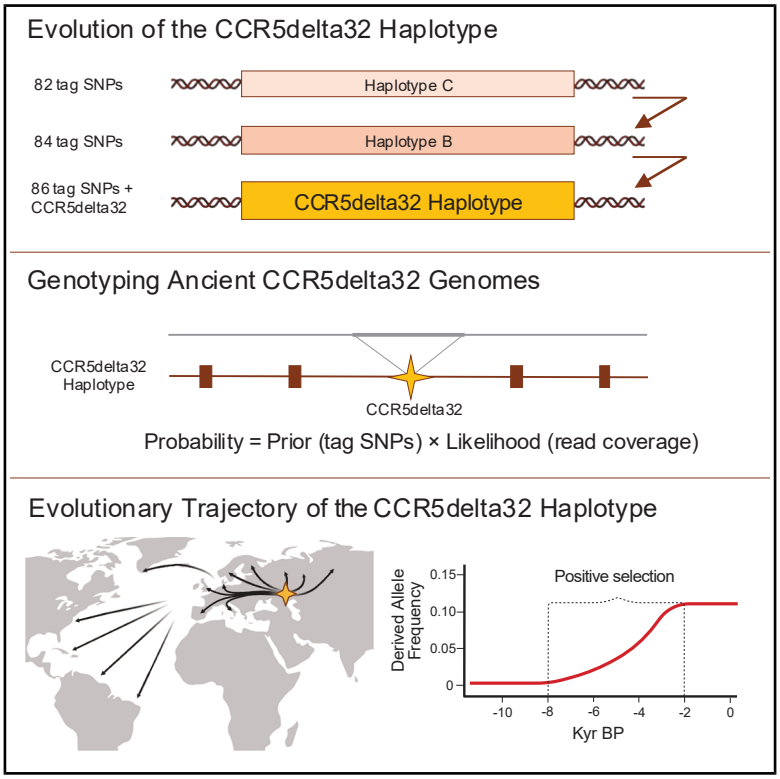
Cell

Article

Tracing the evolutionary history of the CCR5delta32 deletion via ancient and modern genomes

Graphical abstract



Highlights

- The CCR5delta32 deletion arose on a pre-existing haplotype comprising 84 variants
- The CCR5delta32 haplotype originated in the Western Steppe at least 6,700 years ago
- Positive selection of CCR5delta32 occurred in the Late Neolithic and Bronze Age
- The haplotype places the CCR5delta32 allele in a new medical context

Authors

Kirstine Ravn, Leonardo Cobuccio, Rasa Audange Muktupavela, ..., Morten E. Allentoft, Evan K. Irving-Pease, Simon Rasmussen

Correspondence

evan.irving-pease@sund.ku.dk (E.K.I.-P.), srasmuss@sund.ku.dk (S.R.)

In brief

The CCR5delta32 deletion arose on a pre-existing haplotype of 84 variants over 6,700 years ago in the Western Steppe. Positive selection drove it to high frequency, resulting in a variant of present-day medical importance.

Ravn et al., 2025, Cell *188*, 3679–3695
July 10, 2025
https://doi.org/10.1016/j.cell.2025.04.015

CellPress

Cell

CellPress
OPEN ACCESS

Article

Tracing the evolutionary history of the CCR5delta32 deletion via ancient and modern genomes

Kirstine Ravn,[1,2,10] Leonardo Cobuccio,[1,2,10] Rasa Audange Muktupavela,[3,10] Jonas Meisner,[1,2,4] Lasse Schnell Danielsen,[2] Michael Eriksen Benros,[4] Thorfinn Sand Korneliussen,[5,6] Martin Sikora,[5,6] Eske Willerslev,[5,6,7,8] Morten E. Allentoft,[5,9] Evan K. Irving-Pease,[3,5,11,*] and Simon Rasmussen[1,2,11,12,*]

[1]Novo Nordisk Foundation Center for Protein Research, Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen, Denmark
[2]Novo Nordisk Foundation Center for Basic Metabolic Research, Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen, Denmark
[3]Section for Molecular Ecology and Evolution, Globe Institute, University of Copenhagen, Copenhagen, Denmark
[4]Copenhagen Research Centre for Mental Health, Mental Health Centre Copenhagen, Copenhagen University Hospital, Copenhagen, Denmark
[5]Lundbeck Foundation GeoGenetics Centre, Globe Institute, University of Copenhagen, Copenhagen, Denmark
[6]Centre for Ancient Environmental Genomics, University of Copenhagen, Copenhagen, Denmark
[7]Department of Genetics, University of Cambridge, Cambridge, UK
[8]MARUM, University of Bremen, Bremen, Germany
[9]Trace and Environmental DNA (TrEnD) Laboratory, School of Molecular and Life Sciences, Curtin University, Perth, WA, Australia
[10]These authors contributed equally
[11]Senior author
[12]Lead contact
*Correspondence: evan.irving-pease@sund.ku.dk (E.K.I.-P.), srasmuss@sund.ku.dk (S.R.)
https://doi.org/10.1016/j.cell.2025.04.015

SUMMARY

The chemokine receptor variant CCR5delta32 is linked to HIV-1 resistance and other conditions. Its evolutionary history and allele frequency (10%–16%) in European populations have been extensively debated. We provide a detailed perspective of the evolutionary history of the deletion through time and space. We discovered that the CCR5delta32 allele arose on a pre-existing haplotype consisting of 84 variants. Using this information, we developed a haplotype-aware probabilistic model to screen 934 low-coverage ancient genomes and traced the origin of the CCR5delta32 deletion to at least 6,700 years before the present (BP) in the Western Eurasian Steppe region. Furthermore, we present strong evidence for positive selection acting upon the CCR5delta32 haplotype between 8,000 and 2,000 years BP in Western Eurasia and show that the presence of the haplotype in Latin America can be explained by post-Columbian genetic exchanges. Finally, we point to complex CCR5delta32 genotype-haplotype-phenotype relationships, which demand consideration when targeting the CCR5 receptor for therapeutic strategies.

INTRODUCTION

Humans have been exposed to pathogens over the course of our evolutionary history, and adaptations to them have left numerous signatures in our genomes.[1–3] In recent years, evidence for selection has been found in genes involved in the development of tolerance against intracellular pathogens and in the inflammatory response against extracellular microbes.[4–6] These include, for example, the Toll-like receptor (TLR)6-TLR1-TLR10 cluster of Toll-like receptors, which are crucial components of innate immunity against pathogens, and were likely under positive selection in anatomically modern humans after introgression from archaic hominin groups.[7,8] More recently, Domínguez-Andrés et al.[9] showed that alleles associated with cytokine profiles reflecting immune tolerance were under selection during the transition to farming in the Neolithic period, as sedentarism and population density increased, enabling the development of pathogen reservoirs in newly domesticated animals.

Perhaps one of the most intensively debated immune-associated loci previously posited to have been under selection in humans is a 32-bp deletion (CCR5delta32, rs333), which introduces a premature stop codon in the C-C chemokine receptor 5 gene (*CCR5*).[10–17] CCR5 is a member of the G protein-coupled receptor family of proteins, and upon activation by the C-C chemokines, CCL3, CCL4, and CCL5, it plays a critical function in regulating the inflammatory response by facilitating

communication between immune cells and the environment.[18–20] Thus, CCR5 can act as a regulator of the host's immune response. In 1996, CCR5 was identified as a necessary co-receptor for the macrophage-tropic HIV strains,[21,22] and it was subsequently reported that CCR5delta32 could provide HIV-1 infection resistance to individuals carrying this allele in homozygous form.[23] CCR5 is now an important target in preventing and treating HIV infection, using various therapeutic strategies.[24,25] As an example, a female patient with HIV-1 was recently potentially cured for both HIV-1 and acute myeloid leukemia through a CCR5delta32/delta32 haplo-cord transplant,[26] a method that has successfully cured two similar cases using unrelated donor stem cells with the same genetic modification.[27,28]

Besides the significant effect on HIV infection, the CCR5delta32 allele has also been associated with other pathological conditions including infection by other viral organisms (such as SARS-CoV-2 that causes COVID-19), immune-related diseases, neurological disorders, and various types of cancer.[20,29–41] Together, these studies indicate that the CCR5delta32 allele is pleiotropic and can act as a modulator of a given phenotype expression, with both advantages and disadvantages, depending on the medical context. In this perspective, serious concerns have been raised by the scientific community about possible clinical side effects on the so-called "CCR5delta32" CRISPR babies, whose genomes have been edited to confer lifetime HIV immunity.[42–45]

The evolutionary history of the CCR5delta32 deletion has been widely debated, with conflicting research results.[10,12,13,15–17,46–54] Today, the CCR5delta32 allele frequency (AF) is between 0.10 and 0.16 in Northern European populations and less than 0.08 in Southern and South East Europe.[12,55] Outside of Europe, the deletion is found only in populations with European ancestry.[56–59] Past studies have estimated the age of the CCR5delta32 allele with divergent results ranging from ~700 and ~3,400 to >5,000 years ago.[10–12,15,17] Positive selection, negative selection, balancing selection, and genetic drift have each been proposed as an explanation for the distribution of present-day allele frequencies.[10,12,13,15–17,46–54,60,61]

The few studies conducted on the CCR5delta32 deletion in ancient individuals have been constrained by a limited geographic scope and small sample sizes, leading to the possibility of biasing the results by familial relations. So far, the oldest CCRdelta32 alleles have been detected in a 4,900-year-old individual belonging to the Yamnaya culture[62] and in several Swedish individuals dating to the Neolithic period (5,250–1,690 BCE).[48] However, the latter study raised concerns about allelic dropout during the assaying process, which could lead to genotype misclassification. Two studies conducted on ancient genomes from individuals in central and northern Germany revealed no significant change in the frequency of the CCR5delta32 variant over the past millennium, including during the Black Death pandemic.[46,49] In contrast, a study conducted in Poland reported a nearly doubled frequency of the CCR5delta32 variant from the late medieval period to the present day.[63] Finally, a recent preprint by Le et al.[61] used a single variant, suggested to tag the CCR5delta32 allele, to investigate 1,291 ancient samples with an in-solution capture of 1.2 million markers, finding no evidence of positive selection.

The emergence of paleogenomics has shed light on our understanding of human population history, but evidence from large ancient genomic datasets has been missing in the debate on the CCR5delta32 allele. Due to the degraded nature of ancient DNA, ancient genomic datasets tend to be characterized by short-read lengths and postmortem DNA damage,[64] impairing the ability to identify insertions or deletions (indels) like the CCR5delta32 deletion. Moreover, mapping efforts in the CCR5delta32 region are particularly challenging because the breakpoint's flanking regions contain repeated sequences.

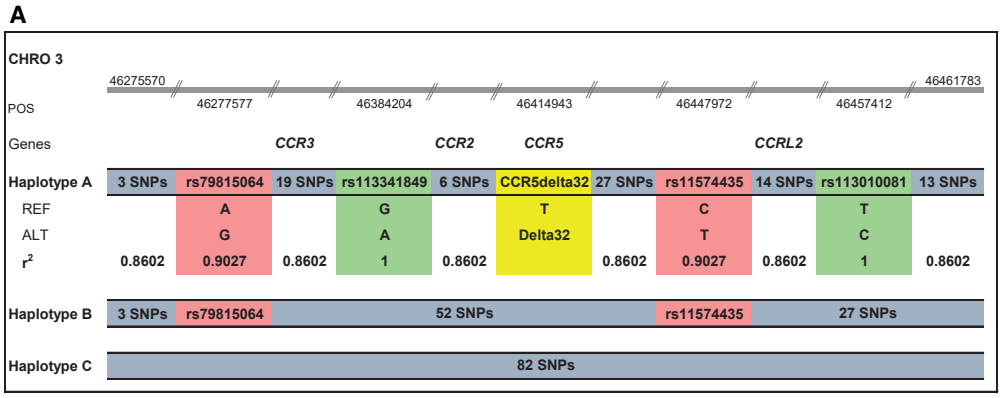
In this study, we trace the evolutionary trajectory of the CCR5delta32 deletion by analyzing 934 ancient and 2,504 present-day genomes. We discovered that the CCR5delta32 allele emerged on a pre-existing haplotype containing 84 variants, and this led us to develop a probabilistic model to detect the allele in low-coverage genomes. Applying our model to ancient genomes, we found evidence that the CCR5delta32 deletion likely originated more than 6,700 years ago in the Western Eurasian Steppe. Our results show that the allele underwent selection between 8,000 and 2,000 years before the present (BP) and later spread to Latin America through historical events such as the Columbian Exchange. This study offers the first comprehensive picture of the CCR5delta32 allele's evolutionary history.

RESULTS

Identification of three CCR5 haplotypes in Europe

By re-analyzing the European individuals from the 1000 Genomes Project Phase 3 (1KGP3)[65] data, we discovered that the CCR5delta32 allele was located on a haplotype with up to 107 variants in high linkage disequilibrium (LD): $r^2 > 0.8$ (for details see Table S1). The longest haplotype was identified in the Finnish in Finland (FIN) population, spanning 107 variants including 76 variants with $r^2 = 1$. We note that 86 of the 107 variants were also found to be in high LD in the CEU panel (Utah residents with ancestry from northern and western Europe), including two variants with $r^2 = 1$ (rs113341849 and rs113010081) (Figure 1A). In contrast, in the Toscani in Italy (TSI, $r^2 > 0.8$, 3 SNPs), Iberian in Spain (IBS, $r^2 > 0.8$, 3 SNPs), and British in England and Scotland (GBR, $r^2 > 0.8$, 2 SNPs) panels, we could only identify a few variants in high LD (Table S1). However, these variants were among those with highest LD ($r^2 > 0.9$) to the deletion in the CEU population. We termed the CEU CCR5delta32 haplotype "Haplotype A" (Figure 1A; Table S1) and identified it in all the 112 CCR5delta32 carriers of the 505 1KGP3 European individuals (AF = 0.111), including three carriers with homologous recombinants of the haplotype (Table S2).

Given the strong correlation between the deletion and the variants of Haplotype A in the CEU population, it was surprising that the LD was weaker in Southern and Western Europe (Figure S1A). Upon further analysis, we found that this weaker LD was due to the presence of additional haplotypes and multiple recombination events across the EUR populations beyond the CEU individuals. Specifically, we discovered an additional haplotype, "Haplotype B," which shared 84 of 86 tag variants with Haplotype A but differed at two specific variants and lacked the CCR5delta32 deletion and the two SNPs in complete LD with



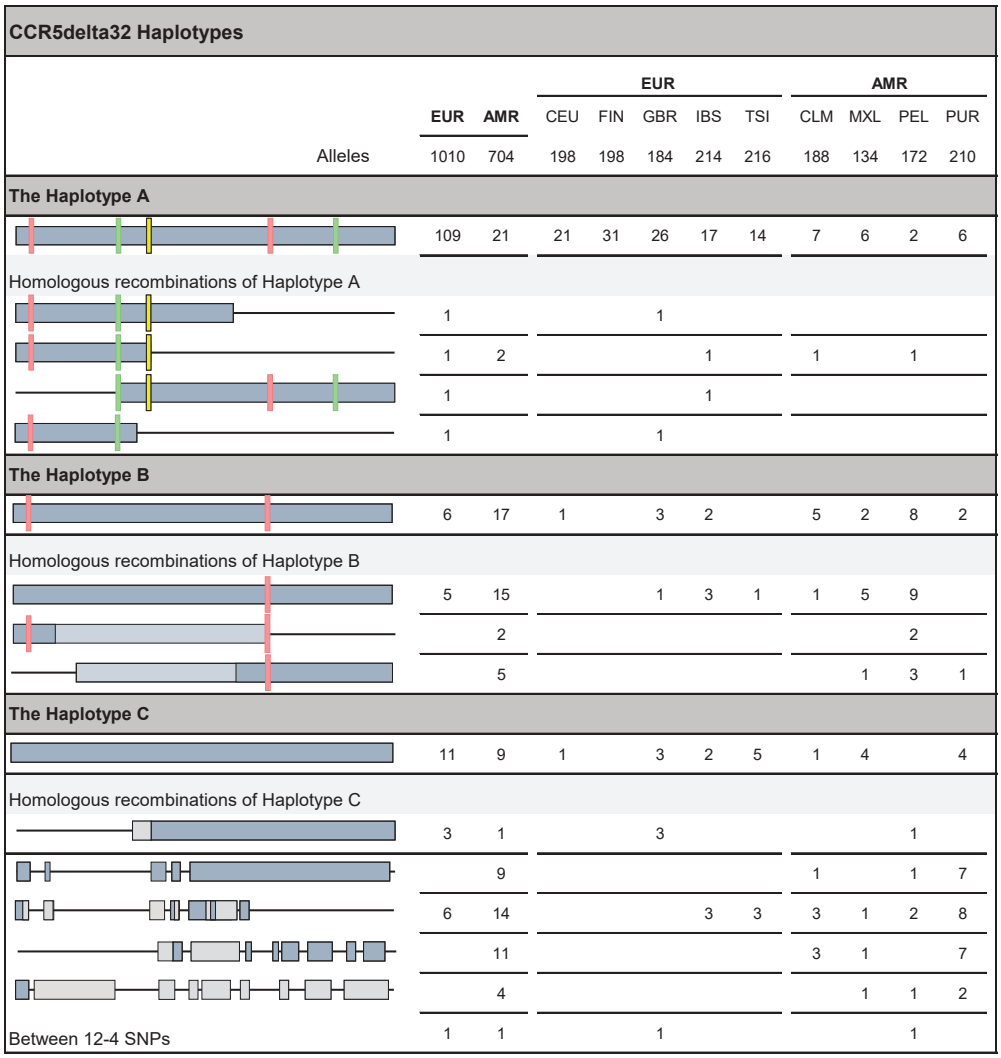


Figure 1. Schematic view of CCR5delta32 and the associated Haplotypes A, B, and C

(A) Haplotype A consists of CCR5delta32 and 86 tag variants, including two SNPs with an r^2 value of 1 (rs113341849 and rs113010081, green), two SNPs with an r^2 value of 0.9027 (rs79815064 and rs1157443, pink), and 82 variants with an r^2 value of 0.8602 (gray). All r^2 values are related to the CEU population (Table S1A). The haplotype is located on chr3 3p21.31, spans > 0,19 Mb, and encompasses several genes: *CCR3*, *CCR2*, *CCR5*, and *CCRL2*. Detailed information on the genomic locations of the genes, CCR5delta32, and the 86 tag variants are provided in Figure S1B.

(legend continued on next page)

the deletion in the CEU population (Figure 1A). We detected Haplotype B in 6 of the 505 1KGP3 European individuals (frequency = 0.006), with only one individual from the CEU population. Additionally, we identified a third haplotype, "Haplotype C," which included 82 of the 84 tag SNPs from Haplotype B, found in 10 European individuals (frequency = 0.01), again with only one individual from the CEU population. Beyond Haplotypes A, B, and C, we identified homologous recombinations and recurrent LD blocks involving Haplotypes A, B, and C, specifically restricted to the GRB, IBS, and TSI populations (frequency = 0.031) (Figure 1B; Table S2). These distinct haplotype patterns and recombination events likely explain the weaker LD of Haplotype A observed in Southern and Western Europe.

The three haplotypes span > 0.18 Mb (chromosome 3 [chr3]:46275570–46461783), including several cytokine receptor genes such as C-C motif chemokine receptor 3, 2, and 5 (*CCR3*, *CCR2*, *CCR5*), and C-C chemokine receptor-like 2 (*CCRL2*) (Figure S1B). To understand the potential functional effect of the variants carried by these haplotypes, we used the Ensembl Variant Effect Predictor (VEP),[66] and found that none of the tag SNPs could be annotated with clinical significance, as assigned by ClinVar.[67] However, from the genome-wide association study (GWAS) catalog,[68] the tag SNPs with $r^2 > 0.9$ have been previously associated with complex traits and diseases, such as insulin-dependent diabetes mellitus (IDDM), inflammatory bowel disease (IBD), and Alzheimer's disease (AD)[69–71] (Table S1G). Querying the Phenoscanner database[72,73] showed that 82 of the 86 tag variants of Haplotype A (including all the tag variants with $r^2 > 0.9$) were linked to many of the same phenotypic traits that were already associated with CCR5delta32s' multiple phenotypes (Table S1H). Notably, as the CCR5delta32 deletion is not detectable in traditional SNP-based GWAS analyses, some of these GWAS associations might be caused by LD with the CCR5delta32 allele.

Local admixture analysis revealed the European origin of CCR5delta32

We then expanded the analysis to the entire 1KGP3 dataset[65] (2,504 individuals from 26 populations), where we detected 35 individuals having the CCR5delta32 deletion outside of the EUR superpopulation panel, primarily in populations that have European ancestry (Table S2A). In Latin America, we could identify a homologous recombination of Haplotype A in two individuals from Colombian in Medellin (CLM) and Puerto Rican in Puerto Rico (PUR), which we also previously had detected in an individual from Spain (Figure 1B). To further investigate local admixture around the CCR5delta32 locus, we applied HaploNet[74] to all individuals of the 1KGP3 (Figure 2A). Here, we found evidence of a European sequence segment in 138 out of 141 individuals who harbored at least one allele of the deletion, while the remaining three individuals from Punjabi in Lahore, Pakistan (PJL) carried insufficient European ancestry proportions for HaploNet to distinguish fine-scale ancestry signals. Additionally, the complete Haplotype B (frequency = 0.024) and shorter homologous recombinants (frequency = 0.031) (Figure 1B) were found in significantly higher proportions in Latin Americans than in European populations (chi-squared test $p = 0.002616$ and $4.048e{-5}$, respectively). Likewise, we also observed the same pattern for homologous recombinants of Haplotype C (frequency = 0.057, chi-squared test $p = 2.115e{-6}$) (Figure 1B). Among the 1,008 individuals with African ancestry originating from the African continent (excluding African ancestry in SW USA [ASW] and African Caribbean in Barbados [ACB]), we did not detect any of the three haplotypes. We did, however, identify precursor SNPs for Haplotype C (Figure 2B). Out of the 82 variants of Haplotype C, 38 had a higher AF in the African population, compared with the European population (Figure S1C). Therefore, the increased AF of certain haplotype blocks in Latin American populations could be explained by admixture with individuals of African ancestry.

Haplotype A originated from Haplotype B in Europe

In general, older haplotypes are expected to be both more frequent and more diverse, compared with younger haplotypes. However, positive selection can decouple the expected relationship between haplotype age, frequency, and diversity, leading to a young, high frequency haplotype with low levels of diversity.[75,76] The presence of only four haplotype recombinants (HRs) of Haplotype A in the EUR population (Figure 1B), along with the high frequency of Haplotype A, indicates that this haplotype is much younger than Haplotypes B and C and/or that the CCR5delta32 deletion may have been subject to selection favoring its spread within EUR populations. Hence, based on present-day data alone, this suggests that at some point in the history of present-day Europeans, the two variants rs113341849 and rs113010081 (both with $r^2 = 1$) and the CCR5delta32 deletion emerged on Haplotype B, leading to Haplotype A, which is now present at substantial frequencies in present-day Europeans and in certain Latin American individuals, due to post-Columbian admixture (Figure 2B).

A probabilistic framework for calling the CCR5delta32 allele in low-coverage aDNA genomes

To trace the evolution of the CCR5delta32 allele through time, we aimed at identifying ancient individuals carrying the deletion. To achieve this, we developed a haplotype-aware probabilistic model for indels (HAPI), which allowed us to identify the deletion in low-coverage ancient genomes (see STAR Methods). In HAPI, we utilized the information from the four tag SNPs having the highest pairwise LD with the CCR5delta32 deletion ($r^2 > 0.90$, Table S1) as a prior for the presence of the deletion, and we modeled the information from the reads mapping to the *CCR5* deletion region in the form of a likelihood function. To mitigate reference bias and improve CCR5delta32 mapping detection, we used both the standard reference sequence (GRCh37) and a reference sequence where we added the deletion, hereinafter referred to as canonical and collapsed references, respectively.

(B) Detailed mapping of Haplotypes A, B, and C and their homologous recombinations among the individuals from 1KGP EUR and AMR populations. The light-gray blocks indicate deviations from different combinations of haplotype blocks. In the Latin American population, a specific homologous recombination of Haplotype A was identified in two individuals from CLM and PUR, which had also been previously detected in an individual from Spain.
See also Figure S1.

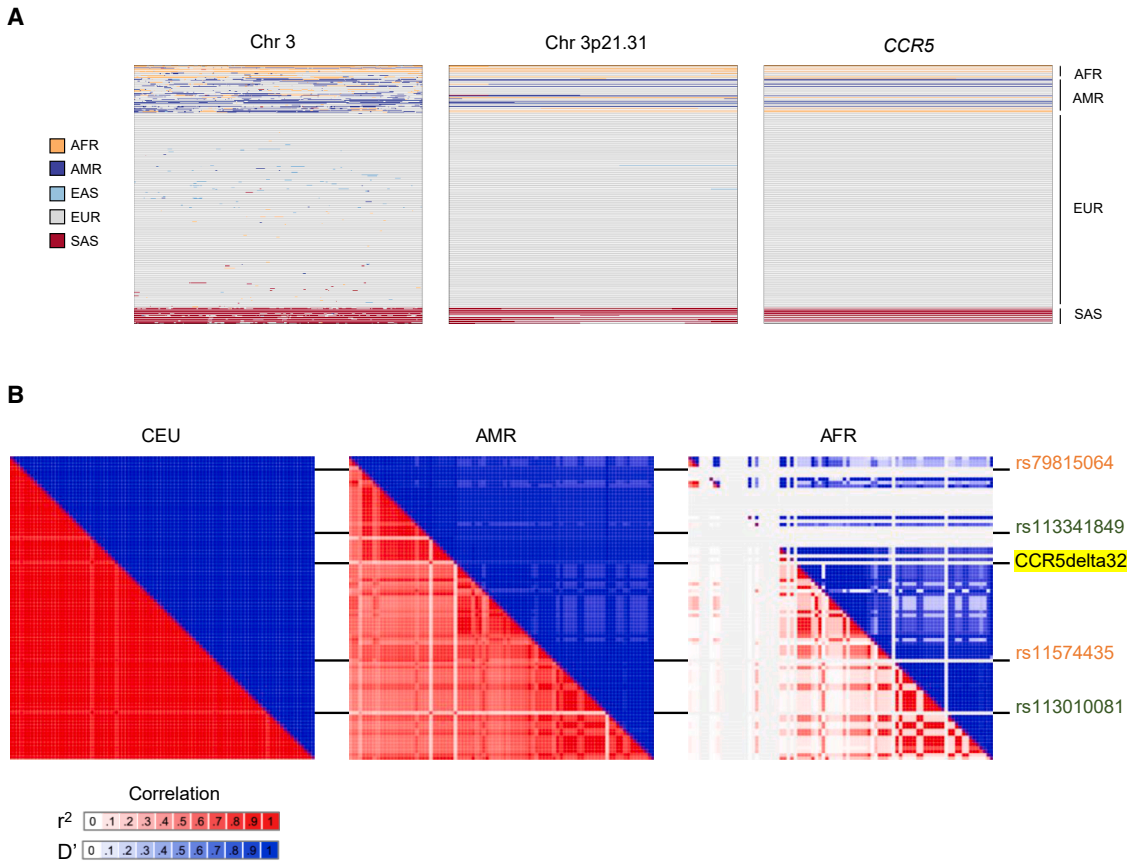


Figure 2. CCR5delta32 locus and Haplotype A patterns of LD in different populations

(A) Identification of a European CCR5delta32 locus in individuals with the deletion. Each line corresponds to a locus of an individual. The segments are colored according to HaploNet population estimation. From the HaploNet analysis, a European sequence segment was identified in 138 out of 141 1KG3 individuals, all genotyped with at least 1 allele of the deletion (Table S2).

(B) Heatmap matrices of pairwise LD statistics from Haplotype A in the CEU, AMR, and AFR populations. The strong LD pattern from Haplotype A in the CEU population becomes weaker in Latin America, owing to the significantly higher homologous recombination rates we observe from Haplotypes B and C in Latin America. These higher recombination rates may be explained by post-Columbian admixture among three groups, i.e., African, European, and Native American, as the AFR populations also harbor precursor variations for Haplotype C. The r^2 values are in shades of red, while the D′ values are in shades of blue. Darker values indicate a higher degree of pairwise LD.

See also Figure S1.

The full HAPI model thus used information from both the reads aligning to the CCR5delta32 region (in both the canonical and collapsed references) and from the reads mapping to the top four tagging SNPs (HAPI with informed prior) (Figure 3A). In addition to this, we also tested a HAPI model where we used only the information from the reads mapping to the deletion by using a uniform prior for the genotype calls (HAPI with uniform prior). We first tested HAPI on 15 genotyped CCR5delta32 genomes from the 1KGP3 and found that it correctly classified all of them (see STAR Methods). Since ancient genomes typically have much lower coverage, compared with modern genomes, we evaluated HAPI's performance at low coverages by simulating a dataset containing 144 ancient genomes with coverages from 0.3× to 10× (Figure S2A). First, we found that HAPI with the haplotype-informed prior performed better, compared with HAPI with the uniform prior, with an increase of Matthews correlation coefficient (MCC) from 0.79 to 0.97 (Figure 3B; Table S3). We then benchmarked HAPI versus the commonly used Genome Analysis Toolkit (GATK) HaplotypeCaller[77] and a graph-genome-based model variation graph (*vg*).[78] Each of the methods classified different numbers of samples because of the inherent differences in how the algorithms work, which influence their capacity to resolve ambiguous deletion genotypes. This is particularly evident when information is insufficient, which leads the algorithms to assign evenly distributed probabilities (0.33) across all deletion genotypes (RR, RD, and DD, i.e., homozygous for the reference, heterozygous, or homozygous for the deletion), thereby failing to make confident classifications (Figure 3E). In particular, we found that HAPI could classify 129 genomes out of 144 with an MCC of 0.97, compared with only 79 by the GATK HaplotypeCaller (MCC 0.47), an increase in genomes recovered by 63% (Figures 3B–3E). In comparison to *vg*, which integrates genetic variations like the CCR5delta32 deletion into the genome graph, HAPI delivered superior results, as *vg* called

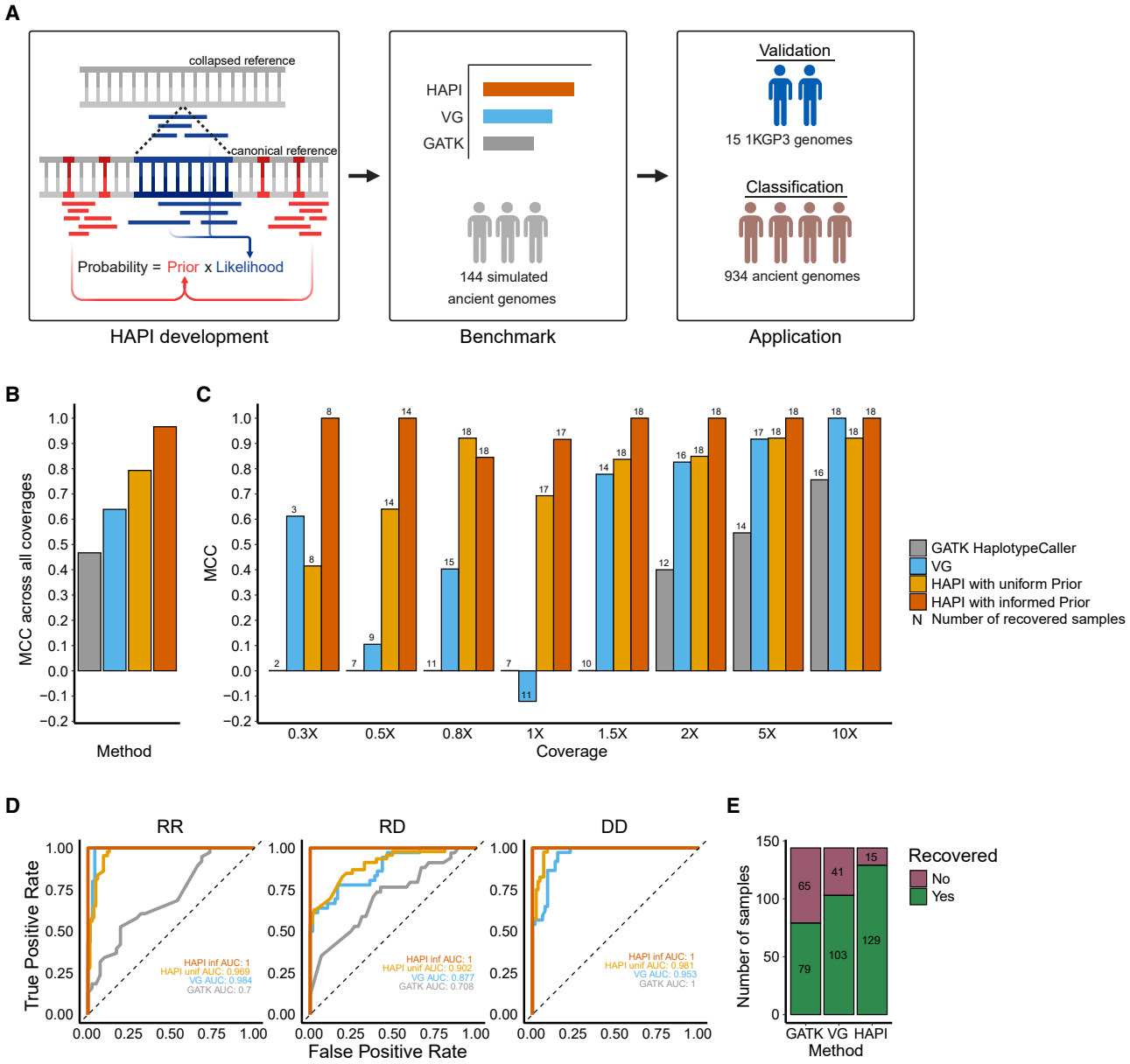


Figure 3. HAPI performances on simulated data

(A) Schematic overview of the information used in HAPI and of the steps of benchmarking on simulated data, validation on 1KGP3 samples, and testing on ancient genomes.

(B) Comparison of the performance of GATK HaplotypeCaller (gray), variation graph (*vg*) (blue), HAPI with uniform prior (yellow), and HAPI with informed prior (orange) in calling CCR5delta32 on 144 ancient simulated genomes. Performance is shown as MCC, using the simulated genomes called by each method across different coverages (0.3×–10×) and across all three deletion genotypes (RR, RD, and DD).

(C) MCC performance in calling CCR5delta32 by sequencing coverage, across all three deletion genotypes (RR, RD, and DD). The number of samples called by each method is indicated by the number at the top of the bars. This is because MCC is calculated using called samples and not where the methods did not report a genotype. The performance should be assessed using both MCC and the number of samples called.

(D) Performance shown as ROC-AUC using all simulated genomes stratified by CCR5delta32 genotype, across all coverages. HAPI provides a higher degree of reliability and accuracy in genotyping the deletion, compared with *vg* and GATK HaplotypeCaller.

(E) Number of samples called (recovered) using GATK Haplotypecaller, *vg*, and HAPI (with informed prior).

See also Figures S2 and S3.

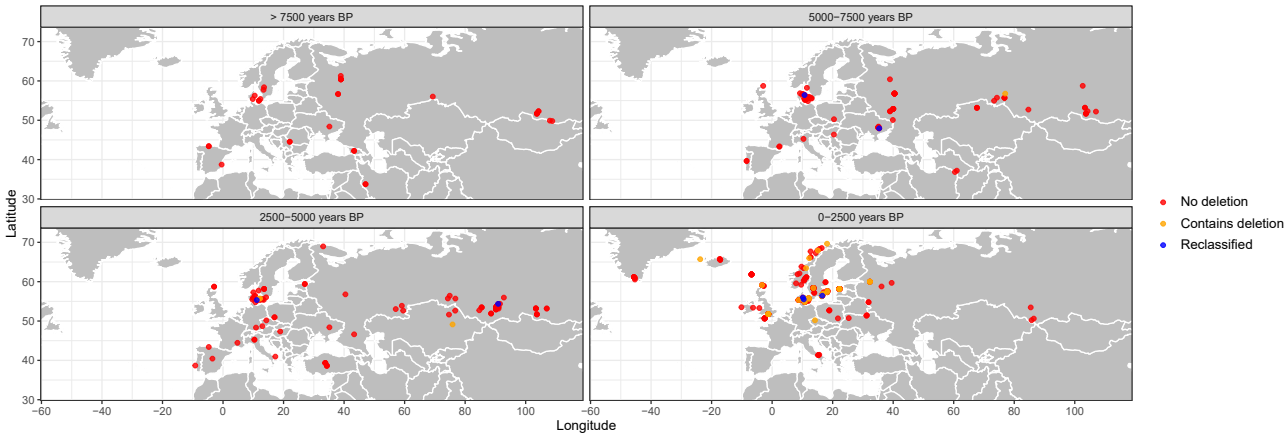


Figure 4. Geographical locations of the ancient genomes genotyped for the CCR5delta32

Map of distribution of ancient genomes genotyped with the permissive filter applied to the HAPI output, faceted by four time periods and colored based on the absence (red) or presence (orange) of the CCR5delta32. Blue dots correspond to reclassified genomes, i.e., genomes that are classified as having the deletion in the permissive filter genotype call set but not having the deletion according after applying the strict filter. The affected genomes are NEO300, NEO590, NEO855, RISE509, VK316, and VK342; see Table S5.

See also Figure S4.

only 103 out of 144 samples with an MCC of 0.64. Therefore, HAPI was not only able to call more genomes, but it was also much more precise, compared with both GATK and *vg*. The increased performance was most pronounced for the subset of genomes that had coverages between 0.5× and 1×, where we found that HAPI could correctly classify 49 of 54 genomes (90%) (Figure 3C). For coverage at 0.3×, 8 of 18 genomes (33%) could be classified by HAPI. Across these very low-coverage genomes ($\leq$1×) HAPI with informed prior had an MCC $\geq$ 0.84 (Figure 3C). On the contrary, GATK HaplotypeCaller and *vg* had problems when average genome coverage was below 2× and 1.5×, respectively (Figure 3C). Among the three deletion genotypes (RR, RD, and DD), the difference between the methods was even more pronounced for the RR and RD ones. Here, HAPI with informed prior had an average increase in performance in terms of ROC-AUC (receiver operating characteristic area under the curve) of 0.30 and 0.07, compared with GATK and *vg*, respectively (Figure 3D). For the DD genotype, the performance differences between the methods were less pronounced (Figure 3D). However, it should be highlighted that the reported performances are calculated only from the classified samples, potentially exaggerating the success of methods that leave more samples unclassified. Thus, a complete evaluation should weigh both classification performance and the ability to retrieve a high number of samples. Considering both these aspects, our HAPI model was highly specific for identifying the CCR5delta32 allele, even in the heterozygous form and with as little as 0.3× coverage, in addition to being the model that could classify the most samples.

Applying HAPI to ancient dataset

We then applied HAPI to our extensive ancient DNA dataset, which consisted of 934 genomes[79–82] from various regions across Eurasia, including a dense sampling collection in Northern Europe, specifically in Denmark (Table S4). The final dataset encompasses consecutive historical eras, ranging from the early Mesolithic and Neolithic periods to the Bronze Age and extending into the Viking Age. To take into account the complexity of the haplotype and the damaged nature of ancient DNA, we applied two curation steps to the results of the model: a "permissive filter" to reclassify genomes that had artifacts typical of ancient DNA damage and a "strict filter" to reclassify genomes that were likely harboring the Haplotype B (see STAR Methods). Across the ancient DNA dataset, we found that 418 genomes had at least one read mapping to the *CCR5* region from either the canonical or collapsed reference and at least 6 bases overlapping the CCR5delta32 breakpoint (the geographical locations of these ancient genomes are provided in Figure S4A). Using this approach, we identified the CCR5delta32 allele in 46 and 39 individuals using the permissive and strict filters, respectively (Figure 4; Table S5). From the Allentoft et al.[79] dataset spanning the Mesolithic and Neolithic, 135 of the 317 genomes passed the inclusion criteria for HAPI (see STAR Methods). Among these Mesolithic and Neolithic samples, three individuals were identified with the deletion using the strict filter and eight individuals were classified with Haplotype B across the different output schemes from HAPI (Table S5). From the Botai dataset[81] of the Early Bronze Age, a single genome was identified with the CCR5delta32 allele. Only 31 out of 101 genomes from the Bronze Age[82] met the criteria for the analysis by HAPI, and although the sample pool was small, we detected one sample with the CCR5delta32 deletion, using the strict filter, and two samples carrying Haplotype B (Table S5). From the Viking dataset,[80] 252 of 442 genomes passed the HAPI inclusion criteria (Table S5). From these, 34 genomes were detected to have the CCR5delta32 deletion with the strict filter (Haplotype A, AF = 0.067), and two genomes were identified as having Haplotype B (AF = 0.003). Furthermore, we observed 22 genomes containing sections of Haplotype C (>20 proxy SNPs). To show temporal changes in CCR5delta32 AF in our whole ancient dataset, we calculated the derived allele frequencies (DAFs) for CCR5delta32 in 1,000-year bins, fitting binomial regressions and 95% confidence intervals

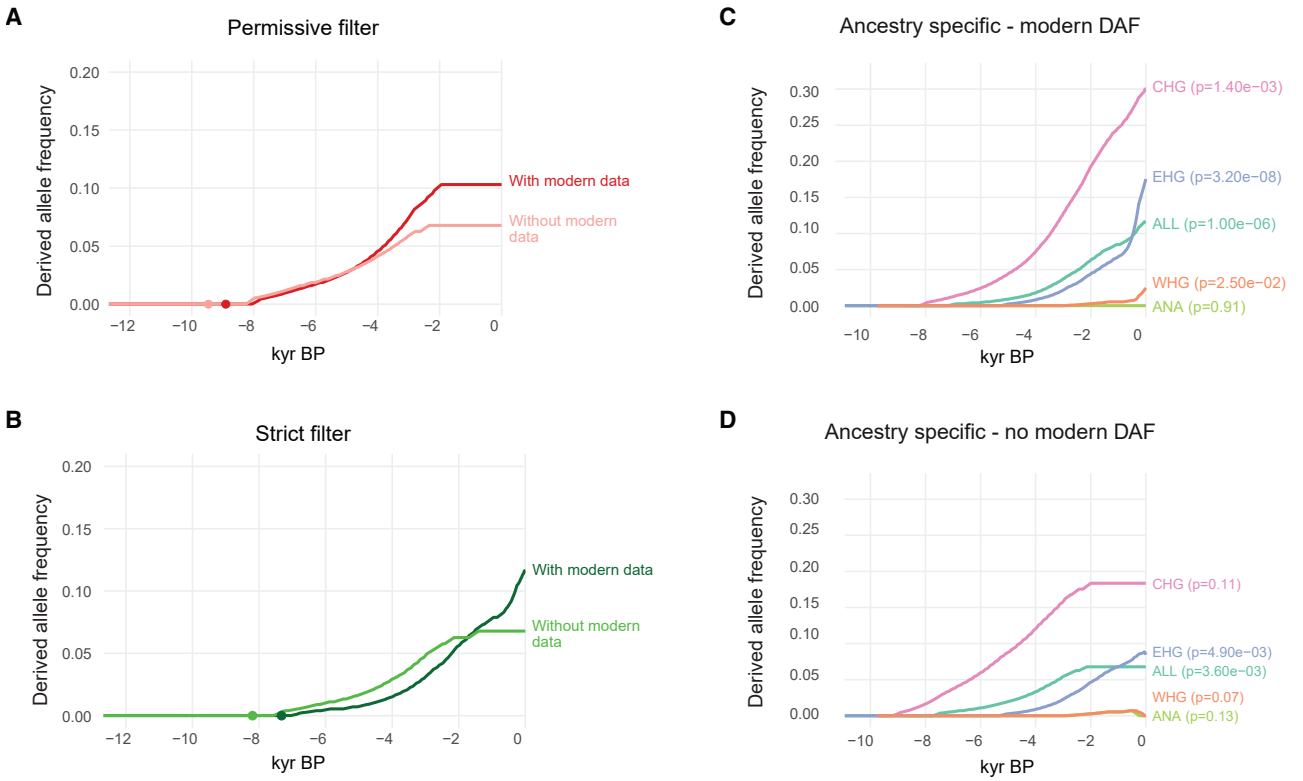


Figure 5. CCR5delta32 AF trajectory

Maximum likelihood trajectory of the CCR5delta32 estimated using CLUES.

(A) Results obtained using HAPI calls and using permissive filter.

(B) Results obtained using HAPI calls and strict filter. The dots in each figure represent the age estimate of the variant either with or without conditioning on modern ascertainment.

(C) Ancestry-specific trajectory analysis using modern DAF ascertainment and consensus calling of CCR5delta32 based on imputation of the four tag SNPs.

(D) As (C), but without using modern DAF. The ancestries tested were ALL (population wide), Caucasus hunter-gatherer (CHG), Eastern hunter-gatherer (EHG), Western hunter-gatherer (WHG), and Anatolian Neolithic ancestry (ANA). The *p* values are the result of the CLUES selection test.

See also Figures S5–S7.

for each bin (Figure S5). A detailed view of the location of the ancient DNA genomes analyzed here is provided in Figures S4A and S4B and Table S5.

Geographic distribution of Haplotypes A, B, and C in Mesolithic and Neolithic Eurasia

In our dataset from the Mesolithic and Neolithic, the oldest known sample with the CCR5delta32 allele was identified at Protoka, Russia, dating back approximately 5,800 years (NEO309: 5,824 calibrated years before the present [cal. BP]). For this sample, we identified 3 reads covering the canonical reference sequence (GRCh37) and 2 reads covering the deletion in the collapsed reference, as well as 54 proxy SNPs linked to Haplotype A, including 3 of the 4 highest-ranked tag SNPs utilized in the HAPI model. Another significant sample harboring the CCR5delta32 allele was discovered in Karagash, Kazakhstan, dating nearly 1,000 years younger (Yamnaya: 4,903 cal. BP). This genome had high coverage (26×), allowing us to identify all 86 tag SNPs associated with Haplotype A, as well as 14 reads covering the CCR5delta32 deletion. Owing to the extensive collection of ancient genomes from Danish individuals (>100 genomes), we were able to identify the presence of a CCR5delta32 allele (Haplotype A) in Denmark over 4,000 years ago (NEO878: 4,038 cal. BP). We also detected the deletion in a Bronze Age skeleton from Hove Å (NEO946: 3,145 cal. BP) (Figures 4 and S4). For Haplotypes B and C, we found them to be widespread across Eurasia, with signal found in genomes predating Haplotype A. For instance, we identified traces of Haplotype B in Ukraine (NEO300: 6,678 cal. BP), Sweden (NEO27: 9,693 cal. BP), and Portugal (NEO631: 7,135 cal. BP). The oldest evidence of Haplotype C was found in Russia, dating to 10,853 cal. BP (NEO202: 69/82 tag variants). Similarly, we identified a sample from northwest Spain (NEO646), dated to 8,274 cal. BP, with 35 tag variants of Haplotype C. Together, these results indicate that despite a deep genetic divide between Western and Eastern Eurasian populations,[79] both groups carried fragments of the three haplotypes (Figure S4B; data provided in Table S5).

Positive selection operated on CCR5delta32 from 8,000 to 2,000 years BP

Based on these results, we aimed to model the spatiotemporal frequency dynamics of the CCR5delta32 allele across West

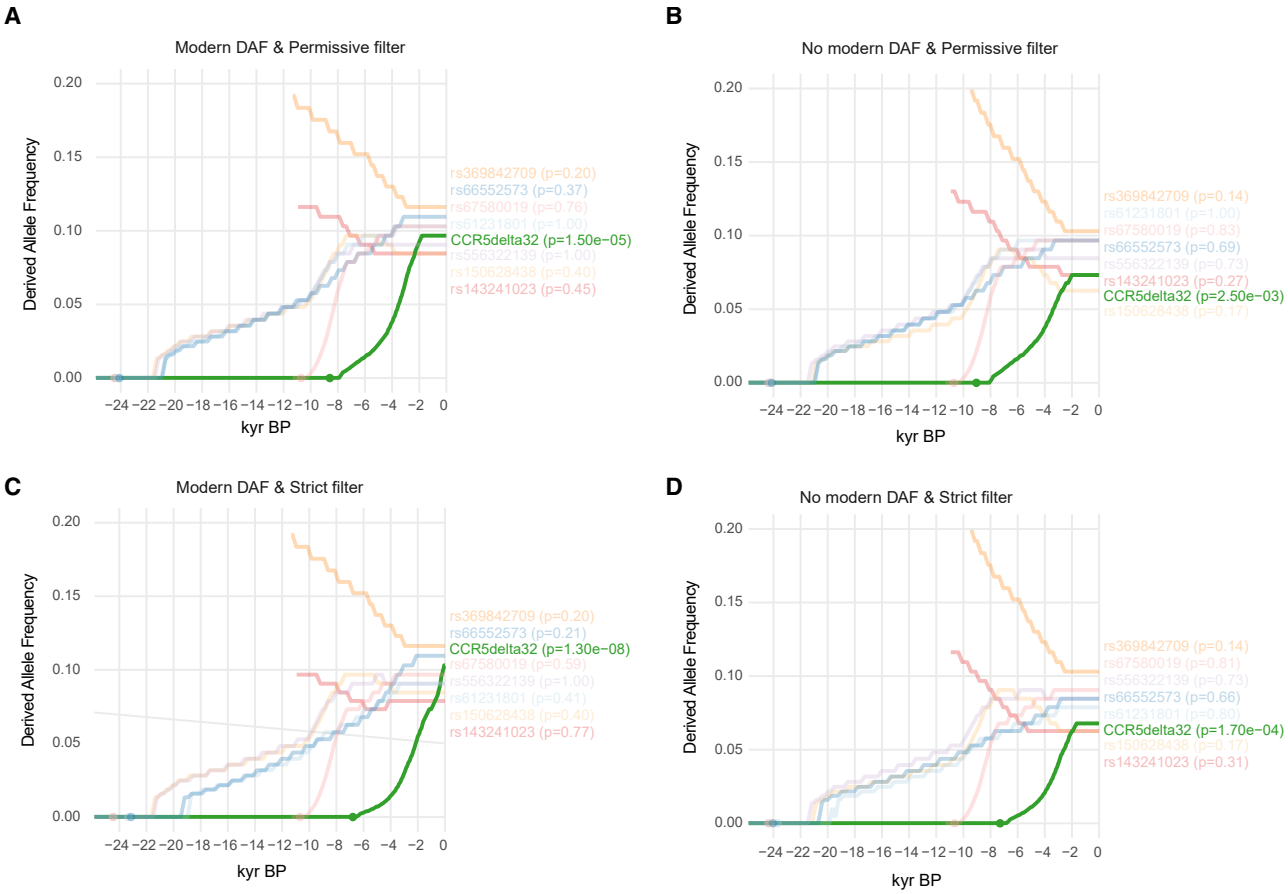


Figure 6. No evidence for selection using seven control deletions

(A–D) AF trajectories for CCR5delta32 and seven randomly ascertained control deletions with similar characteristics to the CCR5delta32 deletion (chr3, 32 ± 5 bp in length and a MAF of 0.1098 ± 0.02) and in LD with variants exhibiting $r^2 > 0.8$. We applied HAPI and CLUES as used for CCR5delta32 and found that none of the 7 deletions showed evidence of selection (minimum $p = 0.17$). (A and B) Data from permissive filter, and (C and D) data from strict filter.

Eurasia to reconstruct the evolutionary history of this allele and to investigate the evidence in favor of positive selection at the locus. We used a modified version of CLUES[83,84] to infer AF trajectories over time, using ancient genomes. In addition to our different genotype call sets (strict filter and permissive filter), we evaluated the trajectories if conditioned on allele frequencies observed in present-day European populations (Figures 5A, 5B, and S6). Across these analyses, we observed a rapid rise in the CCR5delta32 frequency until 2,000 years BP, followed by a stabilization of the frequency until the present. When using the strict filter call set and modern ascertainment (Figures 5B and S6), however, we observed a very recent uptick in frequency. This was likely an artifact of under-calling the ancient genomes under this filtering scheme, causing the model to reach present-day frequencies very quickly. We found significant evidence for positive selection acting on the CCR5delta32 allele in the ancient past, using both strict and permissive filters against a neutral model (with p values of 1.7e−4 and 2.5e−3, respectively, Table S6). In addition, when conditioning on present-day frequency, we obtained even greater significance levels in favor of selection (with p values of 1.3e−8 and 1.5e−5 for strict and permissive filters, respectively. We estimated that a large selection coefficient ($s > 0.01$) better explained the initial AF rise. When using the strict filter calls, the coefficient was predicted to be larger ($s = 0.033$ with conditioning, $s = 0.026$ without conditioning) than when using the permissive filter deletion calls ($s = 0.021$ with conditioning, $s = 0.017$ without conditioning). The posterior estimates for the age of the CCR5delta32 deletion from the CLUES analysis were between 8,988 and 6,748 years BP. To ensure that our methodology was not biased toward detecting positive selection, we repeated the analysis on deletions with characteristics similar to those of CCR5delta32. We scanned chr3 for a set of control deletions, which were not linked to any known phenotypic trait, and identified seven deletions matching CCR5delta32 in length (32 ± 5 bp), minor AF (MAF = 0.1098 ± 0.02), and tag SNP number (ranging from 2 to 203 tag SNPs at $r^2 > 0.8$) within the 1KGP EUR superpopulation (Tables S7 and S8). We then repeated the HAPI and CLUES analysis for each of the control deletions and found that none had significant evidence of selection in any of the four models (minimum $p = 0.17$) (Figure 6). Taken together,

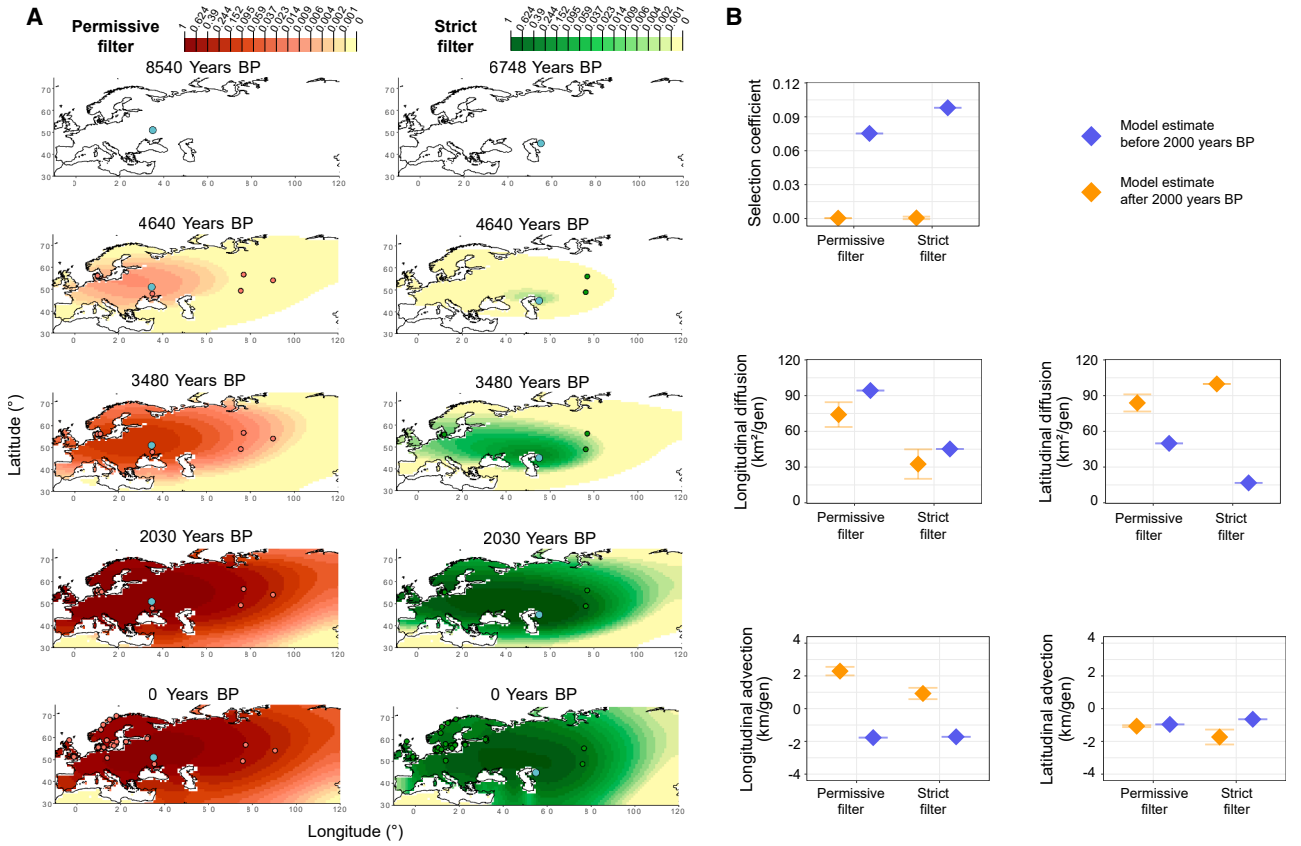


Figure 7. CCR5delta32 AF dynamics across West Eurasia

(A) Spatial AF dynamics inferred by the diffusion-advection method. Left: permissive filter, right: strict filter. The green and red dots are genomes containing the deletions that are at least as old as the year indicated in each corresponding time slice. The light blue dot corresponds to the inferred origin of the allele.

(B) Parameter estimates from the spatiotemporal diffusion analysis used to generate AF dynamic maps along with 95% confidence intervals. Results are shown for permissive and strict filter genotype call sets for time periods before and after 2,000 years before present. The selection coefficient estimates indicate that selection likely operated early in the history of the allele, during the Late Neolithic and Bronze Age. The bars show 95% confidence intervals.

these findings suggest that CCR5delta32 has been under positive selection from approximately 8,000 to 2,000 years BP.

CCR5delta32 was positively selected in Eastern and Caucasus hunter-gatherers

Because our findings of selection could be confounded by admixture between ancient populations with different AFs frequencies, we performed an ancestry-stratified CLUES analysis using local ancestry tracts from Irving-Pease et al.[84] We stratified the analysis by Western hunter-gatherers (WHGs), Caucasus hunter-gatherers (CHGs), Eastern hunter-gatherers (EHGs), and Anatolian Neolithic ancestry (ANA) ancestry tracts. We found that selection on CCR5delta32 was significant for the two ancestry paths contributing to the Steppe pastoralist expansion in the Early Bronze Age (Figures 5C and 5D; Table S6). Specifically, EHG was significant both when conditioning with (p = 3.2e−8) and without modern-day AF (p = 4.9e−3), and CHG was significant when conditioning on modern-day AF (p = 1.4e−3). These results indicate that positive selection for the CCR5delta32 allele likely began in the EHG and CHG populations.

Spatiotemporal AF dynamics

To investigate the spread of the allele, we fitted a two-dimensional diffusion-advection method that integrates present-day and ancient human genomes to infer AF dynamics across space and time.[85] The method infers parameters associated with how fast the allele spreads across the landscape and how fast it increases in frequency locally due to positive selection. It also estimates the likely geographic origin of the allele, given the data. Because the CLUES analysis indicated that the AF dynamics changed before and after 2,000 years BP, we partitioned our spatial inference framework into these two periods, allowing the method to find two separate selection coefficients and diffusion parameters for each period. For the permissive filter, we estimated s = 0.0753 before 2,000 years BP and s = 0.0003 after 2,000 years BP. For the strict filter, the corresponding selection coefficients were s = 0.098 before 2,000 years BP and s = 0.0005 after 2,000 years BP. The selection coefficient estimates were higher using the strict filter compared with the permissive filter, likely due to the younger allele age estimate (Figure 7B; Table S7). We inferred the allele's origin to be in the Western Eurasian Steppe region, using both filtering schemes, with the

strict filter placing the allele further east, compared with the permissive filter (Figure 7A). This was followed by rapid longitudinal expansion during the earlier period toward the west. Regardless of the call set, the selection coefficient was inferred to be higher in the period before 2,000 years BP (Figure 7B), consistent with the CLUES analysis (Table S6), suggesting that selection likely operated early in the history of the allele (i.e., during the Late Neolithic and Bronze Age).

Caution is required when using tag variants for CCR5delta32 due to complex haplotypes

Previous work by Le et al. analyzed data from 1,291 ancient samples using an in-solution hybridization capture assay to test for selection of the CCR5delta32 allele.[61] Since the dataset was generated using a capture technique, the CCR5delta32 allele was not directly measured, and their selection analysis relied on the proxy variant rs73833033. However, we found that this variant was not a reliable proxy for the CCR5delta32 deletion, as it also occurred in Haplotypes B and C, making it unsuitable as a tag SNP for CCR5delta32. Notably, rs73833033 had an AF = 0.25–0.30 in the AFR population from 1KGP3, where the CCR5delta32 allele was absent (Figure S7B), suggesting a possible non-European origin. When we analyzed this allele with CLUES analyses on our dataset, we did not find significant evidence of selection (Figures S7C–S7F), reinforcing that the evolutionary patterns of rs73833033 and CCR5delta32 should not be directly equated.

DISCUSSION

This study provides fundamental insights into the evolutionary history of the CCR5delta32 allele. Our discovery and mapping of the Haplotypes A, B, and C in present-day genomes led us to develop a probabilistic model, HAPI, to investigate the CCR5delta32 allele in ancient genomes. The model allowed us to call the CCR5delta32 allele in genomes with as little as 0.3× coverage. Based on this, we date the deletion to be at least 6,700 years old, arising among peoples occupying the Western Eurasian Steppe region in the Neolithic. We also show that the CCR5delta32 allele was exposed to positive selection during the Late Neolithic and Bronze Age, followed by stability in the AF until the present day.

Applying the knowledge of Haplotype A, combined with the evidence from HaploNet, we show that the CCR5delta32 allele originated on the Western Eurasian Steppe, while previous work suggested a European origin.[12,55] The Columbian Exchange, which is considered to have facilitated genetic admixture among three groups, i.e., African, European, and Native American,[86,87] can account for the apparent higher recombination rates we observed from Haplotypes B and C in Latin America, compared with European populations, along with the higher AF from some of the variants including in Haplotype C (Figures 2B and S1C). Thus, we can propose to include the CCR5delta32 allele, together with the variants rs113341849 and rs113010081, as European ancestry-informative markers. Furthermore, the CCR5delta32 deletion can be reliably imputed from SNP arrays using the two $r^2 = 1$ tag SNPs (rs113341849 and rs113010081), as they are located on each side of the CCR5delta32 allele.

Previous studies investigating the evolutionary history of the CCR5delta32 allele have either been restricted to contemporary individuals,[12] used very few ancient genomes from limited geographic areas,[46,48,49,63] or relied on a tag SNP in low LD with the deletion.[61] Here, we present results obtained using a large comprehensive set of ancient genomes (>900 genomes) combined with modern genomes. The CLUES analysis revealed that the allele rose quickly in frequency in the period before 2,000 years BP, followed by a period of stabilization, over which the allele remained at around 10% frequency. This agreed with findings from Bouwman et al.[46] and Hummel et al.,[49] who posited a period of recent allele stability over the past millennium in central and North Germany. Based on our data, the allele had an origin in the Western Steppe and a fast rapid diffusion eastward and westward early in its history, partly coinciding with the eastward movements from the Steppe during the Bronze Age.[82,88] We note, though, that the origin of the allele inferred by the model is highly dependent on the first instances of the allele in the data, and it is thus highly dependent on the mode of deletion calling. Under the curated calling schemes, the lower inferred counts of the allele during the Mesolithic and Neolithic led the model to estimate a fast longitudinal diffusion, as the most likely AF surface rapidly shifts from complete absence to widespread presence of the allele in distant regions across the continent. The rapid longitudinal spread of the allele is consistent with previous evidence for long-distance dispersal of the allele,[12] although our ancient data suggest that this dispersal occurred earlier than previously thought. Our estimated age of the allele is consistent with a more ancient origin as postulated in Sabeti et al.,[15] rather than a recent origin as suggested in other studies.[11,17] All age estimates we obtained were older than 6,700 years BP (posterior estimates between 8,988 and 6,748 years BP).

We found significant evidence of positive selection acting on the CCR5delta32 allele in the ancient past, when fitting the data to the CLUES model. When we conditioned the CLUES trajectories on reaching the frequencies observed in present-day data, they resulted in stronger rises in frequency, compared with using ancient data alone, which in turn resulted in more significant p values in the rejection of neutrality. This likely indicates an undercounting of the allele in the more ancient time periods. Regardless of the calling scheme, we found significantly large selection coefficients when deploying the spatiotemporal spread model, particularly in the early time period, but no evidence for selection after 2,000 years BP. Of note, however, is that the spatiotemporal model is deterministic, and thus necessarily underestimates the amount of AF stochasticity that occurs during the period under study, so the selection coefficient inferred under this model may be an overestimate. To contextualize the finding of positive selection of the CCR5delta32 allele, we scanned chr3 for a set of control deletions (i.e., not associated with any known trait and thus likely unaffected by selection). From this scan, we identified seven deletions associated with haplotypes similar in characteristics to the CCR5delta32 Haplotype A. Using our HAPI, we called the deletions and applied a CLUES selection test to each. None of the seven deletions

showed significant evidence of selection, strongly supporting our findings of the CCR5delta32 deletion and its clear signal of positive selection. A recent study by Le et al.[61] found no evidence for the selection of the CCR5delta32 allele during ancient times. However, as a proxy for the CCR5delta32 they used the SNP rs73833033, which is present in all three haplotypes (A, B, and C) rather than being specific to Haplotype A that contains the CCR5delta32 allele, and this likely led to an overestimation of the allele's distribution in their study (Figure S7).

The notable increase in CCR5delta32 AF prior to the Iron Age implies that the high frequencies of this allele in modern-day Europe cannot be caused by the Medieval Plague as hypothesized previously.[46,49] Instead, the selection signature may have resulted from pressures exerted by previous outbreaks or other pathogens that existed in the past.[14,16,89] The observed distribution and our age estimation of the CCR5delta32 allele were also not consistent with the Viking-spread hypothesis.[50] Instead, the rapid longitudinal spread that we infer (approx. 60–100 km^2 per generation (Figure 7) and the rapid rise in frequency observed during the Bronze Age suggest a possible spread associated with the Late Neolithic and Early Bronze Age expansion of Steppe-related ancestry into Europe.[82,88] This is further supported by our ancestry-stratified selection test, which accounts for changes in admixture proportions over time. Our analysis shows that the CCR5delta32's AF was the highest in the two ancestral populations contributing to the Steppe pastoralist expansion during the Early Bronze Age, specifically the EHG and CHG (Figures 5C and 5D). This reinforces that our selection results are not biased by gene flow effects.

Today, immunological genetic signatures by selection and/or adaptation through admixture can be observed in the human genome.[2,3] Our data show that the CCR5delta32 allele (Haplotype A) could very well be among these genetic signatures. We cannot point out a direct cause for the increase of CCR5delta32's AF during the Neolithic and Early Bronze Age, but it is clear that Haplotype B did not undergo the same evolutionary trajectory. The key to understanding the driving forces for the CCR5delta32 deletion is challenged by the immune system redundancy and the immune gene pleiotropy.[90,91] One hypothesis could be that the CCR5delta32 allele with the 86 tag variants is associated with cytokine/chemokine profiles reflecting immune tolerance, which has been shown to be under selection during the Neolithic age.[9]

Finally, the fact that individuals bearing the CCR5delta32 allele also harbor a defined haplotype widens the complexity of the deletion effects. The CCR5delta32 deletion has been studied extensively for nearly three decades, especially due to its strong link to HIV-1 infection resistance and thereby the potential to target CCR5 for HIV treatment and for HIV pre/post-exposure prophylaxis medicine.[25,27,92,93] These therapeutic approaches include gene editing techniques like CRISPR, CCR5 blockade using antibodies or antagonists, or combinations of both.[25,94] Furthermore, the CCR5delta32 allele can be viewed as a pleiotropic variant, owing to its influence on multiple phenotypic traits, e.g., autoimmune and inflammatory diseases, cardiovascular diseases, neurodegenerative disease, and cancer.[20,29,31,44,95] It is possible that some of the CCR5delta32 tag SNPs contribute to the pleiotropic nature of CCR5delta32, although *in silico* analysis shows no direct clinical significance. More precisely, the gene expression of cytokine receptors (*CCR3*, *CCR2*) and *CCRL2* might be affected by one or more of the tag SNPs (Haplotype A), leading to modulation of chemokine-chemokine receptor signal transduction.[19,95,96] This calls for further studies to elucidate these possible direct or indirect effects. Thus, the tag SNPs should be considered when analyzing causes of the CCR5delta32 pleiotropic effects and when developing therapeutic approaches, based on mimicking the naturally occurring CCR5delta32 genotype-phenotype correlations. Therefore, our results point in the direction of complex CCR5delta32 genotype-haplotype-phenotype relationships, which demand consideration when targeting the CCR5 receptor for therapeutic strategy.

Limitations of the study

Working with ancient DNA introduces several inherent limitations. Our whole-genome ancient dataset is characterized by low-coverage genomes of fragmented DNA sequences, which may include misincorporated nucleotides. This may limit the accuracy and reliability of our findings. To assess our ability to accurately genotype the CCR5delta32 allele in ancient genomes, we benchmarked HAPI using simulated data, but we cannot guarantee that these capture the full spectrum of existing CCR5 recombinant haplotypes. Consequently, differences in recombinant haplotypes between ancient and contemporary individuals may pose a risk of overestimating the accuracy of HAPI. Although we implemented various filtering steps to control for artifacts typical of ancient DNA reads and account for diverse haplotypes, misclassifications of the CCR5delta32 genotype may still occur. Additionally, uneven spatiotemporal sampling, in particular the underrepresentation of key periods, can affect the generalizability of our conclusions and may bias results toward specific populations or time frames.

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources should be directed to and will be fulfilled by the lead contact, Professor Simon Rasmussen (srasmuss@sund.ku.dk).

Materials availability

This study did not generate new unique reagents.

Data and code availability

The modern data were downloaded from the 1KGP3 (ftp://ftp.1000genomes.ebi.ac.uk/.uk). The Mesolithic and Neolithic data were obtained from Allentoft et al.[79] The Bronze Age and Iron/Viking Age datasets were obtained from de Barros Damgaard et al.,[81] Allentoft et al.,[82] and Margaryan et al.,[80] respectively. The ancient DNA and simulated sequence data that mapped to the GRCh37 and the collapsed reference, restricted to the CCR5delta32 region, are available at https://doi.org/10.17894/ucph.b42e9ebf-23b6-4908-a69e-7afb1c3d4ad0. The HAPI model is available as a pip package at https://pypi.org/project/hapi-pyth/, and instructions on how to install and run it are available at https://github.com/RasmussenLab/HAPI. This same repository contains the bioinformatics code used for pre- and post-processing the samples. The code for the CLUES analysis and the calls for rs73833033 are available at https://github.com/ekirving/ccr5_paper. The code for reproducing the spatiotemporal diffusion analysis can be found at https://github.com/RasaMukti/ccr5delta32_analysis. Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

ACKNOWLEDGMENTS

K.R., L.C., and S.R. were supported by the Novo Nordisk Foundation (grants NNF14CC0001 and NNF23SA0084103). E.K.I.-P. was supported by the Lundbeck Foundation (grant R302-2018-2155) and the Novo Nordisk Foundation (grant NNF18SA0035006). E.K.I.-P. and R.A.M. were additionally supported by a Villum Young Investigator grant and a Novo Nordisk Ascending Investigator grant given to Fernando Racimo (project nos. 00025300 and NNF22OC0076816, respectively). We would like to express our sincere gratitude to Fernando Racimo for his invaluable insights and expertise contributed to this research. While he has chosen not to be listed as an author, his input was fundamental in enhancing the quality of this work.

AUTHOR CONTRIBUTIONS

Conceptualization, M.E.A., E.W., K.R., L.C., and S.R.; methodology, K.R., L.C., and S.R.; data curation, M.S., M.E.A., and E.W.; investigation, K.R., L.C., R.A.M., J.M., and E.K.I.-P.; software, K.R., L.C., L.S.D., R.A.M., J.M., T.S.K., and E.K.I.-P.; formal analysis, M.S.; writing – original draft, K.R., L.C., R.A.M., and S.R.; writing – review & editing, all authors; resources, M.E.A. and E.W.; supervision, E.K.I.-P. and S.R.

DECLARATION OF INTERESTS

S.R. is the founder and owner of the Danish company BioAI and has performed consulting for Sidera Bio ApS.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
- METHOD DETAILS
  - Data
  - Identification of the CCR5delta32 deletion and the haplotypes
  - Analysis of proxy variant rs73833033
  - Simulation of ancient CCR5delta32 genomes
  - Processing of simulated and ancient genomes
  - Processing of the 15 human genomes from the 1KGP3 dataset
  - Applying HAPI to the ancient genomes and filtering of calls
  - Benchmarking HAPI versus GATK HaplotypeCaller and VG
  - Identification of seven ascertained control deletions and their haplotypes
  - Local ancestry of individuals harboring CCR5delta32 deletion in the 1KGP3
- QUANTIFICATION AND STATISTICAL ANALYSIS
  - Haplotype-Aware Probabilistic model for Indels (HAPI)
  - Determining minimum overlapping length
  - Optimizing the HAPI model
  - CLUES analysis
  - Estimating the age of the CCR5delta32 deletion
  - Method for modeling the spatiotemporal diffusion of the deletion allele

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.cell.2025.04.015.

Received: August 16, 2023
Revised: January 31, 2025
Accepted: April 7, 2025
Published: May 5, 2025

REFERENCES

1. Morens, D.M., and Fauci, A.S. (2020). Emerging Pandemic Diseases: How We Got to COVID-19. Cell *182*, 1077–1092. https://doi.org/10.1016/j.cell.2020.08.021.
2. Quintana-Murci, L. (2019). Human Immunology through the Lens of Evolutionary Genetics. Cell *177*, 184–199. https://doi.org/10.1016/j.cell.2019.02.033.
3. Kerner, G., Neehus, A.-L., Philippot, Q., Bohlen, J., Rinchai, D., Kerrouche, N., Puel, A., Zhang, S.-Y., Boisson-Dupuis, S., Abel, L., et al. (2023). Genetic adaptation to pathogens and increased risk of inflammatory disorders in post-Neolithic Europe. Cell Genomics *3*, 100248. https://doi.org/10.1016/j.xgen.2022.100248.
4. Karlsson, E.K., Kwiatkowski, D.P., and Sabeti, P.C. (2014). Natural selection and infectious disease in human populations. Nat. Rev. Genet. *15*, 379–393. https://doi.org/10.1038/nrg3734.
5. Barreiro, L.B., and Quintana-Murci, L. (2010). From evolutionary genetics to human immunology: how selection shapes host defence genes. Nat. Rev. Genet. *11*, 17–30. https://doi.org/10.1038/nrg2698.
6. Casadó-Llombart, S., Velasco-de Andrés, M., Català, C., Leyton-Pereira, A., Lozano, F., and Bosch, E. (2021). Contribution of Evolutionary Selected Immune Gene Polymorphism to Immune-Related Disorders: The Case of Lymphocyte Scavenger Receptors CD5 and CD6. Int. J. Mol. Sci. *22*, 5315. https://doi.org/10.3390/ijms22105315.
7. Deschamps, M., Laval, G., Fagny, M., Itan, Y., Abel, L., Casanova, J.-L., Patin, E., and Quintana-Murci, L. (2016). Genomic Signatures of Selective Pressures and Introgression from Archaic Hominins at Human Innate Immunity Genes. Am. J. Hum. Genet. *98*, 5–21. https://doi.org/10.1016/j.ajhg.2015.11.014.
8. Quach, H., Rotival, M., Pothlichet, J., Loh, Y.-H.E., Dannemann, M., Zidane, N., Laval, G., Patin, E., Harmant, C., Lopez, M., et al. (2016). Genetic Adaptation and Neandertal Admixture Shaped the Immune System of Human Populations. Cell *167*, 643–656.e17. https://doi.org/10.1016/j.cell.2016.09.024.
9. Domínguez-Andrés, J., Kuijpers, Y., Bakker, O.B., Jaeger, M., Xu, C.-J., Van der Meer, J.W., Jakobsson, M., Bertranpetit, J., Joosten, L.A., Li, Y., et al. (2021). Evolution of cytokine production capacity in ancient and modern European populations. eLife *10*, e64971. https://doi.org/10.7554/eLife.64971.
10. Faure, E., and Royer-Carenzi, M. (2008). Is the European spatial distribution of the HIV-1-resistant CCR5-Δ32 allele formed by a breakdown of the pathocenosis due to the historical Roman expansion? Infect. Genet. Evol. *8*, 864–874. https://doi.org/10.1016/j.meegid.2008.08.007.
11. Libert, F., Cochaux, P., Beckman, G., Samson, M., Aksenova, M., Cao, A., Czeizel, A., Claustres, M., De La Rúa, C., Ferrari, M., et al. (1998). The Δccr5 mutation conferring protection against HIV-1 in Caucasian populations has a single and recent origin in Northeastern Europe. Hum. Mol. Genet. *7*, 399–406. https://doi.org/10.1093/hmg/7.3.399.
12. Novembre, J., Galvani, A.P., and Slatkin, M. (2005). The geographic spread of the CCR5 Δ32 HIV-resistance allele. PLoS Biol. *3*, e339. https://doi.org/10.1371/journal.pbio.0030339.
13. Galvani, A.P., and Novembre, J. (2005). The evolutionary history of the CCR5-Δ32 HIV-resistance mutation. Microbes Infect. *7*, 302–309. https://doi.org/10.1016/j.micinf.2004.12.006.
14. Galvani, A.P., and Slatkin, M. (2003). Evaluating plague and smallpox as historical selective pressures for the CCR5-Δ32 HIV-resistance allele. Proc. Natl. Acad. Sci. USA *100*, 15276–15279. https://doi.org/10.1073/pnas.2435085100.
15. Sabeti, P.C., Walsh, E., Schaffner, S.F., Varilly, P., Fry, B., Hutcheson, H.B., Cullen, M., Mikkelsen, T.S., Roy, J., Patterson, N., et al. (2005). The Case for Selection at Ccr5-Δ32. PLoS Biol. *3*, e378. https://doi.org/10.1371/journal.pbio.0030378.

16. Duncan, S.R., Scott, S., and Duncan, C.J. (2005). Reappraisal of the historical selective pressures for the CCR5-Δ32 mutation. J. Med. Genet. *42*, 205–208. https://doi.org/10.1136/jmg.2004.025346.

17. Stephens, J.C., Reich, D.E., Goldstein, D.B., Shin, H.D., Smith, M.W., Carrington, M., Winkler, C., Huttley, G.A., Allikmets, R., Schriml, L., et al. (1998). Dating the origin of the CCR5-Delta32 AIDS-resistance allele by the coalescence of haplotypes. Am. J. Hum. Genet. *62*, 1507–1515. https://doi.org/10.1086/301867.

18. Oppermann, M. (2004). Chemokine receptor CCR5: insights into structure, function, and regulation. Cell. Signal. *16*, 1201–1210. https://doi.org/10.1016/j.cellsig.2004.04.007.

19. Hughes, C.E., and Nibbs, R.J.B. (2018). A guide to chemokines and their receptors. FEBS Journal *285*, 2944–2971. https://doi.org/10.1111/febs.14466.

20. Ellwanger, J.H., Kaminski, V. de L., Rodrigues, A.G., Kulmann-Leal, B., and Chies, J.A.B. (2020). CCR5 and CCR5Δ32 in bacterial and parasitic infections: Thinking chemokine receptors outside the HIV box. Int. J. Immunogenet. *47*, 261–285. https://doi.org/10.1111/iji.12485.

21. Deng, H., Liu, R., Ellmeier, W., Choe, S., Unutmaz, D., Burkhart, M., Di Marzio, P.D., Marmon, S., Sutton, R.E., Hill, C.M., et al. (1996). Identification of a major co-receptor for primary isolates of HIV-1. Nature *381*, 661–666. https://doi.org/10.1038/381661a0.

22. Dragic, T., Litwin, V., Allaway, G.P., Martin, S.R., Huang, Y., Nagashima, K.A., Cayanan, C., Maddon, P.J., Koup, R.A., Moore, J.P., et al. (1996). HIV-1 entry into CD4+ cells is mediated by the chemokine receptor CC-CKR-5. Nature *381*, 667–673. https://doi.org/10.1038/381667a0.

23. Samson, M., Libert, F., Doranz, B.J., Rucker, J., Liesnard, C., Farber, C. M., Saragosti, S., Lapoumeroulie, C., Cognaux, J., Forceille, C., et al. (1996). Resistance to Hiv-1 Infection in Caucasian Individuals Bearing Mutant Alleles of the Ccr-5 Chemokine Receptor Gene. Nature *382*, 722–725. https://doi.org/10.1038/382722a0.

24. Latinovic, O.S., Reitz, M., and Heredia, A. (2019). CCR5 Inhibitors and HIV-1 Infection. J. Aids HIV Treat. *1*, 1–5. https://doi.org/10.33696/AIDS.1.001.

25. Mohamed, H., Gurrola, T., Berman, R., Collins, M., Sariyer, I.K., Nonnemacher, M.R., and Wigdahl, B. (2021). Targeting CCR5 as a Component of an HIV-1 Therapeutic Strategy. Front. Immunol. *12*, 816515. https://doi.org/10.3389/fimmu.2021.816515.

26. Hsu, J., Van Besien, K., Glesby, M.J., Pahwa, S., Coletti, A., Warshaw, M. G., Petz, L., Moore, T.B., Chen, Y.H., Pallikkuth, S., et al. (2023). HIV-1 remission and possible cure in a woman after haplo-cord blood transplant. Cell *186*, 1115–1126.e8. https://doi.org/10.1016/j.cell.2023.02.030.

27. Hütter, G., Nowak, D., Mossner, M., Ganepola, S., Müssig, A., Allers, K., Schneider, T., Hofmann, J., Kücherer, C., Blau, O., et al. (2009). Long-term control of HIV by CCR5 Delta32/Delta32 stem-cell transplantation. N. Engl. J. Med. *360*, 692–698. https://doi.org/10.1056/NEJMoa0802905.

28. Gupta, R.K., Abdul-Jawad, S., McCoy, L.E., Mok, H.P., Peppa, D., Salgado, M., Martinez-Picado, J., Nijhuis, M., Wensing, A.M.J., Lee, H., et al. (2019). HIV-1 remission following CCR5Δ32/Δ32 haematopoietic stem-cell transplantation. Nature *568*, 244–248. https://doi.org/10.1038/s41586-019-1027-4.

29. Ellwanger, J.H., Kulmann-Leal, B., Kaminski, V. de L., Rodrigues, A.G., Bragatte, M.A. de S., and Chies, J.A.B. (2020). Beyond HIV infection: Neglected and varied impacts of CCR5 and CCR5Δ32 on viral diseases. Virus Res. *286*, 198040. https://doi.org/10.1016/j.virusres.2020.198040.

30. Mehlotra, R.K. (2020). New Knowledge About CCR5, HIV Infection, and Disease Progression: Is "Old" Still Valuable? AIDS Res. Hum. Retroviruses *36*, 795–799. https://doi.org/10.1089/AID.2020.0060.

31. Rautenbach, A., and Williams, A.A. (2020). Metabolomics as an Approach to Characterise the Contrasting Roles of CCR5 in the Presence and Absence of Disease. Int. J. Mol. Sci. *21*, 1472. https://doi.org/10.3390/ijms21041472.

32. Weissberg, O., Gorohovski, A., Shay, D.R., and Frenkel-Morgenstern, M. (2021). Significant Effects of CCR5delta32 Polymorphism on Alzheimer'S Disease, Neurological Disorders, Cancer, Diabetes and Viral Infection in the Worldwide Population. Am. J. Biomed. Sci. Res. *13*, 177.

33. Bernas, S.N., Baldauf, H., Wendler, S., Heidenreich, F., Lange, V., Hofmann, J.A., Sauter, J., Schmidt, A.H., and Schetelig, J. (2021). CCR5Δ32 mutations do not determine COVID-19 disease course. Int. J. Infect. Dis. *105*, 653–655. https://doi.org/10.1016/j.ijid.2021.02.108.

34. Chua, R.L., Lukassen, S., Trump, S., Hennig, B.P., Wendisch, D., Pott, F., Debnath, O., Thürmann, L., Kurth, F., Völker, M.T., et al. (2020). COVID-19 severity correlates with airway epithelium–immune cell interactions identified by single-cell analysis. Nat. Biotechnol. *38*, 970–979. https://doi.org/10.1038/s41587-020-0602-4.

35. Cuesta-Llavona, E., Gómez, J., Albaiceta, G.M., Amado-Rodríguez, L., García-Clemente, M., Gutiérrez-Rodríguez, J., López-Alonso, I., Hermida, T., Enríquez, A.I., Hernández-González, C., et al. (2021). Variant-genetic and transcript-expression analysis showed a role for the chemokine-receptor CCR5 in COVID-19 severity. Int. Immunopharmacol. *98*, 107825. https://doi.org/10.1016/j.intimp.2021.107825.

36. Gómez, J., Cuesta-Llavona, E., Albaiceta, G.M., García-Clemente, M., López-Larrea, C., Amado-Rodríguez, L., López-Alonso, I., Hermida, T., EnrÍquez, A.I., Gil, H., et al. (2020). The CCR5-delta32 variant might explain part of the association between COVID-19 and the chemokine-receptor gene cluster. Preprint at medRxiv. https://doi.org/10.1101/2020.11.02.20224659.

37. Hubacek, J.A., Dusek, L., Majek, O., Adamek, V., Cervinkova, T., Dlouha, D., Pavel, J., and Adamkova, V. (2021). CCR5Delta32 deletion as a protective factor in Czech first-wave COVID-19 subjects. Physiol. Res. *70*, 111–115. https://doi.org/10.33549/physiolres.934647.

38. Panda, A.K., Padhi, A., and Prusty, B.A.K. (2020). CCR5 Δ32 minorallele is associated with susceptibility to SARS-CoV-2 infection and death: An epidemiological investigation. Clin. Chim. Acta *510*, 60–61. https://doi.org/10.1016/j.cca.2020.07.012.

39. Patterson, B.K., Seethamraju, H., Dhody, K., Corley, M.J., Kazempour, K., Lalezari, J., Pang, A.P.S., Sugai, C., Mahyari, E., Francisco, E.B., et al. (2021). CCR5 inhibition in critical COVID-19 patients decreases inflammatory cytokines, increases CD8 T-cells, and decreases SARS-CoV2 RNA in plasma by day 14. Int. J. Infect Dis. *103*, 25–32. https://doi.org/10.1016/j.ijid.2020.10.101.

40. Starcevic Cizmarevic, N., Kapovic, M., Roncevic, D., and Ristic, S. (2021). Could the CCR5-Delta32 mutation be protective in SARS-CoV-2 infection? Physiol. Res. *70*, S249–S252. https://doi.org/10.33549/physiolres.934725.

41. Hachim, M.Y., Hachim, I.Y., Naeem, K.B., Hannawi, H., Al Salmi, I., and Hannawi, S. (2020). C-C chemokine receptor type 5 links COVID-19, rheumatoid arthritis, and Hydroxychloroquine: in silico analysis. Transl. Med. Commun. *5*, 14. https://doi.org/10.1186/s41231-020-00066-x.

42. Xu, M. (2020). CCR5-Δ32 biology, gene editing, and warnings for the future of CRISPR-Cas9 as a human and humane gene editing tool. Cell Biosci. *10*, 48. https://doi.org/10.1186/s13578-020-00410-6.

43. Pieczynski, J.N., and Kee, H.L. (2021). "Designer babies?!" A CRISPR-based learning module for undergraduates built around the CCR5 gene. Biochem. Mol. Biol. Educ. *49*, 80–93. https://doi.org/10.1002/bmb.21395.

44. Li, T., and Shen, X. (2019). Pleiotropy Complicates Human Gene Editing: CCR5Δ32 and Beyond. Front. Genet. *10*, 669. https://doi.org/10.3389/fgene.2019.00669.

45. (2019). Act now on CRISPR babies. Nature *570*, 137. https://doi.org/10.1038/d41586-019-01786-3.

46. Bouwman, A., Shved, N., Akgül, G., Rühli, F., and Warinner, C. (2017). Ancient DNA investigation of a medieval German cemetery Confirms long-term stability of CCR5-Δ32 allele frequencies in central Europe. Hum. Biol. *89*, 119–124. https://doi.org/10.13110/humanbiology.89.2.02.

47. Vargas, A.E., Cechim, G., Correa, J.F., Gomes, P.A., Macedo, G. de S., de Medeiros, R.M., Perotoni, G., Rauber, R., Villodre, E.S., and Chies, J.A.B. (2009). Pros and cons of a missing chemokine receptor–comments on "Is the European spatial distribution of the HIV-1-resistant CCR5-D32 allele formed by a breakdown of the pathocenosis due to the historical Roman expansion?" by Eric Faure and Manuela Royer-Carenzi (2008). Infect. Genet. Evol. *9*, 387–389. https://doi.org/10.1016/j.meegid.2009.01.001.

48. Lidén, K., Linderholm, A., and Götherström, A. (2006). Pushing it back. Dating the CCR5-Δ 32 bp deletion to the Mesolithic in Sweden and its implications for the Meso Neo transition. Doc. Praehist. *633*, 29–37. https://doi.org/10.4312/dp.33.5.

49. Hummel, S., Schmidt, D., Kremeyer, B., Herrmann, B., and Oppermann, M. (2005). Detection of the CCR5-Δ32 HIV resistance gene in Bronze Age skeletons. Genes Immun. *6*, 371–374. https://doi.org/10.1038/sj.gene.6364172.

50. Lucotte, G. (2001). Distribution of the CCR5 gene 32-basepair deletion in west Europe. A hypothesis about the possible dispersion of the mutation by the vikings in historical times. Hum. Immunol. *62*, 933–936. https://doi.org/10.1016/S0198-8859(01)00292-0.

51. Lucotte, G., and Dieterlen, F. (2003). More about the Viking hypothesis of origin of the Δ32 mutation in the CCR5 gene conferring resistance to HIV-1 infection. Infect. Genet. Evol. *3*, 293–295. https://doi.org/10.1016/j.meegid.2003.07.001.

52. Silva-Carvalho, W.H.V., de Moura, R.R., Coelho, A.V.C., Crovella, S., and Guimarães, R.L. (2016). Frequency of the CCR5-delta32 allele in Brazilian populations: A systematic literature review and meta-analysis. Infect. Genet. Evol. *43*, 101–107. https://doi.org/10.1016/j.meegid.2016.05.024.

53. Hedrick, P.W., and Verrelli, B.C. (2006). "Ground truth" for selection on CCR5-Delta32. Trends Genet. *22*, 293–296. https://doi.org/10.1016/j.tig.2006.04.007.

54. Mecsas, J., Franklin, G., Kuziel, W.A., Brubaker, R.R., Falkow, S., and Mosier, D.E. (2004). Evolutionary genetics: CCR5 mutation and plague protection. Nature *427*, 606. https://doi.org/10.1038/427606a.

55. Solloch, U.V., Lang, K., Lange, V., Böhme, I., Schmidt, A.H., and Sauter, J. (2017). Frequencies of gene variant CCR5-Δ32 in 87 countries based on next-generation sequencing of 1.3 million individuals sampled from 3 national DKMS donor centers. Hum. Immunol. *78*, 710–717. https://doi.org/10.1016/j.humimm.2017.10.001.

56. Buhler, M.M., Craig, M., Donaghue, K.C., Badhwar, P., Willis, J., Manolios, N., Tait, B.D., Silink, M., Bennetts, B.H., and Stewart, G.J. (2002). CCR5 genotyping in an Australian and New Zealand type 1 diabetes cohort. Autoimmunity *35*, 457–461. https://doi.org/10.1080/0891693021000041088.

57. Fahrioglu, U., Ergoren, M.C., and Mocan, G. (2020). CCR5-Δ32 gene variant frequency in the Turkish Cypriot population. Braz. J. Microbiol. *51*, 1711–1717. https://doi.org/10.1007/s42770-020-00352-8.

58. Kulmann-Leal, B., Ellwanger, J.H., and Chies, J.A.B. (2021). CCR5Δ32 in Brazil: Impacts of a European Genetic Variant on a Highly Admixed Population. Front. Immunol. *12*, 758358. https://doi.org/10.3389/fimmu.2021.758358.

59. Bhatnagar, I., Singh, M., Mishra, N., Saxena, R., Thangaraj, K., Singh, L., and Saxena, S. (2009). The Latitude Wise Prevalence of the CCR5-Δ32-HIV Resistance Allele in India. Balk. J. Med. Genet. *12*, 17–27. https://doi.org/10.2478/v10034-010-0001-0.

60. Bamshad, M.J., Mummidi, S., Gonzalez, E., Ahuja, S.S., Dunn, D.M., Watkins, W.S., Wooding, S., Stone, A.C., Jorde, L.B., Weiss, R.B., et al. (2002). A strong signature of balancing selection in the 5′ cis-regulatory region of CCR5. Proc. Natl. Acad. Sci. USA *99*, 10539–10544. https://doi.org/10.1073/pnas.162046399.

61. Le, M.K., Smith, O.S., Akbari, A., Harpak, A., Reich, D., and Narasimhan, V.M. (2022). 1,000 ancient genomes uncover 10,000 years of natural selection in Europe. Preprint at bioRxiv. https://doi.org/10.1101/2022.08.24.505188.

62. Martiniano, R., Garrison, E., Jones, E.R., Manica, A., and Durbin, R. (2020). Removing reference bias and improving indel calling in ancient DNA data analysis by mapping to a sequence variation graph. Genome Biol. *21*, 250. https://doi.org/10.1186/s13059-020-02160-7.

63. Zawicki, P., and Witas, H.W. (2008). HIV-1 protecting CCR5-Δ32 allele in medieval Poland. Infect. Genet. Evol. *8*, 146–151. https://doi.org/10.1016/j.meegid.2007.11.003.

64. Orlando, L., Allaby, R., Skoglund, P., Der Sarkissian, C., Stockhammer, P.W., Ávila-Arcos, M.C., Fu, Q., Krause, J., Willerslev, E., Stone, A.C., et al. (2021). Ancient DNA analysis. Nat. Rev. Methods Primer *1*, 14. https://doi.org/10.1038/s43586-020-00011-0.

65. Auton, A., Brooks, L.D., Durbin, R.M., Garrison, E.P., Kang, H.M., Korbel, J.O., Marchini, J.L., McCarthy, S., McVean, G.A., et al.; 1000 Genomes Project Consortium (2015). A global reference for human genetic variation. Nature *526*, 68–74. https://doi.org/10.1038/nature15393.

66. McLaren, W., Gil, L., Hunt, S.E., Riat, H.S., Ritchie, G.R.S., Thormann, A., Flicek, P., and Cunningham, F. (2016). The Ensembl Variant Effect Predictor. Genome Biol. *17*, 122. https://doi.org/10.1186/s13059-016-0974-4.

67. Landrum, M.J., Lee, J.M., Riley, G.R., Jang, W., Rubinstein, W.S., Church, D.M., and Maglott, D.R. (2014). ClinVar: public archive of relationships among sequence variation and human phenotype. Nucleic Acids Res. *42*, D980–D985. https://doi.org/10.1093/nar/gkt1113.

68. Buniello, A., MacArthur, J.A.L., Cerezo, M., Harris, L.W., Hayhurst, J., Malangone, C., McMahon, A., Morales, J., Mountjoy, E., Sollis, E., et al. (2019). The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. Nucleic Acids Res. *47*, D1005–D1012. https://doi.org/10.1093/nar/gky1120.

69. Kauwe, J.S.K., Bailey, M.H., Ridge, P.G., Perry, R., Wadsworth, M.E., Hoyt, K.L., Staley, L.A., Karch, C.M., Harari, O., Cruchaga, C., et al. (2014). Genome-wide association study of CSF levels of 59 alzheimer's disease candidate proteins: significant associations with proteins involved in amyloid processing and inflammation. PLoS Genet. *10*, e1004758. https://doi.org/10.1371/journal.pgen.1004758.

70. Liu, J.Z., van Sommeren, S., Huang, H., Ng, S.C., Alberts, R., Takahashi, A., Ripke, S., Lee, J.C., Jostins, L., Shah, T., et al. (2015). Association analyses identify 38 susceptibility loci for inflammatory bowel disease and highlight shared genetic risk across populations. Nat. Genet. *47*, 979–986. https://doi.org/10.1038/ng.3359.

71. Vistnes, M., Tapia, G., Mårild, K., Midttun, Ø., Ueland, P.M., Viken, M.K., Magnus, P., Berg, J.P., Gillespie, K.M., Skrivarhaug, T., et al. (2018). Plasma immunological markers in pregnancy and cord blood: A possible link between macrophage chemo-attractants and risk of childhood type 1 diabetes. Am. J. Reprod. Immunol. *79*, e12802. https://doi.org/10.1111/aji.12802.

72. Kamat, M.A., Blackshaw, J.A., Young, R., Surendran, P., Burgess, S., Danesh, J., Butterworth, A.S., and Staley, J.R. (2019). PhenoScanner V2: an expanded tool for searching human genotype-phenotype associations. Bioinformatics *35*, 4851–4853. https://doi.org/10.1093/bioinformatics/btz469.

73. Staley, J.R., Blackshaw, J., Kamat, M.A., Ellis, S., Surendran, P., Sun, B.B., Paul, D.S., Freitag, D., Burgess, S., Danesh, J., et al. (2016). PhenoScanner: a database of human genotype-phenotype associations. Bioinformatics *32*, 3207–3209. https://doi.org/10.1093/bioinformatics/btw373.

74. Meisner, J., and Albrechtsen, A. (2022). Haplotype and population structure inference using neural networks in whole-genome sequencing data. Genome Res. *32*, 1542–1552. https://doi.org/10.1101/gr.276813.122.

75. McVean, G. (2007). The Structure of Linkage Disequilibrium Around a Selective Sweep. Genetics *175*, 1395–1406. https://doi.org/10.1534/genetics.106.062828.

76. Slatkin, M. (2008). Linkage disequilibrium — understanding the evolutionary past and mapping the medical future. Nat. Rev. Genet. *9*, 477–485. https://doi.org/10.1038/nrg2361.

77. Van der Auwera, G.A., and O'Connor, B. (2020). Genomics in the Cloud: Using Docker, GATK, and WDL in Terra, First Edition (O'Reilly Media).

78. Garrison, E., Sirén, J., Novak, A.M., Hickey, G., Eizenga, J.M., Dawson, E.T., Jones, W., Garg, S., Markello, C., Lin, M.F., et al. (2018). Variation graph toolkit improves read mapping by representing genetic variation in the reference. Nat. Biotechnol. *36*, 875–879. https://doi.org/10.1038/nbt.4227.

79. Allentoft, M.E., Sikora, M., Refoyo-Martínez, A., Irving-Pease, E.K., Fischer, A., Barrie, W., Ingason, A., Stenderup, J., Sjögren, K.-G., Pearson, A., et al. (2024). Population genomics of post-glacial western Eurasia. Nature *625*, 301–311. https://doi.org/10.1038/s41586-023-06865-0.

80. Margaryan, A., Lawson, D.J., Sikora, M., Racimo, F., Rasmussen, S., Moltke, I., Cassidy, L.M., Jørsboe, E., Ingason, A., Pedersen, M.W., et al. (2020). Population genomics of the Viking world. Nature *585*, 390–396. https://doi.org/10.1038/s41586-020-2688-8.

81. de Barros Damgaard, P., Martiniano, R., Kamm, J., Moreno-Mayar, J.V., Kroonen, G., Peyrot, M., Barjamovic, G., Rasmussen, S., Zacho, C., Baimukhanov, N., et al. (2018). The first horse herders and the impact of early Bronze Age steppe expansions into Asia. Science *360*, eaar7711. https://doi.org/10.1126/science.aar7711.

82. Allentoft, M.E., Sikora, M., Sjögren, K.-G., Rasmussen, S., Rasmussen, M., Stenderup, J., Damgaard, P.B., Schroeder, H., Ahlström, T., Vinner, L., et al. (2015). Population genomics of Bronze Age Eurasia. Nature *522*, 167–172. https://doi.org/10.1038/nature14507.

83. Stern, A.J., Wilton, P.R., and Nielsen, R. (2019). An approximate full-likelihood method for inferring selection and allele frequency trajectories from DNA sequence data. PLoS Genet. *15*, e1008384. https://doi.org/10.1371/journal.pgen.1008384.

84. Irving-Pease, E.K., Refoyo-Martínez, A., Barrie, W., Ingason, A., Pearson, A., Fischer, A., Sjögren, K.-G., Halgren, A.S., Macleod, R., Demeter, F., et al. (2024). The selection landscape and genetic legacy of ancient Eurasians. Nature *625*, 312–320. https://doi.org/10.1038/s41586-023-06705-1.

85. Muktupavela, R.A., Petr, M., Ségurel, L., Korneliussen, T., Novembre, J., and Racimo, F. (2022). Modeling the spatiotemporal spread of beneficial alleles using ancient genomes. eLife *11*, e73767. https://doi.org/10.7554/eLife.73767.

86. Jordan, I.K. (2016). The Columbian Exchange as a source of adaptive introgression in human populations. Biol. Direct *11*, 17. https://doi.org/10.1186/s13062-016-0121-x.

87. Norris, E.T., Wang, L., Conley, A.B., Rishishwar, L., Mariño-Ramírez, L., Valderrama-Aguirre, A., and Jordan, I.K. (2018). Genetic ancestry, admixture and health determinants in Latin America. BMC Genomics *19*, 861. https://doi.org/10.1186/s12864-018-5195-7.

88. Haak, W., Lazaridis, I., Patterson, N., Rohland, N., Mallick, S., Llamas, B., Brandt, G., Nordenfelt, S., Harney, E., Stewardson, K., et al. (2015). Massive migration from the steppe was a source for Indo-European languages in Europe. Nature *522*, 207–211. https://doi.org/10.1038/nature14317.

89. Novembre, J. (2015). Human evolution: Ancient DNA steps into the language debate. Nature *522*, 164–165. https://doi.org/10.1038/522164a.

90. Brinkworth, J.F. (2017). Infectious Disease and the Diversification of the Human Genome. Hum. Biol. *89*, 47–65. https://doi.org/10.13110/humanbiology.89.1.03.

91. Dyer, D.P., Medina-Ruiz, L., Bartolini, R., Schuette, F., Hughes, C.E., Pallas, K., Vidler, F., Macleod, M.K.L., Kelly, C.J., Lee, K.M., et al. (2019). Chemokine Receptor Redundancy and Specificity Are Context Dependent. Immunity *50*, 378–389.e5. https://doi.org/10.1016/j.immuni.2019.01.009.

92. Allers, K., and Schneider, T. (2015). CCR5Δ32 mutation and HIV infection: basis for curative HIV therapy. Curr. Opin. Virol. *14*, 24–29. https://doi.org/10.1016/j.coviro.2015.06.007.

93. Gupta, R.K., Peppa, D., Hill, A.L., Gálvez, C., Salgado, M., Pace, M., McCoy, L.E., Griffith, S.A., Thornhill, J., Alrubayyi, A., et al. (2020). Evidence for HIV-1 cure after CCR5Δ32/Δ32 allogeneic haemopoietic stem-cell transplantation 30 months post analytical treatment interruption: a case report. Lancet HIV *7*, e340–e347. https://doi.org/10.1016/S2352-3018(20)30069-2.

94. Jasinska, A.J., Pandrea, I., and Apetrei, C. (2022). CCR5 as a Coreceptor for Human Immunodeficiency Virus and Simian Immunodeficiency Viruses: A Prototypic Love-Hate Affair. Front. Immunol. *13*, 835994. https://doi.org/10.3389/fimmu.2022.835994.

95. Ellwanger, J.H., Kaminski, V. de L., and Chies, J.A. (2020). What we say and what we mean when we say redundancy and robustness of the chemokine system - how CCR5 challenges these concepts. Immunol. Cell Biol. *98*, 22–27. https://doi.org/10.1111/imcb.12291.

96. Kleist, A.B., Getschman, A.E., Ziarek, J.J., Nevins, A.M., Gauthier, P.-A., Chevigné, A., Szpakowska, M., and Volkman, B.F. (2016). New paradigms in chemokine receptor signal transduction: moving beyond the two-site model. Biochem. Pharmacol. *114*, 53–68. https://doi.org/10.1016/j.bcp.2016.04.007.

97. R Core Team (2020). R: A Language and Environment for Statistical Computing (R Foundation for Statistical Computing). https://www.R-project.org/.

98. Machiela, M.J., and Chanock, S.J. (2015). LDlink: a web-based application for exploring population-specific haplotype structure and linking correlated alleles of possible functional variants. Bioinformatics *31*, 3555–3557. https://doi.org/10.1093/bioinformatics/btv402.

99. Machiela, M.J., and Chanock, S.J. (2018). LDassoc: an online tool for interactively exploring genome-wide association study results and prioritizing variants for functional investigation. Bioinformatics *34*, 887–889. https://doi.org/10.1093/bioinformatics/btx561.

100. Renaud, G., Hanghøj, K., Willerslev, E., and Orlando, L. (2017). Gargammel: A sequence simulator for ancient DNA. Bioinformatics *33*, 577–579. https://doi.org/10.1093/bioinformatics/btw670.

101. Schubert, M., Lindgreen, S., and Orlando, L. (2016). AdapterRemoval v2: rapid adapter trimming, identification, and read merging. BMC Res. Notes *9*, 88. https://doi.org/10.1186/s13104-016-1900-2.

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

103. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R.; 1000 Genome Project Data Processing Subgroup (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics *25*, 2078–2079. https://doi.org/10.1093/bioinformatics/btp352.

104. Broad Institute. (2019). Picard toolkit. GitHub. http://broadinstitute.github.io/picard/.

105. Mölder, F., Jablonski, K.P., Letcher, B., Hall, M.B., Tomkins-Tinch, C.H., Sochat, V., Forster, J., Lee, S., Twardziok, S.O., Kanitz, A., et al. (2021). Sustainable data analysis with Snakemake. F1000Res *10*, 33. https://doi.org/10.12688/f1000research.29032.2.

106. Kopanos, C., Tsiolkas, V., Kouris, A., Chapple, C.E., Albarca Aguilera, M., Meyer, R., and Massouras, A. (2019). VarSome: the human genomic variant search engine. Bioinformatics *35*, 1978–1980. https://doi.org/10.1093/bioinformatics/bty897.

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

108. Speidel, L., Forest, M., Shi, S., and Myers, S.R. (2019). A method for genome-wide genealogy estimation for thousands of samples. Nat. Genet. *51*, 1321–1329. https://doi.org/10.1038/s41588-019-0484-x.

109. Sirén, J., Monlong, J., Chang, X., Novak, A.M., Eizenga, J.M., Markello, C., Sibbesen, J.A., Hickey, G., Chang, P.-C., Carroll, A., et al. (2021).

Pangenomics enables genotyping of known structural variants in 5202 diverse genomes. Science *374*, abg8871. https://doi.org/10.1126/science.abg8871.

110. Hickey, G., Heller, D., Monlong, J., Sibbesen, J.A., Sirén, J., Eizenga, J., Dawson, E.T., Garrison, E., Novak, A.M., and Paten, B. (2020). Genotyping structural variants in pangenome graphs using the vg toolkit. Genome Biol. *21*, 35. https://doi.org/10.1186/s13059-020-1941-7.

111. Li, H. (2011). A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics *27*, 2987–2993. https://doi.org/10.1093/bioinformatics/btr509.

112. Schubert, M., Ginolhac, A., Lindgreen, S., Thompson, J.F., Al-Rasheid, K.A.S., Willerslev, E., Krogh, A., and Orlando, L. (2012). Improving ancient DNA read mapping against modern reference genomes. BMC Genomics *13*, 178. https://doi.org/10.1186/1471-2164-13-178.

113. Dabney, J., Meyer, M., and Pääbo, S. (2013). Ancient DNA damage. Cold Spring Harb. Perspect. Biol. *5*, a012567. https://doi.org/10.1101/cshperspect.a012567.

114. Knaus, B.J., and Grünwald, N.J. (2017). vcfr: a package to manipulate and visualize variant call format data in R. Mol. Mol. Ecol. Resour. *17*, 44–53. https://doi.org/10.1111/1755-0998.12549.

115. Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A.R., Bender, D., Maller, J., Sklar, P., de Bakker, P.I.W., Daly, M.J., et al. (2007). PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. Am. J. Hum. Genet. *81*, 559–575. https://doi.org/10.1086/519795.

116. McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., et al. (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. *20*, 1297–1303. https://doi.org/10.1101/gr.107524.110.

117. Irving-Pease, E.K., Refoyo-Martínez, A., Ingason, A., Pearson, A., Fischer, A., Barrie, W., Sjögren, K.-G., Halgren, A.S., Macleod, R., Demeter, F., et al. (2024). The Selection Landscape and Genetic Legacy of Ancient Eurasians. Nature *625*, 312–320. https://doi.org/10.1038/s41586-023-06705-1.

118. Bélisle, C.J.P. (1992). Convergence Theorems for a Class of Simulated Annealing Algorithms on Rd. J. Appl. Probab. *29*, 885–895. https://doi.org/10.2307/3214721.

119. Byrd, R.H., Lu, P., Nocedal, J., and Zhu, C. (1994). A LIMITED MEMORY ALGORITHM FOR BOUND CONSTRAINED OPTIMIZATION. SIAM J. Sci. Comput. *16*, 1190–1208.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
Human reference genome NCBI build 37, GRCh37	Genome Reference Consortium	https://www.ncbi.nlm.nih.gov/grc/human
Modern dataset (1000 Genomes Project Phase 3)	1000 Genomes Project Consortium et al.[65]	ftp.1000genomes.ebi.ac.uk/vol1/ftp/phase3/data/
Ancient dataset	Allentoft et al., 2024,[79] 2015[82]; Margaryan et al.[80]; de Barros Damgaard et al.[81]	The Neolithic Age dataset: https://www.ebi.ac.uk/ena/browser/view/PRJEB64656 Bronze Age dataset: https://www.ebi.ac.uk/ena/browser/text-search?query=PRJEB9021 Viking Age dataset: https://www.ebi.ac.uk/ena/browser/view/PRJEB37976 Botai dataset: https://www.ebi.ac.uk/ena/browser/view/PRJEB26349
Software and algorithms		
R (v.4.0.3)	R Core Team[97]	https://www.r-project.org/
LDLink 3.0 web tool	Machiela and Chanock, 2015,[98] 2018[99]	https://ldlink.nci.nih.gov/
Ensembl Variant Effect Predictor (VEP)	McLaren et al.[66]	https://www.ensembl.org/info/docs/tools/vep/index.html
GWAS catalog	Buniello et al.[68]	https://www.ebi.ac.uk/gwas/
PhenoScanner V2	Kamat et al.[72] and Staley et al.[73]	https://github.com/phenoscanner
gargammel (v.1.1.2)	Renaud et al.[100]	https://github.com/grenaud/gargammel
FastaAlternateReferenceMaker, GATK (v.4.1.8.1)	Van der Auwera and O'Connor[77]	https://gatk.broadinstitute.org/hc/en-us
AdapterRemoval (v.2.1.3)	Schubert et al.[101]	https://github.com/MikkelSchubert/adapterremoval
bwa aln (v.7.16a)	Li and Durbin[102]	https://github.com/lh3/bwa
samtools (v.1.9)	Li et al.[103]	https://samtools.sourceforge.net/
Picard MarkDuplicates (v1.128)	Broad Institute[104]	https://broadinstitute.github.io/picard/
GATK (v.3.3.0)	Van der Auwera and O'Connor[77]	https://gatk.broadinstitute.org/hc/en-us
Snakemake (v.5.12.0)	Mölder et al.[105]	https://github.com/snakemake/snakemake
VarSome	Kopanos et al.[106]	https://varsome.com/
pysam (v.0.16.0.1)	Danecek et al.[107]	https://github.com/pysam-developers/pysam
HaplotypeCaller from GATK (v.4.1.9.0)	Van der Auwera and O'Connor[77]	https://gatk.broadinstitute.org/hc/en-us
bcftools norm (v.1.10.2)	Danecek et al.[107]	https://samtools.github.io/bcftools/bcftools.html
CLUES	Stern et al.[83]; Irving-Pease et al.[84]	https://github.com/standard-aaron/clues
Relate (v.1.1)	Speidel et al.[108]	https://myersgroup.github.io/relate/
Haplonet	Meisner and Albrechtsen[74]	https://github.com/Rosemeis/HaploNet
vg (1.53.0)	Garrison et al.[78]; Sirén et al.[109]; Hickey et al.[110]	https://github.com/vgteam/vg
Other		
HAPI analysis	This paper	https://github.com/RasmussenLab/HAPI
CLUES analysis	This paper	https://github.com/ekirving/ccr5_paper
Spatiotemporal analysis	This paper	https://github.com/RasaMukti/ccr5delta32_analysis

EXPERIMENTAL MODEL AND SUBJECT DETAILS

All ancient human genomes analyzed in this study were unearthed and sequenced in four previous papers.[79–82] All details can be found in Table S4. All modern genomes were obtained from the 1000 Genomes Project Phase 3 (1KGP3, https://www.internationalgenome.org/). No experimental models were used.

METHOD DETAILS

Data

The modern dataset constituted of whole-genome sequencing data of 2,504 individuals from 26 populations which were generated by the 1000 Genomes Project Phase 3 (1KGP3, https://www.internationalgenome.org/), assigned to the following 5 super populations: African (AFR), admixed from the Americas (AMR), East Asian (EAS), South Asian (SAS), and European (EUR).[65] The ancient dataset consisted of 934 shotgun-sequenced genomes from various regions across Eurasia, spanning the Stone Age, Bronze Age, and Viking Age in a continuous temporal sequence. The Stone Age dataset (11,000–3,000 BP) focuses on the Mesolithic and Neolithic periods, containing 317 genomes from archaeological sites across Europe, with particularly dense sampling in northern Europe, including 100 samples from Denmark (ENA Project ID: PRJEB64656).[79] The Early Bronze Age dataset (3,000–2,000 BP), referred to as BOTAI, includes 74 genomes from the Eurasian steppes (ENA Project ID: PRJEB26349),[81] alongside 101 genomes from Europe and Central Asia representing the broader Bronze Age (ENA Project ID: PRJEB9021).[82] The Viking Age dataset (2450 BP–1600 CE) comprises 442 genomes, predominantly from northern Europe, but also from archaeological sites in Greenland (ENA Project ID: PRJEB37976).[80] The geographic locations and chronological categorization of our final ancient genomes dataset, passing quality control filter,[79] represented in Table S4. The geographic locations of the genomes called by HAPI can be found in Figure S4A.

Identification of the CCR5delta32 deletion and the haplotypes

We used the LDLink 3.0 web tool, which includes the LDmatrix and LDpair modules,[98] to identify the CCR5delta32 proxy SNPs within the European (EUR) population of the 1KGP3 dataset (Table S1). All D'prime and r^2 values were obtained using LDLink 3.0. These results were then explored in additional 1KGP3 populations. The haplotypes were called with samtools mpileup[111] (Table S2) using the region (chr3:46200000-46800000) from all available 1KGP3 whole-genome bam files. To determine the effect of the 86 tag SNPs belonging to Haplotype A, we employed Ensembl Variant Effect Predictor (VEP),[66] while the GWAS catalog[68] and PhenoScanner V2[72,73] were used to evaluate possible genotype-phenotype associations of tag SNPs (Tables S1G and S1H). All annotations refer to the Human reference genome NCBI build 37, GRCh37.

Analysis of proxy variant rs73833033

We used the 1KGP3 dataset to identify samples that had called rs73833033 and used these to calculate the allele frequencies of this proxy variant in the 1KGP populations. We called rs73833033 in the ancient genomes by restricting the analysis to sequences passing the same quality filters used by Le et al.[61] Briefly, we used bcftools mpileup (v. 1.10.2)[107] with region '3:43914944-48914944', 'min-BQ 20', 'min-MQ 10' to extract the reference and alternate counts at the rs73833033's genomic position, and called the variant using and bcftools call '-m -Ob –gvcf 1'.

Simulation of ancient CCR5delta32 genomes

We used gargammel (v.1.1.2)[100] to simulate a total of 144 ancient genomes at 8 different coverages (0.3X, 0.5X, 0.8X, 1X, 1.5X, 2X, 5X, 10X), using empirical read length distributions and post-mortem damage derived from 6 real ancient genomes (NEO78, NEO79, NEO752, VK287, VK543, VK526) from our dataset (Figure S2A). We simulated 48 genomes for each genotype (RR, RD, DD) using combinations of two versions of the GRCh37 human genome reference: a canonical, and one in which we manually added the CCR5delta32 deletion and the 86 variants from the haplotype (here referred to as "collapsed reference") using the tool FastaAlternateReferenceMaker, GATK (v.4.1.8.1).[77] The reads were simulated from HiSeq 2500 Illumina single-end runs with a length of 81 base pairs including adapters.

Processing of simulated and ancient genomes

For the simulated genomes we used AdapterRemoval (v.2.1.3)[101] with parameters "–mm 3 –minlength 30 –minquality 2" to trim the reads from the simulated genomes at a length of at least 30 bp and to remove bases with quality 2 or less. We used bwa aln (v.7.16a)[102] to map the adaptor-trimmed reads to both the canonical and the collapsed human reference genome (GRCh37) with seed disabled (parameter "-l 1024") to allow for higher sensitivity in ancient DNA.[112] We sorted the resulting alignments with samtools (v.1.9),[102] removed duplicates with Picard MarkDuplicates (v.1.128)[104] and realigned the reads using GATK (v.3.3.0)[77] with Mills and 1000G gold-standard insertions and deletions. Finally, the alignment files were converted to cram with samtools view and indexed with samtools index. The ancient genomes were aligned to both the canonical and the collapsed human genome reference using the same pipeline as the simulated genomes except that the read groups were first merged to the library level, then duplicates were removed using Picard (v.1.128),[104] and then the files were merged to sample level. Sample level bam files were subsequently

realigned using GATK (v.3.3.0)[77] and then converted and indexed to cram format. The workflows were implemented using Snakemake (v.5.12.0).[105]

Processing of the 15 human genomes from the 1KGP3 dataset

We used 15 genomes from the 1KGP3 to benchmark the model, 5 for each deletion genotype: 5 RR (HG00179, HG00185, HG01500, HG01510, HG00159), 5 RD (HG00171, HG00267, HG01537, HG01605, HG00264), and 5 DD (HG00320, HG00323, HG01684, HG01762, HG00137). The reads were aligned to both the canonical and the collapsed human genome reference with the same pipeline used for the alignment of the ancient genomes.

Applying HAPI to the ancient genomes and filtering of calls

To be analyzed by the model, we applied a filter requiring genomes to have at least one read mapping to the CCR5delta32 deletion region, i.e. in either the canonical or collapsed reference, with a minimum overlapping length δ of 6. We ran the model on the simulated, 1KGP3, and ancient genomes and we classified them as being RR, RD, DD (homozygous for the reference, heterozygous, or homozygous for the deletion, respectively) based on the highest posterior probability among the three, with a classification threshold of 0.5. To take into account the fact that the flanking regions of the deletion include repeated nucleotides, and that two of the 4 variants used for calculating the haplotype-informed prior resemble ancient DNA damage, we applied two curation steps. In the first one "permissive filter" we re-classified genomes if they had only 1 SNP called, and if this same SNP appeared to be as a potential aDNA damage artifact (apparent G to A, and C to T substitutions).[113] In the second curation step "strict filter" we re-classified genomes if the mapped reads only covered the reference allele on the canonical reference, and not the CCR5delta32 deletion on the collapsed one. The scripts used to genotype the CCR5delta32 allele and that applied the different filter schemes are available at the GitHub repository (See key resources table: HAPI analysis).

Benchmarking HAPI versus GATK HaplotypeCaller and VG

The simulated 144 ancient and 15 genomes selected from the 1KGP3 population were processed using HaplotypeCaller from GATK (v.4.1.9.0)[77] to produce VCF files with SNV and indels calls using the following options: –intervals 3:46277577-46457412 –interval-padding 100 -stand-call-conf 30.0 -ERC BP_RESOLUTION. The VCF files were left aligned and normalized using bcftools norm (v.1.10.2),[107] and then processed in R (v.4.0.3)[97] with the package vcfR (v.1.15.0).[114] Similarly, we used *vg* (v.1.53.0)[78] to process the 144 simulated ancient genomes. We used *vg* giraffe[109] to map the trimmed reads to the hprc-v1.1-mc.grch38 graph genome (.gbz,.snarls) generating gam-files and then vg pack –Q 5 to compute read support from the gam. Finally, we used *vg* call -a using the pack-files as input to call variants.[110] We extracted the CCR5delta32 deletion from the VCFs and combined the variants into a single VCF using bcftools merge (v.1.10.2). The code is available at the GitHub repository (See key resources table: HAPI analysis).

Identification of seven ascertained control deletions and their haplotypes

We scanned chromosome 3 from the 1KGP3 data (ALL.chr3.phase3_shapeit2_mvncall_integrated_v5a.20130502.genotypes.vcf) for small deletions having an architecture similar to the CCR5delta32's one. The screening criteria included deletions of 32 bp $\pm$ 5 bp and a MAF of 0.1098 $\pm$ 0.02 in the EUR population. Possible haplotypes ($r^2 > 0.8$) were identified using PLINK,[115] and the deletions were further evaluated with the LDLink 3.0 web tool,[98] specifically using the LDmatrix and LDpair modules. This process led to the identification of seven deletions associated with haplotypes. We configured HAPI with parameters specific to each deletion, including proxy SNPs and overlap criteria similar to those used for CCR5delta32, and applied it to the ancient genome datasets. The edited chromosome 3 positions of the eight deletions engineered in the collapsed reference sequence are provided in Table S7. For three deletions (rs556322139, rs150628438, rs369842709), the permissive filter was equivalent to the strict filter due to the nature of their haplotypes. Further details are provided in Table S7. Scripts for classifying the different genotype filters are provided at the GitHub repository (See key resources table: HAPI analysis).

Local ancestry of individuals harboring CCR5delta32 deletion in the 1KGP3

We used HaploNet[74] on the full 1KGP3 dataset to generate haplotype cluster likelihoods in windows along the genome with default parameters of "haplonet train" besides "–x_dim 512", such that the genomic windows had a fixed size of 512 SNPs. We used the haplotype cluster likelihoods to estimate ancestry proportions with an assumption of K=5 ancestral populations, representing the 5 super populations of 1KGP3, using the "haplonet admix" command. The haplotype cluster likelihoods and ancestry proportions were then finally used to infer local ancestry for all genomic windows in the individuals with the CCR5delta32 locus using the "haplonet fatash" command.

QUANTIFICATION AND STATISTICAL ANALYSIS

Haplotype-Aware Probabilistic model for Indels (HAPI)

We developed a probabilistic model to combine the information from the 4 variants in the highest pairwise LD with the deletion (rs113341849, rs113010081, rs11574435, and rs79815064, $r^2 > 0.90$, CEU) as a Prior, and the information from the reads mapping the CCR5delta32 deletion region as a Likelihood. The variants were called using samtools mpileup as implemented in pysam

(v.0.16.0.1) in python. For a detailed overview of the algorithm see Methods S1. For each sample, we calculated the posterior probability for each deletion genotype (RR, RD, DD) as

$$P(G|D) = \frac{P(G)P(D|G)}{P(D)} \quad \text{(Equation 1)}$$

where $P(G)$ is the prior probability of the given deletion genotype calculated using the information from the 4 variants (see Equation 2 below), $P(D|G)$ is the likelihood of the deletion genotype based on the reads mapping to the CCR5delta32 deletion region (either canonical or collapsed reference) (see Equation 3 below) and $P(D)$ is the marginal probability of the data. For the prior, we calculated the posterior probability of each deletion genotype using a simple bayesian genotyper based on the one developed by McKenna et al.[116] as

$$P(G|r) = P(G)\,P(r|G)P(r) \quad \text{(Equation 2)}$$

where G is the given SNP genotype $(ref|ref, ref|alt, or\ alt|alt)$ and r is the data (the read base pileups mapping to each variant). We assume a uniform prior distribution for $P(G)$, $P(r)$ is the marginal probability of the data, and $p(r|G) = p(b|G)$, where b represents each base covering the target locus. The probability of each base given the SNP genotype, considering only alleles from the reference and deletion genotype, is defined as

$$p(b|G) = p(b|\{ref, alt\}) = \frac{1}{2}p(b|ref) + \frac{1}{2}p(b|alt) \quad \text{(Equation 3)}$$

when the genotype G is decomposed into its two alleles. For simplicity, here we assumed that a sample having the genotype RR, RD, or DD also carries each of the four variants in the SNP genotype ref|ref, ref|alt, or alt|alt, respectively. The probability of observing a base given an allele is

$$p(b|A) = \begin{cases} \frac{e}{3} : b \neq A \\ 1 - e : b = A' \end{cases} \quad \text{(Equation 4)}$$

where e is the reversed *phred* scaled quality score at the base. At this point, each of the four variants has a posterior probability $P(G|r)$ for each deletion genotype (RR, RD, DD). We scaled the posterior of each variant by the LD r^2 value it has to the deletion in the CEU population. For each deletion genotype, we calculated the prior of Equation 1 as the joint probability (calculated with the specific multiplication rule, assuming each variant to be independent for simplicity) of the posteriors of the four variants, and we finally normalized them between 0 and 1 (subtracting by max and dividing by the sum). To calculate the Likelihood, we mapped the reads of each sample against two reference genomes: the canonical, and the collapsed one. The reads mapping to the canonical and collapsed references, together with their minimum overlapping lengths δ, were used to compute the Likelihood of each deletion genotype RR, RD, DD as follows. For each of the two references, and for each read, we calculated the probability of observing the read given the specific reference with

$$p(r|R) = \begin{cases} 1 - \left(\frac{1}{\delta}\right)^2 : r = R \\ \frac{1 - \left(1 - \left(\frac{1}{\delta}\right)^2\right)}{2} : r \neq R \end{cases} \quad \text{(Equation 5)}$$

an adaptation from McKenna et al.,[116] where R is the specific reference used for the mapping, i.e. canonical or collapsed. We then calculated the probability of observing the reads given the deletion genotype with

$$p(r|G) = p(r|ref, del) = \frac{1}{2}p(r|ref) + \frac{1}{2}p(r|del) \quad \text{(Equation 6)}$$

The genotype likelihood for each reference was then calculated with $p(D|G) = p(r|G)$. The final genotype likelihood for each deletion genotype was computed as the joint probability of the likelihoods for the individual references (canonical and collapsed) with $p(D|G) = p(D|G)$. Finally, for each sample, HAPI outputs three posterior probabilities for each deletion genotype RR, RD, DD, summing up to 1.

Determining minimum overlapping length

We assigned a minimum overlapping length (δ) to each read mapping the deletion region, either on the canonical or collapsed reference. The δ represents the minimum number of nucleotides the reads overlap either the 5' or the 3' of the locus coordinates (See Figure S2B for a detailed clarification example). The CCR5delta32 has 4 equivalent representations, each with its own coordinates

(Varsome, https://varsome.com/variant/hg19/rs333?annotation-mode=germline). Thus, for each read mapping to the canonical reference, we calculated its minimum overlapping length δ by averaging across the δs calculated for each of the representations' coordinates. A value of $\delta = 32$ was assigned to the reads overlapping both the starting and ending coordinates of the canonical reference. For the collapsed reference, we calculated δ based on the coordinate 3:46414943 (GRCh37). For all the reads mapped, only those having a value of δ equal or greater than 6 were kept. The reads mapping to both deletion regions (from the canonical and collapsed references) were assigned to the reference to which they mapped with the lowest number of mismatches. This was done because, during the alignment of the simulated ancient DNA genomes, we observed that reads originating from the canonical deletion region mapped to the collapsed deletion region with a significantly higher number of mismatches compared to when they mapped to the canonical deletion region (and vice-versa) (signed test, p-value < 0.0001) (Figure S3B). Reads mapping to both references with the same number of mismatches were assigned to the reference to which they mapped with the highest δ. Reads mapped to both references with the same number of mismatches and the same δ were discarded. The read mappings were analyzed using pysam (v. 0.16.0.1).

Optimizing the HAPI model

During the developmental stage, we explored different approaches to optimize the model. To investigate how the minimum overlapping length of the reads across the deletion region influences the performance of HAPI, we ran the model using 10 different δ thresholds, from 1 to 10, on the simulated data. As expected, increasing the δ threshold resulted in an increase in the performance of the model from an MCC of 0.75 with δ=1 to a value of 0.873 with δ=10 (Figure S3C), but at the expense of having less reads satisfying the threshold and thus less genomes recovered (121 with δ=1 and 107 with δ=10) having at least one read mapping to the deletion region. We arbitrarily selected the δ threshold of 6 (corresponding to 6 nucleotides flanking each side of the breakpoint) because we found it to be a good compromise between performance and the number of genomes recovered (MCC = 0.81, genomes recovered 116). Additionally, we investigated rescaling the bam files to account for DNA damage and excluding reads without a perfect match in the alignment. Here, we found that rescaling did not have any significant effect on the performance of the model and that using only perfect match reads improved the performance of the model but at the expense of losing 22 genomes. These strategies were therefore not included in the final model.

CLUES analysis

To reconstruct the allele frequency trajectory of the CCR5delta32 deletion, we used a modified version of the software CLUES, adapted for time-series data.[83,117] We converted the output of HAPI into hard called genotypes, using the outputs from the permissive and strict filters. We then conditioned the inference of the trajectories on the present-day EUR frequency of 0.1237 and an estimate of the effective population size history inferred from genomes in the Finnish (FIN), British (GBR), and Tuscan (TSI) populations from the 1KGP3, using the software Relate (v.1.1).[108] For the ancestry stratified CLUES analyses we used ancestry paths from Irving-Pease et al.,[84] and used the four strongest tag SNPs to make a consensus call for the deletion. i.e. we called the deletion if an imputed haplotype had three or more of the four tag SNPs. Then for each haplotype we made a consensus call by assigning the most common ancestry observed across the four SNPs. We performed the analysis with and without modern ascertainment and used five ancestry models, ALL (pan-ancestry), ANA (Anatolian Farmers), CHG (Caucasus hunter-gatherer), WHG (Western hunter-gatherer) and EHG (Eastern hunter-gatherer). For the analysis of rs73833033[61] we used the genotype calls from bcftools as described above. The code to reproduce these analyses is available in the GitHub repository (See key resources table: CLUES analysis).

Estimating the age of the CCR5delta32 deletion

To infer an estimate for the age of the CCR5delta32 deletion, we extracted the time series of posterior probability densities from all the CLUES models. As CLUES does not have an explicit mutational model, we approximated the temporal origin of the CCR5delta32 deletion by finding the most recent time-point in which the majority of the posterior density was assigned to the two lowest frequency bins – i.e., the time point at which the model estimates that there is a greater than 50% probability that the allele is at the lower limit of possible frequency values. For each genotype call set, we averaged the approximated allele ages inferred from CLUES in the models with and without conditioning on the present-day frequency from 1KGP3 and used the resulting average as an input parameter for the spatiotemporal model.

Method for modeling the spatiotemporal diffusion of the deletion allele

To model the diffusion of the CCR5delta32 allele across space and time, we use a method described in Muktupavela et al.[85] and available from: https://github.com/RasaMukti/ccr5delta32_analysis. We adapted the method so that the input genotype calls for each individual corresponded to the genotype with the highest posterior probability obtained from HAPI. To do this, we modified the Equation 5 from Muktupavela et al.[85]:

$$L(d_i, a_i) = \sum_{h=0}^{2} P[d_i, a_i | g_i = h] P[g_i = h | p(x_i, y_i, t_i)] \quad \text{(Equation 7)}$$

Here L is the likelihood of the observed data for individual i, a_i and d_i represent the number of reads carrying ancestral or derived alleles, respectively, $g_i \in \{0, 1, 2\}$ is the genotype of the individual at the particular locus, (x_i, y_i) represent the coordinates of the sampling location for that individual and t_i is the estimated sample age. $P[d_i, a_i | g_i = h|]$ is the likelihood for genotype g_i and $P[g_i = h|p(x_i, y_i, t_i)|]$ corresponds to binomial distribution, where $p(x_i, y_i, t_i)$ is the solution to a reaction-diffusion partial differential equation and it represents the allele frequency distribution across a two-dimensional (x, y) landscape at a time point t:

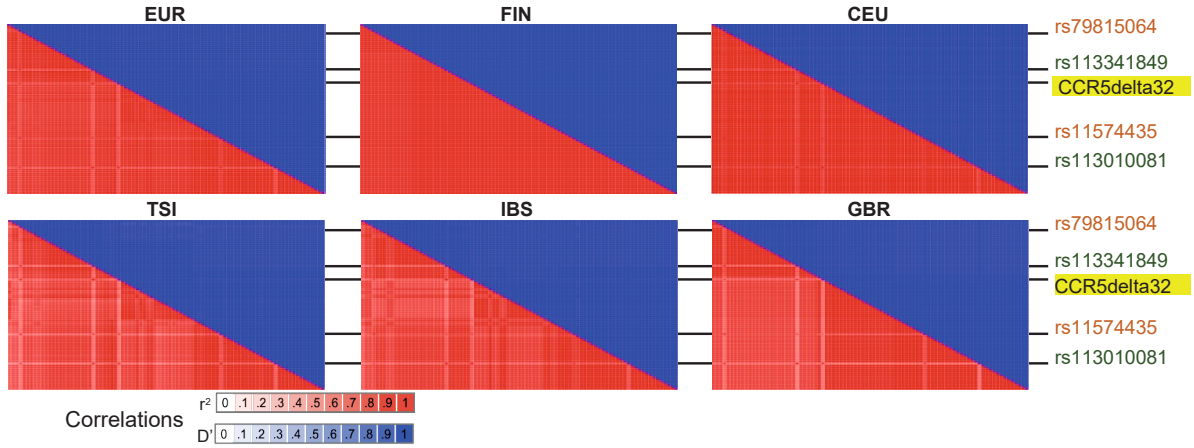
$$\frac{\partial p}{\partial t} = \frac{1}{2}\sigma_x^2\frac{\partial^2 p}{\partial x^2} + \frac{1}{2}\sigma_y^2\frac{\partial^2 p}{\partial y^2} + v_x\frac{\partial p}{\partial x} + v_y\frac{\partial p}{\partial y} + ps(1 - p) \quad \text{(Equation 8)}$$

where σ_x, σ_y are the longitudinal and latitudinal diffusion coefficients, respectively, v_x and v_y represent the longitudinal and latitudinal advection coefficients, respectively, and s is the selection coefficient. We modified the equation so that the likelihood of the genotype g_i is equal to 1 if the genotype corresponds to the genotype with the highest posterior probability and 0 otherwise. We applied the method to the different deletion call datasets, combining them with the present-day geographically-spread deletion calls compiled in Novembre et al.[12] (Figure S4C). We removed genomes that were from individuals outside of the geographic area bounded latitudinally by 30°N and 75°N and longitudinally by 30°W and 120°E. Maximum likelihood optimization was carried out by initializing 50 points in the multi-parameter space and using a first round of simulated annealing[118] followed by a run of the L-BFGS-B algorithm[119] to refine the optimization. The code to reproduce the analysis is available in the GitHub repository (See key resources table: Spatio-temporal analysis).

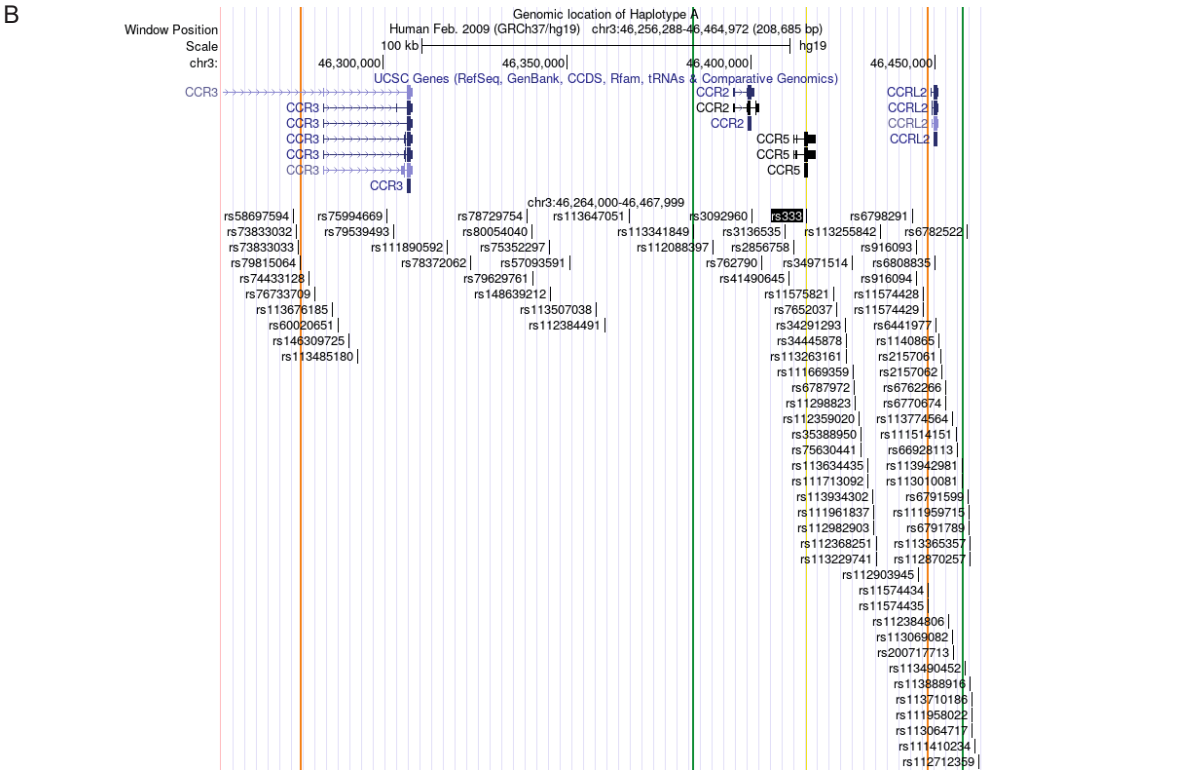
CellPress
OPEN ACCESS

Supplemental figures

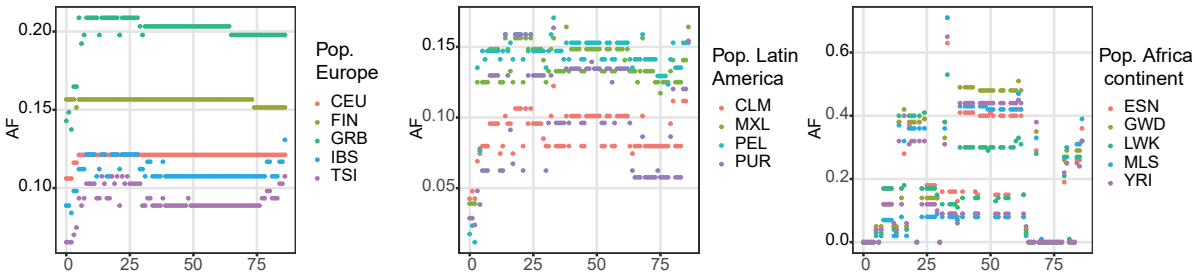
A



B



C



(legend on next page)

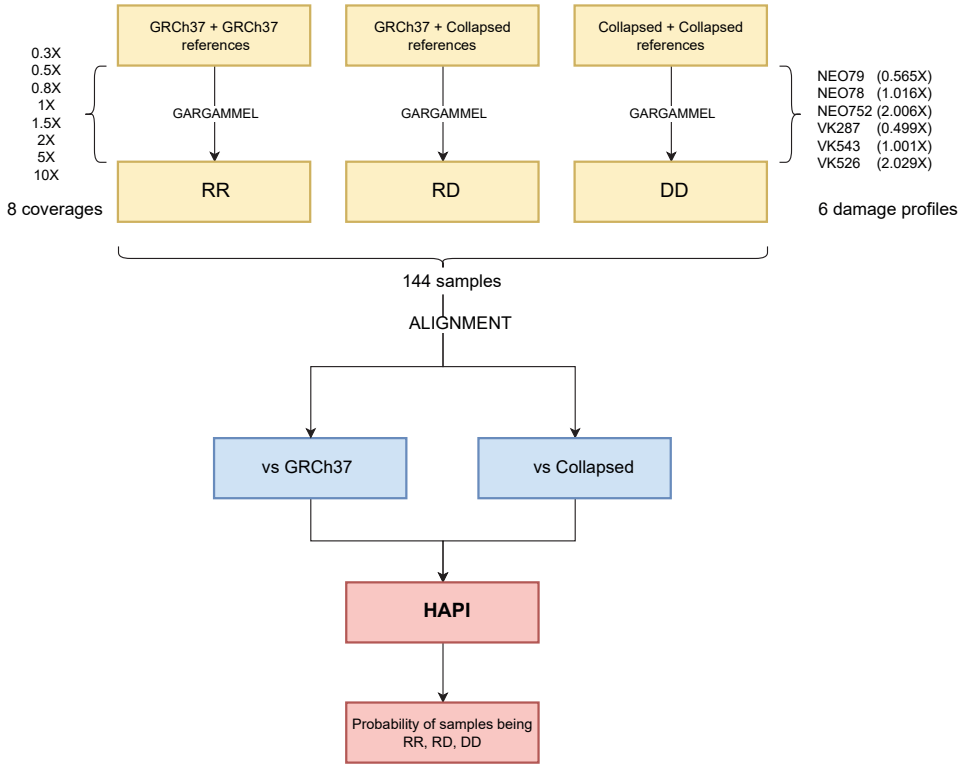
Figure S1. Detailed information of Haplotype A: LD statistics, genomic location, and the AF of the proxy variants, related to Figures 1 and 2

(A) Heatmap matrix of pairwise LD statistics from Haplotype A in the five EUR populations: EUR (FIN, CEU, TSI, IBS, and GBR), followed by each EUR population separately: r^2 values are in shades of red, while D′ values are in shades of blue, wherewith darker colors indicate a higher degree of LD. The strong LD pattern from Haplotype A is observed in the FIN and CEU populations, whereas the pattern becomes weaker in Southern and Western Europe. The weaker LD patterns are caused by the homologous recombinations of Haplotype A and the more frequent presence of Haplotypes B and C and their homologous recombinations.

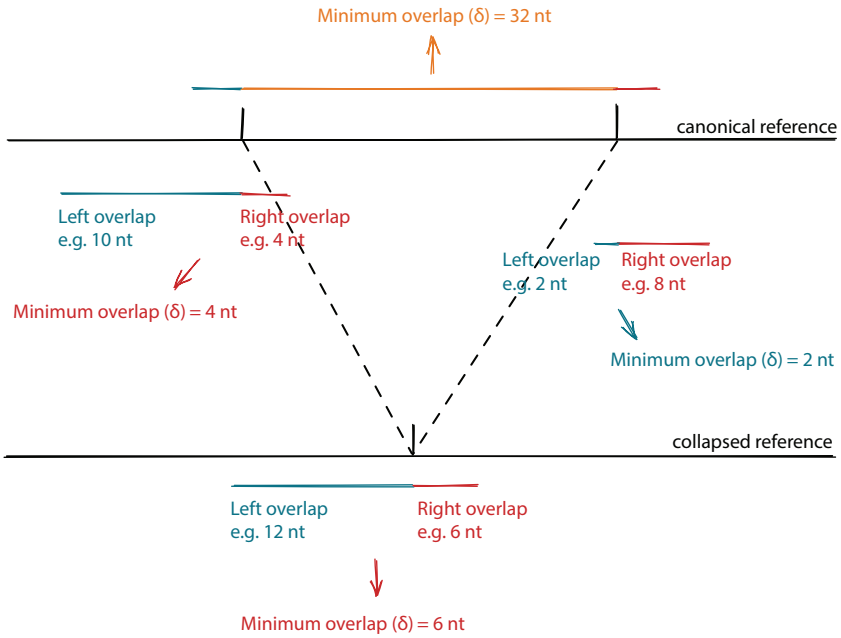
(B) The UCSC genome browser (https://genome.ucsc.edu/) displays the location of Haplotype A on 3p21.31. The CCR5delta32 allele (rs333) is highlighted in yellow, and the two SNPs with $r^2 = 1$ are marked in green, while the two SNPs with $r^2 = 0.903$ are marked in orange. The haplotypes span ≈ 0.19 Mb and encompass CCR3, CCR2, CCR5, and CCRL2. Detailed information on the tag variants is included in Table S1.

(C) AF of CCR5delta32 and the 86 tag variants, in different populations and continents. The *x* axis corresponds to Haplotype A, position 0 = CCR5delta32, and the 86 tag variants are ranked according to their r^2 values. The *y* axis shows the AF obtained from the 1KGP3. Populations from Europe and Latin America all have individuals carrying the CCR5delta32 allele/Haplotypes A, B, and C, whereas none of the individuals from the African continent carried any of the three haplotypes. However, precursor variants for Haplotype C exist, where 38 of the variants have a higher AF in the African population compared with the European population.

A



B



(legend on next page)

Figure S2. Workflow of the data analysis on the simulated ancient samples and details of the overlapping lengths, related to Figure 3

(A) The software Gargammel was used to simulate a total of 144 ancient samples, using damage profiles derived from 6 real ancient samples (right), at 8 different coverages (left). After the simulation, the reads were aligned to the canonical GRCh37 and to the collapsed reference using bwa. Finally, HAPI was used to calculate the probability of a sample having each of the three deletion genotypes (RR, RD, and DD).

(B) Schema of the reads mapping to the two references. Each read mapping to the canonical and the collapsed references is assigned a minimum overlapping length δ, which represents the minimum number of nucleotides with which it overlaps the deletion coordinates. In order for a genome to be analyzed by HAPI, it needs to have at least one read mapping to either the canonical or collapsed reference, with a minimum overlapping length of 6.

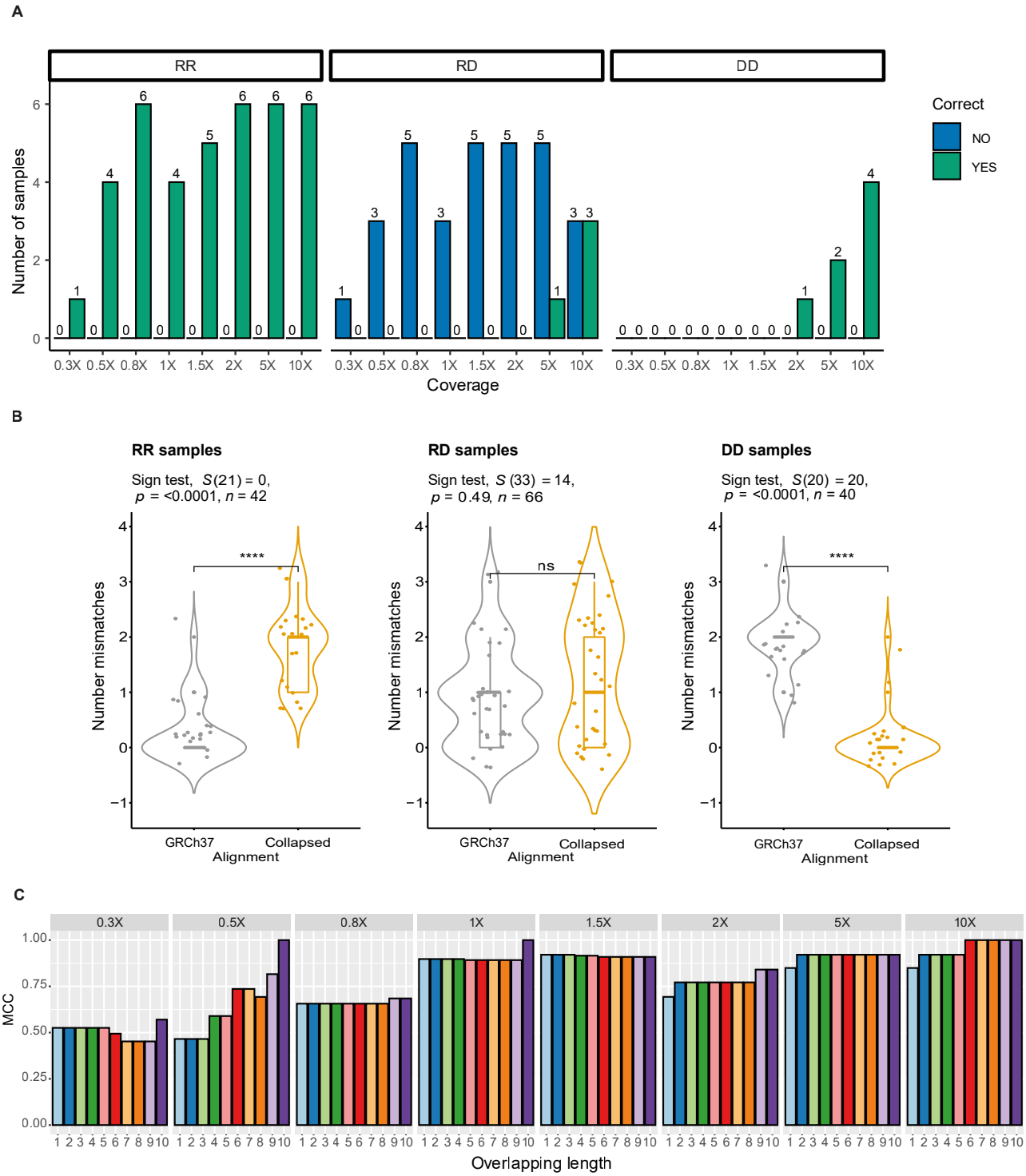


Figure S3. Assessing GATK HaplotypeCaller, mismatch rates, and HAPI performance at different overlapping lengths, related to Figure 3

(A) Number of simulated samples classified by GATK HaplotypeCaller. Bars represent the number of ancient simulated samples correctly (green) or incorrectly (blue) classified by GATK HaplotypeCaller, stratified by coverage (from 0.3× to 10×) and by deletion genotype (RR, RD, and DD). A considerable number of samples were not classified by GATK HaplotypeCaller as it failed to detect any reads; thus their columns are 0.

(B) Results of the sign test for the number of mismatches of reads originating from simulated individuals carrying the RR, RD, and DD deletion genotypes when aligned to the canonical reference (GRCh37) or the collapsed reference (collapsed). We can see that reads originating from individuals with DD genotype, and thus

(legend continued on next page)

having the deletion, mapped to the reference genome with a higher number of mismatches, compared with the collapsed reference (plot on the right). The opposite effect was seen in the plot on the left, while no significant difference was shown for reads originating from simulated individuals with RD genotype. (C) MCCs of HAPI at different values for the parameter "overlapping length." The performance of HAPI in MCC is shown at different values of the minimum overlapping length for the reads mapping to the deletion region. The MCC increases with higher overlapping lengths values but at the expense of having less reads satisfying the requirements and thus less samples analyzed. A minimum overlapping length value of 6 was chosen (see STAR Methods and Figure S2).

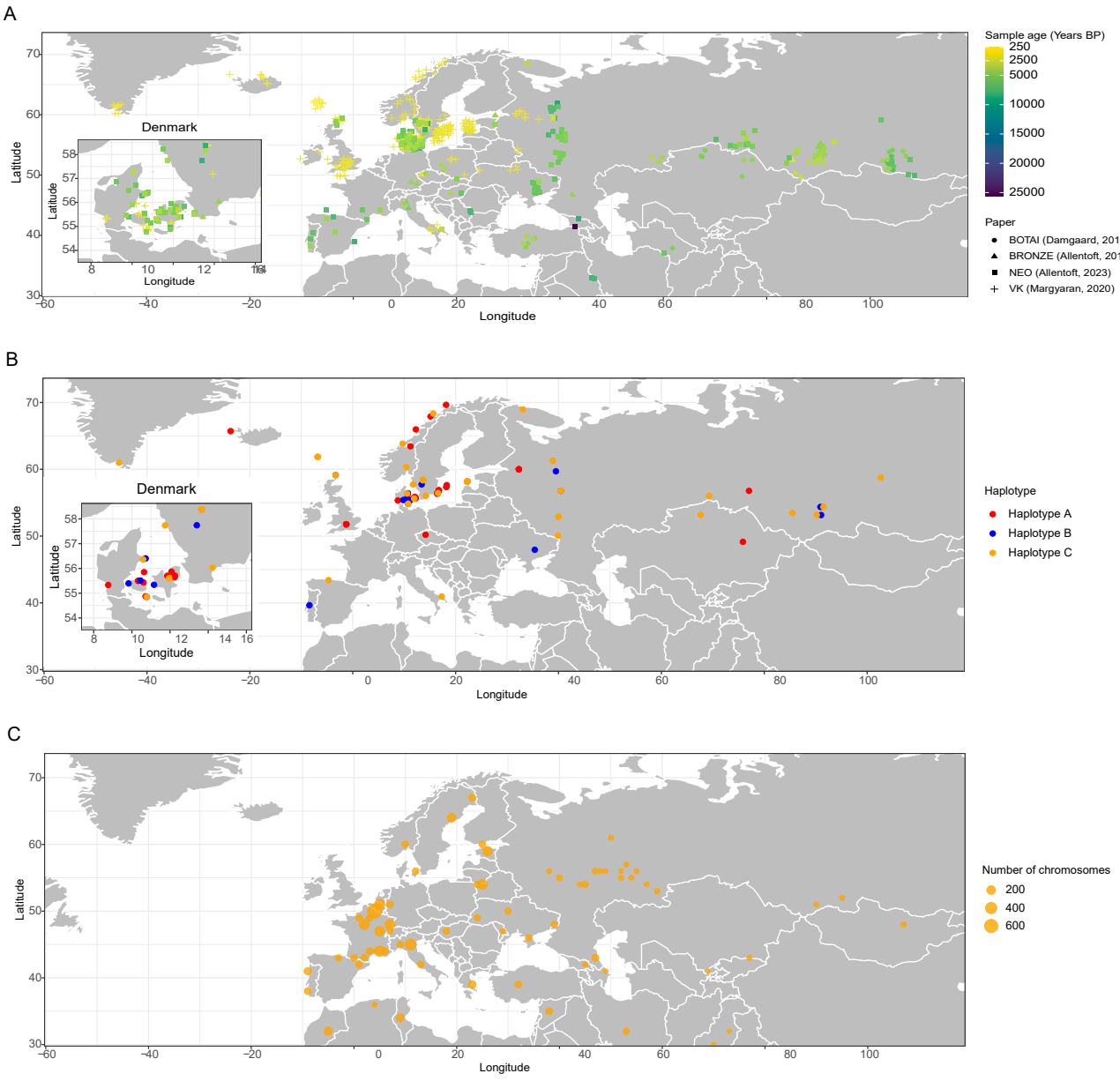


Figure S4. Geographic and chronological coordinates of the ancient and the modern genomes, related to Figure 4

(A) Geographic locations of the ancient DNA genomes analyzed by HAPI and used in the spatiotemporal analysis.[80–83] Each point represents a genome, and the shape corresponds to the original paper that published the sample. The points of the overall plot have been jittered to better show overlapping genomes. The points in the insert of Denmark have not been jittered. Precise coordinates of each genome are available in Table S4.

(B) Ancient genomes carrying either Haplotype A after applying the strict filter or traces of either Haplotype B or Haplotype C. More details are available in Table S5.

(C) Present-day samples used in the spatiotemporal analysis. The data are compiled in Novembre et al.[12]

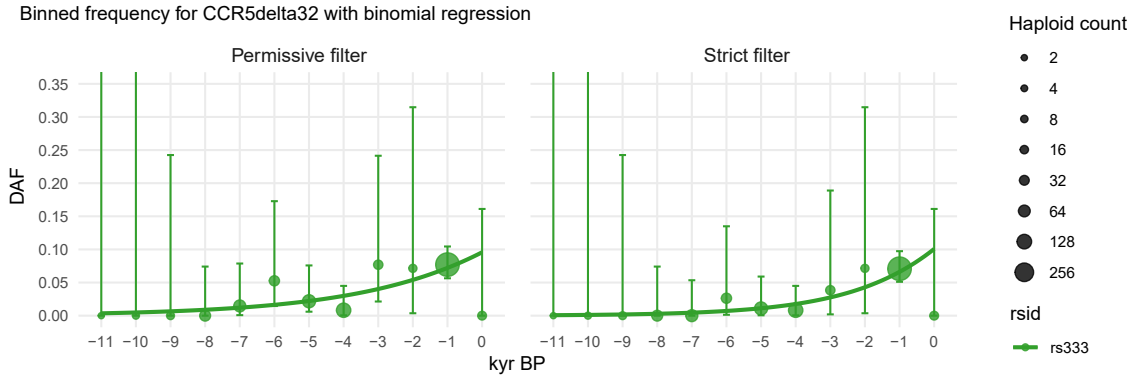


Figure S5. Scatterplot of the DAF for CCR5delta32 deletion (rs333) calculated in 1,000-year bins, related to Figure 5

The left-hand panel shows DAF calculated from the permissive filter, and the right-hand panel shows DAF calculated from the strict filter. Solid lines depict binomial regressions to the observed DAF estimates in each time bin, and error bars show the 95% confidence interval of the DAF estimate in each bin.

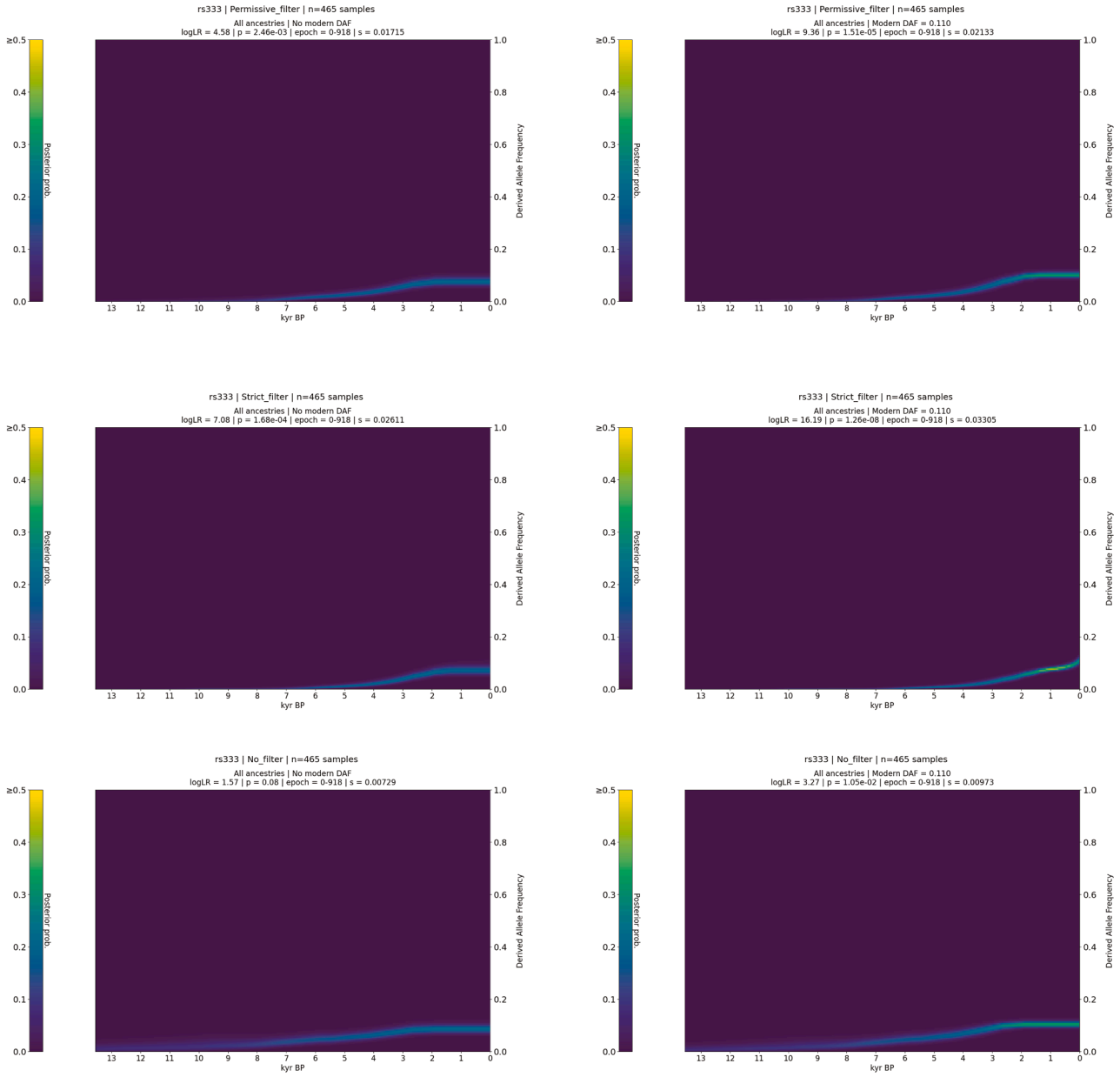


Figure S6. AF trajectory inferred by CLUES, related to Figure 5

AF trajectory inferred by CLUES. Upper row shows results using permissive filter genotype call set, middle, using strict filter, and bottom row shows results using HAPI classifications. Left column shows results using ancient data only, and the right column corresponds to results when ancient data are combined with modern ascertainment from 1KGP3. In each figure the line represents posterior probability density. The *p* value indicates evidence for rejecting a neutral model, and we also provide the most likely selection coefficient inferred by CLUES.

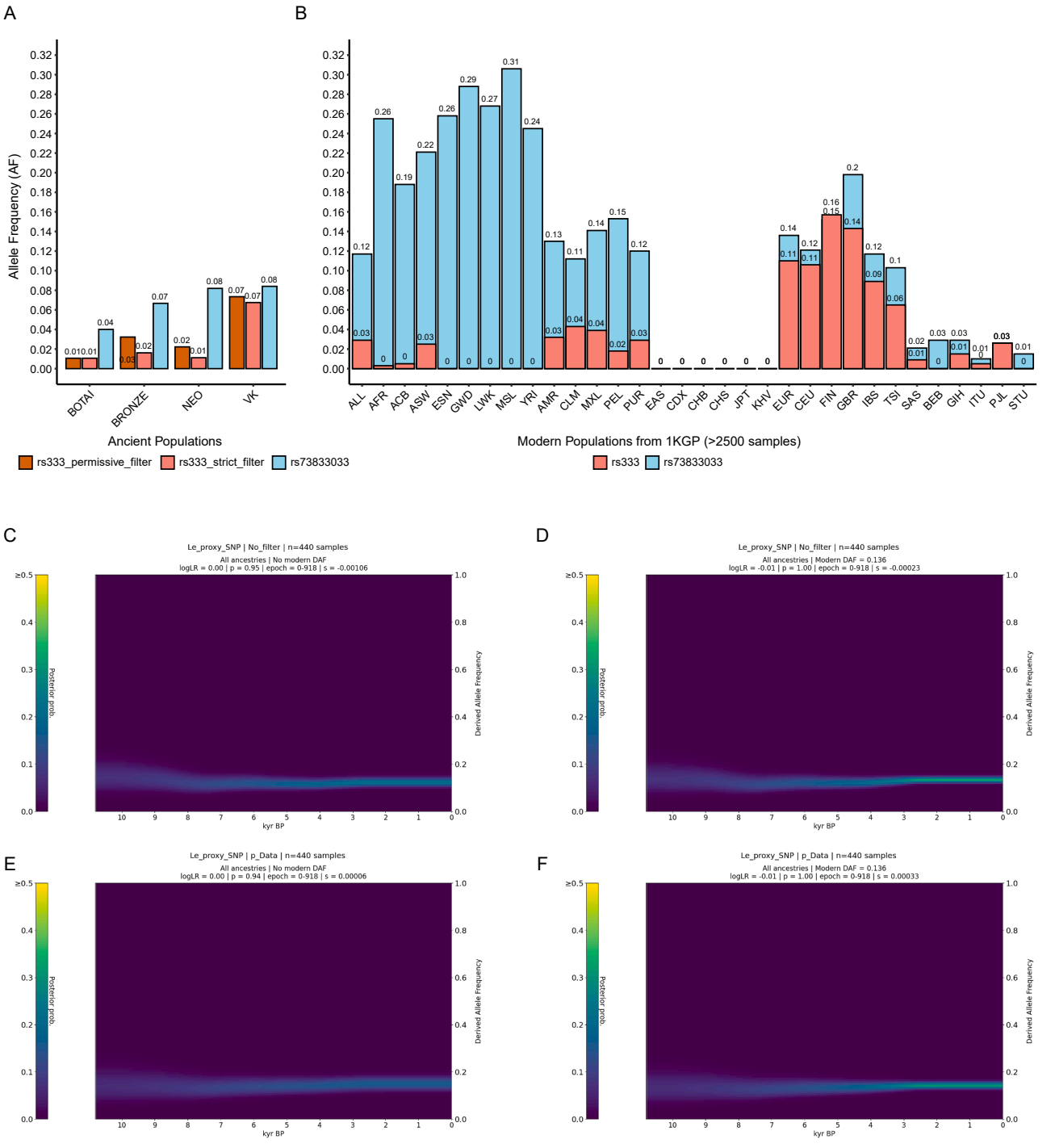


Figure S7. Comparison of the Le et al. proxy variant (rs73833033) to the CCR5delta32 variant (rs333), related to Figure 5

(A) AF of rs333 using permissive (red) and strict (orange) filters and rs73833033 (blue) in the ancient genome datasets showing that rs73833033 has a higher AF than rs333 in Neolithic and Bronze Age datasets. The rs73833033 variant was called using bcftools, as done in the Le et al. paper, and the rs333 deletion was called using HAPI.

(B) AF of rs333 and rs73833033 in the modern dataset, retrieved from 1KGP3. As in (A), this shows striking different AFs for the Le et al.[61] proxy variant (blue), compared with CCR5delta32 (orange).

(C–F) AF trajectory inferred by CLUES for rs73833033 with (C) using genotype classifications and no modern ascertainment, (D) using genotype classifications and modern ascertainment, (E) using genotype probabilities and no modern ascertainment, and (F) using genotype probabilities and modern ascertainment. In all cases, there is no evidence of selection on rs73833033.