Distinct haplogroups with star-like diversification in the APOBEC3 regulatory region indicate ancient viral pandemics before and during the Out-of-Africa migration of modern humans

Naoko T Fujito[1], Revathi Devi Sundaramoorthy[1], Yoko Satta[2], and Ituro Inoue[1]

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

[2] School of Advanced Science, SOKENDAI (The Graduate University for Advanced Studies), Hayama, Kanagawa, Japan

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

Abstract

APOBEC3s are powerful human anti-viral proteins that inhibit viral replication and proliferation by inducing mutations in viral genomes. We found that in the regulatory region of the APOBEC3 cluster on human chromosome 22, there exist three distinct haplogroups that began to have diverged nearly 1 million years ago and have since been maintained in both African and non-African populations today. Despite the long persistence time of the haplogroups, one of which is shared by archaic hominins, we also found that the nucleotide diversity within each haplogroup is extremely low. With a simulation-based method, the observed low diversity within haplogroups and large divergence between haplogroups suggested that these haplogroups have been maintained by special forms of positive Darwinian selection. Repeated occurrence of such selection entails the enhanced substitution rate in this regulatory region in both the human and chimpanzee lineages. The causal events of such positive selection, which are most likely viral epidemics, are shown to have occurred at least three times in the modern human population: one before the Out-of-Africa migration and another two in the common ancestors of non-Africans.

Introduction

Infectious diseases are one of the most substantial selective pressures affecting human genome evolution (Fumagalli et al. 2011). When a host population is affected by clustered infections from a pathogen, adaptations occur at host loci that interact with the pathogen (Quintana-Murci and Clark 2013; Karlsson et al. 2014). Thus, epidemics in the past could be inferred by detecting signatures of adaptive evolution on host immune-related loci (Enard and Petrov 2020). There are many known examples, where variants, which confer resistance to specific pathogens, show signatures of positive selection, suggesting outbreaks of the same or related pathogens in the past (Kwiatkowski 2005; Karlsson et al. 2013; Klunk et al. 2022). The evolutionary analyses of selection signatures on human immune-related loci would be important for elucidating the dynamics of the human immune system and help to understand current differences in susceptibility to infections among individuals. It would even make it possible to predict the susceptibility of individuals to infectious diseases and the amount of risk that a certain pathogen has in causing future epidemics.

Among pathogens, viruses are regarded as a dominant driver of adaptation (Enard et al. 2016; Enard and Petrov 2020). Enrichment analyses in selective sweeps at virus-interacting proteins suggested that viruses, particularly RNA viruses, have been a significant selective pressure in the past 50,000 years of human evolution (Enard and Petrov 2020). In the present study, we focused on one of the most powerful human anti-

virus factors, APOBEC3 (apolipoprotein B mRNA editing enzyme, catalytic polypeptide-like 3, A3). APOBEC3 proteins are a family of cytidine deaminases and constitute a front line of defense against retroviruses, endogenous retro-elements, and DNA viruses in the innate immune system. They catalyze cytidine deamination of single-stranded DNA or RNA, leading to cytidine to uridine (C > U) mutations in viral genomes. This C to U hypermutation disrupts the genomic information of viruses, inhibiting their replication and proliferation, which then leads to degradation via the UNG2 (uracil DNA glycosylase)-dependent pathway. It is also known that APOBEC3s can inhibit viral replication through deaminase-independent mechanisms (Newman et al. 2005; Wang et al. 2012; Pollpeter et al. 2018; Hakata et al. 2019). The APOBEC3 family consists of seven enzymes (A3A/B/C/DE/F/G/H), and their genes are clustered within a 200-kb region on chromosome 22 (the APOBEC3 cluster, Figure 1a). Each of the seven members has distinct cellular localization, tissue-specificity in expression, substrate-specificity (single-stranded DNA or RNA), and viral restriction properties, but their functions overlap each other. For example, APOBEC3G acts as the most powerful anti-HIV factor, yet APOBEC3D, 3F, and 3H also contribute to restriction of HIV. APOBEC3A, 3B, 3C, 3F, and 3G restrict DNA viruses, including hepatitis B virus, human papillomavirus, adeno-associated virus, herpes simplex-1, and Epstein-Barr virus [Reviewed in] (Stavrou and Ross 2015; Revathidevi et al. 2021).

Because of their mutagenic characters, APOBEC3s potentially harm the host genomes as well. A3A and A3B are known as the primary sources of C to T somatic mutations in many tumors (Alexandrov et al. 2013; Roberts et al. 2013; Alexandrov et al. 2020). One well-known variant associated with cancer risks in APOBEC3 genes is a deletion polymorphism in *A3B*. This deletion spans the 29.5-kb region from intron 4 of *A3A* to exon 8 of the adjacent *A3B*, resulting in a complete loss of the protein-coding region of *A3B*. Subsequently, a chimeric transcript fusing the *A3A* coding sequence with the 3' UTR of *A3B* is generated, and the fused-gene transcript has been reported to be more stable than the original transcript of *A3A* (Caval et al. 2014; Nik-Zainal et al. 2014). The *A3B* deletion polymorphism is rare in Africans (frequency of 0.9 %) and Europeans (6 %) and more common in East Asians (36.9 %) (Kidd et al. 2007; Long et al. 2013; Middlebrooks et al. 2016). The possibility of positive selection on the deletion polymorphism was tested in a previous study, but no evidence of positive selection was found in East Asians or Europeans. Only a weak signal of selection was detected in the African (Yoruba) population there (Kidd et al. 2007).

While searching for selection signals in the APOBEC3 cluster, we found that a non-coding 16-kb region, which is adjacent to the deletion polymorphism of *A3B*, showed three distinct star-like bursts in its phylogenetic tree, suggestive of positive Darwinian selection acting on each haplogroup. The 16-kb region does not include any coding sequence, but harbors possible regulatory sequences. We tested selective neutrality of nucleotide differences within and between haplogroups using a simulation-based method. We performed network analyses to clarify the order of the onsets of selective events and

population splits. At the same time, we estimated the time of expansions of selected haplogroups using intra-allelic diversity. Frequency trajectories of haplogroups during the past 40,000 years were also examined using genome data from ancient individuals. The functional differences between the haplogroups in regulating APOBEC3 genes were evaluated according to eQTL analyses from the GTEx database. Further, we examined the long-term evolution of this regulatory region in the human and chimpanzee lineages. Finally, we tested whether recent positive selection is acting on the *A3B* deletion. Our results inferred viral outbreaks in ancient times and demonstrated the importance of the APOBEC3 regulatory region in human evolution.

Results

Phylogenetic analyses of the 16-kb APOBEC3 regulatory region showing three distinct star-shaped phylogenies

We constructed phylogenetic trees based on the genomic sequence of the 16-kb non-coding region in the APOBEC3 cluster. The 16-kb region is located between *A3B* and *A3C* and is defined as a region between the deletion polymorphism of *A3B* and a recombination hotspot (Figure 1(a-b), Figure S1- S2). It harbors possible regulatory sequences, such as a promoter, three enhancers, and two CTCF binding sites, annotated by Ensemble Regulatory Build (Figure 1(c)). We found a unique structure in the neighbor-joining trees, where haplotype sequences are clustered into three major groups supported by high bootstrap values (Figure 2(a-e)). These groups are termed Haplogroups A, B, and C. Interestingly, these three haplogroups are conserved in African and non-African populations (the results for the Japanese (JPT), northern European-ancestry (CEU), Punjabi (PJL), Puerto Rican (PUR), and Luhya (LWK) populations are shown as examples for East Asian, European-ancestry, South Asian, American, and African populations, respectively (Figure 2(a-e)). In each population, Haplogroups A shows clear star-like phylogeny (i.e., multifurcation with many short branches connected at the internal node), which often reflects the action of positive selection. Haplogroup B is supported by a long branch but separated into two sub-clades, B1 and B2, each of which shows star-like bursts in non-Africans (Figure 2(a-d)). Haplogroup C also shows a star-like burst in non-Africans (Figure 2(a-d)), though it is less clear in East Asians (Figure 2(a)). On the other hand, in Africans, Haplogroups B and C showed bifurcating structures (Figure 2(e)). The frequencies of these three haplogroups are high in all 24 populations (Figure 2(f)); ~63%, ~50%, and ~29% for Haplogroups A, B, and C, respectively). The African populations show higher proportions of sequences that do not belong to any of the three haplogroups (Figure 2(f)). The order of the splits among the haplogroups is unclear from the phylogenetic analyses alone, even after adding the sequence of chimpanzees as an outgroup (Figure S3).

High divergence among the haplogroups and low diversity within haplogroups in the APOBEC3 regulatory region

The divergence between haplogroups was measured from the averaged nucleotide differences over pairs of sequences between haplogroups for East Asian populations (JPT and Han Chinese (CHB)), European populations (CEU and British (GBR)), and African populations (LWK and Yoruba (YRI)) (Table S1). The divergence between Haplogroups A and B is the highest, estimated to be $1.8 \times 10^{-3} \sim 1.9 \times 10^{-3}$ per site. The divergence between Haplogroups B and C and between Haplogroups A and C is lower, $1.4 \times 10^{-3} \sim 1.6 \times 10^{-3}$ per site.

Compared with the high divergences among haplogroups, the intra-haplogroup diversities are extremely low (Figure 2(g), Table S1). Nucleotide diversity (π) within Haplogroup A is of the order of 10^{-5} in all populations. The π values for Haplogroups B (B1) and C in non-Africans are in the same range as Haplogroup A, whereas those in Africans are 10 times higher. The combination of the high divergence among haplogroups and the low diversity within a haplogroup suggested that the haplogroups were possibly under positive selection.

Positive selection on the haplogroups in the APOBEC3 regulatory region

To test whether selection has been acting on the APOBEC3 regulatory haplogroups, we first examined whether positive selection worked on a single haplogroup in that region. We calculated F_c (Fujito et al. 2018; Fujito et al. 2018) for each haplogroup in JPT, CEU, and LWK populations (Table S3). This statistic detects positive selection on a single haplogroup, though it would not capture it if signatures of positive selection on distinct haplogroups overlap in a single genomic region (Discussion). As expected, any haplogroups in non-Africans does not show considerably low F_c (Table S3), while in the African population (LWK), F_c is significantly reduced for Haplogroup A. We confirmed that four out of the five African populations consistently show significantly low F_c values for Haplogroup A (Table S4; p = 0.0002, 0.0244, 0.0044, 0.0024, and 0.0016 for Haplogroup A in LWK, MSL, YRI, ESN, and GWD, respectively). These results suggest that positive selection has been acting on Haplogroup A in Africans.

Next, we assessed selection signals on multiple haplogroups in the APOBEC3 regulatory region in non-Africans. To do that, we examined whether the observed occurrence of low intra-haplogroup diversity and high inter-haplogroup divergence could be explained under neutrality. We simulated the coalescent process in a population for a given demographic model, collecting 10000 replications where three allelic groups were found in the same sizes (i.e., number of chromosomes) as Haplogroups A, B, and C observed in the population. Then, we compared the nucleotide differences within and between haplogroups observed in the population (Table S1) to the simulated values to evaluate statistical significance. In the evaluation, we tested the rarity of the occurrence of low intra-haplogroup diversity and high inter-haplogroup divergence in four different comparisons. In each comparison, we combined nucleotide differences within and between

i) Haplogroups A, B, and C, ii) Haplogroups A and B, iii) Haplogroups A and C, and iv) Haplogroups B and C. In non-Africans, when we include all three haplogroups in the comparison, all simulations fail to explain the observed pattern (Table S5; p-values: 0.0002 for CEU, 0.0006 for GBR, 0.0003 for JPT, and 0.0001 for CHB), suggesting that such three allelic groups are seldom found together under neutrality and that positive selection has been involved in evolution of the APOBEC regulatory haplogroups. When we include only Haplogroups A and B in the comparison, statistical significances are observed for East Asian populations, but not for European-ancestry populations (Table S5; p-values: 0.0185 for CEU, 0.0222 for GBR, 0.0087 for JPT, and 0.0087 for CHB). When only Haplogroups A and C or only Haplogroups B and C are included in the comparison, no significance is observed in any population in non-Africans. These results suggest that Haplogroups A, B, and C have been targeted by positive selection in Europeans, while in East Asians, positive selection is detected only in Haplogroups A and B. In African populations, positive selection is detected on Haplogroup A as described above (Table S3), and this strong signature of selection results in low p-values in all the comparisons where Haplogroup A is included (Table S5).

Timing of the expansions of the haplogroups and population splits

A plausible explanation for the shared signals of positive selection on the same haplogroups among different populations is that the selection event started before the population split. To test this hypothesis, we performed network analysis, by which we could see the order of a population split and the onset of positive selection. If the most frequent sequence for a haplogroup was shared between populations, it means that an expansion of the haplogroup occurred first before a split between populations.

In the resulting network, Haplogroups A, B, and C are separated with long branches (Figure 3). The sequences that do not belong to any of the three haplogroups formed another two clades, one of which clusters with the Neanderthal and Denisovan sequences. The other clades consist mainly of sequences from African populations. These clusters of non-A/B/C sequences are also confirmed in rooted and unrooted phylogenetic trees as well (Figure S3 and S4), and the cluster including the sequences of ancient hominins seems to be comprised of ancestral haplotypes (Figure S3(c)); however, the order of these clusters and the three haplogroups is not clear.

The most frequent sequence in Haplogroup A is found at the highest frequency in all five meta-populations in the 1000 Genomes Project data, that is, Africans, Europeans, South Asians, East Asians, and Americans (Figures 3 and S5(a)), indicating that the expansion of Haplogroup A started before the population split between Africans and non-Africans. On the other hand, the most frequent sequences in Haplotypes B1 and B2 are rarely found in Africans but are shared among non-Africans (Figures 3 and S5(b)), suggesting that the positive selection on Haplogroup B started in the common ancestor of non-Africans after the split between Africans and non-Africans.

As for Haplogroup C, the most frequent sequence is shared among non-Africans and is also quite commonly found in Africans (frequency of 0.28, Figures 3 and S5(c)). However, the difference between the top two sequences in Africa is small, and they are within the range of frequency with non-A/B/C sequences (Figure S5(d)). Taken together with the fact that Haplogroup C in Africans does not show star-like phylogeny in the phylogenetic analyses (Figures S3(c), S4(c), and S9(c)), the most likely explanation might be that Haplogroup C expanded in the common ancestors of non-Africans from the original sequence that was also widespread in Africa.

Time estimation for the expansions of the haplogroups

We estimated the time of expansions of the three haplogroups by two methods: i) calculating the mean coalescent time within haplogroups and ii) calculating the height of the genealogy of haplogroups. Consistent with the network analyses, the estimates of the times of the expansion of Haplogroup A are quite similar between Africans and non-Africans, being accumulated around 50~80 kya (Figure 4, Table S6) when a mutation rate of 0.5 x 10^{-9}/site/year was assumed(Scally 2016). This roughly overlaps with the time of the Out-of-Africa migration of modern humans (60-70 kya). For Haplogroup B, the estimates of the time are around 40~80 kya in non-Africans, whereas in Africans, they are much older at 200~350 kya (Figure 4, Table S6). A similar pattern is observed for Haplogroup C; 30~80 kya for non-Africans and 100~180 kya for Africans (Figure 4, Table S6). In summary of these results, the estimated times of expansions of the three haplogroups support the view that the expansion of Haplogroup A occurred before the Out-of-Africa migration and that the expansions of Haplogroups B and C had started in the common ancestor of non-Africans after the split between Africans and non-Africans.

Long-term evolution and divergence times of the haplogroups of the APOBEC3 regulatory region

We investigated long-term evolution and divergence times of the haplogroups. The high divergence among the haplogroups observed above (Table S1) suggests rapid allelic turnover caused by recurrent positive selection on the haplogroups since they diverged. To test this hypothesis, we first calculated a substitution rate for this region. If the regulatory region has been under recurrent selection over long-term evolution, the substitution rate should be enhanced. Indeed, the average nucleotide differences between human haplogroups and chimpanzees is 266.75, which is 38 % higher than the expectation from the genome-wide average (193 substitutions based on the sequence difference of 1.2 % between humans and chimpanzees) (Table S7). To know how many of these differences were derived from mutations in each of the human and chimpanzee lineages, we constructed a phylogenetic tree rooted with another primate species. Because the sequence from gorillas contains a large number of gaps, we instead utilize the orangutan sequence for rooting. In the resulting NJ tree (Figure S6), the average number of nucleotide differences between humans and chimpanzees is 262.75 after removing all positions containing gaps.

Among them, 123 substitutions occur in the human lineage and 140 substitutions occurred in the chimpanzee lineage (Figure S6), suggesting that positive selection has been acting not only in the human lineage but also in the chimpanzee lineage in this regulatory region. We calculated the substitution rate for both lineages. The species divergence time between humans and chimpanzees is estimated as 6-7.5 MYA, while the most recent common ancestor of orthologous DNA sequences can be substantially older(Hara et al. 2012). If we assume that the sequence divergence time is equal to the species divergence time of 6-7.5 MYA, the substitution rate in the human and chimpanzee lineages becomes as high as 1.0 x 10^{-9} - 1.3 x 10^{-9} /site/year and 1.2 x 10^{-9}-1.5 x 10^{-9} per site per year, respectively. If, on the other hand, we assume the sequence divergence time is 9 MYA, the substitution rate becomes 0.85 x 10^{-9}/site/year in humans and 0.97 x 10^{-9} per site per year in chimpanzees. These are still higher than the expectations under the *de novo* mutation rate (0.427 x 10^{-9}/site/year for humans(Jónsson et al. 2017) and 0.62 x 10^{-9}/site/year for chimpanzees(Tatsumoto et al. 2017)), suggesting that the substitution rate in the regulatory region has been enhanced in both the human and chimpanzee lineages.

To be conservative, we used this elevated substitution rate (1.0 x 10^{-9} -1.3 x 10^{-9} per site per year) to estimate the divergence times between human haplogroups. The divergence time between Haplogroups A and B becomes as old as 0.8-1.1 MYA, whereas the divergence times between A and C and between B and C are as old as 0.6-0.8 MYA and 0.7-0.9 MYA, respectively. If we compare the number of mutations that occurred on the stem of Haplogroups A, B, and C (Figure S6; 14.3, 14.5, and 9.1 for Haplogroups A, B1, and C, respectively) to the expected number under neutrality (8 mutations for each haplogroup, assuming a mutation rate of 0.5 x 10^{-9}/site/year and multi-furcation with a divergence time of 1 MYA), the observed numbers are elevated by 13~81 %, suggesting rapid allelic turnover in each of the haplogroups due to recurrent positive selection since the divergence of the haplogroups. Although these estimates are necessarily crude, the antiquity of the three haplogroups are also supported by the data of ancient hominins. In fact, all of the four pseudo-haplotypes from Neanderthal and Denisovan genomes used in this study carry Haplogroup A-linked alleles at 3 out of 14 SNPs on the stem of Haplogroup A. This strongly suggests that when archaic hominins and modern humans diverged (550 KYA (Prüfer et al. 2014)), Haplogroups A and B had already been differentiated from each other.

Frequency trajectory of the three haplogroups

To see how the frequencies of the three haplogroups have changed over time, we utilized the genome-wide data of prehistoric humans. We searched Allen Ancient Data Resource for the index SNPs for the three haplogroups, and found that rs2142833 (r^2=0.93 with Haplogroup A in the global population), rs9607601 (r^2=1 with Haplogroup B in the global population), and rs11089916 (r^2=1 with Haplogroup C in CEU and JPT) were informative. Because this analysis is limited by the small number of samples, we divided the data into different sizes of time windows (400~20,000 years). Although the oldest time window,

between 20,000 to 40,000 years ago, is very wide and has a very limited number of samples in it, we could see that the three haplogroups have been maintained at similar frequencies only with moderate changes both in Europe and East Asia during the past 40,000 years (Figure 5). We could not examine the data for Africans because the sample size was limited for separation into time windows.

Functional differences among the haplogroups of the APOBEC3 regulatory region

We searched the GTEx eQTL data from 52 tissues for variants that might affect the transcriptional regulation of the seven APOBEC3 genes, and found that the associated variants are accumulated around the APOBEC3 cluster (Figure S7(c-i)). In the 16-kb regulatory region, significant associations for all the APOBEC3 genes except for APOBEC3H are found (Figure S7(c-i), Table S8). These are relatively weak associations compared to the downstream region within the APOBEC3 cluster (chr22:39,405,000-39,580,000; Figure S7). However, the downstream region is separated from the regulatory region by a recombination hotspot (Figures 1 and S7(a)), and in this downstream region, haplotypes associated with a smaller amount of mutation signatures from APOBEC3s on pro-viral HIV-1 sequence exist (Gourraud et al. 2011). Since a selective sweep is detected on one of such haplotypes in APOBEC3H (Fujito et al. unpublished data), the downstream region seems to be under influence of another process of evolution in terms of an arms race with viruses. The 16-kb regulatory region seems to be the only possible region where natural selection can act simply on the transcriptional regulation of the APOBEC3 genes. The effects of the three haplogroups on the APOBEC3 genes are summarized in Table S9. Taking the above findings together, we found that the haplogroups show clear functional differences from each other in the expression of APOBEC3 genes.

APOBEC3B deletion polymorphism

We further investigated the evolutionary impact of the deletion polymorphism of A3B, which is adjacent to the 16-kb regulatory region. We identified a deletion variant in the 1000 Genomes Phase3 structural variant data, YL_CN_STU_4456, as the A3B deletion (Supplementary Text). Using its linkage information, we found that the deletion is mainly linked to Haplogroup A. Some sequences in Haplogroups B, C, and non-A/B/C haplotypes also carry the deletion (Figures S8 and S9(a-e)), which is suggestive of an ancient origin of the deletion. Only the deletions linked to Haplogroup A form star-like phylogenies, whereas those linked to non-Haplogroup A sequences show no sign of expansion (Figure S9(a-e)). Within Haplogroup A, the deletion is commonly linked to the most frequent sequence in all five major populations. It is considered that the deletion may have been under positive selection because of its high frequency in Asia (～35 % in East Asians and ~15 % in South Asians) and low frequency in Europe and Africa (~6% and ~0.9 %, respectively). In this study, as described above, Haplogroup A was suggested to be under positive selection (Table S5). We further investigated whether the deletion conferred an additional selective advantage to the linked sequences in Haplogroup A. Within a selected sequence group,

equally advantageous sequences appear to be under neutrality. However, if some of them are more beneficial over other sequences in the same selected group, then the selective sweep within the selected group has a chance to be observed. We calculated F_c for the sequences linked to the deletion within Haplogroup A in the Japanese population, but the resulting F_c value was not significantly low (p=0.49). The most parsimonious explanation for the above observations is that the deletion increased its frequency along with the expansion of Haplogroup A. The deletion itself does not seem to have any difference in selective advantage, and it seems to have been lost by genetic drift later in Europe and Africa.

Discussion

Genomic regions with tandem gene duplications often experience duplications of surrounding regions and suffer from inaccuracy in sequencing caused by mis-mapped reads. To ensure the accuracy of read-mapping, we additionally analyzed the data generated by a study from the Human Genome Structural Variant Consortium (HGSVC) (Ebert et al. 2021), which consists of phased genome data from 35 diverse human samples based on long-read PacBio WGS data. Using this data, we constructed a phylogenetic tree of the APOBEC3 regulatory region and confirmed that the tree is consistent with our results (Figure S10), showing Haplogroups A, B, and C. We also constructed a haplotype network with the high-coverage dataset of 3,202 samples in the 1000 Genome Project resource, which were sequenced to a depth of 30X (Byrska-Bishop et al. 2022), confirming that the haplotype sharing that we observed among populations is replicated (Figure S11).

We detected selective sweeps acting on multiple haplogroups in the APOBEC3 regulatory region in non-African populations. The form of positive selection may be categorized as balancing selection, if it is defined as a mode of selection where multiple haplotypes are maintained for a long time. Considering the APOBEC3 regulatory region as an example of balancing selection would then raise a question about the specific mechanism involved. Two mechanisms of balancing selection, overdominance and frequency-dependent selection with rare-allele advantage, have been well-studied. A third mechanism, fluctuating selection (spatio-temporal variation in selection pressure), has also been proposed (Charlesworth 2006). The genetic signatures left by these three mechanisms overlap with each other, and it is very challenging, often impossible, to distinguish underlying mechanisms from genetic signatures (Spurgin and Richardson 2010; Fijarczyk and Babik 2015). In the case of the APOBEC3 regulatory region, the decreased sequence diversity within each of the three haplogroups can eliminate the possibility of overdominance, because it would maintain the diversity within an allele or a haplogroup not to a small extent (Takahata and Nei 1990). Considering the function of the haplogroups in regulating the expression levels of APOBEC3 anti-viral factors, it is likely that the

balancing selection described here is an example of frequency-dependent selection with episodic minority advantage.

We used a simulation-based method to detect positive selection on the APOBEC3 regulatory haplogroups, because the existing detection methods are not appropriate in the present case. Such methods as EHH-based methods (Voight et al. 2006), SDS (Field et al. 2016) and F_c (Fujito et al. 2018) use summary statistics that directly or indirectly measure the ratio of diversity in the focal sequence group to the overall sequence diversity. This ratio would be extremely low when a focal group is under selection. However, the ratio cannot be very low if another group besides the focal group is also under positive selection, because the denominator would become small.

Since APOBEC3s are anti-viral factors standing at the forefront of host defense, the signatures of positive selection on the APOBEC3 regulatory region are most likely to be footprints left by viral epidemics. The estimated time of expansion of each haplogroup corresponds to the respective time of onset of epidemics. We have shown that the expansion of Haplogroup A occurred before the population split between Africans and non-Africans. Considering that the Out-of-Africa migration occurred at 60~70 KYA, we inferred a viral epidemic around 70~80 KYA as the selective pressure. Today, human dispersals are prevented by effective barriers of the Sahara and Arabian deserts. However, a recent study using computer simulations showed the existence of resource-rich corridors in the Arabian and Sinai peninsulas during three time windows: 130 to 118 KYA, 106 to 94 KYA, and 89 to 73 KYA (Timmermann and Friedrich 2016). The establishment of these corridors thereby made waves of human migrations possible. Our findings support that a viral epidemic on the third time window may have been one of the possible pressures for humans to migrate out of Arica.

The expansions on Haplogroups B1 and C have begun in the common ancestors of non-Africans. The onset times of the selection events are estimated as 30~70 KYA and 30~90 KYA for Haplogroups B1 and C, respectively (Table S6). The primary bifurcation of eastern and western Eurasian populations was inferred as 40-45 KYA(Lipson and Reich 2017; Vallini et al. 2022). Before that, it is believed that the common ancestor of non-Africans stayed in West Asia for more or less 15,000 years. The positive selection on Haplogroups B1 and C might suggest that multiple viral epidemics attacked the common ancestor of non-Africans around 45-60 KYA in West Asia.

In African populations, only Haplogroup A is suggested to be under positive selection. This raises a question about the environments surrounding African populations. There are two possibilities: First, the viral environments in Africa differed from those in other continents, and Haplogroups B and C have never been advantageous in Africa. Second, positive selection also occurred on Haplogroups B and C in Africa, but such selection occurred on standing variations and did not leave obvious signatures of selection in the genomes of Africans. The recurrent positive selection, suggested by the high divergence among the haplogroups, supports the second scenario. Even after the three haplogroups diverged, they underwent recurrent positive selection. It means that viral

epidemics where one of the three haplogroups became advantageous happened many times in Africa before the epidemics detected around the time of Out-of-Africa migration. APOBEC3s act in the innate immune system and their viral specificity seems to be very broad. Wide range of viruses can likely be the selective pressure for the haplogroups.

We showed haplotype frequency trajectories based on ancient genomes, where the three haplogroups have been maintained in high frequency during the past 40,000 years. These moderate changes in frequencies seem to be explained by genetic drift under neutrality after the selection events on each haplogroup. This phase might continue until the next selective event happens, where some or all of the haplotypes become advantageous.

We demonstrated a functional difference among the haplogroups in regulating the APOBEC3 genes using the GTEx eQTL data (Figure S7, Table S8, Table S9). The functional effects of the APOBEC3 regulatory region are relatively small; however, this region is disconnected from the downstream coding genes by a recombination hotspot and thus can be a target of natural selection for gene regulatory function, free from functional constraints or positive selection on the downstream genes. Additionally, we searched the GWAS Catalog for variants located in the APOBEC3 regulatory region. We found that rs5995653, which is linked to Haplogroup B, is associated with asthma exacerbations in admixed children (Hispanics/Latinos and African Americans) treated with inhaled corticosteroids (p=5.0E-6; OR= 1.52), and such association is also found in Europeans (p=7.52E-3) (Hernandez-Pacheco et al. 2019). In the same study, another SNP, rs6001366, the index SNP for sub-Haplogroup B2, is also associated with the same trait (p=3.0E-6, OR= 2.13). These results suggest the impact of Haplogroup B on the overall immune state.

We did not identify any evidence of positive selection acting specifically on the A3B deletion despite its association with cancer risks and large impacts on the expression levels of A3A and A3B (Figure S7, Table S9). It seems that the functional effects of the deletion have been neutral during the process of evolution so far, and the deletion increased its frequency simply due to the positive selection on Haplogroup A, which is tightly linked to the deletion. It remains possible that the sequences linked to the deletion are more advantageous among sequences in Haplogroup A, but that the decrease in sequence diversity caused by the additional positive selection acting on the deletion would be too small to be observed inside Haplogroup A, where the sequence diversity is already extremely low due to the selective sweep acting on this haplogroup. However, even in such a case, the deletion may need to be linked to Haplogroup A in order to be advantageous, since the deletion linked to other sequences shows no sign of expansion (Figure S9).

We found that the substitution rate in the chimpanzee lineage has also been accelerated in the APOBEC3 regulatory region, suggesting the functional importance of this region in primates. Similar selection events, which happened on the three human haplogroups, must have also occurred in the chimpanzee lineage.

There are some issues that should be clarified in the future. First, the exact regulatory functions that confer a selective advantage to each haplogroup need to be elucidated (i.e., which A3 gene is upregulated or downregulated as the core function for each particular haplogroup). This is difficult to determine in the current study because research in APOBEC3s is still ongoing, and specific immune cell types which are important for anti-viral activities of particular A3s are not known. We utilized the publicly available eQTL data to examine the functional differences among the haplogroups, but currently, eQTL data for each immune cell type are not separated and mixed in whole blood. Second, we could not specify the viral targets of each haplogroup. Further elucidation of the functions of each APOBEC3, especially single-cell eQTL analysis for immune cells, will help to connect each of the haplogroups and its target viruses. These will help to clarify the detailed selective advantages of each haplogroup and make it possible to evaluate the susceptibility of individuals to viral infections.

Materials and Methods

Whole genome sequence data for modern humans

We used phased whole-genome sequence (WGS) data from 1000 Genomes Project Phase 3(1000 Genomes Project Consortium et al. 2015). Samples from two populations, ASW (African Ancestry in Southeast US) and ACB (African Caribbean in Barbados), were removed because of possible admixture, and the remaining 4694 samples from 24 populations were used [216 for YRI (Yoruba in Ibadan, Nigeria), 198 for LWK (Luhya in Webuye, Kenya), 226 for GWD (Gambian in Western divisions in the Gambia), 170 for MSL (Mende in Sierra Leone), 198 for ESN (Esan in Nigeria) in Africans; 198 for CEU (Utah residents with Northern and Western European ancestry), 214 for TSI (Toscani in Italy), 198 for FIN (Finnish in Finland), 182 for GBR (British in England and Scotland), and 214 for IBS (Iberian population in Spain) in Europeans (EUR); 206 for GIH (Gujarati Indian from Houston, Texas, USA), 192 for PJL (Punjabi from Lahore, Pakistan), 172 for BEB (Bengali from Bangladesh), 204 for STU (Sri Lankan Tamil from the UK), and 204 for ITU (Indian Telugu from the UK) in South Asians (SAS); 206 for CHB (Han Chinese in Beijing, China), 208 for JPT (Japanese in Tokyo, Japan), 210 for CHS (Southern Han Chinese, China), 186 for CDX (Chinese Dai in Xishuangbanna, China), and 198 for KHV (Kinh in Ho Chi Minh City, Vietnam) in East Asians (EAS); and 128 for MXL (Mexican ancestry in Los Angeles, California), 208 for PUR (Puerto Rican in Puerto Rico), 188 for CLM (Colombian in Medellin, Colombia), and 170 for PEL (Peruvian in Lima, Peru) in Americans (AMR)]. The haplotypes in the APOBEC3 regulatory region were constructed by including SNPs in the region chr22: 39388993 to 39405510 (GRCh37). After removing indels using VCFtools (Danecek et al. 2011), 588 SNPs remained in that region. For time estimation, we obtained ancestral allele information from 1000 Genomes Project Phase 3 and used 584 SNPs that are bi-allelic and have ancestral allele information. The 16-kb APOBEC3B regulatory region excludes the APOBEC3B deletion (as well as the identical 350-bp sequences flanking both ends of the deletion) (Figure S2). Moreover, we checked the sequence similarity within the 200kb-APOBEC3 cluster using Dotplot (Figure S2) (Schwartz et al. 2000). There is a cluster of short fragments (22 ~1668 bp) with high sequence similarity (~98 %) between the upstream and downstream areas of the deletion. We confirmed that this cluster of sequences was not included in the 16-kb regulatory region analyzed in this study (Figure S2).

We used a deletion polymorphism, "YL_CN_STU_4456", in the 1000 Genomes Phase3 structural variant data as the linkage information of the APOBEC3B deletion (Supplementary text). We downloaded the regulatory features from the Ensemble webpage ("homo_sapiens.GRCh38.Regulatory_Build.regulatory_features.20220201.gff") and lifted it over to GRCh37 using the UCSC webpage.

To confirm our results, we also utilized WGS data from the Human Genome Structural Variation Consortium (HGSVC)(Ebert et al. 2021), which is based on long-read PacBio WGS data and Strand-Seq data from 32 diverse human samples, and the high-coverage (30X) WGS data of the 1000 Genome Project resource (Byrska-Bishop et al. 2022), which includes data from 3202 samples.

Data for ancient hominins and chimpanzees

The WGS data for ancient hominins were downloaded from http://cdna.eva.mpg.de/neandertal/Vindija/. We used the data for the Altai Neanderthal and the Denisova Pinky for the phylogenetic and network analyses. The sequence data for chimpanzees (NC_036901.1) was downloaded from the NCBI webpage and the sequence for orangutans (Susie_PABv2) was downloaded from the Ensemble webpage.

Phylogenetic analysis

Neighbor-Joining trees were constructed with MEGA 11 software(Saitou and Nei 1987; Tamura et al. 2021) based on the haplotype sequences in the 16-kb region. The evolutionary distances were computed using the number of differences method. The data for JPT, CEU, PJL, PUR, YRI, and LWK populations were used as representative populations of East Asians, Europeans, South Asians, Americans, and Africans, respectively.

Network analysis

Median-joining networks were constructed with the default option (epsilon=0) using PopART software (http://popart.otago.ac.nz). To perform network analysis for the global population, we constructed haplotypes after removing singletons and used haplotypes that were observed more than once. For networks for each of the metapopulations, we included all SNPs (including singletons).

Identification of haplogroups and recombinants

Based on the topology of the NJ trees, one of the SNPs on the branch that leads each of the haplogroups was defined as an index SNP: rs4303815 for Haplogroup A, rs35043274 for subgroup B1, rs6001366 for subgroup B2, and rs111468681 for Haplogroup C.

We identified and removed recombinants from time estimation and measuring haplogroup diversity. The number of nucleotide differences for all pairs of sequences within a haplogroup was calculated to identify recombinants. We then plotted their distribution and identified outlier pairs. Sequences that caused outliers were regarded as recombinants and removed from the analyses.

Estimation of the time of expansions of the haplogroups

The time of the expansion of each haplotype group was estimated using intra-allelic diversity (IAV) with two different methods. The first method is an estimation of mean coalescent time within a haplotype group. We took the average over the number of nucleotide differences across all pairs of sequences within a haplotype. Then, we rescaled the average with the length of the sequences to obtain π values (per site nucleotide diversity). We converted the π values to years with mutation rate (using $pi = 2t\mu$, where t is the time since the common ancestor of the sequences in a haplogroup and μ is the

mutation rate per site per year), assuming that the mutation rate is 0.5 x 10^{-9}/site/year. Standard error estimates were obtained with by a bootstrap procedure using MEGA 11.

The second method was to calculate the height of the genealogy of each haplogroup (Satta et al. 2019). Based on the ancestral allele information, we identified sites, where the derived allele is segregating in a haplogroup of interest and the ancestral allele is fixed outside of it. Using such sites, we counted the number of derived alleles accumulated within a haplogroup divided by the number of sequences to obtain the height of the haplogroup, which we converted to years as described above. The variance of height was calculated using the formula described by Satta et al. 2018. For the calculation of the tree heights, we removed the sites at which i ($m-2 \leq i < m$, where m is the size of the haplotype group) derived alleles are found in the haplotype of interest, and no derived alleles are found outside of it. This is because these sites are known to be the results of recombination by which a few ancestral alleles are flown into the haplotype of interest (Satta et al. 2019).

Inference of selection

To detect balancing selection in the APOBEC3 regulatory region, where extremely low nucleotide diversity and long distance among haplogroups were observed, we took a simple simulation-based method. We measured the mean nucleotide difference within and between haplogroups A, B (B1), and C in JPT, CEU, and LWK populations. Then, we investigated the rarity of observing the combination of low diversity within each of the haplogroups and high divergence between the haplogroups at the same time under neutrality (i.e. reject neutrality if the combination is less than 1%). Simulations were carried out using ms (Hudson 2002) to examine the pattern of nucleotide diversity under the demographic model proposed for East Asians, Europeans, and Africans (Schaffner et al. 2005). The specified (observed) number of segregating sites was placed on a coalescent tree. The command lines we used are shown in Supplementary Table 1. We retrieved 10,000 replications whereby the three allelic groups showed the same size as observations (e.g., 103, 76, and 19 chromosomes for Haplogroups A, B, and C in JPT, respectively), and calculated nucleotide differences within and between clades. Using these simulated values, we compared within and between nucleotide differences between simulated and observed data for: 1) Haplogroups A, B, and C; 2) Haplogroups A and B only; 3) Haplogroups A and C only; and 4) Haplogroups B and C only.

Estimation of divergence time among the haplogroups

We estimated the divergence time between two haplogroups using the number of nucleotide differences. Specifically, we calculated the average number of nucleotide differences across all pairs of sequences between haplotype groups, and rescaled it with the length of the sequences to obtain π values (per site nucleotide diversity). To convert π to years, we calculated the mutation rate for long-term evolution by obtaining a substitution rate for the human lineage, because positive selection on the APOBEC3 regulatory region accelerates the substitution rate on the long stem branch of each haplogroup. To do that,

we calculated the number of differences between chimpanzees and human haplogroups. We determined the number of mutations that occurred in each of the human and chimpanzee lineages, using an NJ tree of the APOBEC3 regulatory region including the human haplogroup sequences and a sequence from chimpanzees, with a sequence from orangutans as a root. By dividing the number of mutations that occurred in the human lineage by the divergence time between humans and chimpanzees (6-7.5 MYA), we obtained the average substitution rate for the human lineage and used it as a mutation rate for long-term evolution.

Functional analysis of the haplogroups

We utilized GTEx eQTL data and downloaded the list of variants that affect the transcriptional regulation of the seven APOBEC3 genes. From the data, we searched for the index SNPs of the three haplogroups and the SNPs that were linked to them ($r^2 \geq 0.8$). We also checked for the APOBEC3B deletion. Since the deletion was not included in the data, we used rs12628403 as a tag SNP for the deletion (Middlebrooks et al. 2016).

Frequency trajectories over time

We downloaded genotype data from Allen Ancient Genome Resource (Mallick et al. 2023; Anon) (version 52.2, downloaded from https://reich.hms.harvard.edu/allen-ancient-dna-resource-aadr-downloadable-genotypes-present-day-and-ancient-dna-data), and used genotypes from the samples from Europe (4624 samples) and East Asia (898 samples) from 100 – 40,000 years ago. Because the index SNPs for Haplogroups A, B, and C were not in this data, we instead utilized rs2142833, rs9607601, and rs11089916, respectively, as tag SNPs.

Acknowledgments

We thank Dr. Naoyuki Takahata for useful discussion and critical comments that greatly improved this study. We also thank Dr. Toshiyuki Hayakawa for useful discussion and helpful comments. We thank Dr. Naruya Saitou for his critical reading on this manuscript. This research was funded by Research Organization of Information and Systems (ROIS; https://www.rois.ac.jp). Computations were partially performed on the NIG supercomputer at ROIS National Institute of Genetics. This manuscript was English edited by Quintin Lau.

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* 526:68–74.

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* 578:94–101.

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* 500:415–421.

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

Byrska-Bishop M, Evani US, Zhao X, Basile AO, Abel HJ, Regier AA, Corvelo A, Clarke WE, Musunuri R, Nagulapalli K, et al. 2022. High-coverage whole-genome sequencing of the expanded 1000 Genomes Project cohort including 602 trios. *Cell* 185:3426-3440.e19.

Caval V, Suspène R, Shapira M, Vartanian J-P, Wain-Hobson S. 2014. A prevalent cancer susceptibility APOBEC3A hybrid allele bearing APOBEC3B 3'UTR enhances chromosomal DNA damage. *Nat. Commun.* 5:5129.

Charlesworth D. 2006. Balancing selection and its effects on sequences in nearby genome regions. *PLoS Genet.* 2:e64.

Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE, Lunter G, Marth GT, Sherry ST, et al. 2011. The variant call format and VCFtools. *Bioinformatics* 27:2156–2158.

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

Enard D, Cai L, Gwennap C, Petrov DA. 2016. Viruses are a dominant driver of protein adaptation in mammals. *Elife* [Internet] 5. Available from: http://dx.doi.org/10.7554/eLife.12469

Enard D, Petrov DA. 2020. Ancient RNA virus epidemics through the lens of recent adaptation in human genomes. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 375:20190575.

Field Y, Boyle EA, Telis N, Gao Z, Gaulton KJ, Golan D, Yengo L, Rocheleau G, Froguel P, McCarthy MI, et al. 2016. Detection of human adaptation during the past 2000 years. *Science* 354:760–764.

Fijarczyk A, Babik W. 2015. Detecting balancing selection in genomes: limits and prospects. *Mol. Ecol.* 24:3529–3545.

Fujito NT, Satta Y, Hane M, Matsui A, Yashima K, Kitajima K, Sato C, Takahata N, Hayakawa T. 2018. Positive selection on schizophrenia-associated ST8SIA2 gene in post-glacial Asia. *PLoS One* 13:e0200278.

Fujito NT, Satta Y, Hayakawa T, Takahata N. 2018. A new inference method for detecting an ongoing selective sweep. *Genes Genet. Syst.* 93:149–161.

Fumagalli M, Sironi M, Pozzoli U, Ferrer-Admetlla A, Pattini L, Nielsen R. 2011. Signatures of environmental genetic adaptation pinpoint pathogens as the main selective pressure through human evolution. *PLoS Genet.* 7:e1002355.

Gourraud PA, Karaouni A, Woo JM, Schmidt T, Oksenberg JR, Hecht FM, Liegler TJ, Barbour JD. 2011. APOBEC3H haplotypes and HIV-1 pro-viral vif DNA sequence diversity in early untreated human immunodeficiency virus-1 infection. *Hum. Immunol.* 72:207–212.

Hakata Y, Li J, Fujino T, Tanaka Y, Shimizu R, Miyazawa M. 2019. Mouse APOBEC3 interferes with autocatalytic cleavage of murine leukemia virus Pr180gag-pol precursor and inhibits Pr65gag processing. *PLoS Pathog.* 15:e1008173.

Hara Y, Imanishi T, Satta Y. 2012. Reconstructing the demographic history of the human lineage using whole-genome sequences from human and three great apes. *Genome Biol. Evol.* 4:1133–1145.

Hernandez-Pacheco N, Farzan N, Francis B, Karimi L, Repnik K, Vijverberg SJ, Soares P, Schieck M, Gorenjak M, Forno E, et al. 2019. Genome-wide association study of inhaled corticosteroid response in admixed children with asthma. *Clin. Exp. Allergy* 49:789–798.

Hudson RR. 2002. Generating samples under a Wright–Fisher neutral model of genetic variation. *Bioinformatics* 18:337–338.

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* 549:519–522.

Karlsson EK, Harris JB, Tabrizi S, Rahman A, Shlyakhter I, Patterson N, O'Dushlaine C, Schaffner SF, Gupta S, Chowdhury F, et al. 2013. Natural selection in a bangladeshi population from the cholera-endemic ganges river delta. *Sci. Transl. Med.* 5:192ra86.

Karlsson EK, Kwiatkowski DP, Sabeti PC. 2014. Natural selection and infectious disease in human populations. *Nat. Rev. Genet.* 15:379–393.

Kidd JM, Newman TL, Tuzun E, Kaul R, Eichler EE. 2007. Population stratification of a common APOBEC gene deletion polymorphism. *PLoS Genet.* 3:e63.

Klunk J, Vilgalys TP, Demeure CE, Cheng X, Shiratori M, Madej J, Beau R, Elli D, Patino MI, Redfern R, et al. 2022. Evolution of immune genes is associated with the Black Death. *Nature* 611:312–319.

Kwiatkowski DP. 2005. How malaria has affected the human genome and what human genetics can teach us about malaria. *Am. J. Hum. Genet.* 77:171–192.

Lipson M, Reich D. 2017. A Working Model of the Deep Relationships of Diverse Modern Human Genetic Lineages Outside of Africa. *Mol. Biol. Evol.* 34:889–902.

Long J, Delahanty RJ, Li G, Gao Y-T, Lu W, Cai Q, Xiang Y-B, Li C, Ji B-T, Zheng Y, et al. 2013. A common deletion in the APOBEC3 genes and breast cancer risk. *J. Natl. Cancer Inst.* 105:573–579.

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

Middlebrooks CD, Banday AR, Matsuda K, Udquim K-I, Onabajo OO, Paquin A, Figueroa JD, Zhu B, Koutros S, Kubo M, et al. 2016. Association of germline variants in the APOBEC3 region with cancer risk and enrichment with APOBEC-signature mutations in tumors. *Nat. Genet.* 48:1330–1338.

Newman ENC, Holmes RK, Craig HM, Klein KC, Lingappa JR, Malim MH, Sheehy AM. 2005. Antiviral function of APOBEC3G can be dissociated from cytidine deaminase activity. *Curr. Biol.* 15:166–170.

Nik-Zainal S, Wedge DC, Alexandrov LB, Petljak M, Butler AP, Bolli N, Davies HR, Knappskog S, Martin S, Papaemmanuil E, et al. 2014. Association of a germline copy number polymorphism of APOBEC3A and APOBEC3B with burden of putative APOBEC-dependent mutations in breast cancer. *Nat. Genet.* 46:487–491.

Pollpeter D, Parsons M, Sobala AE, Coxhead S, Lang RD, Bruns AM, Papaioannou S, McDonnell JM, Apolonia L, Chowdhury JA, et al. 2018. Deep sequencing of HIV-1 reverse transcripts reveals the multifaceted antiviral functions of APOBEC3G. *Nat Microbiol* 3:220–233.

Prüfer K, Racimo F, Patterson N, Jay F, Sankararaman S, Sawyer S, Heinze A, Renaud G, Sudmant PH, de Filippo C, et al. 2014. The complete genome sequence of a Neanderthal from the Altai Mountains. *Nature* 505:43–49.

Quintana-Murci L, Clark AG. 2013. Population genetic tools for dissecting innate immunity in humans. *Nat. Rev. Immunol.* 13:280–293.

Revathidevi S, Murugan AK, Nakaoka H, Inoue I, Munirajan AK. 2021. APOBEC: A molecular driver in cervical cancer pathogenesis. *Cancer Lett.* 496:104–116.

Roberts SA, Lawrence MS, Klimczak LJ, Grimm SA, Fargo D, Stojanov P, Kiezun A, Kryukov GV, Carter SL, Saksena G, et al. 2013. An APOBEC cytidine deaminase mutagenesis pattern is widespread in human cancers. *Nat. Genet.* 45:970–976.

Saitou N, Nei M. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* 4:406–425.

Satta Y, Zheng W, Nishiyama KV, Iwasaki RL, Hayakawa T, Fujito NT, Takahata N. 2019. Two-dimensional site frequency spectrum for detecting, classifying and dating incomplete selective sweeps. *Genes Genet. Syst.* 94:283–300.

Scally A. 2016. The mutation rate in human evolution and demographic inference. *Curr. Opin. Genet. Dev.* 41:36–43.

Schaffner SF, Foo C, Gabriel S, Reich D, Daly MJ, Altshuler D. 2005. Calibrating a coalescent simulation of human genome sequence variation. *Genome Res.* 15:1576–1583.

Schwartz S, Zhang Z, Frazer KA, Smit A, Riemer C, Bouck J, Gibbs R, Hardison R, Miller W. 2000. PipMaker—A Web Server for Aligning Two Genomic DNA Sequences. *Genome Res.* 10:577–586.

Spurgin LG, Richardson DS. 2010. How pathogens drive genetic diversity: MHC, mechanisms and misunderstandings. *Proc. Biol. Sci.* 277:979–988.

Stavrou S, Ross SR. 2015. APOBEC3 Proteins in Viral Immunity. *J. Immunol.* 195:4565–4570.

Takahata N, Nei M. 1990. Allelic genealogy under overdominant and frequency-dependent selection and polymorphism of major histocompatibility complex loci. *Genetics* 124:967–978.

Tamura K, Stecher G, Kumar S. 2021. MEGA11: Molecular Evolutionary Genetics Analysis Version 11. *Mol. Biol. Evol.* 38:3022–3027.

Tatsumoto S, Go Y, Fukuta K, Noguchi H, Hayakawa T, Tomonaga M, Hirai H, Matsuzawa T, Agata K, Fujiyama A. 2017. Direct estimation of de novo mutation rates in a chimpanzee parent-offspring trio by ultra-deep whole genome sequencing. *Sci. Rep.* 7:1–12.

Timmermann A, Friedrich T. 2016. Late Pleistocene climate drivers of early human migration. *Nature* 538:92–95.

Vallini L, Marciani G, Aneli S, Bortolini E, Benazzi S, Pievani T, Pagani L. 2022. Genetics and Material Culture Support Repeated Expansions into Paleolithic Eurasia from a Population Hub Out of Africa. *Genome Biol. Evol.* 14:evac045.

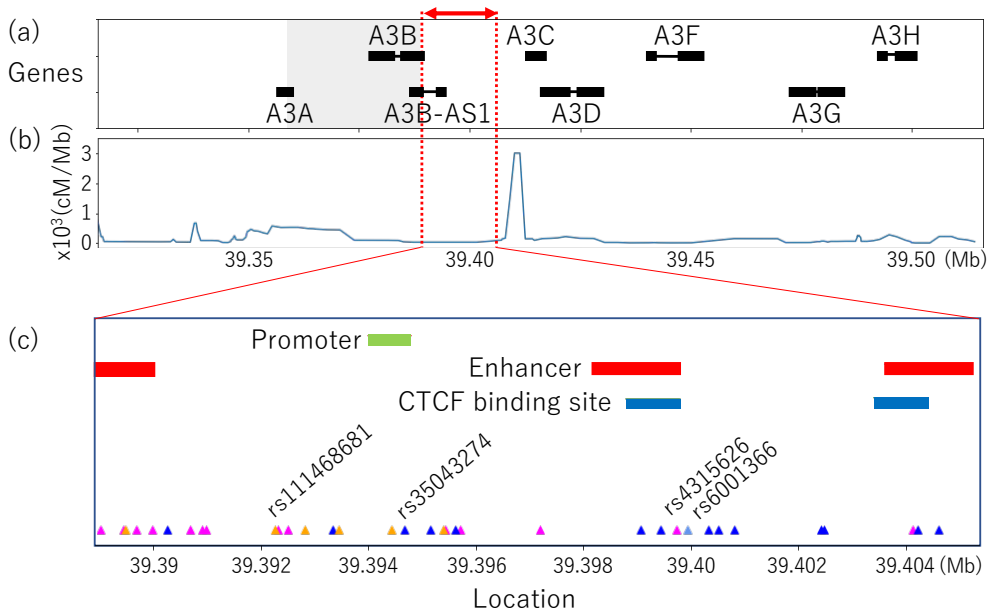
Voight BF, Kudaravalli S, Wen X, Pritchard JK. 2006. A map of recent positive selection in the human genome. *PLoS Biol.* 4:e72.

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.* 86:3777–3786.

1 **Figures**

2

3



4

5 **Figure 1.**

6 The APOBEC3 cluster including the regulatory region.

7 The APOBEC3 cluster region, which spans 200 kb surrounding the APOBEC3
8 genes, is illustrated (a). A3A, A3B, A3C, A3D, A3F, A3G, and A3H are
9 abbreviations for APOBEC3A, APOBEC3B, APOBEC3C, APOBEC3D,
10 APOBEC3F, APOBEC3G, and APOBEC3H, respectively. A long non-coding
11 RNA, APOBEC3B antisense RNA 1 (A3B-AS1), is also locating in this region.
12 The 16-kb regulatory region is represented by the double-headed red arrow and
13 the APOBEC3B deletion polymorphism is indicated by the shaded square. The
14 recombination rate in the APOBEC3 cluster region was plotted based on the
15 HapMap phase II genetic map (b). Annotated regulatory sequences in the
16 Ensemble Regulatory Build are shown (c). The SNPs in LD ($r^2 > 0.8$) with the
17 index SNPs for Haplogroups A (magenta), B1 (blue), B2 (pale blue), and C
18 (orange) are depicted with triangles.

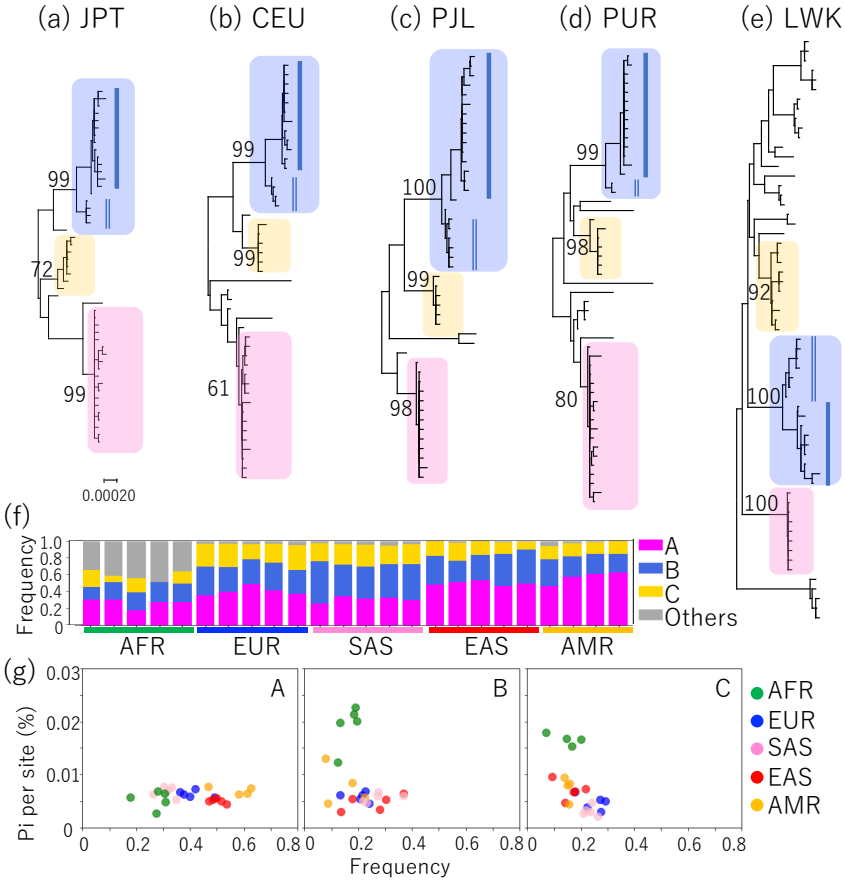


Figure 2.

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 shown only for the three haplogroups: Haplogroups A (pink), B (blue), and C (orange). Haplogroup B is separated into two sub-groups: B1 (blue bars) and B2 (double lines). 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 number of base substitutions per site. (f) Proportion of the three haplogroups in each population in the 1000 Genomes project data. (g) Nucleotide diversity (π per site) within a haplogroup is plotted against its frequency in a population for Haplogroups A (left), B (middle), and C (right).

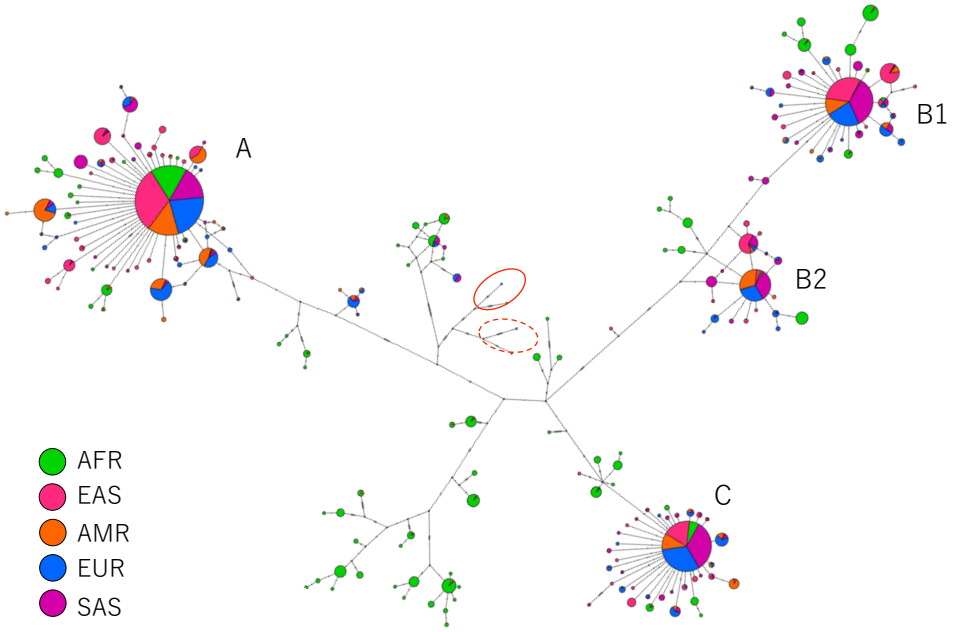


Figure 3.

Median-joining network of the 16-kb APOBEC3 regulatory region. Sequences from 24 populations in the 1000 Genomes Project data were used with the pseudo-haplotype sequences from Neanderthal and Denisovan. Haplotypes were constructed after removing singletons. Haplotypes observed only once were excluded, and the remaining 215 haplotype sequences were used for this analysis. The pie chart size reflects the number of chromosomes, and the largest pie chart in the network analysis represents the most frequent sequence in a haplogroup. Colors represent proportions of sequences from the five super-populations and sequences from Neanderthals and Denisovans are indicated by red and dashed circles, respectively.

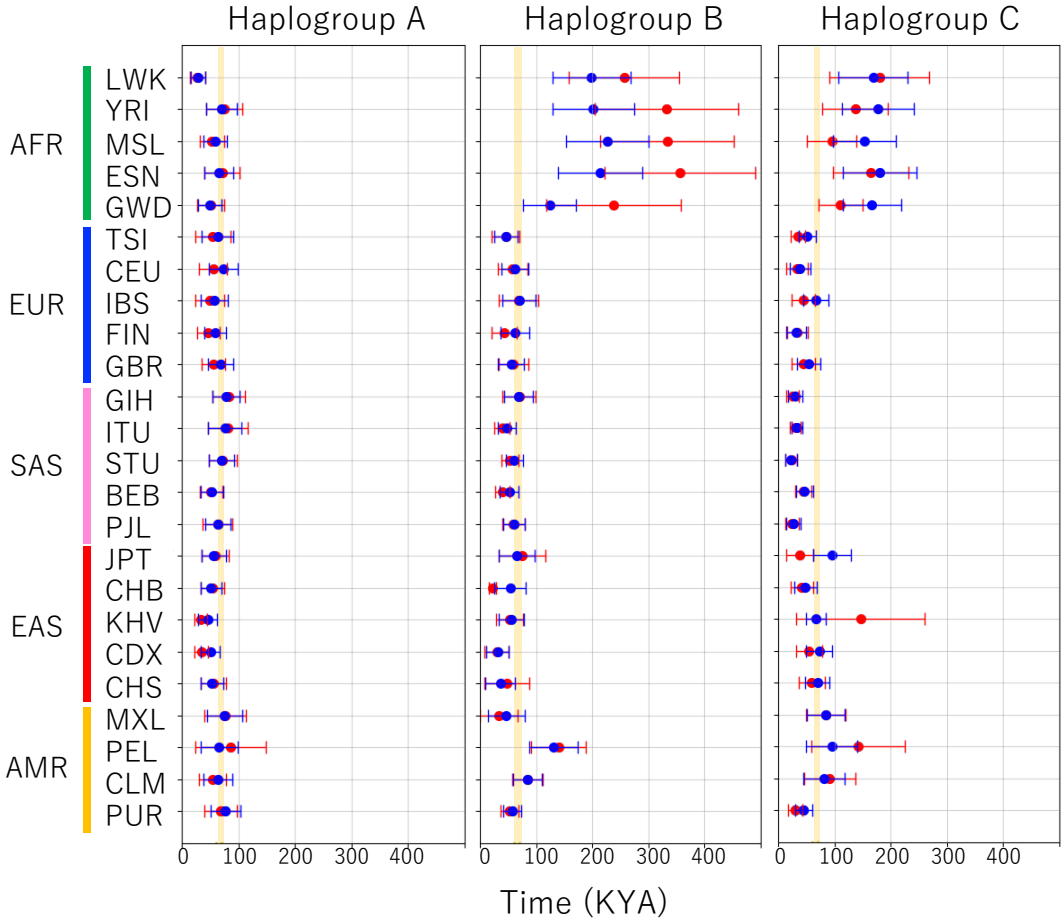


Figure 4.

Estimated time of the expansions of Haplotype groups A, B (B1), and C. The mean coalescent time estimated from nucleotide diversity within a haplogroup (red) and the time of the most recent common ancestor estimated from the height of each haplogroup (blue) are plotted for each population. The time range of the Out-of-Africa migration of modern humans (60 – 70 kya) is colored in yellow.

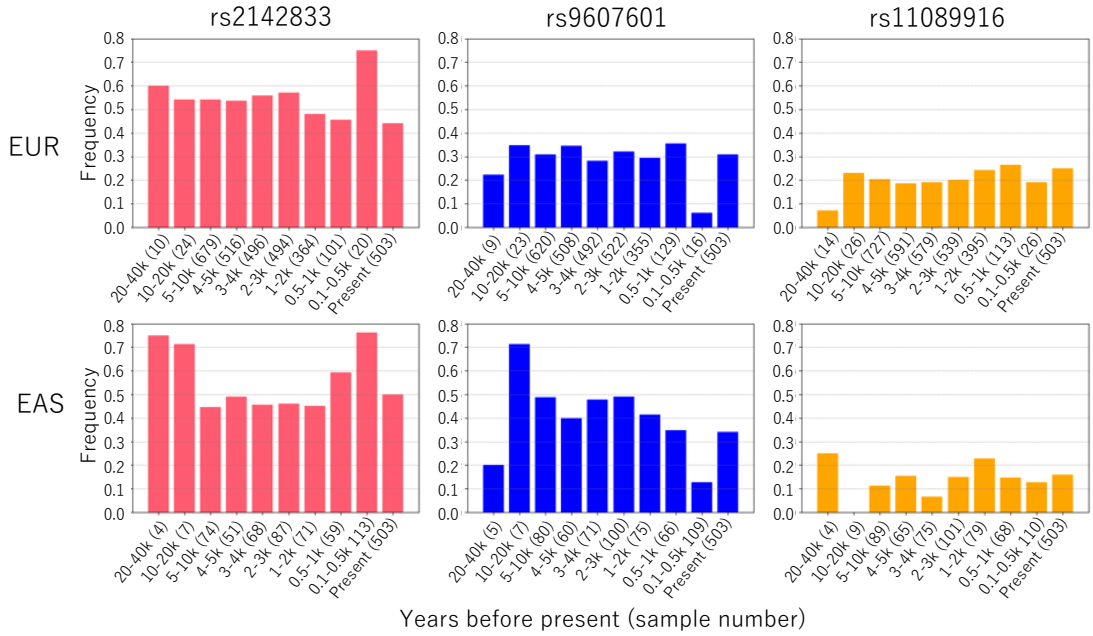


Figure 5.

Frequencies of the three haplogroups in modern humans across time. Samples from Europe (top) and East Asia (bottom) are from compiled ancient data and current data from the 1000 Genome Project. Frequency of the allele linked to each haplogroup was calculated at the informative SNPs (rs2142833, rs9607601, and rs11089916 for Haplogroups A, B, and C, respectively). Since frequency was calculated after removing the samples for which genotypes were not available at each of the informative SNPs, the sum of the frequencies of the three informative SNPs is not necessarily equal to 1. The number of samples is shown in parentheses. Time periods are indicated in years before present.