

1 Tracing the evolutionary path of the CCR5delta32 deletion 2 via ancient and modern genomes

3

4 Authors

5 Kirstine Ravn^{1,*}, Leonardo Cobuccio^{1,*}, Rasa Audange Muktupavela^{2,*}, Jonas Meisner^{1,3},
6 Michael Eriksen Benros³, Thorfinn Sand Korneliusen⁴, Martin Sikora⁴, Eske Willerslev^{4,5,6,7},
7 Morten E. Allentoft^{4,8}, Evan K. Irving-Pease^{2,4}, Fernando Racimo^{2,4}, Simon Rasmussen^{1,9}

8

9 Affiliations

- 10 1. Novo Nordisk Foundation Center for Protein Research, Faculty of Health and Medical
11 Sciences, University of Copenhagen, Copenhagen, Denmark
12 2. Section for Molecular Ecology and Evolution, Globe Institute, University of Copenhagen,
13 Copenhagen, Denmark
14 3. Copenhagen Research Centre for Mental Health, Mental Health Centre Copenhagen,
15 Copenhagen University Hospital, Copenhagen, Denmark
16 4. Lundbeck Foundation GeoGenetics Centre, Globe Institute, University of Copenhagen,
17 Copenhagen, Denmark
18 5. GeoGenetics Group, Department of Zoology, University of Cambridge, Cambridge, UK
19 6. Wellcome Sanger Institute, Wellcome Genome Campus, Cambridge, UK
20 7. MARUM, University of Bremen, Bremen, Germany.
21 8. Trace and Environmental DNA (TrEnD) Laboratory, School of Molecular and Life
22 Sciences, Curtin University, Perth, Australia.
23 9. The Novo Nordisk Foundation Center for Genomic Mechanisms of Disease, Broad
24 Institute of MIT and Harvard, Cambridge, MA 02142, USA.

25

26 Author List Footnotes

27 Correspondence to: Fernando Racimo (fracimo@sund.ku.dk) and Simon Rasmussen
28 (simon.rasmussen@cpr.ku.dk).

29 *: Joint first authors

30

31 **Abstract**

32 The chemokine receptor variant CCR5delta32 is linked to HIV-1 infection resistance and
33 other pathological conditions. In European populations, the allele frequency ranges from 10-
34 16%, and its evolution has been extensively debated throughout the years. We provide a
35 detailed perspective of the evolutionary history of the deletion through time and space. We
36 discovered that the CCR5delta32 allele arose on a pre-existing haplotype consisting of 84
37 variants. Using this information, we developed a haplotype-aware probabilistic model to
38 screen for this deletion across 860 low-coverage ancient genomes and we found evidence
39 that CCR5delta32 arose at least 7,000 years BP, with a likely origin somewhere in the
40 Western Eurasian Steppe region. We further show evidence that the CCR5delta32
41 haplotype underwent positive selection between 7,000-2,000 BP in Western Eurasia and
42 that the presence of the haplotype in Latin America can be explained by post-Columbian
43 genetic exchanges. Finally, we point to new complex CCR5delta32 genotype-haplotype-
44 phenotype relationships, which demand consideration when targeting the CCR5 receptor for
45 therapeutic strategies.

46

47 **Keywords**

48 CCR5delta32, HIV-1 infection resistance, pleiotropy, haplotype, recombination, ancient
49 genomes, evolution, positive selection, immune genes, immune response.

50

51 Introduction

52 Humans have been exposed to pathogens over the course of our evolutionary history, and
53 adaptations to them have left numerous signatures in our genomes¹⁻³. In recent years,
54 evidence for selection has been found in genes involved in the development of tolerance
55 against intracellular pathogens and in the inflammatory response against extracellular
56 microbes⁴⁻⁶. These include, for example, the TLR6-TLR1-TLR10 cluster of toll-like
57 receptors, which are crucial components of innate immunity against pathogens, and were
58 likely under positive selection in anatomically modern humans after introgression from
59 archaic hominin groups^{7,8}. More recently, Domínguez-Andrés et al, 2021⁹ showed that
60 alleles associated with cytokine profiles reflecting immune tolerance were under selection
61 during the transition to farming in the Neolithic period, as sedentarism and population density
62 increased, enabling the development of pathogen reservoirs in newly domesticated animals.

63 Perhaps one of the most intensively debated immune-associated loci previously posited to
64 have been under selection in humans is a 32-bp deletion (CCR5delta32, rs333), which
65 introduces a premature stop codon in the C-C chemokine receptor 5 gene
66 (ENSG00000160791:CCR5)¹⁰⁻¹⁷. CCR5 is a member of the G-protein-coupled receptor
67 family of proteins and, upon activation, by the CC chemokines CCL3, CCL4, and CCL5,
68 CCR5 and its ligands, plays a critical function in regulating the inflammatory response by
69 facilitating communication between immune cells and the environment¹⁸⁻²⁰. Thus, CCR5 can
70 act as a regulator of the host's immune response. In 1996, CCR5 was identified as a
71 necessary co-receptor for the macrophage-tropic HIV strains^{21,22} and it was subsequently
72 reported that CCR5delta32 could provide HIV-1 infection resistance to individuals carrying
73 this allele in homozygous form²³. CCR5 is now an important target in preventing and
74 treating HIV infection, using various therapeutic strategies^{24,25}. A female patient with HIV-1
75 was recently potentially cured for both HIV-1 and acute myeloid leukemia through a
76 CCR5delta32/delta32 haplo-cord transplant²⁶, a method that has successfully cured two
77 similar cases using unrelated donor stem cells with the same genetic modification^{27,28}.

78 Besides the significant effect on HIV infection, the CCR5delta32 allele has also been
79 associated with other pathological conditions including infection by other viral organisms
80 (including SARS-CoV-2 which causes COVID-19), immune-related diseases, neurological
81 disorders, and various types of cancer^{20,29-41}. Together these studies indicate that the
82 CCR5delta32 allele is pleiotropic and can act as a modulator of a given phenotype
83 expression, with both advantages and disadvantages, depending on the medical context. In
84 this perspective, serious concerns have been raised by the scientific community about

85 possible clinical side effects on “CCR5delta32” CRISPR babies, whose genomes have been
86 edited to confer lifetime HIV immunity ^{42–45}.

87 The evolutionary history of the CCR5delta32 deletion has been widely debated and with
88 conflicting research results ^{10,12,13,15–17,46–54}. Today the CCR5delta32 allele frequency (AF) is
89 between 0.10-0.16 in Northern European populations and less than 0.08 in South- and
90 South-East Europe ^{12,55}. Outside of Europe the deletion is found only in populations with
91 European ancestry ^{56–59}. Past studies have estimated the age of the CCR5delta32 allele with
92 divergent results ranging from ~700, ~3,400, and >5,000 years ago ^{10–12,15,17}. Positive
93 selection, negative selection, balancing selection, and genetic drift have each been
94 proposed as an explanation for the distribution of present-day gene frequencies <sup>10,12,13,15–
95 17,46–54,60,61</sup>.

96 The few studies conducted on the CCR5delta32 deletion in ancient individuals have been
97 constrained by a limited geographic scope and small sample sizes, leading to the possibility
98 of biasing the results by familial relations. So far, the oldest CCRdelta32 alleles have been
99 detected in a 4,900-years-old individual belonging to the Yamnaya culture ⁶² and in several
100 Swedish individuals dating to the Neolithic period (5,250-1,690 BCE) ⁴⁸. However, the latter
101 study raised concerns about allelic dropout during the assaying process, which could lead to
102 genotype misclassification. Two studies conducted on ancient genomes from individuals in
103 central and northern Germany revealed no significant change in the frequency of the
104 CCR5delta32 variant over the past millennium, including during the Black Death
105 pandemic^{46,49}. In contrast, a study conducted in Poland reported a nearly doubled frequency
106 of the CCR5delta32 variant from the late medieval period to the present day.⁶³

107 The emergence of paleogenomics has shed light on our understanding of human population
108 history, but evidence from large ancient genomic datasets has been missing in the debate
109 on the CCR5delta32 allele. Due to the degraded nature of ancient DNA, ancient genomic
110 datasets tend to be characterized by short read lengths and post-mortem DNA damage ⁶⁴,
111 impairing the ability to identify indels like the CCR5delta32 deletion. Moreover, mapping
112 efforts in the CCR5delta32 region are particularly challenging because the breakpoint's
113 flanking regions contain repeated sequences.

114 In this study, we describe the evolutionary trajectory of the CCR5delta32 deletion as
115 revealed by ancient and present-day genomes. We discovered that the CCR5delta32 allele
116 emerged on a pre-existing haplotype comprising 84 variants, which prompted us to develop
117 a probabilistic model that allows for reliable detection of CCR5delta32 in low-coverage
118 genomes, improving our ability to study the allele's distribution and impact in space and time.

119 Applying our model to large ancient genomic datasets (>800 genomes), we find evidence in
120 support of a temporal origin of the CCR5delta32 deletion that is at least 7,000 years BP in
121 age, with a likely spatial origin somewhere in the Western Eurasian Steppe. Furthermore,
122 our findings provide evidence that the CCR5delta32 haplotype underwent positive selection
123 in Eurasia between 7,000-2,000 BP. Analyzing the CCR5delta32 haplotype in individuals
124 from Latin America, we determined that the presence of the allele in this region can be
125 attributed to the Columbian Exchange. This study is the first to provide a comprehensive
126 picture of the CCR5delta32 allele's evolutionary history across time and space.

127 **Results**

128 **Identification of three CCR5 haplotypes in Europe**

129 By re-analyzing the European individuals from the 1000 Genomes Project phase 3 (1KGP3)
130 data, we discovered that the CCR5delta32 allele was located on a haplotype with up to 107
131 variants in high linkage disequilibrium (LD): $r^2 > 0.8$ (for details see Table S1). The longest
132 haplotype was identified in the FIN (Finnish in Finland) population, spanning 107 variants
133 including 76 variants with $r^2 = 1$. We note that 86 of the 107 variants were also found to be in
134 high LD in the CEU panel (Utah residents with Northern and Western European ancestry),
135 including two variants with $r^2 = 1$ (rs113341849 and rs113010081) (Figure 1A). In contrast, in
136 the TSI (Toscani in Italy, $r^2 > 0.8$, 3 SNPs), IBS (Iberian in Spain, $r^2 > 0.8$, 3 SNPs), and GBR
137 (British in England and Scotland, $r^2 > 0.8$, 2 SNPs) panels, we could only identify very few
138 variants in high LD (Table S1). However, these variants were among those with highest LD
139 ($r^2 > 0.9$) to the deletion in the CEU population. We termed the CEU CCR5delta32 haplotype
140 'Haplotype A' (Figure 1A, Table S1) and identified it in all the 112 CCR5delta32 carriers of
141 the 505 1KGP3 European individuals (AF = 0.111), including three carriers with homologous
142 recombinants of the haplotype (Table S2A).

143 Given the strong correlation between the deletion and the variants of Haplotype A in the
144 CEU population, it was surprising that the LD was weaker in Southern and Western Europe
145 (**Figure S1A**). We found that this was caused by another haplotype that included 84 of the
146 86 tag variants. This haplotype did not include the CCR5delta32 deletion, nor any of the two
147 tag SNPs that were in complete LD ($r^2=1$) with the deletion in CEU. We termed this
148 'Haplotype B' (**Figure 1A**) and detected it in 6 of the 505 1KGP3 European individuals (AF =
149 0.006). Finally, we discovered a third haplotype ('Haplotype C') that included 82/84 tag SNPs
150 of Haplotype B in 10 European individuals (AF = 0.009) (**Figure 1A**). Besides these three
151 haplotypes, we identified homologous recombinations and recurrent LD blocks of the two
152 haplotypes without the deletion (Haplotype B and C) (AF = 0.012).

153 The three haplotypes span >0.18 Mb (chr3:46275570-46461783), including several cytokine
154 receptor genes such as C-C Motif Chemokine Receptor 3, 2, and 5 (*CCR3*, *CCC2*, *CCR5*),
155 and C-C chemokine receptor-like 2 (*CCRL2*) (**Figure S1B**). To understand the potential
156 functional effect of the variants carried by these haplotypes, we used the Ensembl Variant
157 Effect Predictor (VEP) ⁶⁵, and found that none of the tag SNPs could be annotated with
158 clinical significance, as assigned by ClinVar ⁶⁶. However, from the GWAS catalog ⁶⁷, the tag
159 SNPs with $r^2 > 0.9$ have been previously associated with complex traits and diseases, such
160 as Diabetes Mellitus Insulin-Dependent (IDDM), Inflammatory Bowel Disease (IBD), and
161 Alzheimer's Disease (AD) ⁶⁸⁻⁷⁰ (**Table S3A**). Querying the Phenoscanner database ^{71,72}
162 showed that 82 of the 86 tag variants of Haplotype A (including all the tag variants with $r^2 >$
163 0.9) were linked to many of the same phenotypic traits that were already associated with
164 *CCR5delta32s'* multiple phenotypes (**Table S3B**). Notably, as the *CCR5delta32* deletion is
165 not detectable in traditional SNP-based GWAS analyses, some of these GWAS associations
166 might be caused by the direct linkage to the *CCR5delta32* allele.

167 **Local admixture analysis revealed the European origin of *CCR5delta32***

168 We then expanded the analysis to the entire 1KGP3 dataset (2,535 individuals from 26
169 populations), where we detected 35 individuals having the *CCR5delta32* deletion outside of
170 the EUR super-population panel, primarily in populations that have European ancestry
171 (**Table S2A**). In Latin America, we could identify a homologous recombination of Haplotype
172 A in two individuals from CLM (Colombian in Medellin) and PUR (Puerto Rican in Puerto
173 Rico), which we also previously had detected in an individual from Spain (**Figure 1B**). To
174 further investigate local admixture around the *CCR5delta32* locus, we applied HaploNet ⁷³ to
175 all individuals of the 1KGP3 (**Figure 2A**). Here we found evidence of a European sequence
176 segment in 138 out of 141 individuals who harbored at least one allele of the deletion, while
177 the remaining three individuals from PJJ (Punjabi in Lahore, Pakistan) carried insufficient
178 European ancestry proportions for HaploNet to distinguish fine-scale ancestry signals.

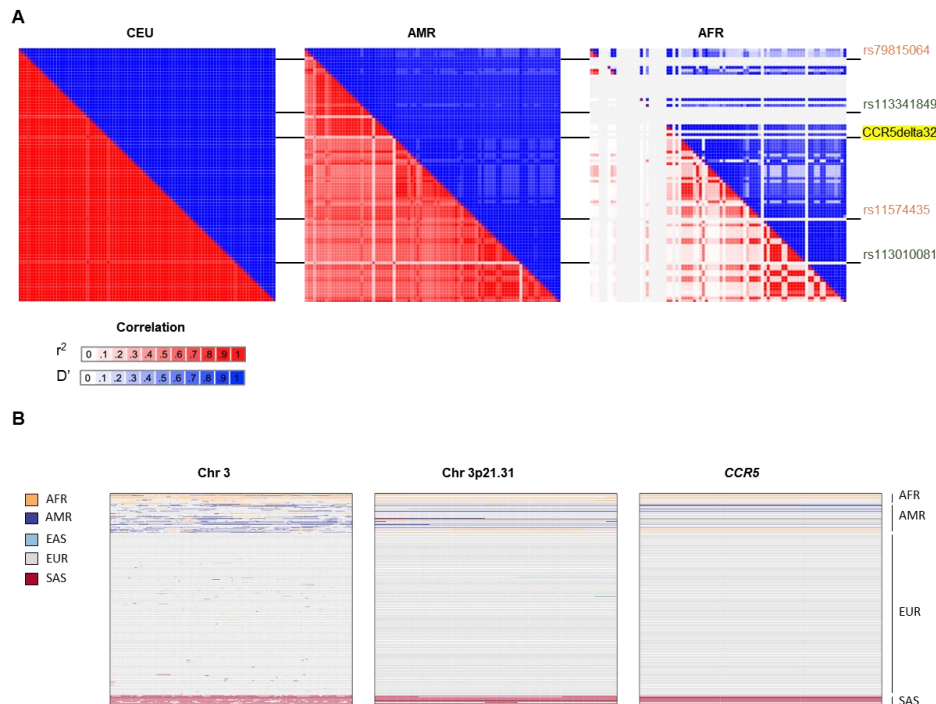


179 **Figure 1: Schematic view of CCR5delta32 and the associated haplotypes: A, B, and C.**

180 **A)** Haplotype A consists of CCR5delta32 and 86 tag variants, including two SNPs with an r^2 value of 1
 181 (rs113341849 and rs113010081, green), two SNPs with an r^2 value of 0.9027 (rs79815064 and rs1157443, pink),
 182 and 82 variants with an r^2 value of 0.8602 (grey). All r^2 values are related to the CEU population (**Table S1A**).

183 The haplotype is located on chromosome 3p21.31, spans > 0,19 Mb, and encompasses several genes: *CCR3*,
184 *CCR2*, *CCR5*, and *CCRL2*. Detailed information on the genomic locations of the genes, *CCR5delta32*, and the
185 86 tag variants is provided in **Figure S1B**. **B**) Detailed mapping of the Haplotype A, B, and C and their
186 homologous recombinations, among the individuals from 1KGP EUR and AMR populations. The light-gray blocks
187 indicate deviations from different combinations of haplotype blocks. In the Latin American population, a specific
188 homologous recombination of Haplotype A was identified in two individuals from CLM and PUR, which had also
189 been previously detected in an individual from Spain.

190 Additionally, the complete Haplotype B (AF = 0.024) and shorter homologous recombinants
191 (HR) (AF = 0.031) (**Figure 1B**) were found in significantly higher proportions in Latin
192 Americans than in European populations (chi-square test p-value = 0.002616 and 4.048e-
193 05, respectively). Likewise, we also observed this same pattern for HR Haplotype C
194 (AF=0.057, chi-square test p-value =2.115e-06) (**Figure 1B**). Among the 1,008 individuals
195 with African ancestry originating from the African continent (excluding African Ancestry in
196 SW USA (ASW), and African Caribbean in Barbados (ACB)), we did not detect any of the
197 three haplotypes. We did, however, identify precursor SNPs for Haplotype C (**Figure 2B**).
198 Out of the 82 variants of Haplotype C, 38 had a higher AF in the African population
199 compared to the European population (**Figure S1C**). Therefore, the increased AF of certain
200 haplotype blocks in Latin American populations could be explained by admixture with
201 individuals of African ancestry.



202

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

204 **A)** Identification of a European CCR5delta32 locus in individuals with the deletion. From the HaploNet analysis, a
205 European sequence segment was identified in 138 out of 141 1KG3 individuals, all genotyped with at least 1

206 allele of the deletion (**Table S2**). **B**) Heatmap matrices of pairwise LD statistics from Haplotype A in the CEU,
207 AMR, and AFR populations. The strong LD pattern from Haplotype A in the CEU population becomes weaker in
208 Latin America, due to the significantly higher homologous recombination rates we observe from haplotypes B and
209 C in Latin America. These higher recombination rates may be explained by post-Columbian admixture among
210 three groups - African, European, and Native American, as the AFR populations also harbor precursor variations
211 for Haplotype C. The r^2 values are in shades of red while the D' values are in shades of blue. Darker values
212 indicate a higher degree of pairwise LD.

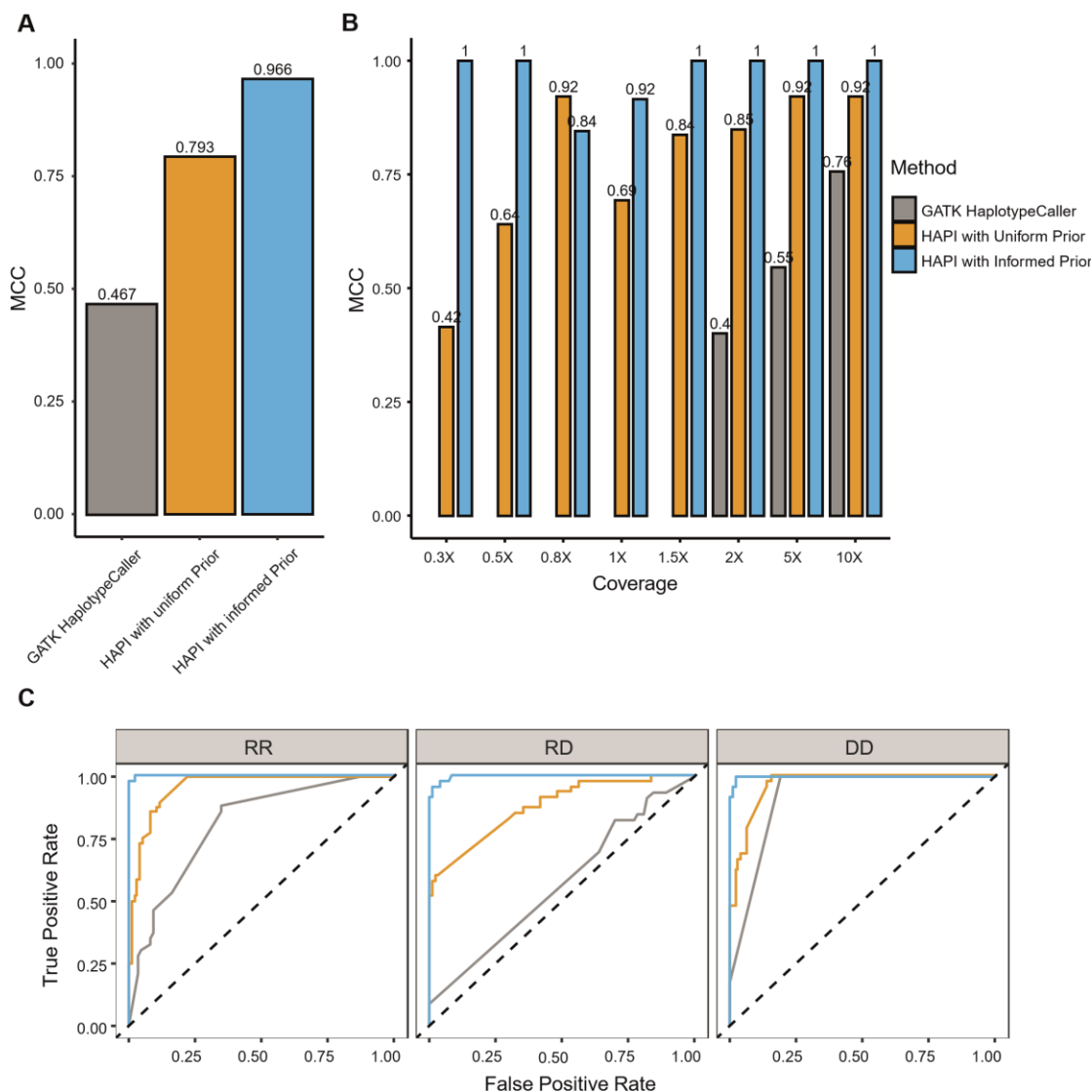
213 **Haplotype A originated from Haplotype B in Europe**

214 The high frequency of Haplotype A, along with the presence of only four haplotype
215 recombinants (HRs) of Haplotype A in the EUR population (**Figure 1B**), indicates that this
216 haplotype is much younger than Haplotype B and C and or that the CCR5delta32 deletion
217 has been exposed to selection in the EUR population. Hence, based on present-day data
218 alone, this suggests that at some point in the history of present-day Europeans, the two
219 variants rs113341849 and rs113010081 and the CCR5delta32 deletion (all with $r^2 = 1$)
220 emerged on Haplotype B, leading to Haplotype A, which is now present at substantial
221 frequencies in present-day Europeans and in certain Latin American individuals, due to post-
222 Columbian admixture (**Figure 2B**).

223 **A probabilistic framework for calling CCR5delta32 allele in low-coverage aDNA** 224 **genomes**

225 To trace the evolution of the CCR5delta32 allele through time, we aimed at identifying
226 ancient individuals carrying the deletion. To achieve this, we developed a Haplotype-Aware
227 Probabilistic model for Indels (HAPI), which allowed us to identify the deletion in low-
228 coverage ancient genomes (see Methods). For this, we utilized the information from the four
229 tag SNPs having the highest pairwise LD with the CCR5delta32 allele ($r^2 > 0.90$, **Table S1**),
230 as a prior for the presence of the deletion and modeled the information from the reads
231 mapping to the *CCR5* region in the form of a likelihood function. To remove reference bias
232 and improve CCR5delta32 mapping detection, we used both the standard reference
233 sequence and a reference sequence including the deletion, hereinafter referred to as
234 canonical and collapsed references, respectively. We first tested HAPI on 15 genotyped
235 CCR5delta32 genomes from the 1KGP3 and showed that it correctly classified all of them
236 (see Methods). To evaluate HAPI's performance at lower coverages we created a simulated
237 dataset containing 144 ancient genomes with coverages from 0.3X to 10X (**Figure S2A**). We
238 found that the haplotype-informed prior model performed better compared to a uniform-prior
239 model, with an increase of Matthews Correlation Coefficient (MCC) from 0.79 to 0.97
240 (**Figure 3A**). Furthermore, we benchmarked the performance of HAPI against the
241 commonly-used GATK HaplotypeCaller⁷⁴. Here, we found that on the simulated dataset,

242 HAPI could correctly classify 129 genomes out of 144 with an MCC of 0.97, compared to
 243 only 79 by the GATK HaplotypeCaller (MCC 0.47), an increase in genomes by 34% (**Figure**
 244 **3A**). Additionally, it was much more precise on the set that was called by both HAPI and
 245 GATK HaplotypeCaller, with MCCs of 0.98 and 0.47, respectively. For the subset of
 246 genomes that had coverages between 0.5X and 1X, we found that HAPI could correctly
 247 classify 45 of 54 genomes (84%) (**Figure 3B**). For coverage at 0.3X, six of 18 genomes
 248 (33%) could be classified by HAPI. Across these very low-coverage genomes ($\leq 1X$) HAPI
 249 had an MCC of ≥ 0.84 (**Figure 3B**). For the genomes with coverage lower than 2X, GATK
 250 HaplotypeCaller could only correctly classify genomes without CCR5delta32 deletion
 251 (**Figure S3A**). The difference between the methods was more pronounced when we
 252 stratified by the performance metrics by deletion genotype (RR, RD, DD, i.e. homozygous for
 253 the reference, heterozygous, or homozygous for the deletion), where GATK HaplotypeCaller,
 254 HAPI with the uniform prior, and HAPI with the informed prior had ROC-AUCs of 0.30, 0.93,
 255 and 0.99, respectively (**Figure 3C**). Taken together our model was highly specific for
 256 identifying CCR5delta32 allele, even in the heterozygous form and with as little as 0.3X
 257 coverage.
 258



259 **Figure 3: Performances on simulated data.**

260 **A)** Comparison of the performance of GATK HaplotypeCaller (grey), HAPI with Uniform Prior (yellow), and HAPI
261 with Informed Prior (blue) on the 144 ancient simulated genomes. Performances are shown as MCC on all the
262 simulated genomes across different coverages. **B)** The same performance comparison but stratified by
263 sequencing coverage. **C)** Performances shown as ROC-AUC on all the simulated genomes stratified by deletion
264 genotype. GATK HaplotypeCaller (grey), HAPI with the Uniform Prior (yellow), and HAPI with the Informed Prior
265 (blue) had ROC-AUCs of 0.30, 0.93, and 0.99, respectively. Further, it is noticeable the precision with which
266 HAPI with the Informed Prior detects the CCR5delta32 allele in heterozygous low-coverage genomes (RD). In
267 these ancient genomes there are often not enough aligned reads to confidently determine the presence of both
268 alleles, which can lead to a biased representation of the genotype in question towards the reference allele. Here
269 HAPI provides a higher degree of reliability and accuracy in genotyping the deletion compared to GATK
270 HaplotypeCaller.

271 **Applying HAPI to ancient datasets**

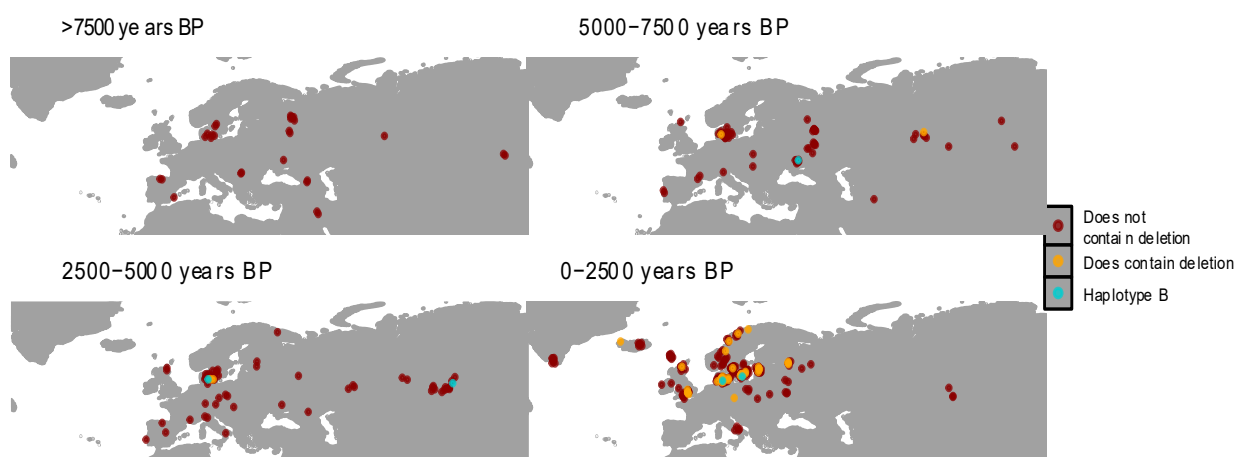
272 We then applied HAPI to our extensive ancient DNA dataset, which consisted of 860
273 genomes⁷⁵⁻⁷⁷ from various regions across Eurasia, including a dense sampling collection in
274 northern Europe, specifically in Denmark. The final dataset encompasses consecutive
275 historical eras, ranging from the early Mesolithic and Neolithic periods to the Bronze Age,
276 and extending into the Viking Age. To take into account the complexity of the haplotype and
277 the damaged nature of ancient DNA, we applied two curation steps to the results of the
278 model: a “permissive filter” to reclassify genomes that had artifacts typical of ancient DNA
279 damage, and a “strict filter” to reclassify genomes which were likely harboring the Haplotype
280 B (see Methods). Across the ancient DNA dataset, we found that 418 genomes had at least
281 one read mapping to the *CCR5* region from either the canonical or collapsed reference and
282 having at least 6 bases overlapping the CCR5delta32 breakpoint. Using this approach, we
283 identified the CCR5delta32 allele in 43 and 38 individuals using the permissive and strict
284 filters, respectively (**Figure 4, Table S4**). From the Allentoft et al. (2022)⁷⁷ dataset, spanning
285 the Mesolithic and Neolithic, four individuals were identified with the deletion using the strict
286 filter and four individuals were classified with Haplotype B across the different output
287 schemes from HAPI (**Table S4**). Only 31 out of 101 genomes from the Bronze Age data⁷⁶
288 met the criteria for the analysis by HAPI and, although the sample pool was small, we
289 detected one sample with the CCR5delta32 deletion using the strict filter, and one sample
290 carrying Haplotype B (**Table S4**). From the Viking dataset⁷⁸, 252 of 442 genomes passed
291 the HAPI inclusion criteria (**Table S4**). From these, 33 genomes were detected to have the
292 CCR5delta32 deletion with the strict filter (Haplotype A, AF = 0.065) and two genomes were
293 identified as having Haplotype B (AF = 0.003). Furthermore, we observed 21 genomes
294 having portions of Haplotype C (>20 proxy SNPs). Detailed view of the location of the
295 ancient DNA genomes is provided in **Figure 4** and **Table S4A-C**.

296 **Haplotype A and B were present in Denmark more than 6,000 years ago**

297 In our Mesolithic and Neolithic dataset⁷⁷ we had an extensive collection of ancient genomes
298 from Danish individuals, totaling 100 genomes. We found evidence that the CCR5delta32
299 allele (Haplotype A) was present in Denmark over 6,000 years ago (NEO855: 6299 cal. BP),
300 as well as in an individual carrying the Haplotype B (NEO683: 7521 cal. BP) (**Figure S4A**).
301 This places the CCR5delta32 allele in the Danish Ertebølle culture, a hunter-gatherer and
302 fisher, pottery-making culture, dating to the end of the Mesolithic period⁷⁹. Both genomes
303 were identified to be of the Western Hunter-gatherers ancestry group (HG_EuropeW), which
304 were the predominant ancestry group in Denmark at the time⁷⁷. In contrast, the Danish
305 individuals detected with the Haplotypes A and B in the later Neolithic and early Bronze Age
306 harbored Steppe-related ancestry (EUR_BA). During the transition from hunter-gatherer to
307 Neolithic and Bronze Age periods, Denmark's population genomic landscape underwent
308 significant changes⁷⁷. These changes involved the replacement of hunter-gatherer
309 populations and the introduction of Steppe-related ancestry during the late Neolithic and
310 Bronze Age periods.

311 **Haplotypes A, B, and C were present in Mesolithic and Neolithic Periods** 312 **across Eurasia**

313 Outside of Denmark, we detected the CCR5delta32 allele in Russia (NEO309: 5824 cal. BP)
314 as well as Haplotype B in Ukraine (NEO300: 6678 cal. BP), Sweden (NEO27: 9693) and
315 Portugal (NEO631: 7135 cal. BP). Interestingly, we detected nine genomes having between
316 19 and 69 variants from the Haplotype C in Russia (Table S4 B and C), with the oldest
317 sample dating 10,853 cal. BP (NEO202: 69/82 variants). Further, a sample (NEO646) from
318 northwest of Spain dated 8,274 cal. BP was also detected with 35 variants from Haplotype
319 C. Together these results show that, although there was a deep genetic divide between the
320 western and the eastern Eurasia populations⁷⁷, both groups carried fragments of the three
321 haplotypes (**Figure S4**, Data provided in **Table S4**).



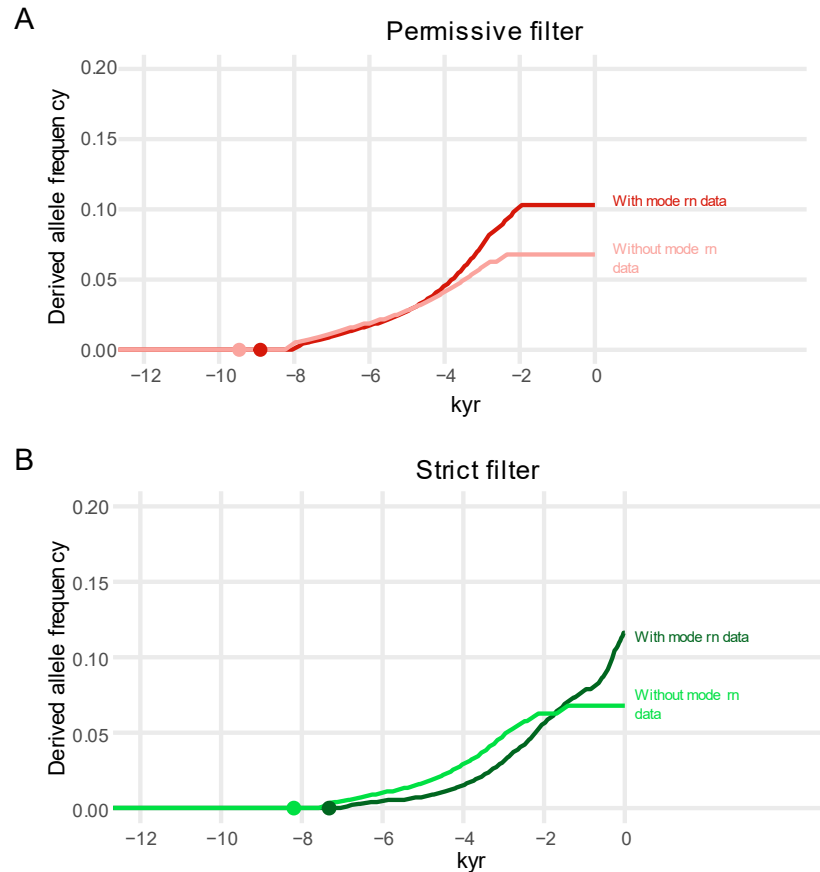
322

323 **Figure 4: Geographical locations of the ancient genomes genotyped for the CCR5delta32**

324 Map of distribution of ancient genomes genotyped with the permissive filter, faceted by four time periods and
325 colored based on the presence (yellow) or absence (red) of the CCR5delta32. Blue dots correspond to genomes
326 that are classified as having the deletion in the permissive filter genotype call set, but not having the deletion
327 according to the strict filter (The affected genomes are NEO300, NEO590, RISE509, VK316, VK342, see **Table**
328 **S4**).

329 **Evidence for ancient selection operating on the deletion**

330 Based on these results, we aimed to model the spatiotemporal frequency dynamics of the
331 CCR5delta32 allele across West Eurasia, to reconstruct the evolutionary history of this allele
332 and investigate the evidence in favor of positive selection at the locus. We used a modified
333 version of CLUES^{80,81} to infer allele frequency trajectories over time using ancient genomes.
334 In addition to our different genotype call set (strict filter and permissive filter), we evaluated
335 the trajectories if conditioned on allele frequencies observed in present day European
336 populations (**Figure 5, Figure S5**). Across these analyses, we observed a rapid rise in the
337 CCR5delta32 frequency until 2,000 years BP, followed by a stabilization of the frequency
338 until the present. When using the strict filter call set and modern ascertainment (**Figure 5B,**
339 **Figure S5**), however, we observed a very recent uptick in frequency. This was likely an
340 artifact of under-calling of the ancient genomes under this filtering scheme, causing the
341 model to reach present-day frequencies very quickly. We found significant evidence for
342 positive selection acting on the CCR5delta32 allele in the ancient past using both strict and
343 permissive filters against a neutral model (with p-values of 2.27e-3 and 9.34e-3,
344 respectively). In addition, when conditioning on present-day frequency, we obtained even
345 greater significance levels in favor of selection (with P values of 2.69e-8 and 3.41e-5 for
346 strict and permissive filters, respectively. We estimated that a large selection coefficient ($s >$
347 0.01) better explained the initial allele frequency rise. When using the strict filter calls, the
348 coefficient was predicted to be smaller ($s = 0.0198$ with conditioning, $s = 0.0152$ without
349 conditioning) than when using the “permissive filter” deletion calls ($s = 0.0327$ with
350 conditioning, $s = 0.0208$ without conditioning). The best posterior estimates for the age of the
351 CCR5delta32 deletion from the CLUES analysis were 9,128 and 7,714 years BP. **Table S5**
352 provides a detailed view of the results.



353

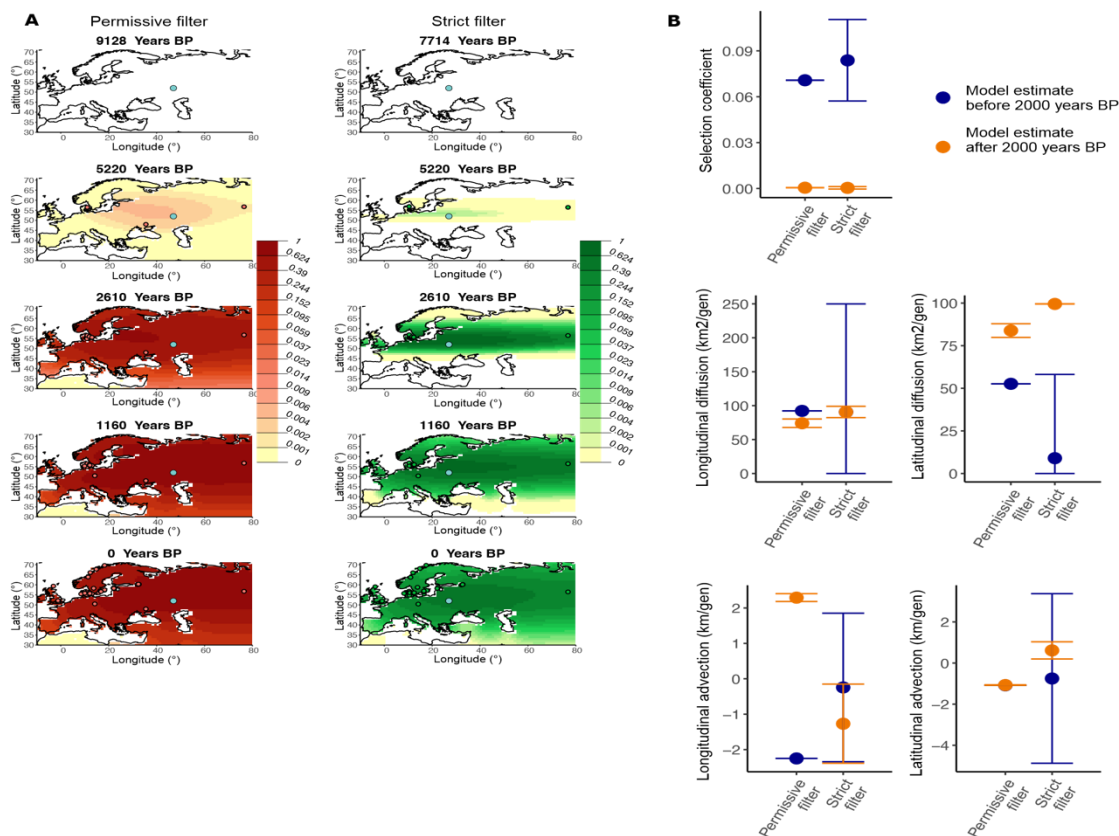
354 **Figure 5: CCR5delta32 allele frequency trajectory**

355 Maximum likelihood trajectory of the CCR5delta32 estimated using CLUES. **A)** Results obtained using
356 permissive filter. **B)** Results obtained using the strict filter. The dots in each figure represent the age estimate of
357 the variant either with or without conditioning on modern ascertainment.

358 **Spatiotemporal allele frequency dynamics**

359 To investigate the spread of the allele, we fitted a two-dimensional diffusion-advection
360 method that integrates present-day and ancient human genomes to infer allele frequency
361 dynamics across space and time⁸². The method infers parameters associated with how fast
362 the allele spreads across the landscape and how fast it increases in frequency locally due to
363 positive selection. It also estimates the likely geographic origin of the allele, given the data.
364 Because the CLUES analysis indicated that the allele frequency dynamics changed before
365 and after 2,000 years BP, we partitioned our spatial inference framework into these two
366 periods, allowing the method to find two separate selection coefficients and diffusion
367 parameters for each period (**Figure 6 and Table S6**). We inferred the allele origin to be in
368 the Western Eurasian Steppe region using both filtering schemes, with the strict filter placing
369 the allele more Eastwards compared to the permissive filter. This was followed by a rapid
370 longitudinal expansion in the earlier time period. Regardless of the call set, the selection
371 coefficient was inferred to be higher in the time period before 2,000 years BP (**Figure 6B**),

372 consistent with the CLUES analysis (**Table S5**), suggesting that selection likely operated
 373 early in the history of the allele (i.e., during the late Neolithic and Bronze Age). The selection
 374 coefficient estimate was higher using the “permissive filter” than when we used the strict
 375 filter, likely due to the younger allele age estimate (**Table S6**).



376

377 **Figure 6: CCR5delta32 allele frequency dynamics across West Eurasia**

378 **A)** Spatial allele frequency dynamics inferred by the diffusion-advection method. Left - permissive filter, right -
 379 strict filter. The green and red dots are genomes containing the deletion that are at least as old as the year
 380 indicated in each corresponding time slice. The light blue dot corresponds to the inferred origin of the allele. **B)**
 381 Parameter estimates from the spatiotemporal diffusion analysis used to generate allele frequency dynamic maps
 382 along with 95% confidence intervals. Results are shown for permissive and strict filter genotype call sets for time
 383 periods before and after 2000 years before present. The selection coefficient estimates indicate that selection
 384 likely operated early in the history of the allele, during the late Neolithic and Bronze Age.

385 **Discussion**

386 This study provides fundamental new insights into the evolutionary history of the
 387 CCR5delta32 allele. Our discovery and mapping of the Haplotypes A, B, and C in present-
 388 day genomes led us to develop a probabilistic model, HAPI, to investigate the CCR5delta32
 389 allele in ancient genomes. The model allowed us to reliably detect CCR5delta32 allele in
 390 genomes with as little as 0.3X coverage. Based on this, we date the deletion to be at least
 391 7,000 years BP in age, possibly arising among peoples occupying the Western Eurasian

392 steppe region in the Neolithic. We also show that the CCR5delta32 allele was exposed to
393 positive selection during the late Neolithic and Bronze Age, followed by stability in the AF
394 until the present day.

395 Applying the knowledge of Haplotype A, combined with the evidence from HaploNet, we can
396 now confirm earlier studies' presumption of a European origin of the CCR5delta32 allele^{12,55}.
397 The Columbian Exchange, which was considered to have facilitated genetic admixture
398 among three groups – African, European, and Native American^{83,84}, can account for the
399 significantly higher recombination rates we observe from Haplotype B and C in Latin
400 America compared to European populations, along with the higher AF from some of the
401 variants including in haplotype C (**Figure 2B, Figure S1C**). Thus, we can propose to include
402 the CCR5delta32 allele, together with the variants rs113341849 and rs113010081, as
403 European ancestry-informative markers. Furthermore, the CCR5delta32 genotype can be
404 reliably imputed from SNP arrays using the two $r^2=1$ tag SNPs (rs113341849 and
405 rs113010081), as they are located on each side of the CCR5delta32 allele and therefore will
406 encounter most possible recombinations of the Haplotype A.

407 Previous studies investigating the evolutionary history of the CCR5delta32 allele have been
408 either restricted to contemporary individuals¹² or used very few ancient genomes from
409 limited geographic areas^{46,49,63}. Here we present results obtained using a large
410 comprehensive set of ancient genomes (>800 genomes) combined with modern genomes.
411 The CLUES analysis revealed that the allele rose quickly in frequency in the period before
412 2,000 years BP, followed by a period of AF stabilization, over which the allele remained at
413 around 10% frequency from 2,000 BP onwards. This agreed with findings from Bouwman et
414 al. 2017⁴⁶ and Hummel et al. 2005⁴⁹, which posited a period of recent allele stability over the
415 past millennium in Central and North Germany. Based on our data, the allele had an origin in
416 the Western steppe and a fast rapid diffusion eastwards and westwards early in its history,
417 partly coinciding with the eastward movements from the Steppe during the Bronze Age^{76,85}.
418 We note, though, that the origin of the allele inferred by the model is highly dependent on the
419 first instances of the allele in the data, and thus is highly dependent on the mode of deletion
420 calling. Under the curated calling schemes, the lower inferred counts of the allele during the
421 Mesolithic and Neolithic lead the model to estimate a fast longitudinal diffusion, as the most
422 likely allele frequency surface rapidly shifts from complete absence to widespread presence
423 of the allele in distant regions across the continent. The rapid longitudinal spread of the allele
424 is consistent with previous evidence for long-distance dispersal of the allele¹² though our
425 ancient data suggests this dispersal occurred earlier than previously thought. Our estimated
426 age of the allele is consistent with a more ancient origin as postulated in Sabeti et al 2005¹⁵,

427 rather than a recent origin as suggested in other studies ^{11,17}. All age estimates we obtained
428 were older than 7,000 years BP (posterior estimates 9,128 and 7,714 years BP).

429 We found significant evidence of positive selection acting on the CCR5delta32 allele in the
430 ancient past, when fitting the data to the CLUES model. When we conditioned the CLUES
431 trajectories on reaching the frequencies observed in present-day data, they result in stronger
432 rises in frequency compared to using ancient data alone, which in turn results in more
433 significant p-values in the rejection of neutrality. This likely indicates an undercounting of the
434 allele in the more ancient time periods. Regardless of the calling scheme, we found
435 significantly large selection coefficients when deploying the spatiotemporal spread model,
436 particularly in the early time period, but no evidence for selection after 2,000 years BP. Of
437 note, however, is that the spatiotemporal model is deterministic, and thus necessarily
438 underestimates the amount of allele frequency stochasticity that occurs during the period
439 under study, so the selection coefficient inferred under this model may be an overestimate.
440 Very recently, Le and colleagues, 2022⁶¹, found no evidence for the selection of the
441 CCR5delta32 allele during ancient times. However, that result was obtained using a
442 CCR5delta32 proxy SNP, rs73833033, that we found to have $r^2 < 0.8$ and therefore the
443 proxy SNP was not included in Haplotype A. Their analysis therefore did not adequately
444 count CCR5delta32 alleles.

445 The notable increase in CCR5delta32 allele frequency prior to the Iron Age implies that the
446 high frequencies of this allele in modern-day Europe cannot be caused by Medieval Plague
447 as hypothesized previously^{46,49}. Instead, the selection signature may have resulted from
448 pressures exerted by previous outbreaks or other pathogens that existed in the past ^{14,16,86}.
449 The observed spread was also not consistent with the Viking-spread hypothesis ⁵⁰. Likewise,
450 our age estimation of the CCR5delta32 allele does not support this hypothesis. Instead, the
451 rapid longitudinal spread that we infer (approx. 60-100 km² per generation, **Table S6**) and
452 the rapid rise in frequency observed during the Bronze Age suggests a possible spread
453 associated with the Late Neolithic and Early Bronze Age expansion of steppe-related
454 ancestry into Europe ^{76,85}.

455 Today, immunological genetic signatures by selection and/or adaptation through admixture
456 can be observed in the human genome ^{2,3}. Our data shows that the CCR5delta32
457 (Haplotype A), could very well be among these genetic signatures. We cannot point out a
458 direct cause for the increase of CCR5delta32's allele frequency during the Neolithic and
459 early Bronze Age, but it is clear that Haplotype B did not undergo the same evolutionary
460 trajectory. The key to understanding the driving forces for the CCR5delta32 deletion is
461 challenged by immune system redundancy and the immune gene pleiotropy ^{87,88}. A

462 hypothesis could be that the CCR5delta32 allele with the 86 tag variants is associated with
463 cytokine/chemokine profiles reflecting immune tolerance, which has been shown to be under
464 selection during the Neolithic age ⁹.

465 Finally, the fact that individuals bearing the CCR5delta32 allele also harbor a defined
466 haplotype widens the complexity of the deletion effects. The CCR5delta32 deletion has been
467 studied extensively for more than two decades, especially due to its strong link to HIV-1
468 infection resistance and thereby the potential to target CCR5 for HIV treatment and for HIV
469 pre/post-exposure prophylaxis medicine ^{25,27,89,90}. These therapeutic approaches include
470 gene-editing techniques like CRISPR, CCR5 blockade using antibodies or antagonists, or
471 combinations of both ^{25,91}. Further, the CCR5delta32 allele can be viewed as a pleiotropic
472 variant, due to its influence on multiple phenotypic traits, e.g. autoimmune and inflammatory
473 diseases, cardiovascular diseases, neurodegenerative disease, and cancer ^{20,29,32,44,92}. It is
474 possible that some of the CCR5delta32 tag SNPs contribute to the pleiotropic nature of
475 CCR5delta32, although *in silico* analysis shows no direct clinical significance. More
476 precisely, the gene expression of cytokine receptors (*CCR3*, *CCR2*) and *CCRL2* might be
477 affected by one or more of the tag SNPs, leading to modulation of chemokine-chemokine
478 receptor signal transduction ^{19,92,93}. This calls for further studies to elucidate these possible
479 direct or indirect effects. Thus, the tag SNPs should be considered when analyzing causes
480 of the CCR5delta32 pleiotropic effects and when developing therapeutic approaches, based
481 on mimicking the naturally occurring CCR5delta32 genotype-phenotype correlations.
482 Therefore, our results point in a direction of new complex CCR5delta32 genotype-haplotype-
483 phenotype relationships, which demand consideration when targeting the CCR5 receptor for
484 therapeutic strategy.

485 **Limitations of the study**

486 We have striven to evaluate our results in light of the challenges encompassed by ancient
487 datasets, such as DNA damage sequencing profiles, low-coverage genomes, and patchy
488 spatiotemporal sampling. We developed and applied the HAPI model with awareness of the
489 three different CCR5 haplotypes. By applying different filter schemes to the CCR5delta32
490 classification of the ancient genomes, we were able to inspect the impact on the ancient
491 sample sizes. Despite these considerations, our results need to be verified through further
492 genome sampling, especially from the European Neolithic period.

494 **Acknowledgments**

495 K.R., L.C. and S.R. were supported by the Novo Nordisk Foundation (grant NNF14CC0001
496 and NNF21SA0072102). E.K.I.P was supported by the Lundbeck Foundation (grant R302-

497 2018-2155) and the Novo Nordisk Foundation (grant NNF18SA0035006). E.K.I.P. and
498 R.A.M. were additionally supported by a Villum Young Investigator grant given to F.R
499 (project no. 00025300).

500 **Author contributions**

501 Conceptualization: K.R., L.C., M.E.A., E.W., F.R., and S.R.; Methodology: K.R., L.C., and
502 S.R.; Data curation: M.S., M.E.A., and E.W.; Investigation: K.R., L.C., R.A.M., J.M., and
503 E.K.I-P.; Software: K.R., L.C., R.A.M., J.M., T.S.K., and E.K.I-P.; Formal Analysis: M.S.;
504 Writing – original draft: K.R., L.C., R.A.M., F.R., and S.R.; Writing – review and editing: all
505 authors.; Resources: E.W. and M.E.A.; Supervision: F.R. and S.R.

506 **Declaration of interests**

507 The authors declare no competing interests.

508 **Supplemental Files**

509 **Supplementary Figures 1-6**

510 **Figure S1:** Details information of Haplotype A: LD statistics, genomic location and the AF of
511 the proxy variants.

512 **Figure S2:** Workflow of the data analysis on the simulated ancient samples and details of
513 the overlapping lengths.

514 **Figure S3:** Assessing GATK HaplotypeCaller, Mismatch Rates, and HAPI performance at
515 different overlapping lengths.

516 **Figure S4:** Ancient and present sample distribution.

517 **Figure S5:** Allele frequency trajectory inferred by CLUES.

518 **Figure S6:** Schema of the algorithm behind HAPI.

519 **Supplementary Tables 1-6, SI_CCR5delta32_Tables.xlsx**

520 **Table S1:** CCR5delta32 proxy variants with an $r^2 > 0.8$, in the five EUR KGP3 populations.
521 The table includes multiple sheets (**A-E**).

522 **Table S2:** Sample ID (1KGP3) for genomes identified with CCR5delta32, Haplotype A, B,
523 and C, and their homologous recombinations. The table includes multiple sheets (**A-F**) and
524 related to **Figure 1B** and **2A**.

525 **Table S3:** Inquiry from CCR5delta32 and Haplotype A's tag SNPs in the GWAS catalog and
526 Phenoscanner database. The table includes two sheets (**A-B**).

527 **Table S4A:** Detailed view of ancient genomes as classified by the HAPI model and the
528 curated filters. The table includes multiple sheets (**A-C**).

529 **Table S5:** Summary of parameter estimates obtained using CLUES.

530 **Table S6:** Parameter estimates obtained using the Spatio-Temporal diffusion model.

531 **Materials and Methods**

532 **Data**

533 The modern dataset is constituted of whole-genome sequencing data of 2,535 individuals
534 from 26 populations which were generated by the 1000 Genomes Project Phase 3 (1KGP3,
535 <http://www.1000genomes.org/>), assigned to the following 5 super populations: African (AFR),
536 admixed from the Americas (AMR), East Asian (EAS), South Asian (SAS), and European
537 (EUR)⁹⁴.

538 The ancient dataset comprises a total of 860 shotgun-sequenced genomes from various
539 regions across Eurasia. The dataset includes genomes from the Stone Age⁷⁷ (NEO samples
540 ID, age: 11,000-3,000 BP), the Bronze Age⁷⁶ (RISE samples ID age: 11,000-3,000 BP), and
541 the Viking Age⁷⁵ (VK samples ID, age: 2450 BP - CE 1600). The dataset for the Stone Age
542 targets the Mesolithic and Neolithic Age and includes 317 genomes from archaeological
543 sites across Europe. The sampling collection was particularly dense in northern Europe,
544 involving 100 samples from Denmark (ENA Project ID: to be published soon). For the
545 Bronze Age, 101 genomes were included from archaeological sites across Europe and
546 Central Asia (ENA Project ID: PRJEB9021). Finally, the Viking Age dataset consisted of 442
547 genomes obtained from archaeological sites across Europe and Greenland, with a dense
548 collection in Northern Europe (ENA Project ID: PRJEB37976).

549 **Identification of the CCR5delta32 deletion and the haplotypes:**

550 We used the LDLink 3.0 web tool, which includes the LDmatrix and LDpair modules⁹⁵, to
551 identify the CCR5delta32 proxy SNPs within the European (EUR) population of the 1KGP3
552 dataset (Table S1). These results were then explored in additional 1KGP3 populations. The
553 haplotypes were called with samtools mpileup⁹⁶ (**Table S2**), by using the region
554 (chr3:46200000-46800000) from all available 1KGP3 whole-genome bam files. To determine
555 the effect of the 86 tag SNPs belonging to Haplotype A, we employed Ensembl Variant
556 Effect Predictor (VEP)⁶⁵, while the GWAS catalog⁶⁷ and PhenoScanner V2^{71,72} were used to
557 evaluate possible genotype-phenotype associations of tag SNPs (**Table S3**). All annotations
558 refer to the human reference genome GRCh37 assembly.

559 **Development of the Haplotype-Aware Probabilistic model for Indels (HAPI)**

560 **Simulations**

561 We used gargammel (v. 1.1.2)⁹⁷ to simulate a total of 144 ancient genomes at 8 different
562 coverages (0.3X, 0.5X, 0.8X, 1X, 1.5X, 2X, 5X, 10X), using empirical read length
563 distributions and post-mortem damage derived from 6 real ancient genomes (NEO78,
564 NEO79, NEO752, VK287, VK543, VK526) from our dataset (**Figure S2A**). We simulated 48
565 genomes for each genotype (RR, RD, DD) using combinations of two versions of the
566 GRCh37 human genome reference: a canonical, and one in which we manually added the
567 CCR5delta32 deletion and the 86 variants from the haplotype (here referred to as “collapsed
568 reference”) using the tool FastaAlternateReferenceMaker from GATK (v.4.1.8.1).⁷⁴ The
569 reads were simulated from HiSeq 2500 Illumina single-end runs with a length of 81 base
570 pairs including adapters.

571 **Processing of simulated and ancient genomes**

572 For the simulated genomes we used AdapterRemoval (v.2.1.3)⁹⁸ with parameters “--mm 3 --
573 minlength 30 --minquality 2” to trim the reads from the simulated genomes at a length of at
574 least 30 bp and to remove bases with quality 2 or less. We used bwa aln (v.7.16a)⁹⁹ to map
575 the adaptor-trimmed reads to both the canonical and the collapsed human reference
576 genome (GRCh37) with seed disabled (parameter “-l 1024”) to allow for higher sensitivity in
577 ancient DNA.¹⁰⁰ We sorted the resulting alignments with samtools (v.1.9)¹⁰¹, removed
578 duplicates with Picard MarkDuplicates (v1.128)¹⁰² and realigned the reads using GATK
579 (v3.3.0)⁷⁴ with Mills and 1000G gold-standard insertions and deletions. Finally, the alignment
580 files were converted to cram with samtools view and indexed with samtools index. The
581 ancient genomes were aligned to both the canonical and the collapsed human genome
582 reference using the same pipeline as the simulated genomes except that the read groups
583 were first merged to the library level, then duplicates were removed using Picard
584 (v.1.128)¹⁰², and then the files were merged to sample level. Sample level bam files were
585 subsequently realigned using GATK (v.3.3.0)⁷⁴ and then converted and indexed to cram
586 format. The workflows were implemented using Snakemake (v.5.12.0).¹⁰³

587 **Processing of the 15 human genomes from the 1KGP3 dataset**

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

594 Haplotype-Aware Probabilistic model for Indels (HAPI)

595 We developed a probabilistic model to combine the information from the 4 variants in the
596 highest pairwise LD with the deletion (rs113341849, rs113010081, rs11574435, and
597 rs79815064, $r^2 > 0.90$, CEU) as a Prior, and the information from the reads mapping the
598 CCR5delta32 deletion region as a Likelihood. The variants were called using samtools
599 mpileup as implemented in pysam (v.0.16.0.1) in python. For a detailed overview of the
600 algorithm see **Figure S6**. For each sample, we calculated the posterior probability for each
601 deletion genotype (RR, RD, DD) as

$$602 \quad \quad \quad P(G|D) = P(G)P(D|G)P(D) \quad \quad \quad (Eq. 1)$$

605 where $P(G)$ is the prior probability of the given deletion genotype calculated using the
606 information from the 4 variants (see eq. 2 below), $P(D|G)$ is the likelihood of the deletion
607 genotype based on the reads mapping to the CCR5delta32 deletion region (either canonical
608 or collapsed reference) (see eq. 3 below) and $P(D)$ is the marginal probability of the data.
609 For the prior, we calculated the posterior probability of each deletion genotype using a
610 simple bayesian genotyper based on the one developed by Mckenna et al, 2010¹⁰⁴ as

$$611 \quad \quad \quad P(G|r) = P(G) P(r|G)P(r) \quad \quad \quad (Eq. 2)$$

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

$$619 \quad \quad \quad p(b|G) = p(b|\{ref, alt\}) = \frac{1}{2}p(b|ref) + \frac{1}{2}p(b|alt) \quad \quad \quad (Eq. 3)$$

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

626

627
$$p(b|A) = \begin{cases} \frac{e}{3} : b \neq A \\ 1 - e : b = A' \end{cases}$$

628 (Eq. 4)

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

642
$$p(r|R) = \begin{cases} 1 - \left(\frac{1}{\delta}\right)^2 : r = R \\ \frac{1 - \left(1 - \left(\frac{1}{\delta}\right)^2\right)}{2} : r \neq R \end{cases}$$

643 (Eq. 5)

644 an adaptation from¹⁰⁴, where R is the specific reference used for the mapping, i.e. canonical
645 or collapsed. We then calculated the probability of observing the reads given the deletion
646 genotype with

647
$$p(r|G) = p(r|ref, del) = \frac{1}{2}p(r|ref) + \frac{1}{2}p(r|del)$$

648 (Eq. 6)

649 The genotype likelihood for each reference was then calculated with $p(D|G) = p(r|G)$. The
650 final genotype likelihood for each deletion genotype was computed as the joint probability of
651 the likelihoods for the individual references (canonical and collapsed) with $p(D|G) =$
652 $p(D|G)$. Finally, for each sample, HAPI outputs three posterior probabilities for each deletion
653 genotype RR, RD, DD, summing up to 1.

654 **Determining minimum overlapping length**

655 We assigned a minimum overlapping length (δ) to each read mapping the deletion region,
656 either on the canonical or collapsed reference. The δ represents the minimum number of

657 nucleotides the reads overlap either the 5' or the 3' of the locus coordinates (See **Figure**
658 **S2B** for a detailed clarification example). The CCR5delta32 has 4 equivalent
659 representations, each with its own coordinates
660 (<https://varsome.com/variant/hg19/rs333?annotation-mode=germline>). Thus, for each read
661 mapping to the canonical reference, we calculated its minimum overlapping length δ by
662 averaging across the δ s calculated for each of the representations' coordinates. A value of δ
663 = 32 was assigned to the reads overlapping both the starting and ending coordinates of the
664 canonical reference. For the collapsed reference, we calculated δ based on the coordinate
665 3:46414943 (GRCh37). For all the reads mapped, only those having a value of δ equal or
666 greater than 6 were kept. The reads mapping to both deletion regions (from the canonical
667 and collapsed references) were assigned to the reference to which they mapped with the
668 lowest number of mismatches. This was done because, during the alignment of the
669 simulated ancient DNA genomes, we observed that reads originating from the canonical
670 deletion region mapped to the collapsed deletion region with a significantly higher number of
671 mismatches compared to when they mapped to the canonical deletion region (and vice-
672 versa) (signed test, p-value < 0.0001) (**Figure S3B**). Reads mapping to both references with
673 the same number of mismatches were assigned to the reference to which they mapped with
674 the highest δ . Reads mapped to both references with same number of mismatches and the
675 same δ were discarded. The read mappings were analyzed using pysam (v. 0.16.0.1).

676 **Optimizing the model**

677 During the developmental stage, we explored different approaches to optimize the model. To
678 investigate how the minimum overlapping length of the reads across the deletion region
679 influences the performance of HAPI, we ran the model using 10 different δ thresholds, from
680 1 to 10, on the simulated data. As expected, increasing the δ threshold resulted in an
681 increase in the performance of the model from an MCC of 0.75 with $\delta=1$ to a value of 0.873
682 with $\delta=10$ (**Figure S3C**), but at the expense of having less reads satisfying the threshold and
683 thus less genomes recovered (121 with $\delta=1$ and 107 with $\delta=10$) having at least one read
684 mapping to the deletion region. We arbitrarily selected the δ threshold of 6 (corresponding to
685 6 nucleotides flanking each side of the breakpoint) because we found it to be a good
686 compromise between performance and the number of genomes recovered (MCC = 0.81,
687 genomes recovered 116). Additionally, we investigated rescaling the bam files to account for
688 DNA damage and excluding reads without a perfect match in the alignment. Here, we found
689 that rescaling did not have any significant effect on the performance of the model and that
690 using only perfect match reads improved the performance of the model but at the expense of
691 losing 22 genomes. These strategies were therefore not included in the final model.

692 **Applying the model to ancient genomes**

693 To be analyzed by the model, a genome must have at least one read mapping to the
694 CCR5delta32 deletion region with a minimum overlapping length δ of 6. The model was run
695 on the simulated, ancient, and 1KGP3 genomes and we classified the genomes as being
696 RR, RD, DD based on the highest posterior probability among the three, with a classification
697 threshold of 0.5. To take into account the fact that the flanking regions of the deletion include
698 repeated nucleotides, and that two of the 4 variants used for calculating the haplotype-
699 informed prior look like ancient DNA damage, we applied two manual curation steps. In the
700 first one “permissive filter” we manually re-classified some genomes if they had only 1 SNP
701 called, and the same SNP looked like aDNA damage (G to A, and C to T). In the second one
702 “strict filter”, we re-classified the genomes which we think have Haplotype B instead of the
703 deletion, because of no reads covering the deletion but only the reference.

704 **Benchmarking using Haplotype Caller**

705 The 15 genomes selected from the 1KGP3 population were processed using
706 HaplotypeCaller from GATK (v. 4.1.9.0)⁷⁴ to produce vcf files with SNP and indels calls using
707 the following options: --intervals 3:46277577-46457412 --interval-padding 100 --stand-call-
708 conf 30.0 -ERC BP_RESOLUTION. The vcf files were left aligned and normalized using
709 bcftools norm (v. 1.10.2)¹⁰⁵ and then processed in R (v.4.0.3)¹⁰⁶.

710 **Local ancestry of individuals harboring CCR5delta32 deletion in the 1KGP3**

711 We used HaploNet⁷³ on the full 1KGP3 dataset to generate haplotype cluster likelihoods in
712 windows along the genome with default parameters of “haplonet train” besides “--x_dim
713 512”, such that the genomic windows had a fixed size of 512 SNPs. We used the haplotype
714 cluster likelihoods to estimate ancestry proportions with an assumption of K=5 ancestral
715 populations, representing the 5 super populations of 1KGP3, using the “haplonet admix”
716 command. The haplotype cluster likelihoods and ancestry proportions were then finally used
717 to infer local ancestry for all genomic windows in the individuals with the CCR5delta32 locus
718 using the “haplonet fatash” command.

719 **CLUES analysis**

720 To reconstruct the allele frequency trajectory of the CCR5delta32 deletion, we used a
721 modified version of the software CLUES, adapted for time-series data^{80,81}. We converted the
722 output of HAPI into hard called genotypes, using the outputs from the permissive and strict
723 filters. We then conditioned the inference of the trajectories on a present-day frequency of
724 0.1237 and an estimate of the effective population size history, inferred from genomes in the
725 Finnish (FIN), British (GBR), and Tuscan (TSI) populations from the 1KGP3⁹⁴, using the

726 software Relate¹⁰⁷. The code to reproduce these analyses is available in the Github
727 repository https://github.com/ekirving/ccr5_paper.

728 **Estimating the age of CCR5delta32**

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

739 **Method for modeling the spatiotemporal diffusion of the deletion allele**

740 To model the diffusion of the CCR5delta32 allele across space and time, we use a method
741 described in Muktupavela et al. 2021⁸² and available from:
742 <https://github.com/RasaMukti/stepadna>. We adapted the method so that the input genotype
743 calls for each individual corresponded to the genotype with the highest posterior probability
744 obtained from HAPI. To do this, we modified the equation (5) from⁸²:

745

$$746 \quad L(d_i, a_i) = \sum_{h=0}^2 P[d_i, a_i | g_i = h] P[g_i = h | p(x_i, y_i, t_i)]$$

747 (Eq. 7)

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

755

$$756 \quad \frac{\partial p}{\partial t} = \frac{1}{2} \sigma_x^2 \frac{\partial^2 p}{\partial x^2} + \frac{1}{2} \sigma_y^2 \frac{\partial^2 p}{\partial y^2} + v_x \frac{\partial p}{\partial x} + v_y \frac{\partial p}{\partial y} + ps(1 - p)$$

757 (Eq. 8)

758 where σ_x , σ_y are the longitudinal and latitudinal diffusion coefficients, respectively, v_x and v_y
759 represent the longitudinal and latitudinal advection coefficients, respectively, and s is the
760 selection coefficient. We modified the equation so that the likelihood of the genotype g_i is
761 equal to 1 if the genotype corresponds to the genotype with the highest posterior probability
762 and 0 otherwise.

763

764 We applied the method to the different deletion call datasets, combining them with the
765 present-day geographically-spread deletion calls compiled in Novembre et al. 2005¹² (**Figure**
766 **S4**). We removed genomes that were outside of the geographic area bounded latitudinally
767 by 30°N and 75°N and longitudinally by 10°W and 80°E.

768 Maximum likelihood optimization was carried out by initializing 50 points in the multi-
769 parameter space and using a first round of simulated annealing¹⁰⁸ followed by a run of the L-
770 BFGS-B algorithm¹⁰⁹ to refine the optimization.

771 **Code availability**

772 The simulated sequence data mapped to the GRCh37 and to the Collapsed reference,
773 restricted to the CCR5Delta32 region, are available at
774 <https://doi.org/10.17894/ucph.a31d9052-546d-4f8f-8e16-e5bd896df67b> together with the
775 results of running HAPI on them. The HAPI model is available as a pip package at
776 <https://pypi.org/project/hapi-pyth/> and instructions on how to install and run it are available at
777 <https://github.com/RasmussenLab/HAPI>. The code for the CLUES analysis is available at
778 https://github.com/ekirving/ccr5_paper. The code for reproducing the spatiotemporal
779 diffusion analysis can be found at https://github.com/RasaMukti/ccr5delta32_analysis. Any
780 additional information required to reanalyze the data reported in this paper is available from
781 the lead contact upon request.

782 **References**

- 783 1. Morens, D. M. & Fauci, A. S. Emerging Pandemic Diseases: How We Got to COVID-
784 19. *Cell* **182**, 1077–1092 (2020).
- 785 2. Human Immunology through the Lens of Evolutionary Genetics. *Cell* **177**, 184–199
786 (2019).
- 787 3. Kerner, G. *et al.* Genetic adaptation to pathogens and increased risk of inflammatory
788 disorders in post-Neolithic Europe. *Cell Genomics* **3**, 100248 (2023).

- 789 4. Karlsson, E. K., Kwiatkowski, D. P. & Sabeti, P. C. Natural selection and infectious
790 disease in human populations. *Nat. Rev. Genet.* **15**, 379–393 (2014).
- 791 5. Barreiro, L. B. & Quintana-Murci, L. From evolutionary genetics to human immunology:
792 how selection shapes host defence genes. *Nat. Rev. Genet.* **11**, 17–30 (2010).
- 793 6. Casadó-Llombart, S. *et al.* Contribution of Evolutionary Selected Immune Gene
794 Polymorphism to Immune-Related Disorders: The Case of Lymphocyte Scavenger
795 Receptors CD5 and CD6. *Int. J. Mol. Sci.* **22**, 5315 (2021).
- 796 7. Deschamps, M. *et al.* Genomic Signatures of Selective Pressures and Introgression
797 from Archaic Hominins at Human Innate Immunity Genes. *Am. J. Hum. Genet.* **98**, 5–
798 21 (2016).
- 799 8. Quach, H. *et al.* Genetic Adaptation and Neandertal Admixture Shaped the Immune
800 System of Human Populations. *Cell* **167**, 643-656.e17 (2016).
- 801 9. Domínguez-Andrés, J. *et al.* Evolution of cytokine production capacity in ancient and
802 modern European populations. *eLife* **10**, e64971 (2021).
- 803 10. Faure, E. & Royer-Carenzi, M. Is the European spatial distribution of the HIV-1-
804 resistant CCR5- Δ 32 allele formed by a breakdown of the pathocenosis due to the
805 historical Roman expansion? *Infect. Genet. Evol.* **8**, 864–874 (2008).
- 806 11. Libert, F. *et al.* The Δ ccr5 mutation conferring protection against HIV-1 in Caucasian
807 populations has a single and recent origin in Northeastern Europe. *Hum. Mol. Genet.* **7**,
808 399–406 (1998).
- 809 12. Novembre, J., Galvani, A. P. & Slatkin, M. The geographic spread of the CCR5 Δ 32
810 HIV-resistance allele. *PLoS Biol.* **3**, 1954–1962 (2005).
- 811 13. Galvani, A. P. & Novembre, J. The evolutionary history of the CCR5- Δ 32 HIV-
812 resistance mutation. *Microbes Infect.* **7**, 302–309 (2005).
- 813 14. Galvani, A. P. & Slatkin, M. Evaluating plague and smallpox as historical selective
814 pressures for the CCR5- Δ 32 HIV-resistance allele. *Proc. Natl. Acad. Sci. U. S. A.* **100**,
815 15276–15279 (2003).
- 816 15. Sabeti, P. C. *et al.* The Case for Selection at Ccr5- Δ 32. *PLOS Biol.* **3**, e378 (2005).

- 817 16. Duncan, SR., Scott, S. & Duncan, C. J. Reappraisal of the historical selective pressures
818 for the CCR5-Δ32 mutation. *J. Med. Genet.* **42**, 205–208 (2005).
- 819 17. Stephens, J. C. *et al.* Dating the origin of the CCR5-Delta32 AIDS-resistance allele by
820 the coalescence of haplotypes. *Am. J. Hum. Genet.* **62**, 1507–1515 (1998).
- 821 18. Oppermann, M. Chemokine receptor CCR5: insights into structure, function, and
822 regulation. *Cell. Signal.* **16**, 1201–1210 (2004).
- 823 19. Hughes, C. E. & Nibbs, R. J. B. A guide to chemokines and their receptors. *Febs J.*
824 **285**, 2944–2971 (2018).
- 825 20. Ellwanger, J. H., Kaminski, V. de L., Rodrigues, A. G., Kulmann-Leal, B. & Chies, J. A.
826 B. CCR5 and CCR5Δ32 in bacterial and parasitic infections: Thinking chemokine
827 receptors outside the HIV box. *Int. J. Immunogenet.* **47**, 261–285 (2020).
- 828 21. Deng, H. *et al.* Identification of a major co-receptor for primary isolates of HIV-1. *Nature*
829 **381**, 661–666 (1996).
- 830 22. Dragic, T. *et al.* HIV-1 entry into CD4+ cells is mediated by the chemokine receptor CC-
831 CKR-5. *Nature* **381**, 667–673 (1996).
- 832 23. Samson, M. *et al.* Resistance to Hiv-1 Infection in Caucasian Individuals Bearing
833 Mutant Alleles of the Ccr-5 Chemokine Receptor Gene. *Nature* **382**, 722–725 (1996).
- 834 24. Latinovic, O. S., Reitz, M. & Heredia, A. CCR5 Inhibitors and HIV-1 Infection. *J. AIDS*
835 *HIV Treat.* **1**, 1–5 (2019).
- 836 25. Mohamed, H. *et al.* Targeting CCR5 as a Component of an HIV-1 Therapeutic Strategy.
837 *Front. Immunol.* **12**, 816515 (2022).
- 838 26. Hsu, J. *et al.* HIV-1 remission and possible cure in a woman after haplo-cord blood
839 transplant. *Cell* **186**, 1115-1126.e8 (2023).
- 840 27. Hütter, G. *et al.* Long-term control of HIV by CCR5 Delta32/Delta32 stem-cell
841 transplantation. *N. Engl. J. Med.* **360**, 692–698 (2009).
- 842 28. Gupta, R. K. *et al.* HIV-1 remission following CCR5Δ32/Δ32 haematopoietic stem-cell
843 transplantation. *Nature* **568**, 244–248 (2019).

- 844 29. Ellwanger, J. H. *et al.* Beyond HIV infection: Neglected and varied impacts of CCR5
845 and CCR5Δ32 on viral diseases. *Virus Res.* **286**, 198040 (2020).
- 846 30. Hachim, M. Y. *et al.* C-C chemokine receptor type 5 links COVID-19, rheumatoid
847 arthritis, and Hydroxychloroquine: in silico analysis. *Transl. Med. Commun.* **5**, 14
848 (2020).
- 849 31. Mehlotra, R. K. New Knowledge About CCR5, HIV Infection, and Disease Progression:
850 Is “Old” Still Valuable? *AIDS Res. Hum. Retroviruses* **36**, 795–799 (2020).
- 851 32. Rautenbach, A. & Williams, A. A. Metabolomics as an Approach to Characterise the
852 Contrasting Roles of CCR5 in the Presence and Absence of Disease. *Int. J. Mol. Sci.*
853 **21**, (2020).
- 854 33. Weissberg, O., Gorohovski, A., Shay, D. R. & Frenkel-Morgenstern, M. Significant
855 Effects of CCR5delta32 Polymorphism on Alzheimer’S Disease, Neurological
856 Disorders, Cancer, Diabetes and Viral Infection in the Worldwide Population. *Am. J.*
857 *Biomed. Sci. Res.* **13**, 177 (2021).
- 858 34. Bernas, S. N. *et al.* CCR5Δ32 mutations do not determine COVID-19 disease course.
859 *Int. J. Infect. Dis. IJID Off. Publ. Int. Soc. Infect. Dis.* **105**, 653–655 (2021).
- 860 35. Chua, R. L. *et al.* COVID-19 severity correlates with airway epithelium–immune cell
861 interactions identified by single-cell analysis. *Nat. Biotechnol.* **38**, 970–979 (2020).
- 862 36. Cuesta-Llavona, E. *et al.* Variant-genetic and transcript-expression analysis showed a
863 role for the chemokine-receptor CCR5 in COVID-19 severity. *Int. Immunopharmacol.*
864 **98**, 107825 (2021).
- 865 37. Gómez, J. *et al.* The CCR5-delta32 variant might explain part of the association
866 between COVID-19 and the chemokine-receptor gene cluster. *medRxiv*
867 2020.11.02.20224659 (2020) doi:10.1101/2020.11.02.20224659.
- 868 38. Hubacek, J. A. *et al.* CCR5Delta32 deletion as a protective factor in Czech first-wave
869 COVID-19 subjects. *Physiol. Res.* **70**, 111–115 (2021).

- 870 39. Panda, A. K., Padhi, A. & Prusty, B. A. K. CCR5 Δ 32 minorallele is associated with
871 susceptibility to SARS-CoV-2 infection and death: An epidemiological investigation.
872 *Clin. Chim. Acta Int. J. Clin. Chem.* **510**, 60–61 (2020).
- 873 40. Patterson, B. K. *et al.* CCR5 inhibition in critical COVID-19 patients decreases
874 inflammatory cytokines, increases CD8 T-cells, and decreases SARS-CoV2 RNA in
875 plasma by day 14. *Int. J. Infect. Dis.* **103**, 25–32 (2021).
- 876 41. Starcevic Cizmarevic, N., Kapovic, M., Roncevic, D. & Ristic, S. Could the CCR5-
877 Delta32 mutation be protective in SARS-CoV-2 infection? *Physiol. Res.* **70**, S249–S252
878 (2021).
- 879 42. Xu, M. CCR5- Δ 32 biology, gene editing, and warnings for the future of CRISPR-Cas9
880 as a human and humane gene editing tool. *Cell Biosci.* **10**, 48 (2020).
- 881 43. Pieczynski, J. N. & Kee, H. L. “Designer babies?!” A CRISPR-based learning module
882 for undergraduates built around the CCR5 gene. *Biochem. Mol. Biol. Educ.* **49**, 80–93
883 (2021).
- 884 44. Li, T. & Shen, X. Pleiotropy Complicates Human Gene Editing: CCR5 Δ 32 and Beyond.
885 *Front. Genet.* **10**, (2019).
- 886 45. Act now on CRISPR babies. *Nature* **570**, 137–137 (2019).
- 887 46. Bouwman, A., Shved, N., Akgül, G., Ruhli, F. & Warinner, C. Ancient DNA investigation
888 of a medieval German cemetery Confirms long-term stability of CCR5- Δ 32 allele
889 frequencies in central Europe. *Hum. Biol.* **89**, 119–124 (2017).
- 890 47. Vargas, A. E. *et al.* Pros and cons of a missing chemokine receptor--comments on ‘Is
891 the European spatial distribution of the HIV-1-resistant CCR5-D32 allele formed by a
892 breakdown of the pathocenosis due to the historical Roman expansion?’ by Eric Faure
893 and Manuela Royer-Carenzi (2008). *Infect. Genet. Evol. J. Mol. Epidemiol. Evol. Genet.*
894 *Infect. Dis.* **9**, 387–389 (2009).
- 895 48. Lidén, K., Linderholm, A. & Götherström, A. Pushing it back. Dating the CCR5- Δ 32 bp
896 deletion to the Mesolithic in Sweden and its implications for the Meso Neo transition.
897 *Doc. Praehist.* **633**, (2006).

- 898 49. Hummel, S., Schmidt, D., Kremeyer, B., Herrmann, B. & Oppermann, M. Detection of
899 the CCR5- Δ 32 HIV resistance gene in Bronze Age skeletons. *Genes Immun.* **6**, 371–
900 374 (2005).
- 901 50. Lucotte, G. Distribution of the CCR5 gene 32-basepair deletion in west Europe. A
902 hypothesis about the possible dispersion of the mutation by the vikings in historical
903 times. *Hum. Immunol.* **62**, 933–936 (2001).
- 904 51. Lucotte, G. & Dieterlen, F. More about the Viking hypothesis of origin of the Δ 32
905 mutation in the CCR5 gene conferring resistance to HIV-1 infection. *Infect. Genet. Evol.*
906 **3**, 293–295 (2003).
- 907 52. Silva-Carvalho, W. H. V., de Moura, R. R., Coelho, A. V. C., Crovella, S. & Guimarães,
908 R. L. Frequency of the CCR5-delta32 allele in Brazilian populations: A systematic
909 literature review and meta-analysis. *Infect. Genet. Evol.* **43**, 101–107 (2016).
- 910 53. Hedrick, P. W. & Verrelli, B. C. 'Ground truth' for selection on CCR5-Delta32. *Trends*
911 *Genet. TIG* **22**, 293–296 (2006).
- 912 54. Meccas, J. *et al.* CCR5 mutation and plague protection. *Nature* **427**, 606–606 (2004).
- 913 55. Solloch, U. V. *et al.* Frequencies of gene variant CCR5- Δ 32 in 87 countries based on
914 next-generation sequencing of 1.3 million individuals sampled from 3 national DKMS
915 donor centers. *Hum. Immunol.* **78**, 710–717 (2017).
- 916 56. Buhler, M. M. *et al.* CCR5 genotyping in an Australian and New Zealand type 1
917 diabetes cohort. *Autoimmunity* **35**, 457–461 (2002).
- 918 57. Fahrioglu, U., Ergoren, M. C. & Mocan, G. CCR5- Δ 32 gene variant frequency in the
919 Turkish Cypriot population. *Braz. J. Microbiol.* **4**, (2020).
- 920 58. Kulmann-Leal, B., Ellwanger, J. H. & Chies, J. A. B. CCR5 Δ 32 in Brazil: Impacts of a
921 European Genetic Variant on a Highly Admixed Population. *Front. Immunol.* **12**, (2021).
- 922 59. Bhatnagar, I. *et al.* The Latitude Wise Prevalence of the CCR5- Δ 32-HIV Resistance
923 Allele in India. *Balk. J. Med. Genet.* **12**, 17–27 (2009).
- 924 60. Bamshad, M. J. *et al.* A strong signature of balancing selection in the 5' cis-regulatory
925 region of CCR5. *Proc. Natl. Acad. Sci. U. S. A.* **99**, 10539–10544 (2002).

- 926 61. Le, M. K. *et al.* 1,000 ancient genomes uncover 10,000 years of natural selection in
927 Europe. 2022.08.24.505188 Preprint at <https://doi.org/10.1101/2022.08.24.505188>
928 (2022).
- 929 62. Martiniano, R., Garrison, E., Jones, E. R., Manica, A. & Durbin, R. Removing reference
930 bias and improving indel calling in ancient DNA data analysis by mapping to a
931 sequence variation graph. *Genome Biol.* **21**, 250 (2020).
- 932 63. Zawicki, P. & Witas, H. W. HIV-1 protecting CCR5- Δ 32 allele in medieval Poland.
933 *Infect. Genet. Evol.* **8**, 146–151 (2008).
- 934 64. Orlando, L. *et al.* Ancient DNA analysis. *Nat. Rev. Methods Primer* **1**, 1–26 (2021).
- 935 65. McLaren, W. *et al.* The Ensembl Variant Effect Predictor. *Genome Biol.* **17**, 122 (2016).
- 936 66. Landrum, M. J. *et al.* ClinVar: public archive of relationships among sequence variation
937 and human phenotype. *Nucleic Acids Res.* **42**, D980-985 (2014).
- 938 67. Buniello, A. *et al.* The NHGRI-EBI GWAS Catalog of published genome-wide
939 association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res.*
940 **47**, D1005–D1012 (2019).
- 941 68. Kauwe, J. S. K. *et al.* Genome-wide association study of CSF levels of 59 alzheimer's
942 disease candidate proteins: significant associations with proteins involved in amyloid
943 processing and inflammation. *PLoS Genet.* **10**, e1004758 (2014).
- 944 69. Liu, J. Z. *et al.* Association analyses identify 38 susceptibility loci for inflammatory
945 bowel disease and highlight shared genetic risk across populations. *Nat. Genet.* **47**,
946 979–986 (2015).
- 947 70. Vistnes, M. *et al.* Plasma immunological markers in pregnancy and cord blood:
948 A possible link between macrophage chemo-attractants and risk of childhood type 1
949 diabetes. *Am. J. Reprod. Immunol. N. Y. N* **79**, (2018).
- 950 71. Kamat, M. A. *et al.* PhenoScanner V2: an expanded tool for searching human
951 genotype-phenotype associations. *Bioinforma. Oxf. Engl.* **35**, 4851–4853 (2019).
- 952 72. Staley, J. R. *et al.* PhenoScanner: a database of human genotype-phenotype
953 associations. *Bioinforma. Oxf. Engl.* **32**, 3207–3209 (2016).

- 954 73. Meisner, J. & Albrechtsen, A. Haplotype and population structure inference using
955 neural networks in whole-genome sequencing data. *Genome Res.* gr.276813.122
956 (2022) doi:10.1101/gr.276813.122.
- 957 74. Van der Auwera, G. A. & O'Connor, B. *Genomics in the Cloud: Using Docker, GATK,
958 and WDL in Terra (1st Edition)*. (2020).
- 959 75. Margaryan, A. *et al.* Population genomics of the Viking world. *Nature* **585**, 390–396
960 (2020).
- 961 76. Allentoft, M. E. *et al.* Population genomics of Bronze Age Eurasia. *Nature* **522**, 167–172
962 (2015).
- 963 77. Allentoft, M. E. *et al.* Population Genomics of Stone Age Eurasia. 2022.05.04.490594
964 Preprint at <https://doi.org/10.1101/2022.05.04.490594> (2022).
- 965 78. Margaryan, A. *et al.* Population genomics of the Viking world. *Nature* **585**, 390–396
966 (2020).
- 967 79. Price, D., Ritchie, K., Gron, K., Gebauer, A. & Nielsen, J. Asnæs Havnebjerg: a late
968 Mesolithic Ertebølle coastal site in western Sjælland, Denmark. *Dan. J. Archaeol.* **7**, 1–
969 22 (2018).
- 970 80. Stern, A. J., Wilton, P. R. & Nielsen, R. An approximate full-likelihood method for
971 inferring selection and allele frequency trajectories from DNA sequence data. *PLOS*
972 *Genet.* **15**, e1008384 (2019).
- 973 81. Irving-Pease, E. K. *et al.* The Selection Landscape and Genetic Legacy of Ancient
974 Eurasians. 2022.09.22.509027 Preprint at <https://doi.org/10.1101/2022.09.22.509027>
975 (2022).
- 976 82. Muktopavela, R. *et al.* Modelling the spatiotemporal spread of beneficial alleles using
977 ancient genomes. 2021.07.21.453231 Preprint at
978 <https://doi.org/10.1101/2021.07.21.453231> (2021).
- 979 83. Jordan, I. K. The Columbian Exchange as a source of adaptive introgression in human
980 populations. *Biol. Direct* **11**, 17 (2016).

- 981 84. Norris, E. T. *et al.* Genetic ancestry, admixture and health determinants in Latin
982 America. *BMC Genomics* **19**, 861 (2018).
- 983 85. Haak, W. *et al.* Massive migration from the steppe was a source for Indo-European
984 languages in Europe. *Nature* **522**, 207–211 (2015).
- 985 86. Novembre, J. Ancient DNA steps into the language debate. *Nature* **522**, 164–165
986 (2015).
- 987 87. Brinkworth, J. F. Infectious Disease and the Diversification of the Human Genome.
988 *Hum. Biol.* **89**, 47–65 (2017).
- 989 88. Dyer, D. P. *et al.* Chemokine Receptor Redundancy and Specificity Are Context
990 Dependent. *Immunity* **50**, 378-389.e5 (2019).
- 991 89. Allers, K. & Schneider, T. CCR5 Δ 32 mutation and HIV infection: basis for curative HIV
992 therapy. *Curr. Opin. Virol.* **14**, 24–29 (2015).
- 993 90. Gupta, R. K. *et al.* Evidence for HIV-1 cure after CCR5 Δ 32/ Δ 32 allogeneic
994 haemopoietic stem-cell transplantation 30 months post analytical treatment interruption:
995 a case report. *Lancet HIV* **7**, e340–e347 (2020).
- 996 91. Jasinska, A. J., Pandrea, I. & Apetrei, C. CCR5 as a Coreceptor for Human
997 Immunodeficiency Virus and Simian Immunodeficiency Viruses: A Prototypic Love-Hate
998 Affair. *Front. Immunol.* **13**, (2022).
- 999 92. Ellwanger, J. H., Kaminski, V. de L. & Chies, J. A. What we say and what we mean
1000 when we say redundancy and robustness of the chemokine system - how CCR5
1001 challenges these concepts. *Immunol. Cell Biol.* **98**, 22–27 (2020).
- 1002 93. Kleist, A. B. *et al.* New paradigms in chemokine receptor signal transduction: moving
1003 beyond the two-site model. *Biochem. Pharmacol.* **114**, 53–68 (2016).
- 1004 94. Auton, A. *et al.* A global reference for human genetic variation. *Nature* **526**, 68–74
1005 (2015).
- 1006 95. Machiela, M. J. & Chanock, S. J. LDlink: a web-based application for exploring
1007 population-specific haplotype structure and linking correlated alleles of possible
1008 functional variants: Fig. 1. *Bioinformatics* **31**, 3555–3557 (2015).

- 1009 96. Li, H. A statistical framework for SNP calling, mutation discovery, association mapping
1010 and population genetical parameter estimation from sequencing data. *Bioinformatics*
1011 **27**, 2987–2993 (2011).
- 1012 97. Renaud, G., Hanghøj, K., Willerslev, E. & Orlando, L. Gargammel: A sequence
1013 simulator for ancient DNA. *Bioinformatics* **33**, 577–579 (2017).
- 1014 98. Schubert, M., Lindgreen, S. & Orlando, L. AdapterRemoval v2: rapid adapter trimming,
1015 identification, and read merging. *BMC Res. Notes* **9**, 88 (2016).
- 1016 99. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler
1017 transform. *Bioinformatics* **25**, 1754–1760 (2009).
- 1018 100. Schubert, M. *et al.* Improving ancient DNA read mapping against modern reference
1019 genomes. *BMC Genomics* **13**, 178 (2012).
- 1020 101. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler
1021 transform. *Bioinformatics* **25**, 1754–1760 (2009).
- 1022 102. Picard toolkit. *Broad Inst. GitHub Repos.* (2019).
- 1023 103. Mölder, F. *et al.* Sustainable data analysis with Snakemake. *F1000Research* **10**, 33
1024 (2021).
- 1025 104. McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for
1026 analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).
- 1027 105. Danecek, P. *et al.* Twelve years of SAMtools and BCFtools. *GigaScience* **10**, giab008
1028 (2021).
- 1029 106. R Core Team. R: A Language and Environment for Statistical Computing.
1030 <https://www.R-project.org/> (2020).
- 1031 107. Speidel, L., Forest, M., Shi, S. & Myers, S. R. A method for genome-wide genealogy
1032 estimation for thousands of samples. *Nat. Genet.* **51**, 1321–1329 (2019).
- 1033 108. Bélisle, C. J. P. Convergence Theorems for a Class of Simulated Annealing Algorithms
1034 on Rd. *J. Appl. Probab.* **29**, 885–895 (1992).
- 1035 109. Byrd, R. H., Lu, P., Nocedal, J. & Zhu, C. A LIMITED MEMORY ALGORITHM FOR
1036 BOUND CONSTRAINED OPTIMIZATION. **25** (1994).

1037