Joint effects of balancing selection and population bottlenecks on the evolution of a regulatory region of human anti-viral *APOBEC3*

Naoko T Fujito[1,3], Sundaramoorthy Revathidevi [1,4], Yoko Satta[2], and Ituro Inoue[1]

[1] Human Genetics Laboratory, National Institute of Genetics, Mishima, Shizuoka, Japan

[2]Reaserch Center for Integrative Evolutionary Science, SOKENDAI (The Graduate University for Advanced Studies), Hayama, Kanagawa, Japan

[3]Department of System Pathology for Neurological Disorder, Brain Research Institute, Niigata University, Niigata, Japan

[4]Department of Genetic Engineering, College of Engineering and Technology, SRM Institute of Science and Technology, Tamil Nadu, India

Corresponding author: Naoko T Fujito (naoko.fujito@gmail.com)

Abstract

Human anti-viral APOBEC3s are crucial components of the innate immune response, playing a key role in inhibiting viral replication and proliferation by inducing mutations in viral genomes. In this study, in a regulatory region of ~16 kb in the *APOBEC3* gene cluster on human chromosome 22, we identified three distinct haplogroups that are maintained at high frequencies in both African and non-African populations today. Despite the long persistence of these haplogroups, one of which is shared by archaic hominins, the nucleotide diversity within each haplogroup was unusually low in non-Africans. The simultaneous observations of low intra-haplogroup diversity and high inter-haplogroup divergence were incompatible with neutrality, as demonstrated by the genome-wide empirical distribution and coalescence-based simulation tests. In addition to the three shared haplogroups, an ancestral sequence group was identified exclusively in African populations. To explain these features and investigate the underlying evolutionary mechanisms, we performed forward simulations to model the joint effects of balancing selection and population bottlenecks. We demonstrated the operation of balancing selection over the past ~1 million years, as well as a hitherto unexplored reduction in intra-haplogroup diversity. The only unexplainable observation was the low diversity within African haplogroup I, which is unaffected by the non-African-specific bottleneck. We hypothesized that an extra haplotype turnover within haplogroup I occurred prior to the Out-of-Africa dispersal of modern humans. Finally, we discussed the functional importance of the

35 *APOBEC3* regulatory region in terms of evolutionary conservation in Hominidae, a target of

36 natural selection in modern humans, and its relevance in GTEx eQTL.

37

Significance statement

This study provides new insights into the joint effects of population bottlenecks and balancing selection on genetic diversity within haplogroups. First, we reveal how balancing selection interacts with demographic history to shape intra-haplogroup diversity, a previously unexplored aspect. Second, our findings are generalizable, as illustrated by the ABO locus─an established example of balancing selection─where low within-haplogroup diversity and high between-haplogroup divergence are also observed in non-African populations. Third, the involvement of balancing selection in a highly conserved regulatory region suggests functional significance in modulating the expression of all seven APOBEC3 genes. While further validation is needed, our findings underscore the evolutionary and functional importance of this region in human innate immunity.

Introduction

APOBEC3 belongs to a family of cytidine deaminases that play essential roles in the innate immune response against retroviruses, endogenous retroelements, and DNA viruses. They catalyze the deamination of cytidine in single-stranded DNA or RNA, converting cytidine-to-uridine (C > U). This C-to-U hypermutation disrupts the genomic information of viruses, inhibiting their replication and proliferation and resulting in their degradation via the uracil DNA glycosylase (UNG2)-dependent pathway. APOBEC3s also inhibit viral replication

57 through deaminase-independent mechanisms (Newman et al. 2005; Wang et al. 2012;
58 Pollpeter et al. 2018; Hakata et al. 2019).

59 The APOBEC3 family consists of seven enzymes (APOBEC3A/B/C/DE/F/G/H), and
60 their genes are clustered within a 200-kb region on chromosome 22 (the *APOBEC3* gene
61 cluster, Figure 1a). Although each member exhibits distinct characteristics, including cellular
62 localization, tissue-specific expression, substrate specificity (e.g., single-stranded DNA or
63 RNA), and viral restriction properties, their functions show mild overlap. For example, while
64 APOBEC3G acts as the most potent anti-human immunodeficiency virus (HIV) factor,
65 APOBEC3D, 3F, and 3H also contribute to HIV restriction. Similarly, APOBEC3A, 3B, 3C,
66 3F, and 3G can restrict DNA viruses, including hepatitis B virus, human papillomavirus,
67 adeno-associated virus, herpes simplex-1, and Epstein-Barr virus (Stavrou and Ross 2015;
68 Revathidevi et al. 2021).

69 In the present study, while initially analyzing patterns of genetic variation around the
70 *APOBEC3* gene cluster, we noted an approximately 16-kb long non-coding region within the
71 cluster exhibits three distinct haplogroups in all 26 populations in the 1000 Genome Project
72 (TGP), each of which is strongly supported by a long branch in the genealogical tree.
73 Although this region does not contain any coding sequences, it harbors several regulatory
74 motifs. Its functional importance is evidenced by evolutionary conservation across the
75 Hominidae family. Interestingly, each haplogroup in this region displays a star-like burst of

closely related lineages. We refer to this pattern as multi-star genealogy and examined the frequency of this observation in the results of coalescence simulation under neutrality, as well as in the empirical distribution obtained through a genome scan. The results from both the simulations and empirical analyses were incompatible with the observed multi-star genealogy, prompting us to reject the null hypothesis of neutrality and to consider an alternative explanation. In doing so, we further identified one important difference between African and non-African genealogical trees: the presence of apparent "ancestral" haplotypes that occur at appreciable frequencies in African populations but are almost absent in non-African populations.

Our alternative hypothesis assumes balancing selection that can maintain multiple alleles in a population for a long time. Two well-known forms of balancing selection are negative frequency-dependent selection and overdominance selection, although the patterns of genetic variability yielded by those mechanisms are often indistinguishable from each other (Takahata and Nei 1990; Charlesworth 2006 and Fijarczyk et al. 2015 for reviews). In recent years, with the advent of next-generation sequencing, numerous studies have rigorously identified candidate loci undergoing balancing selection. There is a consensus that balancing selection primarily affects a limited category of genes, with disease-resistance genes being the most well-known example (Bitarello et al. 2023 for a review). Balancing selection can extend the most recent common ancestor (MRCA) of each allelic group further back in time than

95 genetic drift alone, and it may occasionally result in trans-species polymorphism as observed
96 at major histocompatibility complex (MHC) loci (Takahata 1990; Takahata and Nei 1990).
97 However, as almost all studies on genetic variation under balancing selection have primarily
98 focused on nucleotide divergence among different allelic groups, nucleotide diversity within
99 an allelic group is largely overlooked. Therefore, we performed forward simulations to
100 explore the joint effects of balancing selection and population bottlenecks.
101 We also discussed the functional differences among the three haplogroups by
102 analyzing the distribution of derived alleles within the regulatory region, associating them
103 with Genome-Wide Association Study (GWAS) hits linked to these haplogroups, and
104 incorporating expression Quantitative Trait Loci (eQTL) analyses from the GTEx database.
105 Our results provide insights into the functional importance of the *APOBEC3* regulatory region
106 in human evolution during the Upper Pleistocene.
107

108 ## Results

109 **Non-African and African haplogroups in the regulatory region**

110 The 16-kb non-coding region is located between *APOBEC3B* and *APOBEC3C*, flanked by a
111 recombination hotspot and a deletion polymorphism within *APOBEC3B* (Figures 1a, 1b, S1
112 and S2). We term this region the *APOBEC3* regulatory region as it harbors one promoter,
113 three enhancers, and two CTCF-binding sites, as annotated by the Ensemble Regulatory Build

114 (Figure 1c). When we constructed genealogical trees of haplotypes within this region, it
115 became immediately apparent that in any non-African population in the TGP, three distinct
116 groups of haplotypes (haplogroups, Figure s2a–d and S3) were present. These haplogroups
117 exhibited appreciable frequency ($x > 0.15$), extremely low nucleotide diversity within each
118 haplogroup ($\pi < 0.01\%$), and relatively large nucleotide divergence between a pair of
119 haplogroups ($d > 0.01\%$) (Tables 1 and S1). Using the minimum x, the smallest d, and the
120 highest π values observed among the 15 non-African populations, we defined the three
121 haplogroups I, II, and III (hereafter denoted as HI, HII, and HII, respectively; see Materials
122 and Methods). The divergence time between HI and HII was estimated to be as old as 1.80–
123 1.96 million years ago (MYA), assuming that the *de novo* mutation rate in humans and
124 hominoids is 0.5×10^{-9} site/year (Jónsson et al. 2017; Chintalapati and Moorjani 2020),
125 whereas those between HI and HIII and between HII and HIII were estimated to be 1.36–1.40
126 MYA and 1.43–1.55 MYA, respectively (Table S1). In contrast to these long divergence
127 times, intra-haplogroup diversity is one order of magnitude lower than expected when
128 calculated by multiplying the nucleotide diversity in all pairwise comparisons by the
129 haplogroup frequency (Tables 1 and S1). We term this unique structure, characterized by a
130 genealogical tree exhibiting multiple star-like bursts, “multi-star” genealogy. Hereafter, we
131 mainly aimed to elucidate the evolutionary history and mechanisms reflected in this
132 unprecedented genealogical configuration.

133 To gain insight into the origins of the three haplogroups, we examined their presence
134 in the African populations. All the African populations in the TGP possessed HI, HII, and
135 HIII. However, among the African haplogroups, only HI exhibited extremely low intra-
136 haplogroup diversity, whereas HII and HIII displayed moderate levels of diversity that are
137 consistent with the expectation (Tables 1 and S1). In addition, over 40% of the haplotypes in
138 Africans were unique to their populations and did not belong to any of the three haplogroups
139 (Figure 2e, f). The genealogical tree of this genomic region in African populations indicated
140 that the three haplogroups emerged independently from these African-specific sequences
141 roughly 1 MYA. This finding suggests that the African-specific haplotypes are collectively an
142 ancestral type (Figures S3c and S4, Table S2b).
143 Although the estimates of the above emergence time of the haplogroups are
144 necessarily approximate, their antiquity is supported by genomic data from ancient hominins.
145 Specifically, all four pseudo-haploid sequences from the Altai Neanderthal (Prüfer et al.
146 2014) and Denisovan genomes (Meyer et al. 2012) used in this study harbor HI-linked alleles
147 at three out of 14 SNP sites. These derived alleles are shared by all HI sequences but not by
148 HII or HIII. Although the genealogical clustering of HI with the ancient hominin sequences is
149 only weakly supported by bootstrap probabilities (Figures S3 and S4a, Table S2a), it is
150 suggested that when archaic hominins and modern humans diverged (550 KYA; Prüfer et al.,
151 2014), the HI haplogroup had already been isolated from other haplotypes.

152 The functional differences among the three haplogroups remain unknown. However,
153 a window-based analysis of nucleotide differences revealed that regions with the highest
154 divergence coincided with regulatory sequences, such as promoters and enhancers (Figure
155 S5). This finding suggests potential functional differences among the haplogroups.
156 To investigate changes in the frequencies of the three haplogroups over time, we
157 utilized genome-wide data from prehistoric humans. To this end, we analyzed the Allen
158 Ancient Data Resource (Anon) for index SNPs for the three haplogroups and found that
159 rs2142833 (r^2 = 0.92 with HI in the global population), rs9607601 (r^2 = 1 with HII in the
160 global population, r^2 = 0.87 in the JPT, and r^2 = 0.55 in the CEU), and rs11089916 (r^2 = 1with
161 HIII in CEU and JPT) were informative. Due to the limited number of available ancient
162 samples, we divided the data into different time windows of varying sizes (400–20,000 years).
163 Although we were unable to analyze African data due to an insufficient sample size, we
164 observed the three haplogroups have been maintained at similar frequencies with only
165 moderate changes in both Europe and East Asia during the past 40,000 years (Figure 3).
166
167 **Neutrality test for the regulatory haplogroups**
168 The long-term persistence of the three haplogroups suggests the possibility of non-neutral
169 evolution of the *APOBEC3* regulatory region. We tested the neutrality of the multi-star
170 genealogy using coalescence simulation and found its extreme rarity (Materials and Methods).

171 More specifically, we assessed the statistical rarity of the multi-star genealogy in African and
172 non-African populations by evaluating (1) the frequency of occurrence of three haplogroups
173 with sizes (i.e., the number of chromosomes) comparable to those of HI, HII, and HIII, and
174 (2) the likelihood of co-occurrence of low intra-haplogroup diversity and high inter-
175 haplogroup divergence. First, we searched for a genealogical structure with three distinct
176 haplogroups, each having a size equal to or larger than the observed one. We conducted
177 neutrality simulations by placing a specified (observed) number of segregating sites on a
178 coalescent tree using ms (Hudson 2002). We used two different demographic models. The
179 first is that proposed by Schaffner et al. (2005), which includes the Late Pleistocene Out-of-
180 Africa bottleneck. The second is a modified version of the SFD model, which includes the
181 bottleneck that occurred in Africa immediately before the Chibanian period, around 0.9
182 MYA, as recently inferred by Hu et al. (2023). It should be noted that the latter bottleneck
183 effect cannot be completely erased by the subsequent population expansion with N_e = 20,000
184 (Figure 3), approximately 30,000 generations between the two bottlenecks are not long
185 enough for nucleotide diversity to equilibrate (Charlesworth and Jensen 2021). Of the 10^6
186 coalescence trees we generated, 6159 (SFD model, p =0.62%) or 7628 (HFD model, p
187 =0.76%) exhibited the presence of three distinct allelic groups of comparable sizes in JPT,
188 where the haplogroups showed relatively high frequencies (Table S3; $x =$ 50.0%, 36.5%,
189 and 9.6% for HI, HII, and HIII, respectively). Such occurrence was more frequent when the

190 haplogroup frequencies were low. For CEU, 42916 or 36858 out of 10^6 replicates (p =

191 4.29% with the SFD model and p = 3.69% with the HFD model, respectively) exhibited the

192 presence of three distinct allelic groups with the required sizes (Table S3; $x = 41.9\%$, 21.2%,

193 and 22.2% for HI, HII, and HIII, respectively). These p-values suggest that the likelihood of

194 multi-star genealogy is moderately rare in non-African populations, whereas the three

195 haplogroups with smaller sizes occur rather frequently in African populations (Table S3).

196 Considering that three haplogroups are present, we investigated how often the

197 observed levels of low intra-haplogroup diversity and high inter-haplogroup divergence

198 occur. Using the coalescence simulation again, we collected 10^4 replications where three

199 haplogroups are present with the same sizes as in HI, HII, and HIII. We compared the

200 observed nucleotide differences within and between the haplogroups (Table S1) to simulated

201 ones to evaluate statistical significance. The comparison revealed that only a small proportion

202 of the 10^4 simulation results yielded the observed levels of nucleotide differences, with $p \leq$

203 9.1×10^{-6} for JTP and $p \leq 6.6 \times 10^{-5}$ for CEU (Table S3). This finding suggests that some

204 forms of natural selection have been involved in the emergence of the multi-star genealogy. It

205 is worth emphasizing that the co-occurrence of three low-frequency haplogroups in Africans

206 is not unlikely; however, intra-HI diversity is significantly reduced even in Africans, as in

207 non-Africans (Supplementary Text). This results in $p \leq 1.2 \times 10^{-4}$ for LWK and $p \leq 2.3 \times 10^{-4}$

208 for YRI (Table S3).

209 We also examined genomic data for the multi-star genealogy using a window-based
210 approach. We identified 2,460 regions (45,799 windows) in the JPT population that displayed
211 three haplogroups equal to or larger than those observed, corresponding to 80,464 kb or
212 2.9 % of the total sequence analyzed (2.8 Gb). In the CEU population, we identified 13,507
213 genomic regions (298,169 windows of 489,448 kb length), accounting for 17.5% of the total
214 sequence analyzed. These observations collectively imply that while the occurrence of the
215 multi-star genealogy may be uncommon in human genomes, the likelihood strongly depends
216 on the frequency of haplogroups. To examine the intra-haplogroup diversity, we focused on
217 genomic regions in which each of the three haplogroups occurs with the same frequency as
218 the observed one. We identified 260 and 344 such regions for JPT and CEU, respectively,
219 none of which displayed the observed level of intra-haplogroup diversity.
220 In summary, the structure of genealogy with the three haplogroups in the *APOBEC3*
221 regulatory region is rarely found in coalescence simulation or in actual genomic data. Even
222 when we collected such structures from tens of millions of replications of coalescence
223 simulations or from the whole genome, the likelihood of simultaneous occurrence of low
224 intra-haplogroup diversity and high inter-haplogroup divergence remained exceedingly low.
225 We thus conclude that these haplogroups have unlikely evolved in neutral fashions; rather,
226 their long persistence suggests the involvement of some forms of balancing selection that has
227 persisted in African and non-African populations over the past 1 million years.

228

229 **Balancing selection and population size changes**

230 We then focused on differences in the patterns of genetic diversity between African and non-
231 Africa populations. Although balancing selection was suggested for both populations, the
232 pattern and level of intra-haplogroup diversity were evidently different between these two
233 meta populations (Table 1). In non-African populations, the intra-haplogroup diversity was 10
234 times lower than expected. In contrast, African populations showed high and consistent levels
235 of diversity within HII and HIII, in line with expectations. However, the intra-HI diversity
236 exhibited a substantial reduction not only in non-African populations but also in African
237 populations. Hence, these features require further scrutiny.

238 To explain these differences between African and non-African populations, we
239 developed a forward simulation model that incorporates balancing selection and population
240 size changes with one or two bottlenecks. In this model, we considered four different types of
241 sequences: three representing the three haplogroups and one additional group, hereafter
242 referred to as "ancestral group", representing the African-specific sequences. Given the
243 potential roles of the haplogroups in regulating the expression of the anti-viral APOBEC3s,
244 we first considered negative frequency-dependent selection and hypothesized a scenario in
245 which the ancestral group played a fundamental role in gene regulation, and the new
246 haplogroups that subsequently emerged one after another modified this original regulatory

247 function independently. These functions might include enhancing the expression of certain
248 *APOBEC3* genes or silencing others, thereby conferring robustness against specific viral
249 groups. We thus assumed that each haplogroup is subject to negative frequency-dependent
250 selection with their relative fitness given by $1 - sx$, where x is the frequency of a
251 haplogroup and s is selection coefficient common to the three haplogroups. In contrast, the
252 relative fitness of the ancestral group is defined as $1 - as$, where a is a constant parameter.
253 In our model, the ancestral group is not subject to balancing selection because the actual
254 African-specific sequences are not monomorphic and do not share a common mutation upon
255 which balancing selection could act. The fitness for diploids is assumed to be multiplicative.
256 We used demographic models similar to those used for the coalescence simulation
257 (Figure 4a, b). This time, we included a recent population expansion, though it does not
258 significantly affect genetic diversity within and between sequence groups (Figures 4e-h, S6,
259 and S7). For simplicity, all cases assumed that three haplogroups and an ancestral group
260 emerged simultaneously, with the initial frequency of 0.25 set prior to the pre-Chibanian
261 bottleneck, and are immediately subjected to frequency-dependent selection. Simulations
262 were performed with a range of parameters s and a, and we examined intra- and inter-
263 haplogroup diversity together with frequency changes of the three haplogroups and the
264 ancestral group. Here and thereafter, the value of s is scaled relative to N_e specified in the
265 simulations.

266 The results demonstrated that intra-haplogroup diversity is not affected by the value
267 of s, but sensitively varies depending on the value of a (Figures 4g, 4h, S6a and S7a). This
268 variation arises simply from differences in haplogroup frequency (Figures 4c, 4d, S6b and
269 S7b). In our model, the haplogroup frequency is a function of the constant parameter
270 a (Supplementary Text). The larger the value, the lower the frequency of the ancestral group
271 and the higher the frequency of the three haplogroups. In contrast to intra-haplogroup
272 diversity, inter-haplogroup divergence grows linearly in time even for different sets of
273 parameter values (Figures 4e, 4f, S6c and S7c), as long as selection is effective. When $a =$
274 0.2, the frequency reached the deterministic equilibrium value of 40 % for the ancestral group
275 and 20 % for each haplogroup, as observed in African populations, showing that 0.2 is the
276 proper value for a under negative-frequency dependent selection to explain the observation
277 (Figure 4c, 4d, S6b and S7b). For African populations, the effect of the first bottleneck on
278 intra-haplogroup diversity was somewhat faded at the end of the simulation (corresponding to
279 ~1 million years; Figure 4g). In the absence of the second bottleneck, intra-haplogroup
280 diversity approached 0.02 % (Figures 4g and S6a), which is close to that observed for HII and
281 HIII. However, as mentioned earlier, the observed diversity within HI is considerably small
282 (Table 1) and cannot be explained solely by the present mutation scheme (Materials and
283 Methods). The diversity within the ancestral group reached over 0.03 % (Figure 4i), whereas
284 the observed diversity of the African-specific sequences was significantly higher, reaching

285 0.14 % and 0.15% in the combined African population (AFR) and the YRI population,
286 respectively. This difference arises from the fact that the African-specific sequences are not
287 monophyletic, and substantial genetic diversity had already accumulated prior to the onset of
288 balancing selection. These values are nearly identical to the mean pairwise diversity within
289 AFR and YRI overall, which is 0.15 % in both populations.

290 For non-African populations, the observed state is that all three haplogroups are
291 retained, but the ancestral group is lost during the Out-of-Africa bottleneck. In the simulation,
292 the ancestral group was more frequently lost during the first and more severe bottleneck ($\gamma =$
293 $\frac{t_b}{2N_b} = 0.85$, where t_b and N_b are the duration time in units of generations and the effective
294 size of a bottleneck, respectively), but sometimes lost in the second, Out-of-Africa bottleneck
295 ($\gamma = 0.5$). When we collected replications where the ancestral group was lost in the Out-of-
296 Africa bottleneck, the intra-haplogroup diversity significantly decreased due to the bottleneck,
297 reaching 0.005% (Figure 4h), which is close to that observed for HI, HII, and HIII in non-
298 African populations. With the following recent population expansion, the intra-haplogroup
299 diversity somewhat recovered, but the extent is small enough to be ignored (Figure 4h).

300 To assess whether a single combination of *s* and *a* values could account for the
301 observations in both African and non-African populations, we conducted an additional
302 simulation across a range of *s* and *a* values (Figure 5), using a demographic model that
303 mimics the Out-of-Africa bottleneck (Figure 5a). The initial frequencies of the ancestral group

304 and each haplogroup were 40 % and 20 %, respectively, as observed in African populations.
305 All the sequences were immediately subjected to frequency-dependent selection and the Out-
306 of-Africa bottleneck (Figure 5a, Materials and Methods). We calculated the proportion of
307 simulation replications in which all the haplogroups were retained, but the ancestral group
308 was lost, as observed in non-Africans. When examining the case with $a = 0.2$, the condition
309 where $N_b s = 4$ yielded the observed pattern at a rate as high as 5.3% (Figure 5b).
310 In summary, the extremely low intra-haplogroup diversity and high inter-haplogroup
311 divergence observed in non-African populations can be explained by the combined effects of
312 negative frequency-dependent selection and the Out-of-Africa bottleneck. The loss of the
313 ancestral group in non-Africans and its retention in Africans can occur with the same set of a
314 and s values (a=0.2 and $N_b s$ = 4). For African populations, reasonably high intra-haplogroup
315 diversity of HII and HIII is explained by negative frequency-dependent selection acting
316 similarly to population subdivision. However, the observed intra-haplogroup HI diversity
317 cannot be realized from the present model.
318 To study other possible mechanisms of balancing selection, we examined
319 overdominance or heterozygote advantage with some asymmetry in relative fitness. In our
320 model, the relative fitness of homozygotes is given by $1 - s$ for the three haplogroups and
321 $1 - as$ for the ancestral group. Although we used the same symbol a as in frequency-
322 dependent selection model, the meaning is fundamentally different (Supplementary Text,

323 Figures S8, S9). Under this overdominance model, the value of a that yields the equilibrium

324 frequency of 40 % for the ancestral group and 20 % for each haplogroup was calculated to be

325 0.5 (Supplementary Text). In simulations of the Out-of-Africa bottleneck, we found that with

326 $a = 0.5$, the observed state was reproduced in fewer than 5 % in 10,000 replications,

327 regardless of values of $N_b s$ (Figure 5c). Although the difference in probability was small,

328 these results suggest that balancing selection on the *APOBEC3* regulatory region is more

329 likely to be driven by negative frequency-dependent selection than by overdominance.

330

331 **Turnover of haplogroup I**

332 One possible explanation for the low intra-HI diversity in all populations is a haplotype

333 turnover in HI. This haplogroup was estimated to have emerged around 2 million years ago,

334 and its antiquity is further supported by the presence of HI-linked alleles in archaic hominins,

335 as described above. Therefore, the low intra-haplogroup diversity observed in both African

336 and non-African populations cannot be attributed to a recent origin of HI. In African

337 populations, positive selection for HI can be detected by F_c (Fujito et al. 2018). Other

338 statistics, such as G_{c0} in 2D SFS (Satta et al. 2019), suggest that the haplotype turnover began

339 with a single HI haplotype (Table S4), resembling an ongoing hard sweep within the HI

340 group. This turnover was not detected in non-African populations owing to the

341 simultaneously reduced intra-HII and HIII diversity, mitigating the signal of positive selection

342 on HI. We demonstrated in simulation that when a population bottleneck occurs shortly after
343 the completion of a haplotype turnover, and there is insufficient time for the group carrying
344 this haplotype to increase its intra-haplogroup diversity, the bottleneck does not substantially
345 affect the diversity (Figure S10). This indicates that if the haplotype turnover within HI
346 happened shortly before the Out-of-Africa bottleneck, then the intra-haplogroup diversity
347 should be in a similar range in both African and non-African populations. Several
348 observations support this conjecture. First, the estimated times of expansion of the
349 haplogroups revealed that the TMRCA (time to MRCA) of both HII and HIII significantly
350 differed between African and non-African populations. It is estimated between 30,000 and
351 80,000 years ago in non-Africans, considerably shorter than ~350,000 years ago for HII and
352 180,000 years ago for HIII in Africans (Figure S11, Table S5). These recent TMRCAs in non-
353 African populations are consistent with the Out-of-Africa bottleneck around 70,000 years ago.
354 In contrast, the TMRCA of HI is between 50,000 and 80,000 years ago in both African and
355 non-African populations. As the Out-of-Africa bottleneck cannot affect the TMRCA in
356 African HI, the HI turnover is a likely alternative, although this does not necessarily imply
357 independent operation of some forms of directional selection. Moreover, phylogenetic
358 network analysis revealed that the most frequent HI haplotype appeared across Africans,
359 Europeans, South Asians, East Asians, and Americans in TGP (Figures S12, Supplementary

360 text). Thus, the present-day HI in the world population has expanded from this recent

361 haplotype before African and non-African HI diverged from each other.

362

363 **Deletion polymorphism**

364 The *APOBEC3* regulatory region is located adjacent to a large polymorphic deletion

365 encompassing the entire coding region of *APOBEC3B*. This polymorphism has been

366 repeatedly reported to increase cancer susceptibility (ICGC/TCGA Pan-Cancer Analysis of

367 Whole Genomes Consortium 2020). To investigate whether balancing selection for the

368 regulatory region was influenced by this deletion, we examined its evolutionary history. First,

369 it is important to note that regardless of population, the deletion is predominantly linked to HI

370 and only occasionally to HII, HIII, and the ancestral group (Figure S13), suggesting that it is

371 of ancient single origin. Second, haplotypes with the deletion exhibit star-like genealogy only

372 when it is linked to HI, and no such pattern is observed when linked to non-HI haplotypes

373 (Figure S14a-e). Third, within HI, a large part of the most frequent haplotype carries the

374 deletion in all five metapopulations in TGP (Figure S14). It was previously suspected that the

375 deletion underwent positive selection, primarily based on its high frequency in Asia (~35 % in

376 East Asians and ~15 % in South Asians) and its low frequency in Europe and Africa (~6%

377 and ~0.9%, respectively). However, at that time, no clear conclusion could be drawn (Kidd et

378 al. 2007). In this study, we demonstrated that Asian HI is subject to balancing selection and

379 that the high frequency in Asia is simply due to the hitchhiking effect associated with the
380 expansion of HI.

381 We further studied the possibility that the deletion may confer an additional selective
382 advantage to haplotypes in HI. Equally advantageous sequences are obviously selectively
383 equivalent within that selected group. On the other hand, if some sequences are more
384 beneficial than others even in the same selected group, a nested selective sweep could be
385 observed. We calculated F_c values (Fujito et al. 2018) for the sequences within HI linked to
386 the deletion in the Japanese population. The resulting F_c value was not low (p = 0.49). Thus,
387 in terms of selective advantage, the sequences linked to the deletion do not differ from other
388 HI sequences. It appears that the deletion did not confer its own selective advantage and was,
389 in fact, lost in Europe and Africa due to genetic drift.

390

391 **Discussion**

392 Research on the diversity of alleles under balancing selection has primarily focused on inter-
393 allelic or inter-haplogroup divergence, and, to the best of our knowledge, has paid little
394 attention to intra-allelic or intra-haplogroup diversity. In the present study, our simulation
395 results demonstrated that intra-haplogroup diversity under balancing selection greatly
396 decreases when combined with population bottlenecks. In contrast, inter-haplogroup
397 divergence is less affected by demographic changes, selection strength, haplogroup frequency,

398 or detailed mechanisms of balancing selection. Instead, it is largely proportional to the time

399 elapsed since haplogroups began to co-segregate under balancing selection.

400 The impact of an idealized instantaneous bottleneck can be evaluated using the

401 parameter $\gamma = t_b/(2N_b)$. Irrespective of the presence or absence of selection, t_b remains

402 unchanged. However, the effective population size is influenced by balancing selection in a

403 specific way: balancing selection acts as if the haplotypes are subdivided into m distinct

404 lineages. This situation is analogous to population subdivision, such that under symmetric

405 balancing selection, the effective size of each haplogroup may be reduced to N_b/m. In other

406 words, the extent of intra-haplogroup diversity maintained by balancing selection is subjected

407 to an m-fold severe bottleneck compared to a non-subdivided scenario ($m\gamma$). A caveat,

408 however, is that for large m, as in MHC, it becomes more likely that a certain number of

409 haplogroups are lost due to genetic drift rather than that all haplogroups being retained and

410 responding by reducing their intra-haplogroup diversity. This may still hold true even for

411 $m = 3$. In our Out-of-Africa bottleneck simulations, all three lineages were not always

412 maintained even under strong balancing selection.

413 Our results indicate that intra-haplogroup diversity maintained by balancing selection

414 is lower in non-African populations than in African populations. To assess the generality of

415 this pattern, we analyzed the ABO locus, a well-known example of balancing selection. As

416 expected, the genealogical trees at the ABO locus exhibit lower intra-allelic diversity for the

417 alleles responsible for blood groups A and B in the JPT population than that in the YRI

418 population (Figure S15; π = 0.023% for type A and 0.048% for type B in JPT; π = 0.115%

419 for type A and 0.083% for type B in YRI). Notably, in this analysis, we excluded blood group

420 O given its multiple origins.

421 As mechanisms of balancing selection, both overdominance and frequency-

422 dependent selection with a rare-allele advantage are well-studied. Additionally, fluctuating

423 selection, which involves spatio-temporal variations in selection pressure, has been proposed

424 (Charlesworth 2006). The genetic signatures of these three mechanisms overlap, making it

425 exceedingly challenging, often not feasible, to distinguish the underlying mechanisms from

426 the genetic signatures alone (Spurgin and Richardson 2010; Fijarczyk and Babik 2015). In the

427 present study, using forward simulations, we examined which mechanism, negative

428 frequency-dependent or overdominance selection, is more appropriate to explain the

429 maintenance of the three haplogroups and the ancestral (African-specific) haplogroup. We

430 concluded that negative frequency-dependent selection is more likely from the functional

431 perspective of APOBEC3 and is more parsimonious based on simulations. Under negative

432 frequency-dependent selection, the differing patterns of genetic variation in African and non-

433 African populations can be explained by a single combination of parameter a and parameter

434 s. In contrast, under the overdominance model, the a value, which is assumed based on the

435 ancestral group frequency in the African populations, could not explain the observation in the

436 non-African populations, leading to the necessity of adjusting the parameter between the

437 African and non-African populations.

438 The parameter a defines the relative fitness of the African-specific ancestral group

439 $(1 - as)$ in contrast to the relative fitness of other haplogroups $(1 - sx)$. Under this

440 frequency-dependent selection model, the haplogroup frequency x hardly exceeds the value

441 of a. For example, if $a = 0.2$, the frequency of a haplogroup tends to be restricted to less

442 than 0.2 in a large population. At $x = 0.2$, the haplogroup and the ancestral group possess

443 identical relative fitness. However, at $x > 0.2$, the relative fitness of the haplogroup would

444 fall below that of the ancestral group, resulting in the replacement of the haplogroup by the

445 ancestral group. When a becomes 0.3, the frequency of each of the three haplogroups can

446 increase up to 0.3, which is compensated by the reduced frequency of the ancestral group.

447 Thus, the frequency and diversity of a haplogroup sensitively vary based on the different

448 values of a. We envisage that difference in the value of parameter a reflects differences in

449 the pathogenic viral environment.

450 In the forward simulation, we assumed a single nucleotide site as a target of

451 balancing selection, with no recurrent mutation at this site. Given these assumptions, we

452 cannot expect either haplotype turnovers or enhanced substitution rates that can occur under

453 balancing selection (Maruyama and Nei 1981). As noted by Takahata and Nei (1990), who

454 used the infinite-alleles model, these features are particularly relevant to the evolution of the

455 MHC peptide binding region (Hughes and Nei 1988). In this context, although we
456 emphasized a haplotype turnover in HI prior to the Out-of-Africa bottleneck, it does not
457 necessarily imply the operation of independent directional selection if newly arising
458 mutations are allowed to undergo balancing selection. We note that this may not be restricted
459 to HI and that more general mutation models would be compatible with our sequence data. In
460 conclusion, under different mutation schemes, joint effects of negative frequency-dependent
461 selection and demographic changes with bottlenecks are sufficient to explain the observed
462 pattern and level of nucleotide polymorphism in the *APOBEC3* regulatory region.

463 We relied on a simulation method to detect natural selection on the *APOBEC3*
464 haplogroups, as the existing detection methods are not appropriate in the present case.
465 Approaches for detecting positive selection, such as EHH-based methods (Sabeti et al. 2002;
466 Voight et al. 2006), SDS (Field et al. 2016), and F_c measurements (Fujito et al. 2018), use
467 summary statistics that directly or indirectly measure the ratio of the diversity in the focal
468 sequence group to the overall sequence diversity. This ratio is expected to be low when the
469 focal group is under positive selection. However, the ratio cannot be very low if some
470 remaining haplotypes have also reduced the diversity for any reason. Similarly, other
471 summary statistics, such as β (Siewert and Voight 2017) and $\beta^{(2)}$ (Siewert and Voight 2020),
472 are applied to detect clusters of alleles at similar frequencies around the site targeted by
473 balancing selection. As such a cluster of alleles reflects the accumulation of variations in two

474 allelic lineages maintained by balancing selection through time, the β and $\beta^{(2)}$ statistics gain
475 power for low recombination rate, high mutation rate, and long-term balancing selection.
476 However, the persistence time of the three haplogroups in the *APOBEC3* regulatory region is
477 not long enough to accumulate a sufficiently large number of variants to be detected. Under
478 this scenario, these statistics tools may lose power.

479 We obtained information that supports the existence of functional differences in the
480 *APOBEC3* regulatory region from GTEx eQTL data and the GWAS catalog. Using GTEx
481 eQTL data from 52 tissues, we searched for variants that might affect the transcriptional
482 regulation of the seven *APOBEC3* genes. We found that the regulatory region harbors several
483 tissue-specific eQTLs for the expression of all *APOBEC3* genes, except *APOBEC3H* (Figure
484 S16, Table S6). This includes index SNPs and linked SNPs of the three haplogroups. These
485 eQTL associations are relatively weak compared to those in the downstream region within the
486 *APOBEC3* gene cluster (chr22:39,405,000-39,580,000; Figure S6). However, a recombination
487 hotspot separates the downstream region from the regulatory region (Figures 1a and S16a). As
488 a selective sweep was detected on a haplotype in *APOBEC3H* (Fujito et al., unpublished
489 data), the downstream region appears to be under the influence of another process of
490 evolution in terms of an arms race with viruses. Therefore, the regulatory region appears to be
491 the only region where natural selection can act solely on the transcriptional regulation of
492 *APOBEC3* genes. Additionally, we searched the GWAS catalog for variants in the *APOBEC3*

493 regulatory region. We found that rs5995653, which is linked to HII, is associated with asthma
494 exacerbation in admixed children (Hispanics/Latinos and African-Americans) treated with
495 inhaled corticosteroids ($p = 5 \times 10^{-6}$; OR = 1.52); this association is also found in Europeans (p
496 $= 7.52 \times 10^{-3}$) (Hernandez-Pacheco et al. 2019). These results suggest that HII affects the
497 overall immune state.

498 Owing to their mutagenic characteristics, APOBEC3s can potentially harm host
499 genomes as well. APOBEC3A and APOBEC3B are the primary sources of somatic mutations
500 in many tumors (Alexandrov et al. 2013; Roberts et al. 2013; Alexandrov et al. 2020). The
501 deletion polymorphism of *APOBEC3B* drives the strongest signal for APOBEC3B-like
502 mutagenesis at the pan-cancer level (ICGC/TCGA Pan-Cancer Analysis of Whole Genomes
503 Consortium 2020). Therefore, we examined the possibility that positive selection is triggered
504 by this deletion. However, we found no evidence of positive selection specifically targeting
505 the *APOBEC3B* deletion despite its functional impact. This finding suggests that the deletion
506 confers little to no selective advantage during evolution and that its increased frequency may
507 result simply from genetic hitchhiking of tightly linked HI. Within HI, sequences linked to the
508 deletion may have conferred greater advantages compared to other HI haplotypes. However,
509 the reduction in sequence diversity caused by additional positive selection acting on the
510 deletion is likely too small to be detected. This is because sequence diversity in HI is already
511 extremely low due to the effects of population bottlenecks and haplotype turnover. However,

512 even in such a case, the deletion may need to be linked to HI to have evolved, because when
513 linked to other sequences, it does not exhibit any sign of expansion (Figure S14).
514 Several issues should be addressed in future studies. First, more precise knowledge
515 about regulatory functions that confer a selective advantage to each haplogroup needs to be
516 elucidated (i.e., which *APOBEC3* gene is upregulated or downregulated as the core function
517 of each haplogroup). This was challenging to determine in the present study as research on
518 APOBEC3s is still ongoing, and specific immune cell types where the expression of each A3
519 is critical for the anti-viral activities are not yet known. We utilized the publicly available
520 eQTL data to examine the functional differences among the haplogroups; however, the
521 existing eQTL data are from whole-blood samples, and data for each immune cell type are not
522 available. Second, we could not identify viral targets of each haplogroup. Further elucidation
523 of the functions of each APOBEC3, especially single-cell eQTL analysis of immune cells,
524 might help link haplogroups to its target viruses. This will help to clarify the detailed selective
525 advantages of each haplogroup, enabling evaluation of the susceptibility of individuals to
526 viral infections.
527

528 **Conclusion**

529 We have demonstrated that the "multi-star" genealogy observed in the *APOBEC3*
530 regulatory region in 24 TGP populations is inconsistent with neutral evolution based on

coalescence simulations and genomic data. Most likely, the genealogy has been shaped by the combined effects of negative frequency-dependent selection and a severe reduction in effective population sizes, accompanied by the Out-of-Africa dispersal of modern humans. A relatively recent haplotype turnover within haplogroup I is also consistent with this hypothesis under certain mutation models. The action of such selection in the evolutionary conserved regulatory region underscores its functional importance in regulating the expression of the seven linked *APOBEC3* genes.

Data availability

We utilized whole genome sequence data sets that are publicly available from the 1000 Genomes Project Phase 3 (1000 Genomes Project Consortium et al. 2015) and the HGSVC (Ebert et al. 2021).

Acknowledgments

We thank Dr. Naoyuki Takahata for the useful discussions and critical comments, which greatly improved this study. We also thank Dr. Toshiyuki Hayakawa and Dr. Kohta Yoshida for their helpful discussions and comments. We thank Dr. Naruya Saitou for critically reviewing the manuscript. Computations were partially performed on the NIG supercomputer at ROIS National Institute of Genetics. This manuscript has been edited for English by Dr. Quintin Lau.

Funding

This study was funded by the Research Organization of Information and Systems (ROIS; https://www.rois.ac.jp), by JSPS KAKENHI [Grant Number 23K05870] and Shiseido Female Researcher Science Grant.

Materials and Methods

558 **Whole-genome sequencing (WGS) data for modern humans**

559 We used phased WGS data from the 1000 Genomes Project Phase 3 (1000 Genomes Project
560 Consortium et al. 2015). Samples from two populations, African Ancestry in Southeast US
561 (ASW) and African Caribbean in Barbados (ACB), were excluded owing to possible
562 admixture. The remaining 4694 samples from 24 populations were used. For African
563 populations, 216 samples for YRI (Yoruba in Ibadan, Nigeria), 198 for LWK (Luhya in
564 Webuye, Kenya), 226 for GWD (Gambian in Western divisions in the Gambia), 170 for MSL
565 (Mende in Sierra Leone), and 198 for ESN (Esan in Nigeria) were included. For non-African
566 populations, 198 samples for CEU (Utah residents with Northern and Western European
567 ancestry), 214 for TSI (Toscani in Italy), 198 for FIN (Finnish in Finland), 182 for GBR
568 (British in England and Scotland), and 214 for IBS (Iberian population in Spain) among
569 Europeans (EUR); 206 samples for GIH (Gujarati Indian from Houston, Texas, USA), 192 for
570 PJL (Punjabi from Lahore, Pakistan), 172 for BEB (Bengali from Bangladesh), 204 for STU
571 (Sri Lankan Tamil from the UK), and 204 for ITU (Indian Telugu from the UK) among South
572 Asians (SAS); 206 samples for CHB (Han Chinese in Beijing, China), 208 for JPT (Japanese
573 in Tokyo, Japan), 210 for CHS (Southern Han Chinese, China), 186 for CDX (Chinese Dai in
574 Xishuangbanna, China), and 198 for KHV (Kinh in Ho Chi Minh City, Vietnam) among East
575 Asians (EAS); and 128 samples for MXL (Mexican ancestry in Los Angeles, California), 208
576 for PUR (Puerto Rican in Puerto Rico), 188 for CLM (Colombian in Medellin, Colombia),

577 and 170 for PEL (Peruvian in Lima, Peru) for Americans (AMR) were included. Haplotypes
578 in the *APOBEC3* regulatory region were constructed by including SNPs on chr22: 39388993
579 to 39405510 (GRCh37). After removing indels using VCFtools (Danecek et al. 2011), 588
580 SNPs remained in the region. For time estimation, we obtained ancestral allele information
581 from the 1000 Genomes Project Phase 3 and used 584 biallelic SNPs with ancestral allele
582 information. The 16-kb *APOBEC3* regulatory region excluded the APOBEC3B deletion (as
583 well as identical 350-bp sequences flanking both ends of the deletion) (Figure S2). Moreover,
584 we checked the sequence similarity within the 200-kb *APOBEC3* cluster using a dot plot
585 (Figure S2) (Schwartz et al. 2000). We observed a cluster of short fragments (22–1668 bp)
586 with high sequence similarity (~98%) between the upstream and downstream areas of the
587 deletion. We confirmed that this cluster of sequences was not included in the 16-kb regulatory
588 region analyzed in this study (Figure S2). We also validated read mapping accuracy using
589 different data sets (Supplementary text).

590 We used a deletion polymorphism, "YL_CN_STU_4456," in the 1000 Genomes
591 Phase3 structural variant data as the linkage information of the APOBEC3B deletion
592 (Supplementary text). We retrieved the regulatory features from the Ensemble webpage
593 ("homo_sapiens.GRCh38.Regulatory_Build.regulatory_features.20220201.gff") and
594 subsequently lifted them to the GRCh37 reference genome using the UCSC webpage. To
595 confirm our results, we utilized WGS data from the HGSVC (Ebert et al. 2021), which are

596 based on long-read PacBio WGS data from 32 diverse human samples, and high-coverage

597 (30X) WGS data from the 1000 Genome Project resource (Byrska-Bishop et al. 2022), which

598 include data from 3202 samples (Supplementary text).

599

600 **Definition of the haplogroups**

601 We defined three haplogroups based on the following criteria: a frequency of $f > 0.15$, an

602 average nucleotide difference per site across all pairs between haplogroups of $d > 0.1\%$, and

603 nucleotide diversity within each haplogroup of $\pi < 0.01\%$ in non-African populations. In

604 defining these haplogroups, we excluded American populations, owing to known admixture.

605

606 **Data for ancient hominins and chimpanzees**

607 We retrieved WGS data for ancient hominins from

608 http://cdna.eva.mpg.de/neandertal/Vindija/. We used data from the Altai Neanderthal and

609 Denisova Pinky for the phylogenetic and network analyses. We obtained sequence data for

610 chimpanzees (NC_036901.1) and orangutans (Susie_PABv2) from the NCBI and the

611 Ensemble webpages, respectively.

612

613 **Phylogenetic analysis**

614 We constructed Neighbor-joining trees using MEGA 11 software (Saitou and Nei 1987;
615 Tamura et al. 2021) based on haplotype sequences in the 16-kb region. We computed
616 evolutionary distances using the number of differences method. We considered data for the
617 JPT, CEU, PJL, PUR, YRI, and LWK populations as representative of the East Asian,
618 European, South Asian, American, and African populations, respectively.
619

620 **Identification of haplogroups and recombinants**

621 Based on the topology of the NJ trees, we defined one of the SNPs on the branch that led to
622 each of the haplogroups as an index SNP: rs4303815 for H I, rs35043274 for HII, and
623 rs111468681 for HIII.

624 We identified and removed recombinants from the time estimation and haplogroup
625 diversity measurements (Table S4). We calculated the number of nucleotide differences for all
626 sequence pairs within a haplogroup to identify the recombinants. We then plotted their
627 distributions and identified outlier pairs. Sequences that caused outliers were regarded as
628 recombinant and removed from the analyses.
629

630 **Estimation of divergence time among the haplogroups**

631 We estimated the divergence time between each haplogroup and other sequences using the
632 number of nucleotide differences. Specifically, based on the phylogenetic tree using

633 sequences from African populations (Figure S4a), we grouped four groups of sequences that
634 do not belong to any of the three haplogroups in addition to three haplogroups. We calculate
635 the average number of nucleotide differences from all pairs among the seven sequence groups
636 in total. Using these calculated numbers of differences, we estimated the divergence time of
637 the three haplogroups using the mutation rate of 0.5×10^{-9}/site/year.
638

639 **Frequency trajectories over time**

640 We retrieved genotype data from the Allen Ancient Genome Resource (Mallick et al. 2023;
641 Anon) (version 52.2, downloaded from https://reich.hms.harvard.edu/allen-ancient-dna-
642 resource-aadr-downloadable-genotypes-present-day-and-ancient-dna-data). We also used
643 genotypes of samples from Europe (4624 samples) and East Asia (898 samples) from 100–
644 40,000 years ago. As the index SNPs for HI, HII, and HIII were not included in these data, we
645 utilized rs2142833, rs9607601, and rs11089916, respectively, as tag SNPs.
646

647 **Estimation of the time of expansion of the haplogroups**

648 We estimated the time of expansion for each haplogroup by measuring intra-allelic diversity
649 using two different methods. The first method is estimating the mean coalescence time within
650 a haplotype group. We calculated the average number of nucleotide differences across all
651 sequence pairs within a haplotype; rescaled the average with the length of the sequences to

652 obtain π values (per site nucleotide diversity); and then converted the π values to years

653 with mutation rate (using $\pi = 2t\mu$, where t is the time since the common ancestor of the

654 sequences in a haplogroup and μ is the mutation rate per site per year), assuming that the

655 mutation rate was 0.5×10^{-9}/site/year. Standard error estimates were obtained by a bootstrap

656 procedure using MEGA 11.

657 The second method involved calculating the height of the genealogy of each

658 haplogroup (Satta et al. 2019). Based on the ancestral allele information, we identified sites

659 where the derived allele was segregating in a haplogroup of interest, and the ancestral allele

660 was fixed in populations outside of it. Using these sites, we counted the number of derived

661 alleles accumulated within a haplogroup divided by the number of sequences to obtain the

662 height of the haplogroup, which was converted to years, as described above. We calculated

663 the variance in height using the formula described by Satta et al. (2019). For the calculation of

664 the tree heights, we removed the sites at which i $(k - 2 \leq i < k$, where k is the size of

665 the haplotype group) derived alleles were identified within the haplotype of interest, but no

666 derived alleles were present outside of it. This is because these sites likely result from

667 recombination, which introduces a limited number of ancestral alleles into the haplotype of

668 interest.

669

670 **Inference of selection**

671 To demonstrate the rarity of the phylogenetic structure with three large haplogroups, we
672 searched for a phylogenetic structure with three distinct haplogroups, each of which was
673 comparable to or surpassed the observed structure in haplogroup size (i.e., the number of
674 chromosomes) in both simulated data under neutrality and empirical data. We performed
675 simulations using ms (Hudson 2002) under the two demographic models proposed for East
676 Asians, Europeans, and Africans, which were subsequently referred to as the SFD and HFD.
677 SFD is based on a model proposed by Schaffner et al. (Schaffner et al. 2005). HFD includes
678 an additional population bottleneck around 900,000 years ago, which was recently identified
679 by Hu et al. (Hu et al. 2023). We mapped a specified (observed) number of segregating sites
680 onto a coalescent tree. The command lines used for this process are provided in
681 Supplementary Table S9. For the empirical data, we used a window-based approach (16-kb
682 windows with a step size of 1 kb) with the 1000 Genome data. We screened for an SNP that
683 matched the comparable frequency observed for HI from the 5′- to 3′-ends of both simulated
684 and empirical data, which was then defined as an index SNP of HI. Next, we searched for
685 SNP with the exact frequency observed for HII and derived alleles only on non-HI
686 chromosomes. The closest SNP to the 5′-end was defined as an index SNP for HII. Similarly,
687 the index SNP for HIII was determined using the same approach.
688 Next, we assessed the frequency with which the observed pattern of low intra-
689 haplogroup diversity and high inter-haplogroup divergence co-occurred in the results of

690 neutral simulations. We retrieved 10,000 replications from the simulation described above,
691 whereby the three allelic groups exhibited the same size as the observations (e.g., 103, 76, and
692 19 chromosomes for HI, HII, and HIII in JPT, respectively), and calculated nucleotide
693 differences within and between clades. Subsequently, we investigated the rarity of
694 simultaneously observing low diversity within each of the haplogroups and high divergence
695 between the haplogroups under neutrality (i.e. reject neutrality if the combination is less than
696 1%). The comparison was performed jointly for HI, HII, and HIII.
697
698 **Negative frequency-dependent selection model**
699 We used Monte Carlo forward simulation to explore forms of balancing selection that can
700 consistently explain patterns of genetic diversity for African and non-African populations. We
701 used a demographic model similar to the HSD model, which includes the Out-of-Africa
702 bottleneck and the bottleneck in Africa around 900 KYA. We also accounted for recent
703 population expansions by estimating the harmonic means of exponential growth, based on
704 Gao and Keinan 2016. We assumed that the current population size for both Africans and
705 non-Africans is 500,000 individuals. For African populations, the expansion begins from
706 20,000 individuals at generation 0 and reaches 500,000 individuals at generation 330. In
707 contrast, for non-African populations, it starts from 2,000 individuals at generation 0 and
708 grows to 500,000 individuals by generation 660. We assumed a simple simulation model

where organisms in a diploid population reproduce the next generation under the Wright-Fisher model. Each organism has a chromosome of length $(L + 1)$, which consists of one site under balancing selection and L linked neutral sites without recombination. The selected site undergoes three simultaneous mutations at the start point, defining four sequence groups that mimic three haplogroups and an ancestral sequence group. At the linked neutral sites, mutations occur at the per-site rate of μ, with reverse mutations excluded. The three haplogroups are subject to negative frequency-dependent selection, while the ancestral group possesses constant fitness. The relative fitness for haplogroups is calculated as $1 - sx$, where x is the frequency of each haplogroup and s is the shared selection coefficient. The relative fitness for the ancestral group is defined as $1 - as$, where a is a constant value. To illustrate the shits in within-haplogroup diversity (π) and between-haplogroup divergence and the frequency of sequence groups, we performed simulations with parameters: $L = 160$, and $s = 0.05,\ 0.2$ and $a = 0.2, 0.3$. In this simulation, each neutral site represents a sequence block of 100 bp, effectively mimicking the *APOBEC3* regulatory region (160 kb in length). We selected an effective population size (N_e) and mutation rate such that the production of $4N_e\mu$ corresponds to the averaged π over human genomes (0.1%). To reduce the simulation time, we reduced the population size and the number of generations to one-tenth of their original values, while increasing the mutation rate increased tenfold. The generation time was set to 25 years.

728

729 **Overdominance model**

730 In the overdominance model, we modified the negative frequency-dependent selection model

731 described above by altering the definitions of relative fitness. We assumed that all the four

732 sequence groups are under overdominance. The relative fitness for homozygotes of each

733 haplogroup is calculated as $1 - s$, and as $1 - as$ for homozygotes of the ancestral group.

734 The relative fitness for heterozygotes is assumed to be 1.

735

736

References

1000 Genomes Project Consortium, Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, Korbel JO, Marchini JL, McCarthy S, McVean GA, et al. 2015. A global reference for human genetic variation. *Nature* [Internet] 526:68–74. Available from: http://dx.doi.org/10.1038/nature15393

Alexandrov LB, Kim J, Haradhvala NJ, Huang MN, Tian Ng AW, Wu Y, Boot A, Covington KR, Gordenin DA, Bergstrom EN, et al. 2020. The repertoire of mutational signatures in human cancer. *Nature* [Internet] 578:94–101. Available from: https://www.nature.com/articles/s41586-020-1943-3

Alexandrov LB, Nik-Zainal S, Wedge DC, Aparicio SAJR, Behjati S, Biankin AV, Bignell GR, Bolli N, Borg A, Børresen-Dale A-L, et al. 2013. Signatures of mutational processes in human cancer. *Nature* [Internet] 500:415–421. Available from: https://www.nature.com/articles/nature12477

Anon. The Allen Ancient DNA Resource (AADR): A curated compendium of ancient human genomes. *David Reich Lab Dataverse* [Internet]. Available from: https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/FFIDCW

753 Bitarello BD, Brandt DYC, Meyer D, Andrés AM. 2023. Inferring balancing selection from
754 genome-scale data. *Genome Biol. Evol.* [Internet] 15:evad032. Available from:
755 https://academic.oup.com/gbe/article/15/3/evad032/7055906

756 Byrska-Bishop M, Evani US, Zhao X, Basile AO, Abel HJ, Regier AA, Corvelo A, Clarke
757 WE, Musunuri R, Nagulapalli K, et al. 2022. High-coverage whole-genome sequencing of the
758 expanded 1000 Genomes Project cohort including 602 trios. *Cell* [Internet] 185:3426-
759 3440.e19. Available from: http://dx.doi.org/10.1016/j.cell.2022.08.004

760 Charlesworth B, Jensen JD. 2021. Effects of selection at linked sites on patterns of genetic
761 variability. *Annu. Rev. Ecol. Evol. Syst.* [Internet] 52:177–197. Available from:
762 https://pmc.ncbi.nlm.nih.gov/articles/PMC10120885/

763 Charlesworth D. 2006. Balancing selection and its effects on sequences in nearby genome
764 regions. *PLoS Genet.* [Internet] 2:e64. Available from:
765 http://dx.doi.org/10.1371/journal.pgen.0020064

766 Chintalapati M, Moorjani P. 2020. Evolution of the mutation rate across primates. *Curr. Opin.*
767 *Genet. Dev.* [Internet] 62:58–64. Available from:
768 https://www.sciencedirect.com/science/article/pii/S0959437X20300794

769 Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE, Lunter
770 G, Marth GT, Sherry ST, et al. 2011. The variant call format and VCFtools. *Bioinformatics*
771 [Internet] 27:2156–2158. Available from: http://dx.doi.org/10.1093/bioinformatics/btr330

772 Ebert P, Audano PA, Zhu Q, Rodriguez-Martin B, Porubsky D, Bonder MJ, Sulovari A, Ebler
773 J, Zhou W, Serra Mari R, et al. 2021. Haplotype-resolved diverse human genomes and
774 integrated analysis of structural variation. *Science* [Internet] 372. Available from:
775 http://dx.doi.org/10.1126/science.abf7117

776 Field Y, Boyle EA, Telis N, Gao Z, Gaulton KJ, Golan D, Yengo L, Rocheleau G, Froguel P,
777 McCarthy MI, et al. 2016. Detection of human adaptation during the past 2000 years. *Science*
778 [Internet] 354:760–764. Available from: http://dx.doi.org/10.1126/science.aag0776

779 Fijarczyk A, Babik W. 2015. Detecting balancing selection in genomes: limits and prospects.
780 *Mol. Ecol.* [Internet] 24:3529–3545. Available from: http://dx.doi.org/10.1111/mec.13226

781 Fujito NT, Satta Y, Hayakawa T, Takahata N. 2018. A new inference method for detecting an
782 ongoing selective sweep. *Genes Genet. Syst.* [Internet] 93:149–161. Available from:
783 http://dx.doi.org/10.1266/ggs.18-00008

784 Gao F, Keinan A. 2016. Explosive genetic evidence for explosive human population growth.
785 *Curr. Opin. Genet. Dev.* [Internet] 41:130–139. Available from:
786 https://pubmed.ncbi.nlm.nih.gov/27710906/

787 Hakata Y, Li J, Fujino T, Tanaka Y, Shimizu R, Miyazawa M. 2019. Mouse APOBEC3
788 interferes with autocatalytic cleavage of murine leukemia virus Pr180gag-pol precursor and
789 inhibits Pr65gag processing. *PLoS Pathog.* [Internet] 15:e1008173. Available from:
790 http://dx.doi.org/10.1371/journal.ppat.1008173

791 Hernandez-Pacheco N, Farzan N, Francis B, Karimi L, Repnik K, Vijverberg SJ, Soares P,
792 Schieck M, Gorenjak M, Forno E, et al. 2019. Genome-wide association study of inhaled
793 corticosteroid response in admixed children with asthma. *Clin. Exp. Allergy* [Internet]
794 49:789–798. Available from: http://dx.doi.org/10.1111/cea.13354

795 Hu W, Hao Z, Du P, Di Vincenzo F, Manzi G, Cui J, Fu Y-X, Pan Y-H, Li H. 2023. Genomic
796 inference of a severe human bottleneck during the Early to Middle Pleistocene transition.
797 *Science* [Internet] 381:979–984. Available from:
798 https://www.science.org/doi/10.1126/science.abq7487

799 Hudson RR. 2002. Generating samples under a Wright–Fisher neutral model of genetic
800 variation. *Bioinformatics* [Internet] 18:337–338. Available from:
801 https://academic.oup.com/bioinformatics/article/18/2/337/225783

802 Hughes AL, Nei M. 1988. Pattern of nucleotide substitution at major histocompatibility
803 complex class I loci reveals overdominant selection. *Nature* [Internet] 335:167–170.
804 Available from: https://pubmed.ncbi.nlm.nih.gov/3412472/

ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium. 2020. Pan-cancer analysis of whole genomes. *Nature* [Internet] 578:82–93. Available from: http://dx.doi.org/10.1038/s41586-020-1969-6

Jónsson H, Sulem P, Kehr B, Kristmundsdottir S, Zink F, Hjartarson E, Hardarson MT, Hjorleifsson KE, Eggertsson HP, Gudjonsson SA, et al. 2017. Parental influence on human germline de novo mutations in 1,548 trios from Iceland. *Nature* [Internet] 549:519–522. Available from: http://dx.doi.org/10.1038/nature24018

Kidd JM, Newman TL, Tuzun E, Kaul R, Eichler EE. 2007. Population stratification of a common APOBEC gene deletion polymorphism. *PLoS Genet.* [Internet] 3:e63. Available from: http://dx.doi.org/10.1371/journal.pgen.0030063

Mallick S, Micco A, Mah M, Ringbauer H, Lazaridis I, Olalde I, Patterson N, Reich D. 2023. The Allen Ancient DNA Resource (AADR): A curated compendium of ancient human genomes. *bioRxiv* [Internet]:2023.04.06.535797. Available from: https://www.biorxiv.org/content/10.1101/2023.04.06.535797v1

Maruyama T, Nei M. 1981. Genetic variability maintained by mutation and overdominant selection in finite populations. *Genetics* [Internet] 98:441–459. Available from: https://pubmed.ncbi.nlm.nih.gov/17249094/

822 Meyer M, Kircher M, Gansauge M-T, Li H, Racimo F, Mallick S, Schraiber JG, Jay F, Prüfer
823 K, de Filippo C, et al. 2012. A high-coverage genome sequence from an archaic Denisovan
824 individual. *Science* [Internet] 338:222–226. Available from:
825 https://www.science.org/doi/10.1126/science.1224344

826 Newman ENC, Holmes RK, Craig HM, Klein KC, Lingappa JR, Malim MH, Sheehy AM.
827 2005. Antiviral function of APOBEC3G can be dissociated from cytidine deaminase activity.
828 *Curr. Biol.* [Internet] 15:166–170. Available from:
829 http://dx.doi.org/10.1016/j.cub.2004.12.068

830 Pollpeter D, Parsons M, Sobala AE, Coxhead S, Lang RD, Bruns AM, Papaioannou S,
831 McDonnell JM, Apolonia L, Chowdhury JA, et al. 2018. Deep sequencing of HIV-1 reverse
832 transcripts reveals the multifaceted antiviral functions of APOBEC3G. *Nat Microbiol*
833 [Internet] 3:220–233. Available from: http://dx.doi.org/10.1038/s41564-017-0063-9

834 Prüfer K, Racimo F, Patterson N, Jay F, Sankararaman S, Sawyer S, Heinze A, Renaud G,
835 Sudmant PH, de Filippo C, et al. 2014. The complete genome sequence of a Neanderthal from
836 the Altai Mountains. *Nature* [Internet] 505:43–49. Available from:
837 http://dx.doi.org/10.1038/nature12886

838 Revathidevi S, Murugan AK, Nakaoka H, Inoue I, Munirajan AK. 2021. APOBEC: A
839 molecular driver in cervical cancer pathogenesis. *Cancer Lett.* [Internet] 496:104–116.
840 Available from: http://dx.doi.org/10.1016/j.canlet.2020.10.004

841 Roberts SA, Lawrence MS, Klimczak LJ, Grimm SA, Fargo D, Stojanov P, Kiezun A,
842 Kryukov GV, Carter SL, Saksena G, et al. 2013. An APOBEC cytidine deaminase
843 mutagenesis pattern is widespread in human cancers. *Nat. Genet.* [Internet] 45:970–976.
844 Available from: https://www.nature.com/articles/ng.2702

845 Sabeti PC, Reich DE, Higgins JM, Levine HZP, Richter DJ, Schaffner SF, Gabriel SB, Platko
846 JV, Patterson NJ, McDonald GJ, et al. 2002. Detecting recent positive selection in the human
847 genome from haplotype structure. *Nature* [Internet] 419:832–837. Available from:
848 https://www.nature.com/articles/nature01140

849 Saitou N, Nei M. 1987. The neighbor-joining method: a new method for reconstructing
850 phylogenetic trees. *Mol. Biol. Evol.* [Internet] 4:406–425. Available from:
851 https://academic.oup.com/mbe/article/4/4/406/1029664

852 Satta Y, Zheng W, Nishiyama KV, Iwasaki RL, Hayakawa T, Fujito NT, Takahata N. 2019.
853 Two-dimensional site frequency spectrum for detecting, classifying and dating incomplete
854 selective sweeps. *Genes Genet. Syst.* [Internet] 94:283–300. Available from:
855 http://dx.doi.org/10.1266/ggs.19-00012

856 Schaffner SF, Foo C, Gabriel S, Reich D, Daly MJ, Altshuler D. 2005. Calibrating a
857 coalescent simulation of human genome sequence variation. *Genome Res.* [Internet] 15:1576–
858 1583. Available from: http://genome.cshlp.org/content/15/11/1576.abstract

859 Schwartz S, Zhang Z, Frazer KA, Smit A, Riemer C, Bouck J, Gibbs R, Hardison R, Miller
860 W. 2000. PipMaker—A Web Server for Aligning Two Genomic DNA Sequences. *Genome*
861 *Res.* [Internet] 10:577–586. Available from: http://genome.cshlp.org/content/10/4/577.abstract

862 Siewert KM, Voight BF. 2017. Detecting long-term balancing selection using allele frequency
863 correlation. *Mol. Biol. Evol.* [Internet] 34:2996–3005. Available from:
864 https://academic.oup.com/mbe/article/34/11/2996/3988103

865 Siewert KM, Voight BF. 2020. BetaScan2: Standardized Statistics to Detect Balancing
866 Selection Utilizing Substitution Data. *Genome Biol. Evol.* [Internet] 12:3873–3877. Available
867 from: https://academic.oup.com/gbe/article/12/2/3873/5721358

868 Spurgin LG, Richardson DS. 2010. How pathogens drive genetic diversity: MHC,
869 mechanisms and misunderstandings. *Proc. Biol. Sci.* [Internet] 277:979–988. Available from:
870 http://dx.doi.org/10.1098/rspb.2009.2084

871 Stavrou S, Ross SR. 2015. APOBEC3 Proteins in Viral Immunity. *J. Immunol.* [Internet]
872 195:4565–4570. Available from: http://dx.doi.org/10.4049/jimmunol.1501504

Takahata N. 1990. A simple genealogical structure of strongly balanced allelic lines and trans-species evolution of polymorphism. *Proc. Natl. Acad. Sci. U. S. A.* [Internet] 87:2419–2423. Available from: http://dx.doi.org/10.1073/pnas.87.7.2419

Takahata N, Nei M. 1990. Allelic genealogy under overdominant and frequency-dependent selection and polymorphism of major histocompatibility complex loci. *Genetics* [Internet] 124:967–978. Available from: http://dx.doi.org/10.1093/genetics/124.4.967

Tamura K, Stecher G, Kumar S. 2021. MEGA11: Molecular Evolutionary Genetics Analysis Version 11. *Mol. Biol. Evol.* [Internet] 38:3022–3027. Available from: http://dx.doi.org/10.1093/molbev/msab120

Voight BF, Kudaravalli S, Wen X, Pritchard JK. 2006. A map of recent positive selection in the human genome. *PLoS Biol.* [Internet] 4:e72. Available from: http://dx.doi.org/10.1371/journal.pbio.0040072

Wang X, Ao Z, Chen L, Kobinger G, Peng J, Yao X. 2012. The cellular antiviral protein APOBEC3G interacts with HIV-1 reverse transcriptase and inhibits its function during viral replication. *J. Virol.* [Internet] 86:3777–3786. Available from: http://dx.doi.org/10.1128/JVI.06594-11

891 **Tables**

892 **Table 1. Intra- and inter-haplogroup diversity and divergence of HI, HII, and HIII**

893

Population	Haplogroup	Frequency	π or d (%)	All pairwise π (%)
JPT	HI	0.50	0.0056	0.10
	HII	0.37	0.0065	
	HIII	0.09	0.0096	
	HI - HII	-	0.1821	
	HI - HIII	-	0.1394	
	HII - HIII	-	0.1448	
CEU	HI	0.41	0.0073	0.10
	HII	0.21	0.0062	
	HIII	0.22	0.0039	
	HI - HII	-	0.1803	
	HI - HIII	-	0.1386	
	HII - HIII	-	0.1482	
LWK	HI	0.27	0.0028	0.14
	HII	0.13	0.0198	
	HIII	0.15	0.0168	
	HI - HII	-	0.1894	
	HI - HIII	-	0.1361	
	HII - HIII	-	0.1529	

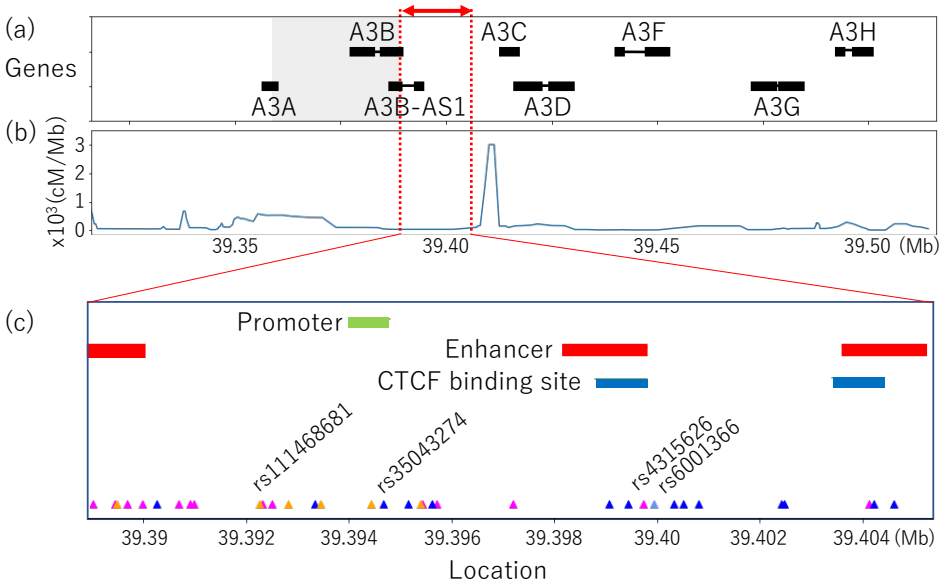
894 Intra-haplogroup diversity (π) and inter-haplogroup divergence (d) for each haplogroup in

895 JPT, CEU, and LWK populations are presented together with haplogroup frequency and

896 overall diversity in the *APOBEC3* regulatory region (all pairwise π).

897 **Figures**

898



899

900 **Figure 1.** The APOBEC3 cluster including the regulatory region.

901 (a). The APOBEC3 cluster region, which spans 200 kb surrounding the APOBEC3 genes, is

902 illustrated. A3A, A3B, A3C, A3D, A3F, A3G, and A3H are abbreviations for APOBEC3A,

903 APOBEC3B, APOBEC3C, APOBEC3D, APOBEC3F, APOBEC3G, and APOBEC3H,

904 respectively. A long non-coding RNA, APOBEC3B antisense RNA 1 (A3B-AS1), is also

905 present in this region. The 16-kb regulatory region is represented by the double-headed red

906 arrow, and the APOBEC3B deletion polymorphism is indicated by the shaded square. (b) The

907 recombination rate in the APOBEC3 cluster region is plotted based on the HapMap phase II

908 genetic map. (c) Annotated regulatory sequences in the Ensemble Regulatory Build. The

909 SNPs in LD (r^2> 0.8) with the index SNPs for HI (magenta), HII (blue), and HIII (orange) are

910 depicted with triangles.

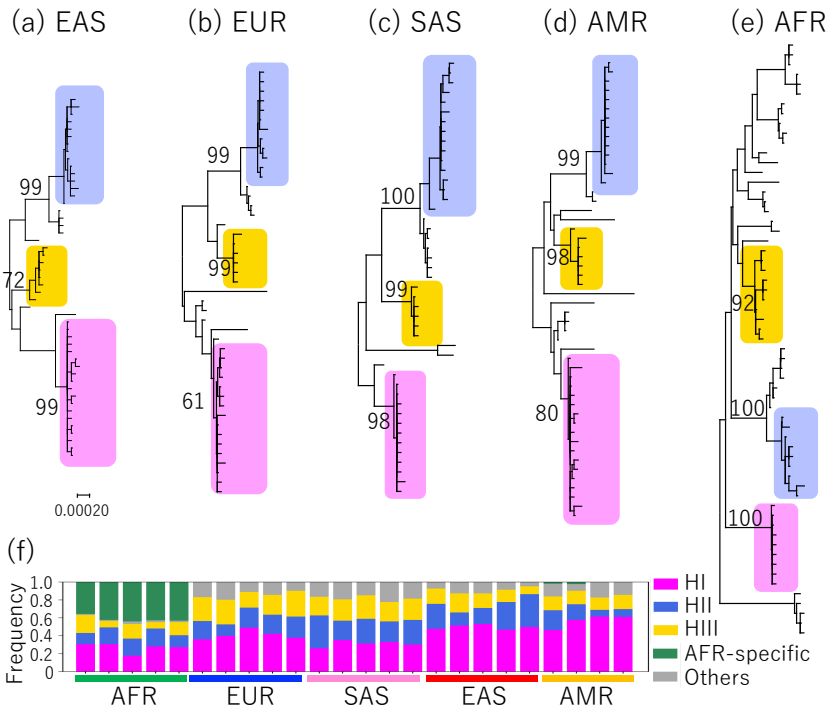


Figure 2. The three haplogroups observed in the APOBEC3 regulatory region. Phylogenetic trees were constructed using the Neighbor-joining method based on haplotype sequences identified in East Asian (a, 49 haplotype sequences from JPT), European (b, 45 haplotype sequences from CEU), South Asian (c, 45 haplotype sequences from PJL), American (d, 50 haplotype sequences from PUR), and African (e, 61 haplotype sequences from LWK) populations. Bootstrap values after 1000 replications are presented only for the three haplogroups: HI (magenta), HII (blue), and HIII (orange). The trees are drawn to scale, with branch lengths in the same units as those of the evolutionary distances used to infer the phylogenetic tree. The evolutionary distances were computed using the Maximum Composite Likelihood method and are displayed in the unit of base substitutions per site. (f) Proportion of the three haplogroups in each population in the 1000 Genomes project data. The sequences

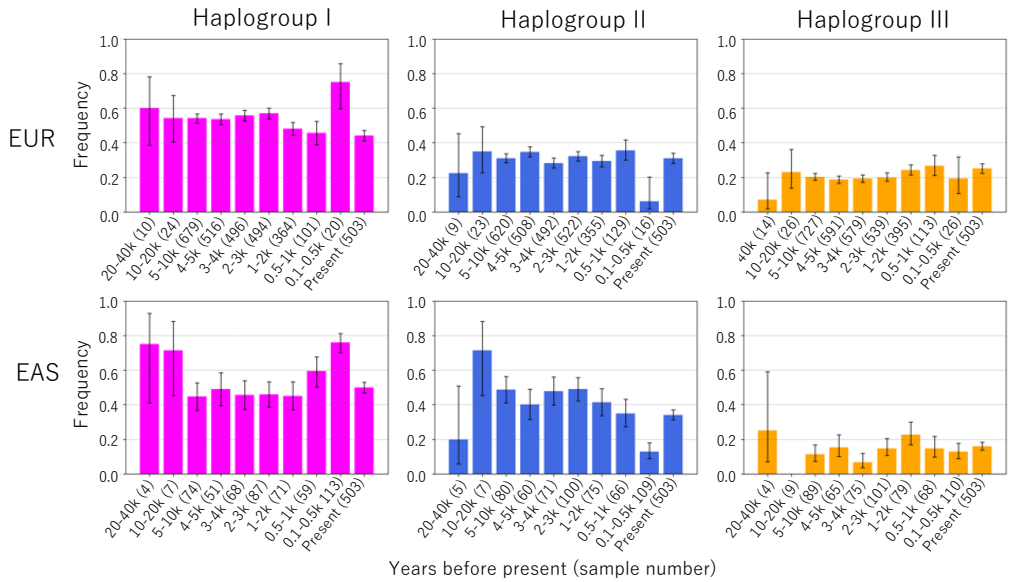
923 that do not belong to any of the three haplogroups are separated into two groups: African-

924 specific sequences (green) and others (gray).

925

926

927



928

929 **Figure 3**. Frequencies of the three haplogroups in modern humans across time. Samples from

930 Europe (top) and East Asia (bottom) are from compiled ancient data from the Allen Ancient

931 Genome Resource and current data from the 1000 Genome Project. Frequency of the allele

932 linked to each haplogroup was calculated at the informative SNPs (rs2142833, rs9607601,

933 and rs11089916 for HI, HII, and HIII, respectively). As frequency was calculated after

934 removing the samples for which genotypes were not available at each of the informative

935 SNPs, the sum of the frequencies of the three informative SNPs is not necessarily equal to 1.

936 The number of samples is shown in parentheses. Time periods are indicated in years before

937 the present. Error bars indicate Wilson's confidence interval (95%).

938

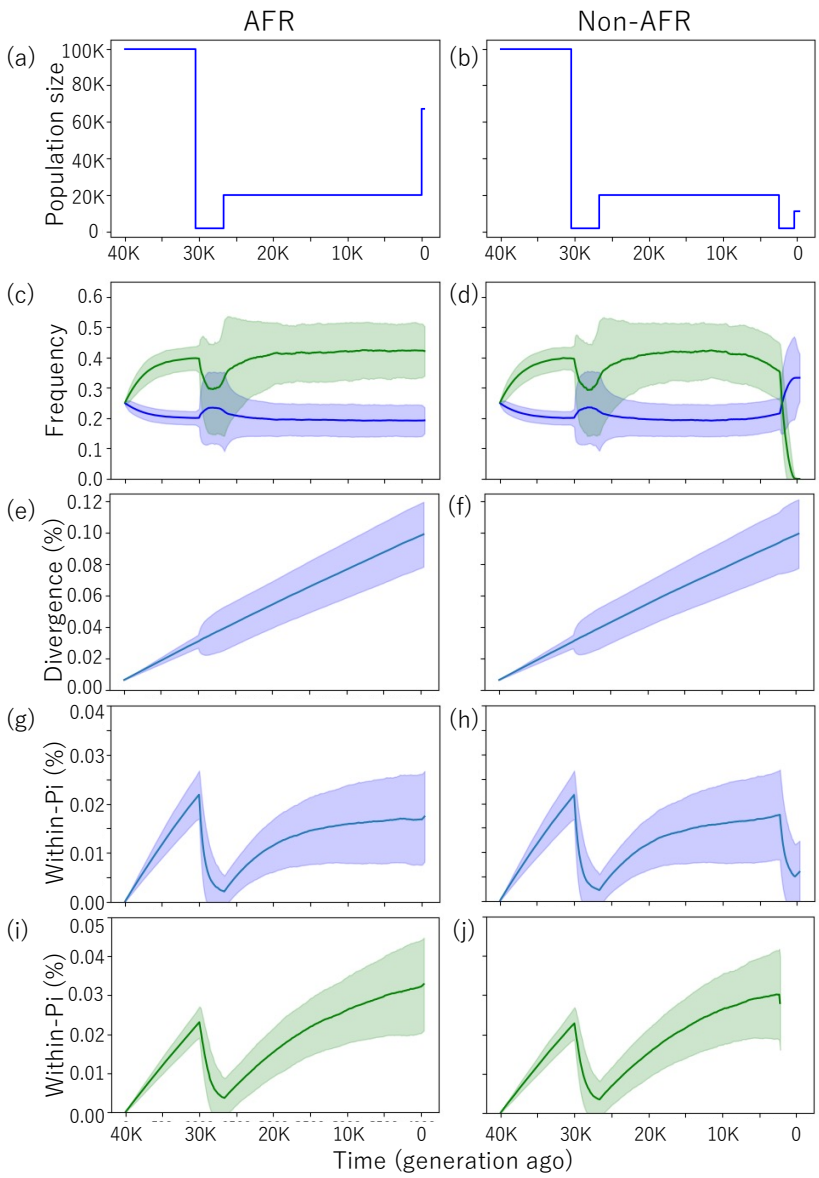


Figure 4. Simulation study using a model of negative frequency-dependent selection with demographic models for African (a) and non-African populations (b). The simulations assumed three haplogroups and one ancestral group, with relative fitness for each haplogroup given by $1 - sx$. The ancestral group fitness was assumed to be $1 - as$. The mutation rate was assumed to be 1.25×10^{-8} per site per generation. Simulations were conducted with population sizes and time duration reduced by a factor of 1/10 and a mutation rate increased by a factor of 10. (For the initial phase, the mutation rate was further increased by a factor of

5 to reduce the computational time.) Shifts in haplogroup frequency, intra-haplogroup diversity, and inter-haplogroup divergence are illustrated for conditions with $s = 0.05, a = 0.2$ for Africans and non-Africans. Only replications in which the observed state was replicated (i.e., the persistence of the three haplogroups and the ancestral group throughout in Africans) were utilized. For non-Africans, replications were considered in which the three haplogroups persisted, while the ancestral group was lost during the Out-of-Africa bottleneck. Lines and shaded areas represent the mean values and standard deviations across 1,000 replications in which the observed sequence groups persisted. Frequency trajectories for haplogroups and the ancestral group are depicted with blue and green lines, respectively, for African (c) and non-African populations (d). Shifts in simulated inter-haplogroup diversity are presented for African (e) and non-African (f) populations. Intra-haplogroup diversity is similarly presented for African (g) and non-African (h) populations. Diversity within the ancestral group was also shown for African (i) and non-African (j) populations.

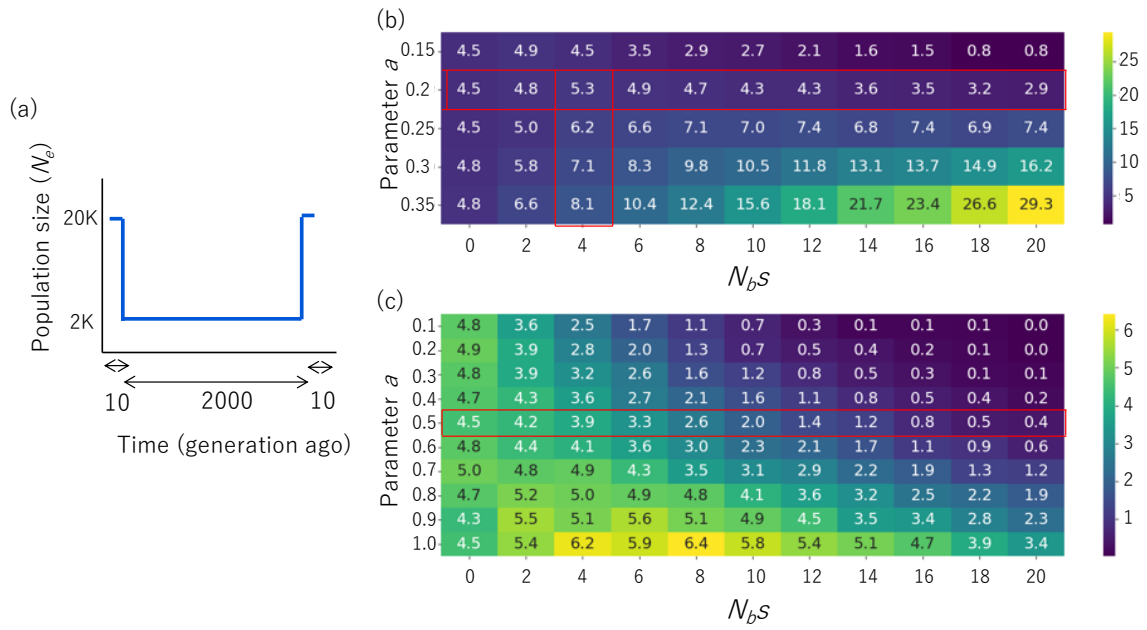


Figure 5. Probability of observed haplotype persistence during the Out-of-Africa bottleneck. The proportions of simulation replicates in which all three haplogroups persisted while the ancestral group was lost were calculated under models of negative frequency-dependent selection (b) and overdominance (c) each based on 10,000 replicates, and are presented as percentages. Panel (a) illustrates the demographic model incorporating a population bottleneck. The initial frequencies of the ancestral sequence group and each of the three haplogroups were 0.4 and 0.2, respectively. The red horizontal bars indicate the probabilities corresponding to the value of *a* that reproduces the initial frequencies under each model. The red vertical bars in (b) highlight the likely value of N_bs that, in combination with the *a* value, produces the observed state with a probability greater than 5 %.