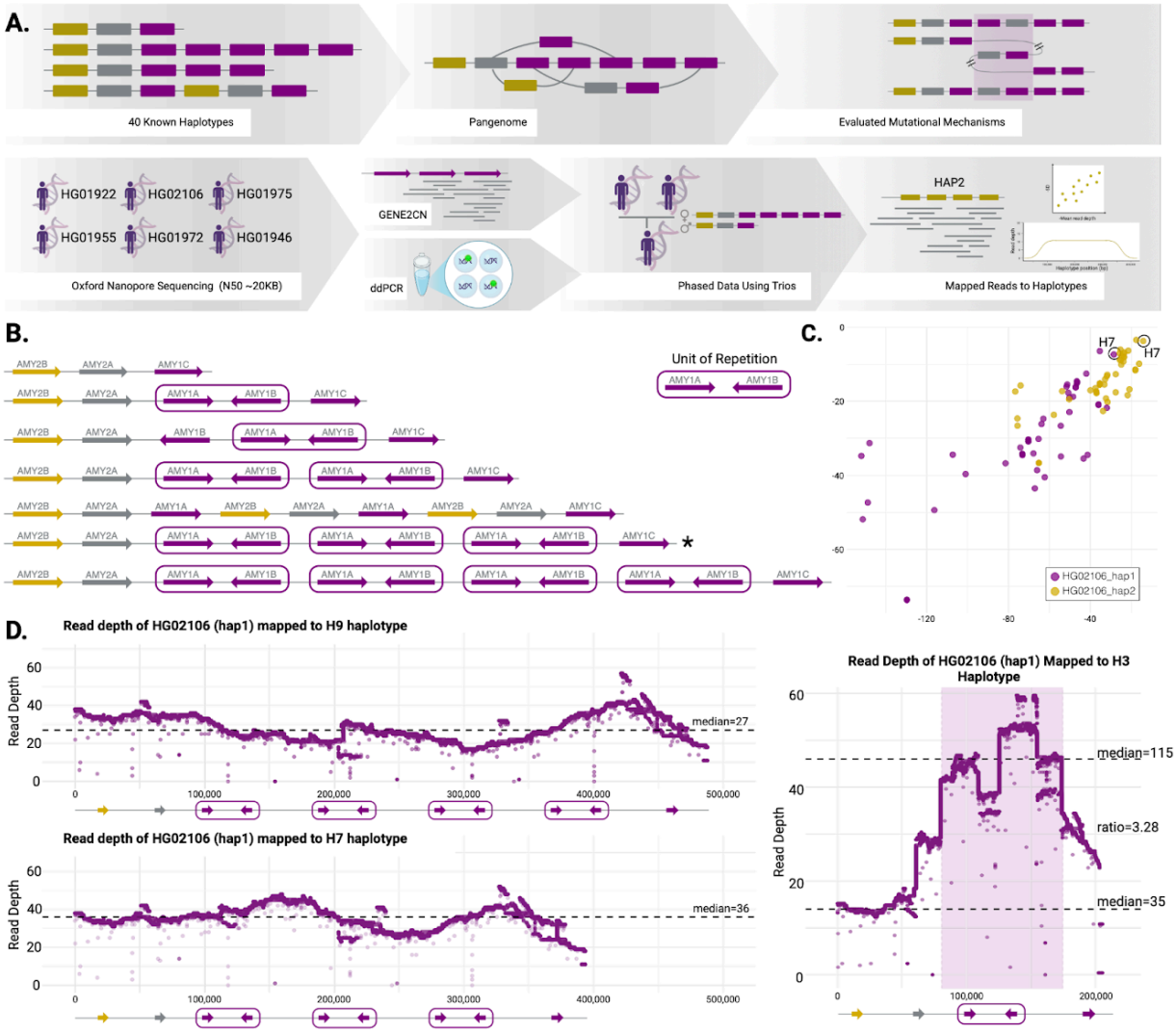

Figure 5: Peruvian populations have uniquely long haplotypes while maintaining the same duplication mechanisms. (**A**) Overview of the analytical approach used to identify and characterize Peruvian haplotypes and their underlying mutational mechanisms. (**B**) Haplotypes found in Peruvian individuals. Known haplotype structures reported in previous studies are shown alongside newly inferred Peruvian haplotypes identified in this study. The asterisk corresponds to the manually curated H7 haplotype present in Peruvians. (**C**) Scatter plot illustrating the relationship between mean read depth (y-axis) and standard deviation (x-axis) for individual HG02106's reads mapped against all known haplotypes. The highlighted data points represent the best-fitting mappings for HG02106's two haplotypes (H7), characterized by their lower standard deviation values. (**D**) The upper panel on the left shows the per-nucleotide read depth for HG02106 when aligned to the H9 haplotype. In contrast, the lower panel shows the alignment to the inferred H7 haplotype, demonstrating higher median read depth and reduced standard deviation, suggesting a better haplotype fit. On the right, the read depth when aligned to an H3 haplotype, shows the ratio between the region with invariable copy number of *AMY2B* and the region comprising the repeat *AMY1A*/*AMY1B*.

Discussion:

The amylase locus has long been discussed in the context of human dietary adaptation. Previous studies showed that the salivary amylase gene (*AMY1)* exhibits remarkable copy number variation, evolving rapidly and showing associations with dietary preferences among human populations (Bolognini et al., 2024; Perry et al., 2007; Yılmaz et al., 2024). However, direct evidence of local selection on *AMY1* has remained elusive. Here, we provide robust evidence for recent selection acting on *AMY1* copy number in Peruvian Andeans. By accurately genotyping *AMY1* copy number variation in 3,723 individuals and reconstructing the structural architecture in six Peruvian genomes, we demonstrate that Andean Peruvians exhibit exceptionally high population differentiation and harbor unusual, derived *AMY1* haplotypes. This supports the hypothesis that positive selection at this locus has enhanced starch digestion in response to a potato rich diet. Furthermore, we show that previously described recombination-based structural mutations—specifically those involving duplications or deletions of two *AMY1* genes in a single mutational step—drive the formation of extremely high-copy *AMY1* haplotypes in Peruvians. Our findings have broad implications for understanding the genetic basis of dietary adaptations to local ecologies and the role of selection in shaping structurally complex loci.

Previous studies have shown that *AMY1* copy number increased in the human lineage prior to agriculture (Mathieson & Mathieson, 2018) and has been shaped by frequent gains and losses. This same increase in amylase copy number has been demonstrated in species consuming a broad range diet that have cohabitated with humans after the agricultural revolution (Arendt et al., 2016; Axelsson et al., 2013; Pajic et al., 2019). Thus, high-copy *AMY1* haplotypes likely existed in populations ancestral to Andean Peruvians. Our results suggest that a soft selective sweep dramatically increased the allele frequency of these high-copy haplotypes in the Andean Peruvian population. This observation aligns with recent work on human adaptive evolution emphasizing the role of local selection on standing variation (Schrider & Kern, 2017), as demonstrated convincingly for skin color (Crawford et al., 2017; A. R. Martin et al., 2017). It also is noteworthy that *AMY1* copy number variation emerges from recurrent NAHR events. The rates of these events are likely influenced by haplotype length where the higher number of *AMY1* gene copies lead to higher probability of NAHR. The selective sweep that led to higher allele frequencies of high-copy haplotypes may lead to additional *de novo AMY1* copy number variation. This *de novo* variation could have been subject to selection alongside pre-existing variation. Therefore, it is plausible that both selection on standing variation and *de novo* mutations contributed to the elevated frequency of amylase haplotypes with more than five haploid *AMY1* copies, leading to an increased proportion of

individuals with ≥ 9 diploid copies in Peruvians (**Figure 3C**). An open question remains as to whether there is an upper limit to *AMY1* copy number expansion and variation. Potential constraints may arise from genomic instability as a function of increased repeated segments or homologous alignment issues caused by copy number mismatches between maternal and paternal chromosomes. Investigating these limitations presents an exciting avenue for future research.

The selection on the amylase locus reported in our study sheds light on the previously suggested starch dietary adaptations among Andeans (Jorgensen et al., 2023; Lindo et al., 2018). Populations living in the Andes domesticated potatoes approximately 6,000-10,000 years ago (Haas Jr & Llave, 2015; Y. Li et al., 2018; Spooner et al., 2005). Since then, their descendants have been consuming an extremely starch-rich diet—a cultural practice that continues today (Berti et al., 2010). Indeed, researchers have found that the *MGAM* gene, which encodes the maltase-glucoamylase enzyme, shows evidence of selection in Andean populations (Jorgensen et al., 2023; Lindo et al., 2018) with our samples deriving from a subset of the individuals in Jorgensen et al. This enzyme functions downstream of amylase in the starch digestion pathway, breaking down oligosaccharides generated by amylase activity on complex carbohydrates. These findings resonate with the general notion that the entire starch digestion pathway may be under adaptive pressure, involving multiple genes—including *AMY1*—as we have demonstrated here.

The observation that starch-digesting pathways are under adaptive pressure in populations with high-starch diets raises questions about the functional impact of high *AMY1* copy numbers. Previous studies have shown that individuals with "high" amylase copy numbers exhibit both higher salivary amylase expression and greater oral microbiome diversity (Poole et al., 2019). Both consequences may be beneficial to metabolic health in the context of high-starch diets. Cooked potatoes have one of the highest glycemic index values among cultivated starches, indicating blood sugar levels rise quickly after consumption (Atkinson et al., 2021). Excessive potato consumption has been associated with negative metabolic health outcomes (Guo et al., 2021), and *AMY1* copy number has been shown to influence metabolic response to starch intake by modulating glycemic response (Atkinson et al., 2018). Specifically, higher salivary amylase levels have been linked to a faster cephalic insulin response (Wiedemann et al., 2020), potentially facilitating better adaptation to high-starch diets. Given that a majority of Peruvians of Andean ancestry (61%) carry “high” *AMY1* copy numbers, their glycemic response to potato consumption may be more rapid, promoting efficient starch digestion and potentially conferring metabolic health benefits. Thus, our findings are concordant with the hypothesis that high *AMY1* copy numbers in Peruvians are adaptive because they help metabolize local potato-starch diets.

We provide a novel approach to investigating local adaptation acting on structurally complex genomic loci. By integrating long-read sequencing-based haplotype reconstruction with accurate genotyping of copy number variation, we explored the evolutionary history of the amylase locus. This integrative approach allows for investigating other similarly complex loci, including structural variations affecting other digestive enzyme-coding genes such as pepsinogen (*PGA*) and lipase (*LPA*). Another interesting aspect of our findings is that adaptive signatures may affect multiple genes within the starch-digestion pathway in Andeans. This observation raises the possibility of co-adaptation of variation in multiple genes, such as amylase and *MGAM,* potentially identifying convergent effects. Future functional studies examining the impact of increased *AMY1* copy number and Andean-specific *MGAM* haplotypes on potato starch digestion could provide key insights into how these genes evolved in response to the unique selective pressures of a high-potato-starch diet, distinguishing it from other starch sources. It would be especially interesting to conduct similar analyses in other populations—for example, in the Akimel O'odham/Pima, where we also found high amylase gene copy numbers but had a small sample size and also in ancient genomes when genome-wide shotgun sequencing with sufficient depth is achieved. Overall, our study exemplifies an exciting avenue for future research integrating novel long-read sequencing and computational tools to investigate the evolution of structurally complex loci at the local scale and provides a framework for broader public health impact of *AMY1* copy number variation in global populations.

Acknowledgments

We are incredibly grateful to the study participants from Lima, Peru, Cerro de Pasco, Peru, and Palenque, Chiapas, Mexico. We are also thankful to Dr. Maria Inclan (Centro de Investigación y Docencia Económicas), Dr Sudipta Ghosh (NEHU University, Shillong, India), Nate Bartman, Jason Howard, Jacqueline Imse, Kevin Heffernan (Syracuse University), Mark Olfert (University of West Virginia), Francisco Villafuerte (Universidad Peruana Cayetano Heredia), Katarina Evans (CUNY), Neha Angel, Thea Spindel, Sarah Burke, Sarita Greer, Adam Conner, Isabel Meda, and Alber Aqil.

This work was supported by NSF grants BCS-1132310 (to A.W.B., T.D.B., and F.L.-V.) and BCS-1638642 (to A.W.B., T.D.B., and F.S.L.), Leakey Foundation Grant (to A.W.B.), an Elizabeth Caroline Crosby Grant from the University of Michigan (to A.W.B.),

Methods

Sample collection

Peruvian Andean study participants self-identified as Quechua and were recruited from Lima (14 m) and Cerro de Pasco (4,338 m), Peru (**Figure 1A**) as part of a larger study investigating high-altitude adaptation (Brutsaert et al., 2019; Childebayeva et al., 2021). It is important to note that both Peruvian Quechua study participant groups were of high-altitude indigenous Andean ancestry. Study participants recruited in Lima, Peru, were born, raised, and resided at low altitudes, but whose parents and both sets of grandparents were of highland Quechua ancestry. Peruvian Quechua recruited in Cerro de Pasco, Peru, were born, raised, and resided at high altitudes. Indigenous Mexican Maya participants spoke Tzeltal, Tzotzil, or Ch'ol and were recruited around or in Palenque, Chiapas, Mexico (13 m). Male and female study participants were unrelated, healthy, non-pregnant/lactating, nonsmokers, and between the ages of 18 and 35. All study participants provided whole blood for DNA extraction. All participants provided written informed consent at enrollment for the study. The study was approved by the institutional review boards at the University of California Los Angeles (IRB # 20-001141), the University of Michigan (IRB #HUM00064387), Syracuse University (IRB#22-364), as well as The Research Ethics Committee of the Universidad Peruana Cayetano Heredia, Lima, Peru and the Centro de Investigación y Docencia Económicas, Mexico City, Mexico.

Short-read genome FASTQ files from the 1000 Genomes Project, HGDP, along with samples from the Simons Genome Diversity Project (SGDP) that belong to populations present in HGDP, were downloaded from the European Nucleotide Archive. We additionally utilized high-coverage short-read sequence data from 38 Tibetan genomes (Lu et al., 2016) as well as high-coverage linked read data from 127 individuals from the Middle East (Almarri et al., 2021).

Gene Copy Number Calling

To calculate the amylase copy number, we employed two different read depth-based gene copy number callers: mrCaNaVaR and GeneToCN (Alkan et al., 2009; Kahveci & Alkan, 2018; Pajuste & Remm, 2023). Both copy number callers have pros and cons. On the one hand, mrCaNaVaR is a promiscuous copy number caller that struggles to differentiate between closely related genes in the same gene families but allows for rapid genome-wide calculations. On the other hand, GeneToCN is quite adept at distinguishing between *AMY1* and *AMY2* genes as well as calculating accurate *AMY1* copy numbers

(Yılmaz et al., 2024), but requires much greater involvement in each different gene calculation. To this end, we have opted to use both in our analyses.

For our mrCaNaVaR analyses, using the provided reference sequences and following the reported methods in (Kahveci & Alkan, 2018) we first aligned each sample FASTQ file to the GRCh38 Human reference genome (CM000663.2) using mrsFAST. We then before calculated the copy number of all genes defined in the reference human_v38.nr.genes.bed file (Hach et al., 2010). We have opted to use the GRCh38 reference genome instead of the newer T2T reference genome because GRCh38 includes three copies of *AMY1* (*AMY1A*, *AMY1B*, and *AMY1C*), which is thought to represent both the ancestral version of the locus and the most common haplotype (Yılmaz et al., 2024), unlike the seven copies present in T2T.

Unlike mrCaNaVaR and most other copy number callers, GeneToCN is an alignment-free method. Instead of calculating read depth across the entire genome, it focuses solely on the read depth of k-mers unique to the *AMY1* genes and conserved flanking regions to determine gene copy number. Following the protocol described in (Pajuste & Remm, 2023), we calculated the amylase copy number for *AMY1* genes, *AMY2B*, and *AMY2A* using 25mers.

Additional Gene Copy Number Testing

Using the *AMY1* copy numbers calculated by GeneToCN (Kaplinski et al., 2015; Pajuste & Remm, 2023) for all individuals, we performed a one-way ANOVA on each designated geographic region using ggplot2, “stat_compare_means” (**Figure 1B**) (Wickham, 2016). We additionally calculated a binomial test of significance using r (**Figure 1E**). This tests whether the proportion of individuals in each population as compared to the total proportion of individuals that have an *AMY1* copy number greater than or equal to 9 is significant.

```
test_results <- GeneToCN_results %>%
  group_by(Pop) %>%
  summarize(
    Total_Count = n(),
    Observed_Count = sum(AMY1 >= 9),
    Proportion = Observed_Count / Total_Count,
    Binom_p_value = binom.test(Observed_Count, Total_Count, p = total_proportion,
    alternative = "greater")$p.value
  ) %>%
```

```
mutate(log_p_value = -log10(Binom_p_value))
```

VST and median copy number analyses

As the reference human_v38.nr.genes.bed file used in our mrCaNaVaR analysis contains other non-coding genetic features such as long non-coding RNAs, we filtered our dataset for all genes that overlapped with the Ensembl annotation database homo sapiens protein-coding genes from GRCh38 (version 97, ID "AH73881"). We then calculated the Vst statistic, $V_{ST} = (V_T - V_S)/V_T$, in R version 4.4.1 using the protocol described in (Redon et al., 2006) to compare the genes between the populations PEL and MXL along with Quechua samples from Cerro de Pasco and Maya from Palenque. We did not include Quechua samples from Lima in this step as the differing sequence coverage (~15X coverage vs. ~30X coverage). Following this step, we calculated the median of every gene in each population, subtracted the difference between the compared populations, and took the absolute value. After filtering for genes that had a median difference or V_{ST} value of 0, we produced a dataset of 1,346 genes for our PEL and MXL comparison and a dataset of 7,163 genes for our Quechua and Maya comparison. We hypothesized that the difference in coverage between the Quechua and Maya genomes and the PEL and MXL genomes (~15X coverage vs. ~30X coverage) might have contributed to a larger number of genes in the Quechua and Maya comparison as fewer reads might have induced greater variance in gene copy number for certain genes that are not in fact copy number variable. We ran an additional 10,000 permutations of both the V_{ST} and median copy number calculations for our Quechua and Maya data, randomizing the population labels with each run.

Admixture analyses

We performed an unsupervised model in ADMIXTURE to determine genome-wide ancestry proportions using 826,848 SNVS after pruning for LD ($r^2 > 0.1$) and 1000 Genomes individuals as ancestral references (Alexander & Lange, 2011). Individuals were clustered with K=4 into the following super-populations of European, Asian, African, and Indigenous American ancestry. Locus-specific ancestry was assigned using RFMix2 and analyzed via chromosomal segments in windows corresponding to the *AMY* region (Maples et al., 2013). The top 100 1KG individuals from each population with the most corresponding ancestry based on ADMIXTURE results were used as ancestral references for RFMix 2: European (IBS), Asian (CHB/CHS), African (YRI), and Indigenous American (MXL/PEL).

Quechua & Maya VCF file creation

Whole-genome data for the Quechua samples from Cerro de Pasco and Maya from Palenque, Chiapas, Mexico were generated using an Illumina NovaSeq with 75 base paired-end reads at 15× depth of coverage at the University of Michigan's DNA Sequencing Core. Whole-genome data for Quechua samples from Lima were sequenced using DIPSEQ-T5 at a higher coverage (~30X). Low-quality reads with a Phred score < 30 were removed following demultiplexing and trimming with Trim Galore! v.0.6.7 (M. Martin, 2011). Alignment was performed using BWA-MEM against hg38 (H. Li & Durbin, 2009), and variant calling was completed with GATK (DePristo et al., 2011). Reads were filtered and removed that had low mapping quality, missing per genotype > 0.10, or missing per individual > 0.10 using both VCFtools and Plink2.0 (Chang et al., 2015; Danecek et al., 2011).

F_{ST} and SNV/AMY1 association

We filtered out Quechua and Maya individuals with greater than 50% non-American ancestry for the amylase region based on the locus-specific ancestry analysis, removing three Quechua individuals from subsequent analyses. Using chromosome 1 data, we filtered for missingness and indels using VCFtools 0.1.16, and retained only biallelic SNVs that had a minor allele frequency greater than 0.05. Following this step, we calculated F_{ST} between Quechua and Maya using VCFtools (Weir & Cockerham, 1984). Information regarding the frequency of the alternative allele in each population was recorded and used to label if the alternative state was present more frequently in Maya or Quechua samples. For each SNV in chromosome 1, we performed a Kruskal-Wallis test in r based on the SNV data and diploid *AMY1* copy number information for all individuals (Kruskal & Wallis, 1952).

```
Kruskal_pvalue <- vapply(snp_data, function(snv) {
  kruskal.test(AMY1_data ~ snv)$p.value
}, numeric(1))
```

We recorded information regarding the median *AMY1* gene copy number for individuals containing homozygous reference, heterozygous, and homozygous alternative states for each SNV and calculated the median gene copy number for each group. Instances where the median *AMY1* copy number of homozygous alternative individuals was greater than either the median of homozygous reference and/or heterozygous individuals, as well as instances where the median *AMY1* copy number of heterozygous individuals was greater than homozygous reference individuals were multiplied by positive one. Instances where the median *AMY1* copy number of homozygous reference individuals was greater than either the

median of homozygous alternative and/or heterozygous individuals, as well as instances where the median *AMY1* copy number of heterozygous individuals was greater than homozygous alternative individuals were multiplied by negative one to show a negative association with *AMY1* copy number and the alternative SNV allele.

We calculated a two-sided Wilcoxon test using "stat_compare_means" with comparisons between each SNV allele state (homozygous reference, heterozygous, and homozygous alternative) for three SNVs of interest (Wilcoxon, 1945).

XP-EHH Analysis

Using the same filtered SNV file from the F_{ST} and Kruskal-Wallis analyses, we calculated XP-EHH between the Quechua and Maya populations using selscan 2.0 (Szpiech, 2024). We created a genetic map file for these SNVs based on human chromosome 1 recombination rates recorded in (Chang et al., 2015). To account for any associations between SNVs on either side of the amylase locus, we extended the haplotype extension to 2 million BP. We additionally applied the Weir and Goudet method in order to account for inbreeding, population structure, or genetic drift that may potentially inflate or distort selection signals (Weir & Goudet, 2017). Finally, we normalized our results across the chromosome.

Phylogeny

We first filtered the phased SNV dataset for missingness, non-American ancestry at the amylase locus, and to only include biallelic SNVs using VCFtools 0.1.16. We then created pseudo haploid FASTA files based on SNVs located between the regions of high recombination surrounding the amylase locus (chr1:103,347,982-103,830,994) for each individual using BCFtools 1.14, SAMtools 1.16.1, and GATK v4.3.0 (Danecek et al., 2021; DePristo et al., 2011). To do this, with each individual, we selected only variants where the individual has at least one copy of the alternative state. We then separated these SNVs based on the haploid genome and used the GRCh38 reference genome to create a haploid FASTA file.

```
gatk SelectVariants \
   -R chr1_hg38.fasta \
   -V${sample}.vcf.gz \
   --select-type-to-include SNP \
   --exclude-non-variants true \
   --O ${sample}_excluded_nonvariants.vcf.gz
```

```
# For haplotype 1
samtools faidx chr1_hg38.fasta ${location} | bcftools consensus -H 1
${sample}_excluded_nonvariants.vcf.gz > ${sample}_region_hap1.fasta
```

After merging all of the FASTA files, we constructed a maximum likelihood tree by testing multiple models, applying robust branch supports (ultrafast bootstrap), and performing an exhaustive NNI tree search using IQ-TREE 2.2.6 (Hoang et al., 2018; Kalyaanamoorthy et al., 2017; Minh et al., 2020). The output .treefile was labeled with the populations and *AMY1* copy numbers in iTOL version 6 (Letunic & Bork, 2024).

```
iqtree -s LWK_Quechua_Maya_no_missing_biallelic.fasta --seqtype DNA -m TESTMERGE -bb
1000 -bnni -allnni -nt AUTO -pre Quechua_Maya_no_missing_biallelic_nni
```

Building pangenome

The pangenome was built using 40 known haplotypes comprising the entire *AMY* locus (Bolognini et al., 2024; Yılmaz et al., 2024). For haplotypes HG02004_pat_H9, HG02293_mat_H5, HG02293_mat_H5, HG02004_mat_H3, HG01978_mat_H3, HG02300_mat_H3, HG02148_pat_H3, HG02293_pat_H1, HG02300_pat_H3, HG01106_mat_H7 (Bolognini et al., 2024) regions comprising the amy locus were extracted by blasting regions upstream and downstream of the locus. Human reference genome hg38 was used to retrieve the upstream and downstream sequences, respectively chr1:103,547,648-103,560,699 and chr1:103,766,322-103,785,201. For haplotypes H1.1, H1.2, H2A0.1, H2A2B2.1, H3r.1, H3r.2, H3r.3, H3r.4, H3r.5, H3r.6, H3r.7, H3B2.1, H3A2.1, H3A3B3.1, H4A0.1, H4A0.2, H4A2.1, H4A2.2, H4A2B2.1, H5.1, H5.2, H5.3, H5.4, H5.5, H5.6, H5.7, H5.8, H5.9, H5.10, and H7.1, data was retrieved from (Yılmaz et al., 2024)) (Supplementary Table 8). All data and scripts used for the pangenome can be found in Github.

We used PGGB (Garrison et al., 2024) to build a local pangenome with the following parameters:

```
pggb -i amylocus_haplotypes.fa  -o amylocus_5k  -n 46 -t 32 -p 99 -s 5k
```

Long-read sequencing and assembly

To understand the Peruvian haplotypes for the amylase locus, we chose six Peruvian samples from the 1000 Genomes Project (Durbin et al., 2010) to sequence (HG01975, HG01972, HG02106, HG0946, HG01922, HG01955). The samples were chosen to represent the diversity of the population, so we chose

samples representing 6, 8, 14, and 18 diploid AMY1 copies, respectively. We generated long-reads on the Oxford Nanopore PromethION sequencing platform using the SQK-LSK114 chemistry and R10.4.1 flow cells for all samples. Because this locus was complicated to assemble correctly, we also generated ultra-long reads for samples HG01975 and HG01946.

We corrected the reads using HERRO (Stanojević et al., 2024) and assembled the genomes using hifiasm (Cheng et al., 2021). We adapted the hifiasm pipeline for the use of corrected ONT reads, as described in (Stanojević et al., 2024).

```
seqkit sliding -s 30000 -W 30000 -g ${cor_reads} > chopped.fasta
seqkit seq -m 10000 chopped.fasta > processed.fasta

#As ultra-long reads, we are using only reads longer than 50kb
seqkit seq -m 50000 $raw > raw_reads_${sample}_50kb.fastq

./hifiasm  -t  32  -o  ${sample}_diploid_50kb  --ul  raw_reads_${sample}_50kb.fastq
--ul-cut 50000 --dual-scaf -1 parent1.yak -2 parent2.yak  processed.fasta
```

To verify the assembly recovery of the amylase locus, we aligned the assemblies to the publicly available genome hg38 **(Figure S11**). Based on the dot plots and manual inspection, we concluded that the assembled haplotypes were not correctly assembled. We then used the contigs that span the region to recover the phased reads using the following approach: 1) we mapped all the contigs to the hg38 reference genome; 2) we retrieved the contigs that mapped to the *AMY* locus (coordinates); 3) we retrieved the reads that mapped to the specific parts of the contigs that mapped to the region of interest.
All scripts used for this are on **Github**.

Mapping to haplotypes and filtering

Because the assembled contigs were incorrect, we mapped the reads to all known distinct haplotypes (n=30) plus previously published Peruvian haplotypes (Bolognini et al., 2024) and used a read depth measure to decide the best fit for each sample. The haplotypes used are on the **Github**, details in **Supplementary Table S9**. We competitively mapped the reads into each haplotype and filtered the mapping for length > 2.5kb. This filtering was done because the region is rich in transposable elements that might skew the analysis.

We hypothesized that the best-fit haplotype will have a lower variance in mapping. Therefore, we calculated the mean and standard deviation for the read depth on the entire haplotype. The best fit for all

haplotypes was the one that showed the lowest standard deviation and a mean read depth of ≥ 1. We also checked that the best-fit haplotype for each individual coincided with the estimated copy number based on ddPCR results (**Table S1 and S11**). Based on this method, we could assign the best fit for 4 individuals (8 phased haplotypes).

However, two individuals (HG02106 and HG01975) were assigned haplotypes discordant with the estimated copy number generated by ddPCR. We manually inspected the best-fit haplotypes, and the mapping was not homogeneous for any of the top haplotypes (H9, H7, H5). The best match for this haplotype was H9, and, therefore, we hypothesized that the best match would be a haplotype similar to H9 but harboring 7 AMY1 copies. We generated a mock H7 haplotype by deleting one duplication unit (AMY1A and AMY1B, see Yilmaz, 2024) from the H9 haplotype (**Supplementary Methods**).

References:

Albers, P. K., & McVean, G. (2020). Dating genomic variants and shared ancestry in population-scale sequencing data. *PLOS Biology*, *18*(1), e3000586. https://doi.org/10.1371/journal.pbio.3000586

Alexander, D. H., & Lange, K. (2011). Enhancements to the ADMIXTURE algorithm for individual ancestry estimation. *BMC Bioinformatics*, *12*(1), 246. https://doi.org/10.1186/1471-2105-12-246

Alkan, C., Kidd, J. M., Marques-Bonet, T., Aksay, G., Antonacci, F., Hormozdiari, F., Kitzman, J. O., Baker, C., Malig, M., Mutlu, O., Sahinalp, S. C., Gibbs, R. A., & Eichler, E. E. (2009). Personalized copy number and segmental duplication maps using next-generation sequencing. *Nature Genetics*, *41*(10), 1061–1067. https://doi.org/10.1038/ng.437

Almarri, M. A., Haber, M., Lootah, R. A., Hallast, P., Turki, S. A., Martin, H. C., Xue, Y., & Tyler-Smith, C. (2021). The genomic history of the Middle East. *Cell*, *184*(18), 4612-4625.e14. https://doi.org/10.1016/j.cell.2021.07.013

Arendt, M., Cairns, K. M., Ballard, J. W. O., Savolainen, P., & Axelsson, E. (2016). Diet adaptation in dog reflects spread of prehistoric agriculture. *Heredity*, *117*(5), 301–306. https://doi.org/10.1038/hdy.2016.48

Atkinson, F. S., Brand-Miller, J. C., Foster-Powell, K., Buyken, A. E., & Goletzke, J. (2021). International tables of glycemic index and glycemic load values 2021: A systematic review. *The American Journal of Clinical Nutrition*, *114*(5), 1625–1632. https://doi.org/10.1093/ajcn/nqab233

Atkinson, F. S., Hancock, D., Petocz, P., & Brand-Miller, J. C. (2018). The physiologic and phenotypic significance of variation in human amylase gene copy number. *The American Journal of Clinical Nutrition*, *108*(4), 737–748. https://doi.org/10.1093/ajcn/nqy164

Axelsson, E., Ratnakumar, A., Arendt, M.-L., Maqbool, K., Webster, M. T., Perloski, M., Liberg, O., Arnemo, J. M., Hedhammar, Å., & Lindblad-Toh, K. (2013). The genomic signature of dog domestication reveals adaptation to a starch-rich diet. *Nature*, *495*(7441), 360–364. https://doi.org/10.1038/nature11837

Berti, P. R., Jones, A. D., Cruz, Y., Larrea, S., Borja, R., & Sherwood, S. (2010). Assessment and characterization of the diet of an isolated population in the Bolivian Andes. *American Journal of Human Biology*, *22*(6), 741–749. https://doi.org/10.1002/ajhb.21075

Bolognini, D., Halgren, A., Lou, R. N., Raveane, A., Rocha, J. L., Guarracino, A., Soranzo, N., Chin, C.-S., Garrison, E., & Sudmant, P. H. (2024). Recurrent evolution and selection shape structural diversity at the amylase locus. *Nature*, *634*(8034), 617–625. https://doi.org/10.1038/s41586-024-07911-1

Brutsaert, T. D., Kiyamu, M., Elias Revollendo, G., Isherwood, J. L., Lee, F. S., Rivera-Ch, M., Leon-Velarde, F., Ghosh, S., & Bigham, A. W. (2019). Association of EGLN1 gene with high aerobic capacity of Peruvian Quechua at high altitude. *Proceedings of the National Academy of Sciences*, *116*(48), 24006–24011. https://doi.org/10.1073/pnas.1906171116

Chang, C. C., Chow, C. C., Tellier, L. C., Vattikuti, S., Purcell, S. M., & Lee, J. J. (2015). Second-generation PLINK: Rising to the challenge of larger and richer datasets. *GigaScience*, *4*(1), s13742-015-0047–0048. https://doi.org/10.1186/s13742-015-0047-8

Chen, J. C., Aldenderfer, M. S., Eerkens, J. W., Langlie, B. S., Llave, C. V., Watson, J. T., & Haas, R. (2024). Stable isotope chemistry reveals plant-dominant diet among early foragers on the Andean Altiplano, 9.0–6.5 cal. Ka. *PLOS ONE*, *19*(1), e0296420. https://doi.org/10.1371/journal.pone.0296420

Cheng, H., Concepcion, G. T., Feng, X., Zhang, H., & Li, H. (2021). Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nature Methods*, *18*(2), 170–175. https://doi.org/10.1038/s41592-020-01056-5

Childebayeva, A., Goodrich, J. M., Leon-Velarde, F., Rivera-Chira, M., Kiyamu, M., Brutsaert, T. D., Dolinoy, D. C., & Bigham, A. W. (2021). Genome-Wide Epigenetic Signatures of Adaptive Developmental Plasticity in the Andes. *Genome Biology and Evolution*, *13*(2), evaa239. https://doi.org/10.1093/gbe/evaa239

Crawford, N. G., Kelly, D. E., Hansen, M. E. B., Beltrame, M. H., Fan, S., Bowman, S. L., Jewett, E.,

Ranciaro, A., Thompson, S., Lo, Y., Pfeifer, S. P., Jensen, J. D., Campbell, M. C., Beggs, W., Hormozdiari, F., Mpoloka, S. W., Mokone, G. G., Nyambo, T., Meskel, D. W., … Tishkoff, S. A. (2017). Loci associated with skin pigmentation identified in African populations. *Science*, *358*(6365), eaan8433. https://doi.org/10.1126/science.aan8433

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., Handsaker, R. E., Lunter, G., Marth, G. T., Sherry, S. T., McVean, G., & Durbin, R. (2011). The variant call format and VCFtools. *Bioinformatics*, *27*(15), 2156–2158. https://doi.org/10.1093/bioinformatics/btr330

Danecek, P., Bonfield, J. K., Liddle, J., Marshall, J., Ohan, V., Pollard, M. O., Whitwham, A., Keane, T., McCarthy, S. A., Davies, R. M., & Li, H. (2021). Twelve years of SAMtools and BCFtools. *GigaScience*, *10*(2), giab008. https://doi.org/10.1093/gigascience/giab008

DePristo, M. A., Banks, E., Poplin, R., Garimella, K. V., Maguire, J. R., Hartl, C., Philippakis, A. A., del Angel, G., Rivas, M. A., Hanna, M., McKenna, A., Fennell, T. J., Kernytsky, A. M., Sivachenko, A. Y., Cibulskis, K., Gabriel, S. B., Altshuler, D., & Daly, M. J. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genetics*, *43*(5), 491–498. https://doi.org/10.1038/ng.806

Durbin, R. M., Altshuler, D., Durbin, R. M., Abecasis, G. R., Bentley, D. R., Chakravarti, A., Clark, A. G., Collins, F. S., De La Vega, F. M., Donnelly, P., Egholm, M., Flicek, P., Gabriel, S. B., Gibbs, R. A., Knoppers, B. M., Lander, E. S., Lehrach, H., Mardis, E. R., McVean, G. A., … The Translational Genomics Research Institute. (2010). A map of human genome variation from population-scale sequencing. *Nature*, *467*(7319), 1061–1073. https://doi.org/10.1038/nature09534

Garrison, E., Guarracino, A., Heumos, S., Villani, F., Bao, Z., Tattini, L., Hagmann, J., Vorbrugg, S., Marco-Sola, S., Kubica, C., Ashbrook, D. G., Thorell, K., Rusholme-Pilcher, R. L., Liti, G., Rudbeck, E., Golicz, A. A., Nahnsen, S., Yang, Z., Mwaniki, M. N., … Prins, P. (2024). Building pangenome graphs. *Nature Methods*, *21*(11), 2008–2012. https://doi.org/10.1038/s41592-024-02430-3

Guo, F., Zhang, Q., Jiang, H., He, Y., Li, M., Ran, J., Lin, J., Tian, L., & Ma, L. (2021). Dietary potato

intake and risks of type 2 diabetes and gestational diabetes mellitus. *Clinical Nutrition*, *40*(6), 3754–3764. https://doi.org/10.1016/j.clnu.2021.04.039

Haas Jr, W. R. H., & Llave, C. V. (2015). Hunter-gatherers on the eve of agriculture: Investigations at Soro Mik'aya Patjxa, Lake Titicaca Basin, Peru, 8000–6700 BP. *Antiquity*, *89*(348), 1297–1312. https://doi.org/10.15184/aqy.2015.100

Hoang, D. T., Chernomor, O., von Haeseler, A., Minh, B. Q., & Vinh, L. S. (2018). UFBoot2: Improving the Ultrafast Bootstrap Approximation. *Molecular Biology and Evolution*, *35*(2), 518–522. https://doi.org/10.1093/molbev/msx281

Jorgensen, K., Garcia, O. A., Kiyamu, M., Brutsaert, T. D., & Bigham, A. W. (2023). Genetic adaptations to potato starch digestion in the Peruvian Andes. *American Journal of Biological Anthropology*, *180*(1), 162–172. https://doi.org/10.1002/ajpa.24656

Kahveci, F., & Alkan, C. (2018). Whole-Genome Shotgun Sequence CNV Detection Using Read Depth. In D. M. Bickhart (Ed.), *Copy Number Variants: Methods and Protocols* (pp. 61–72). Springer. https://doi.org/10.1007/978-1-4939-8666-8_4

Kalyaanamoorthy, S., Minh, B. Q., Wong, T. K. F., von Haeseler, A., & Jermiin, L. S. (2017). ModelFinder: Fast model selection for accurate phylogenetic estimates. *Nature Methods*, *14*(6), 587–589. https://doi.org/10.1038/nmeth.4285

Kaplinski, L., Lepamets, M., & Remm, M. (2015). GenomeTester4: A toolkit for performing basic set operations - union, intersection and complement on k-mer lists. *GigaScience*, *4*(1), 58. https://doi.org/10.1186/s13742-015-0097-y

Kruskal, W. H., & Wallis, W. A. (1952). Use of Ranks in One-Criterion Variance Analysis. *Journal of the American Statistical Association*, *47*(260), 583–621. https://doi.org/10.1080/01621459.1952.10483441

Letunic, I., & Bork, P. (2024). Interactive Tree of Life (iTOL) v6: Recent updates to the phylogenetic tree display and annotation tool. *Nucleic Acids Research*, *52*(W1), W78–W82. https://doi.org/10.1093/nar/gkae268

Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics*, *25*(14), 1754–1760. https://doi.org/10.1093/bioinformatics/btp324

Li, Y., Colleoni, C., Zhang, J., Liang, Q., Hu, Y., Ruess, H., Simon, R., Liu, Y., Liu, H., Yu, G., Schmitt, E., Ponitzki, C., Liu, G., Huang, H., Zhan, F., Chen, L., Huang, Y., Spooner, D., & Huang, B. (2018). Genomic Analyses Yield Markers for Identifying Agronomically Important Genes in Potato. *Molecular Plant*, *11*(3), 473–484. https://doi.org/10.1016/j.molp.2018.01.009

Lindo, J., Haas, R., Hofman, C., Apata, M., Moraga, M., Verdugo, R. A., Watson, J. T., Viviano Llave, C., Witonsky, D., Beall, C., Warinner, C., Novembre, J., Aldenderfer, M., & Di Rienzo, A. (2018). The genetic prehistory of the Andean highlands 7000 years BP though European contact. *Science Advances*, *4*(11), eaau4921. https://doi.org/10.1126/sciadv.aau4921

Lu, D., Lou, H., Yuan, K., Wang, X., Wang, Y., Zhang, C., Lu, Y., Yang, X., Deng, L., Zhou, Y., Feng, Q., Hu, Y., Ding, Q., Yang, Y., Li, S., Jin, L., Guan, Y., Su, B., Kang, L., & Xu, S. (2016). Ancestral Origins and Genetic History of Tibetan Highlanders. *The American Journal of Human Genetics*, *99*(3), 580–594. https://doi.org/10.1016/j.ajhg.2016.07.002

Maples, B. K., Gravel, S., Kenny, E. E., & Bustamante, C. D. (2013). RFMix: A Discriminative Modeling Approach for Rapid and Robust Local-Ancestry Inference. *The American Journal of Human Genetics*, *93*(2), 278–288. https://doi.org/10.1016/j.ajhg.2013.06.020

Martin, A. R., Lin, M., Granka, J. M., Myrick, J. W., Liu, X., Sockell, A., Atkinson, E. G., Werely, C. J., Möller, M., Sandhu, M. S., Kingsley, D. M., Hoal, E. G., Liu, X., Daly, M. J., Feldman, M. W., Gignoux, C. R., Bustamante, C. D., & Henn, B. M. (2017). An Unexpectedly Complex Architecture for Skin Pigmentation in Africans. *Cell*, *171*(6), 1340-1353.e14. https://doi.org/10.1016/j.cell.2017.11.015

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.Journal*, *17*(1), Article 1. https://doi.org/10.14806/ej.17.1.200

Mathieson, S., & Mathieson, I. (2018). FADS1 and the Timing of Human Adaptation to Agriculture. *Molecular Biology and Evolution*, *35*(12), 2957–2970. https://doi.org/10.1093/molbev/msy180

Mejía-Benítez, M. A., Bonnefond, A., Yengo, L., Huyvaert, M., Dechaume, A., Peralta-Romero, J., Klünder-Klünder, M., García Mena, J., El-Sayed Moustafa, J. S., Falchi, M., Cruz, M., & Froguel, P. (2015). Beneficial effect of a high number of copies of salivary amylase AMY1 gene on obesity risk in Mexican children. *Diabetologia*, *58*(2), 290–294. https://doi.org/10.1007/s00125-014-3441-3

Minh, B. Q., Schmidt, H. A., Chernomor, O., Schrempf, D., Woodhams, M. D., von Haeseler, A., & Lanfear, R. (2020). IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. *Molecular Biology and Evolution*, *37*(5), 1530–1534. https://doi.org/10.1093/molbev/msaa015

Pajic, P., Pavlidis, P., Dean, K., Neznanova, L., Romano, R.-A., Garneau, D., Daugherity, E., Globig, A., Ruhl, S., & Gokcumen, O. (2019). Independent amylase gene copy number bursts correlate with dietary preferences in mammals. *eLife*, *8*, e44628. https://doi.org/10.7554/eLife.44628

Pajuste, F.-D., & Remm, M. (2023). GeneToCN: An alignment-free method for gene copy number estimation directly from next-generation sequencing reads. *Scientific Reports*, *13*(1), 17765. https://doi.org/10.1038/s41598-023-44636-z

Perry, G. H., Dominy, N. J., Claw, K. G., Lee, A. S., Fiegler, H., Redon, R., Werner, J., Villanea, F. A., Mountain, J. L., Misra, R., Carter, N. P., Lee, C., & Stone, A. C. (2007). Diet and the evolution of human amylase gene copy number variation. *Nature Genetics*, *39*(10), Article 10. https://doi.org/10.1038/ng2123

Poole, A. C., Goodrich, J. K., Youngblut, N. D., Luque, G. G., Ruaud, A., Sutter, J. L., Waters, J. L., Shi, Q., El-Hadidi, M., Johnson, L. M., Bar, H. Y., Huson, D. H., Booth, J. G., & Ley, R. E. (2019). Human Salivary Amylase Gene Copy Number Impacts Oral and Gut Microbiomes. *Cell Host & Microbe*, *25*(4), 553-564.e7. https://doi.org/10.1016/j.chom.2019.03.001

Redon, R., Ishikawa, S., Fitch, K. R., Feuk, L., Perry, G. H., Andrews, T. D., Fiegler, H., Shapero, M. H., Carson, A. R., Chen, W., Cho, E. K., Dallaire, S., Freeman, J. L., González, J. R., Gratacòs, M., Huang, J., Kalaitzopoulos, D., Komura, D., MacDonald, J. R., … Hurles, M. E. (2006). Global

variation in copy number in the human genome. *Nature*, *444*(7118), Article 7118. https://doi.org/10.1038/nature05329

Roberts, P. J. P., & Whelan, W. J. (1960). The mechanism of carbohydrase action. 5. Action of human salivary α-amylase on amylopectin and glycogen*. *Biochemical Journal*, *76*(2), 246–253. https://doi.org/10.1042/bj0760246

Rose, D., Burgos, G., Bonierbale, M., & Thiele, G. (2009). Understanding the role of potatoes in the Peruvian diet: An approach that combines food composition with household expenditure data. *Journal of Food Composition and Analysis*, *22*(6), 525–532. https://doi.org/10.1016/j.jfca.2008.10.002

Sabeti, P. C., Varilly, P., Fry, B., Lohmueller, J., Hostetter, E., Cotsapas, C., Xie, X., Byrne, E. H., McCarroll, S. A., Gaudet, R., Schaffner, S. F., & Lander, E. S. (2007). Genome-wide detection and characterization of positive selection in human populations. *Nature*, *449*(7164), 913–918. https://doi.org/10.1038/nature06250

Schrider, D. R., & Kern, A. D. (2017). Soft Sweeps Are the Dominant Mode of Adaptation in the Human Genome. *Molecular Biology and Evolution*, *34*(8), 1863–1877. https://doi.org/10.1093/molbev/msx154

Spooner, D. M., McLean, K., Ramsay, G., Waugh, R., & Bryan, G. J. (2005). A single domestication for potato based on multilocus amplified fragment length polymorphism genotyping. *Proceedings of the National Academy of Sciences*, *102*(41), 14694–14699. https://doi.org/10.1073/pnas.0507400102

Stanojević, D., Lin, D., Nurk, S., Sessions, P. F. de, & Šikić, M. (2024). *Telomere-to-Telomere Phased Genome Assembly Using HERRO-Corrected Simplex Nanopore Reads* (p. 2024.05.18.594796). bioRxiv. https://doi.org/10.1101/2024.05.18.594796

Szpiech, Z. A. (2024). selscan 2.0: Scanning for sweeps in unphased data. *Bioinformatics*, *40*(1), btae006. https://doi.org/10.1093/bioinformatics/btae006

Usher, C. L., Handsaker, R. E., Esko, T., Tuke, M. A., Weedon, M. N., Hastie, A. R., Cao, H., Moon, J.

E., Kashin, S., Fuchsberger, C., Metspalu, A., Pato, C. N., Pato, M. T., McCarthy, M. I., Boehnke, M., Altshuler, D. M., Frayling, T. M., Hirschhorn, J. N., & McCarroll, S. A. (2015). Structural forms of the human amylase locus and their relationships to SNPs, haplotypes and obesity. *Nature Genetics*, *47*(8), Article 8. https://doi.org/10.1038/ng.3340

Venkatapoorna, C. M. K., Ayine, P., Parra, E. P., Koenigs, T., Phillips, M., Babu, J. R., Sandey, M., & Geetha, T. (2019). Association of Salivary Amylase (AMY1) Gene Copy Number with Obesity in Alabama Elementary School Children. *Nutrients*, *11*(6), 1379. https://doi.org/10.3390/nu11061379

Weir, B. S., & Cockerham, C. C. (1984). ESTIMATING F-STATISTICS FOR THE ANALYSIS OF POPULATION STRUCTURE. *Evolution; International Journal of Organic Evolution*, *38*(6), 1358–1370. https://doi.org/10.1111/j.1558-5646.1984.tb05657.x

Weir, B. S., & Goudet, J. (2017). A Unified Characterization of Population Structure and Relatedness. *Genetics*, *206*(4), 2085–2103. https://doi.org/10.1534/genetics.116.198424

Wickham, H. (2016). *Ggplot2*. Springer International Publishing. https://doi.org/10.1007/978-3-319-24277-4

Wiedemann, S. J., Rachid, L., Illigens, B., Böni-Schnetzler, M., & Donath, M. Y. (2020). Evidence for cephalic phase insulin release in humans: A systematic review and meta-analysis. *Appetite*, *155*, 104792. https://doi.org/10.1016/j.appet.2020.104792

Wilcoxon, F. (1945). Individual Comparisons by Ranking Methods. *Biometrics Bulletin*, *1*(6), 80–83. JSTOR. https://doi.org/10.2307/3001968

Willerslev, E., & Meltzer, D. J. (2021). Peopling of the Americas as inferred from ancient genomics. *Nature*, *594*(7863), 356–364. https://doi.org/10.1038/s41586-021-03499-y

Yılmaz, F., Karageorgiou, C., Kim, K., Pajic, P., Scheer, K., Human Genome Structural Variation Consortium, Beck, C. R., Torregrossa, A.-M., Lee, C., & Gokcumen, O. (2024). Reconstruction of the human amylase locus reveals ancient duplications seeding modern-day variation. *Science*, *0*(0), eadn0609. https://doi.org/10.1126/science.adn0609