Human Molecular Genetics, 2020, Vol. 29, No. 5 736–744

doi: 10.1093/hmg/ddz304
Advance Access Publication Date: 16 December 2019
General Article One

GENERAL ARTICLE ONE

Genetic susceptibility to severe childhood asthma and rhinovirus-C maintained by balancing selection in humans for 150 000 years

Mary B. O'Neill[1,2,3,4,*], Guillaume Laval[4], João C. Teixeira[4,5], Ann C. Palmenberg[6] and Caitlin S. Pepperell[2,3]

[1]Department of Laboratory of Genetics, University of Wisconsin—Madison, Madison, WI 53706, USA, [2]Department of Medicine, University of Wisconsin—Madison, Madison, WI 53706, USA, [3]Department of Medical Microbiology and Immunology, University of Wisconsin—Madison, Madison, WI 53706, USA, [4]Department of Human Evolutionary Genetics Unit, Institut Pasteur, CNRS UMR2000, Paris 75015, France, [5]Department of Australian Centre for Ancient DNA, The University of Adelaide, Adelaide, South Australia 5005, Australia and [6]Department of Biochemistry, Institute for Molecular Virology, University of Wisconsin—Madison, Madison, WI 53706, USA

*To whom correspondence should be addressed. Tel: +33 (0)144389366; Email: moneill@pasteur.fr

Abstract

Selective pressures imposed by pathogens have varied among human populations throughout their evolution, leading to marked inter-population differences at some genes mediating susceptibility to infectious and immune-related diseases. Here, we investigated the evolutionary history of a common polymorphism resulting in a Y_{529} versus C_{529} change in the cadherin related family member 3 (*CDHR3*) receptor which underlies variable susceptibility to rhinovirus-C infection and is associated with severe childhood asthma. The protective variant is the derived allele and is found at high frequency worldwide (69–95%). We detected genome-wide significant signatures of natural selection consistent with a rapid increase of the haplotypes carrying the allele, suggesting that non-neutral processes have acted on this locus across all human populations. However, the allele has not fixed in any population despite multiple lines of evidence suggesting that the mutation predates human migrations out of Africa. Using an approximate Bayesian computation method, we estimate the age of the mutation while explicitly accounting for past demography and positive or frequency-dependent balancing selection. Our analyses indicate a single emergence of the mutation in anatomically modern humans ∼150 000 years ago and indicate that balancing selection has maintained the beneficial allele at high equilibrium frequencies worldwide. Apart from the well-known cases of the *MHC* and *ABO* genes, this study provides the first evidence that negative frequency-dependent selection plausibly acted on a human disease susceptibility locus, a form of balancing selection compatible with typical transmission dynamics of communicable respiratory viruses that might exploit CDHR3.

Received: August 20, 2019. **Revised:** November 7, 2019. **Accepted:** December 12, 2019

Introduction

There is accumulating evidence to suggest that immunity-related genes are preferential targets of natural selection (1–5), supporting the notion that infectious diseases have been important selective forces on human populations (6). However, for most candidate loci, the mechanisms and phenotypic effects underlying the observed signatures of selection remain elusive.

Human rhinoviruses (RVs) are found worldwide and are the predominant cause of the common cold. While many RV infections cause only minor illness, type C strains of the virus are associated with higher virulence. Unlike RV-A and RV-B, RV-C utilizes the cadherin related family member 3 (CHDR3) receptor to gain entry into host cells (7). A common missense variant in CDHR3 (rs6967330, C529Y) results in differences in virus binding and/or replication; the protein encoded by the derived rs6967330 allele (C_{529}) is displayed at 10-fold lower density on the surfaces of ciliated airway cells than the protein encoded by the ancestral allele (Y_{529}), making CDHR3 less accessible to respiratory viruses seeking the receptor (8,9). Adding clinical support that CDHR3 functions as an RV-C receptor, the ancestral allele at rs6967330 has been found to be associated with an increased risk of respiratory tract illness by RV-C in multiple cohorts (10,11). In accordance with the known role of RV, particularly RV-C, in triggering asthma exacerbations, the allele has also been found to be associated with various severe forms of asthma or asthma-related phenotypes in differing ethnic backgrounds (8,12,13). Taken together, these studies suggest that host genetics mediate differing susceptibility to RV-C infection and asthma by affecting interactions between the virus and its receptor (e.g. whether or not the protein is expressed on the cell surface and is thus visible to the virus).

A recent outbreak of lethal respiratory illness among wild chimpanzees was attributed to human RV-C crossing species boundaries. All members of the infected chimpanzee population were invariant for the ancestral variant at the homologous position corresponding to the human rs6967330 variant. The outbreak resulted in a staggering 8.9% mortality rate (14) and suggests that RV-C infection can result in a high mortality rate in a susceptible population. As respiratory infections, particularly prior to the availability of modern medical interventions, represent significant threats to human health (15–18), CDHR3 and in particular rs6967330 represent a promising candidate target of natural selection. In the present study, we investigated the evolutionary history of this locus for which there already exists strong experimental and clinical data linking genotype with phenotypes that appear to modulate disease susceptibility.

Results

Contemporary diversity patterns of the *CDHR3* locus

The rs6967330 disease susceptibility locus is a common single nucleotide polymorphism (SNP) shared across contemporary worldwide populations. The frequency of the derived G allele ranges from 68.8% ('Mende in Sierra Leone', MSL) to 95.3% ('Peruvians from Lima, Peru', PEL) across the 26 populations of the 1000 Genomes (1000G) Project (19) (Fig. 1). At the super population level, the derived G allele is most common in East Asian populations (EAS, 93.0%), followed by admixed American populations (AMR, 85.7%), South Asian populations (SAS, 80.0%) and European populations (EUR, 79.2%). It is least common in African populations (AFR, 73.5%), albeit also at high frequency.

High linkage disequilibrium (LD) patterns extend from 105 657 078 to 105 659 873 on chromosome 7 in the 1000G populations, with only moderate LD decay extending up to 105 680 022 (Supplementary Material, Fig. S1). Considering only biallelic SNPs with a minor allele frequency (MAF) ≥ 0.01, haplotypes within these blocks were extracted with the Pegas package in R (20,21). Within the smaller 2795 bp region, we identified 19 haplotypes among 16 SNPs, and within the larger 22 945 bp region, we identified 378 haplotypes among 83 SNPs (Supplementary Material, Table S1). Relationships among haplotypes occurring at $\geq$0.5% were inferred using network analysis (Fig. 2). Of the eight haplotypes with frequency $\geq$0.5% in the region of high LD, the majority of individuals in all populations carry the same haplotype with the derived G allele ($n = 3848$); two less frequent haplotypes carrying the derived allele are found in various geographic regions. The ancestral A allele is found in the remaining five haplotypes, which vary in their distributions across regions. Phylogenetic reconstruction

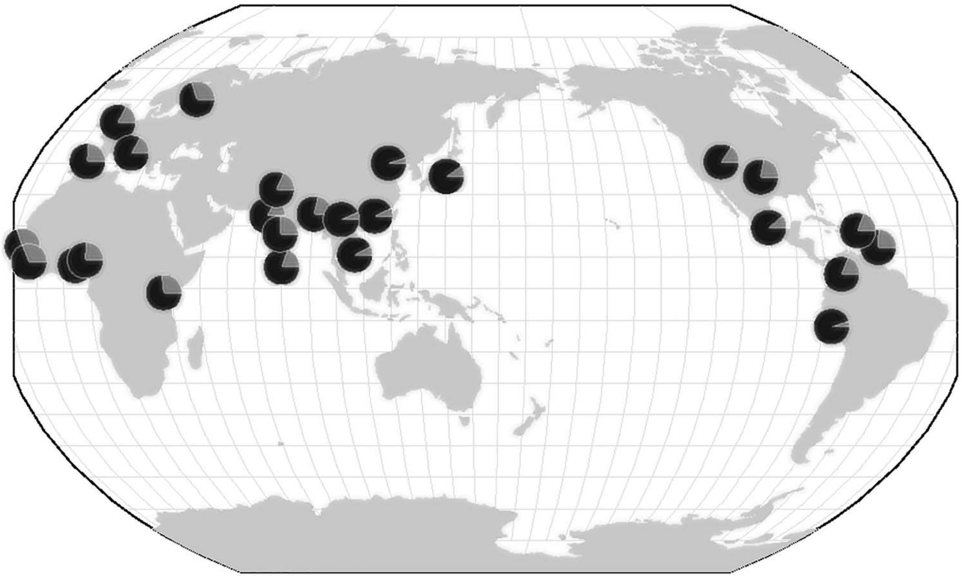

Figure 1. Global distribution of allele frequencies at rs6967330. Pie charts represent the allele frequencies of the ancestral A allele (light gray) and the derived G allele (black) in each of the 26 populations of the 1000 Genomes Project. This image was generated through the *Geography of Genetic Variants* browser (64).

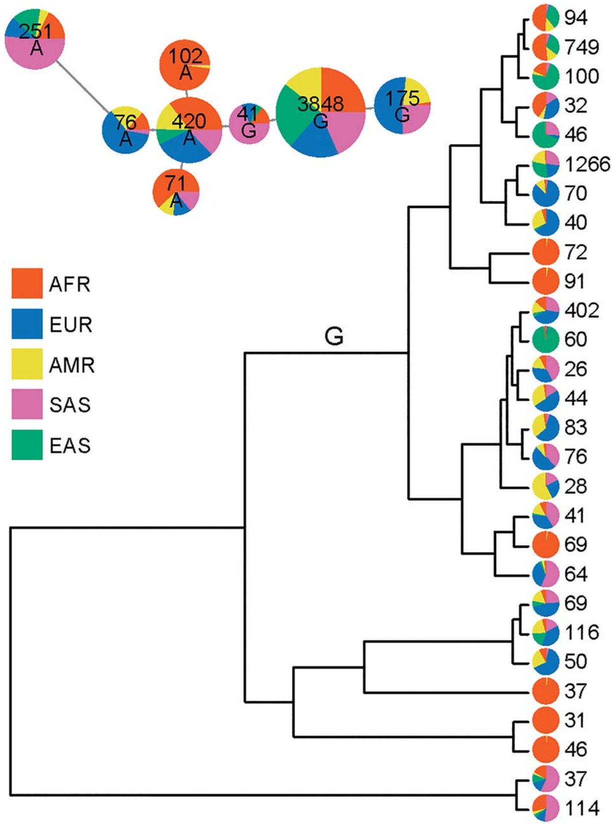

Figure 2. Haplotype structure of the *CDHR3* locus in anatomically modern humans. (Top) Haplotype network from chromosome 7 between 105 657 078 and 105 659 873. Haplotype network was constructed from phased genome sequences of 2504 diploid individuals from the 1000 Genomes Project with variation at 16 biallelic SNPs (MAF $\geq$ 0.01) in the 2795 bp region. Colors reflect super population designation of individuals. (Bottom) Unrooted tree from chromosome 7 between 105 657 078 and 105 680 022. Haplotypes of the same 2504 diploid individuals were derived for the larger genomic region. Again, only haplotypes occurring at >0.5% were analyzed and colors reflect super population designation of individuals. The tree is based on 83 SNPs (68 informative) from the 22 945 bp region.

of the 28 haplotypes with frequency $\geq$0.5% in the larger genomic region with moderate LD resulted in a clear separation of haplotypes carrying the ancestral and derived alleles.

Worldwide selection at the locus

Given the morbidity and mortality associated with viral respiratory infections and severe childhood asthma, we hypothesized that the frequency of the derived allele might have increased more rapidly than under neutrality. We performed genome-wide scans for selection in the 1000G populations using two highly related haplotype-based statistics: the integrated haplotype score (iHS) (22) and the number of segregating sites by length (nS_L) (23). These two statistics are designed to capture rapid increases in haplotypes carrying selected variants and have been applied in the context of detecting recent positive selection (22,24,25). Results of the two statistics are expected to be correlated as both statistics are based on comparing the long-range conservation of haplotypes carrying the derived and the ancestral alleles at polymorphic sites, with the two statistics differing only in how they measure this length. We calculated these neutrality statistics in each population independently (Supplementary Material, Fig. S2). Fourteen populations presented genome-wide significant values of iHS and nS_L ($\geq$95th percentile of the distribution in the population) at rs6967330, with 2 and 5 populations falling in the 99th percentile for the two statistics, respectively.

To investigate whether selection has acted on all populations simultaneously, we implemented a multipopulation (MP) combined statistic approach for iHS and nS_L. For each SNP in each population, we combined the percent ranks of iHS and nS_L (determined from genome-wide distributions in the respective population) across populations into single composite scores, MP-iHS and MP-nS_L. The rationale behind these composite approaches is that neutrality statistics, though expected to be correlated among populations under neutrality, are more strongly correlated for positively selected variants than for neutral variants (26–28). Indeed, under global positive selection, iHS and nS_L tend to become negative in all populations while false positives will only be negative in a few populations. Consequently, candidates genuinely selected in several populations should harbor extreme values for MP-iHS and MP-nS_L. We obtained a MP-iHS score of 156 and a MP-nS_L score of 157 for rs6967330, which is genome-wide significant regardless of whether we include only segregating sites in two or more populations or limited it to those SNPs segregating in all 26 populations examined (P-value < 0.01), indicating that selection has probably acted across all continental populations simultaneously. Collectively, these results suggest that the derived allele at rs6967330 is advantageous and that non-neutral processes have acted on this locus.

Origin of rs6967330

Examination of *CDHR3* in an alignment of 100 vertebrate genomes (29,30) reveals that the locus is highly conserved, with homologs present in 85 species (Supplementary Material, Fig. S3). Tyrosine is the ancestrally encoded amino acid at the homologous position 529 in the human protein sequence, with only humans and a handful of other species encoding alternative protein sequence. Providing further evidence that the locus has been evolutionary constrained, rs6967330 has a positive Genomic Evolutionary Rate Profiling score of 4.39, signifying a deficit of substitutions compared to neutral expectations; typically, scores above 2 are considered constrained with high sensitivity (30,31).

Excluding *Homo sapiens*, sequencing data from the remaining extant species comprising all hominids (great apes) are invariant at the position corresponding to rs6967330 (32). Genotyping of an additional 41 chimpanzees whose community experienced a severe cross-species respiratory outbreak of RV-C in 2013 (8.9% mortality rate) revealed that all individuals were homozygous for the ancestral A allele (14). In examining hominin genomes, we find that Neanderthals and Denisovans carry the ancestral allele, with haplotypes that nest within extant *H. sapiens* diversity (Supplementary Material, Fig. S4). Ancient DNA (aDNA) extracted from a 45 000-year-old *H. sapiens* from western Siberia revealed the man was homozygous for the derived allele (33). In low coverage sequencing data of aDNA extracted from 230 West Eurasian *H. sapiens* estimated to have lived between 6500 and 300 B.C., the derived G allele ranges from 62.1 to 84.3% among the various populations examined (34). Collectively, these data suggest that the derived allele likely arose in the evolutionary branch leading to anatomically modern humans and was not rare in human populations by ~45 000 years ago.

Age of rs6967330

Haplotype-based statistics such as iHS and nS_L are best at detecting positive selection occurring in the recent past (e.g.

$\sim$30 000 years) (35) and have low power to detect selection occurring on standing variation with intermediate frequency (36). While we find signatures of selection at the locus with these methods, the protective variant has not fixed in any of the contemporary populations examined contrary to expectations for a positively selected variant that potentially arose prior to human migrations out of Africa (37). We therefore wondered if balancing selection could explain the lack of fixation observed, as balancing selection can maintain genetic diversity over long periods of time (38). As expected based on our comparative genomic analysis, we did not find a convincing signal of long-term balancing selection (39) (i.e. selection occurring over at least hundreds of thousands of generations (40–42)) (Supplementary Material, Fig. S5). However, balancing selection operating over shorter timescales has clearly played a role in recent human evolution (e.g. the iconic example of the β^S sickle cell mutation (43–45)) and could have shaped the haplotype-based patterns similarly to what we observed in the 1000G populations. Indeed, under both positive selection and short-term balancing selection scenarios, the selected allele rapidly increases in frequency with the allele eventually fixing under positive selection or oscillating around an equilibrium frequency in the case of balancing selection (45).

To estimate the emergence of the derived G allele at rs6967330, we adapted a recently developed simulation-based approximate Bayesian computation (ABC) dating method (45) to jointly model both positive and balancing selection scenarios. Denoting the observed average frequency of the derived G allele across 1000G populations (0.8) as G_{obs}, we simulated haplotypes carrying a selected mutation with the following relative finesses for each genotype: $w_{AA} = 1$, $w_{AG} = 1 + s$ and $w_{GG} = 1 + 2s$, where the selection coefficient, s, was positive when the simulated frequency was lower than G_{obs} or negative when it was higher than G_{obs}, (i.e. the G allele is favorable when the frequency is below G_{obs} and unfavorable when it is above). Hence, when the simulated allele reaches G_{obs} for the first time in the current generation, it is consistent with a classic positive selection scenario (s was positive in every generation), while when the simulated allele oscillates around G_{obs} for many generations, it corresponds to a negative frequency-dependent selection scenario (s was either positive or negative depending on the value of its frequency, with a negative correlation between relative fitness and frequency).

We focused on African (AFR), European (EUR) and East-Asian (EAS) 1000G super populations (Fig. 3) and computed four haplotype-based statistics: iHS, nS_L, the delta of integrated extended haplotype homozygosity (ΔiHH) (26) and the derived intra-allelic nucleotide diversity (DIND) (27) (Supplementary Material, Table S2). ΔiHH measures the difference in the long-range conservation of haplotypes as opposed to the relative ratios (as computed by iHS and nS_L), while DIND compares the nucleotide diversity of haplotypes carrying each of the alleles. We performed our ABC estimations in each continental super population separately, from simulations that closely matched the empirical haplotype-based statistics, allele frequency and associated genetic diversity for each respective population. The simulated DNA regions containing the selected mutation in the middle were generated using population subdivision and demographic parameters previously inferred for AFR, EUR and EAS populations (46) and assuming G_{obs} equals 0.8 (the average of frequencies computed across continental populations). Our ABC age estimates across AFR (143 859 years with 95% CI: [95099–188 777]), EUR (162 387 years with 95% CI: [100666–202 981]) and EAS (155 027 years with 95% CI: [88554–200 636]) super populations were very similar (Fig. 3 and Supplementary Material, Fig. S6). The inferred ages ($\sim$150 000 years ago) and overlapping CIs among continental populations suggest a single emergence of the polymorphism that predates the split of African and non-African populations.

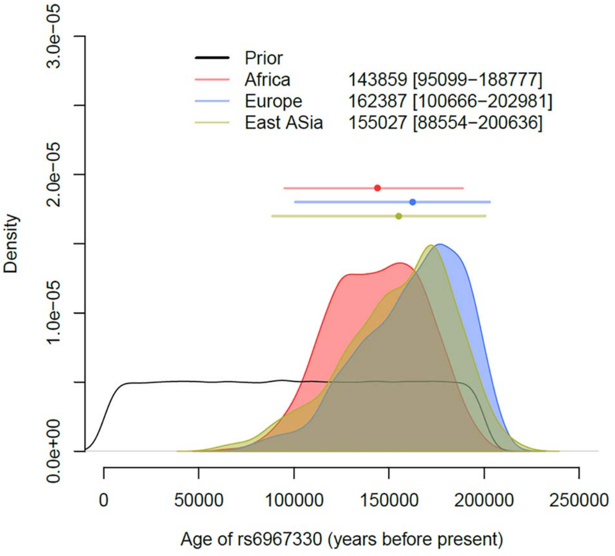

Figure 3. ABC estimations of the age of rs6967330. ABC posterior distributions of the rs6967330 age obtained from a combination of three different ABC methods. Estimations were obtained through the use of haplotype-based statistics computed with 100 kb windows around rs6967330 and five populations per continent merged. We excluded from this analysis the two admixed African populations, ASW and ACB. The posterior average (colored points) and 95% CIs (horizontal colored lines) are plotted above the distributions. The estimations obtained with each ABC method are indicated in Supplementary Material, Fig. S6. Estimations were obtained with simulations performed according to parameters described in the main text. The posterior distributions were each obtained from 200 000 simulations in which the simulated rs6967330 occurred in the respective super population.

To assess the accuracy of our ABC method, we treated simulated data as empirical data for which the ages of the selected mutation are known. The linear correlations between true values and the corresponding ABC estimates indicated that our method should successfully estimate the age of rs6967330 (Supplementary Material, Fig. S7). Moreover, this analysis validated that the 95% credible intervals (CIs) computed from the posterior distributions of age did capture the true values in 95–96% of our simulations. We also note that our estimates of the age of rs6967330 account for the uncertainty associated with the selection coefficient, s, as we simulated a wide range of selection coefficients (s uniformly distributed from 0.001 to 0.1); however, estimates of s were imprecise and thus not shown.

Mode of selection targeting rs6967330

We next sought to determine the likely mode by which the derived allele reached its current frequency (G_{obs}). Recall that when the simulated allele reaches G_{obs} for the first time in our ABC framework, it is consistent with a classic positive selection scenario (s was always positive), while when the simulated allele oscillates around G_{obs} for many generations, it corresponds to a negative frequency-dependent selection scenario (s varies depending on the frequency of G_{obs}). We can therefore quantify the number of simulations that are consistent with each selection scenario by examining the generation at which G_{obs} was first reached. As expected, we found a close relationship between the mode of selection and the age of the mutation, with younger mutations having a higher probability of having

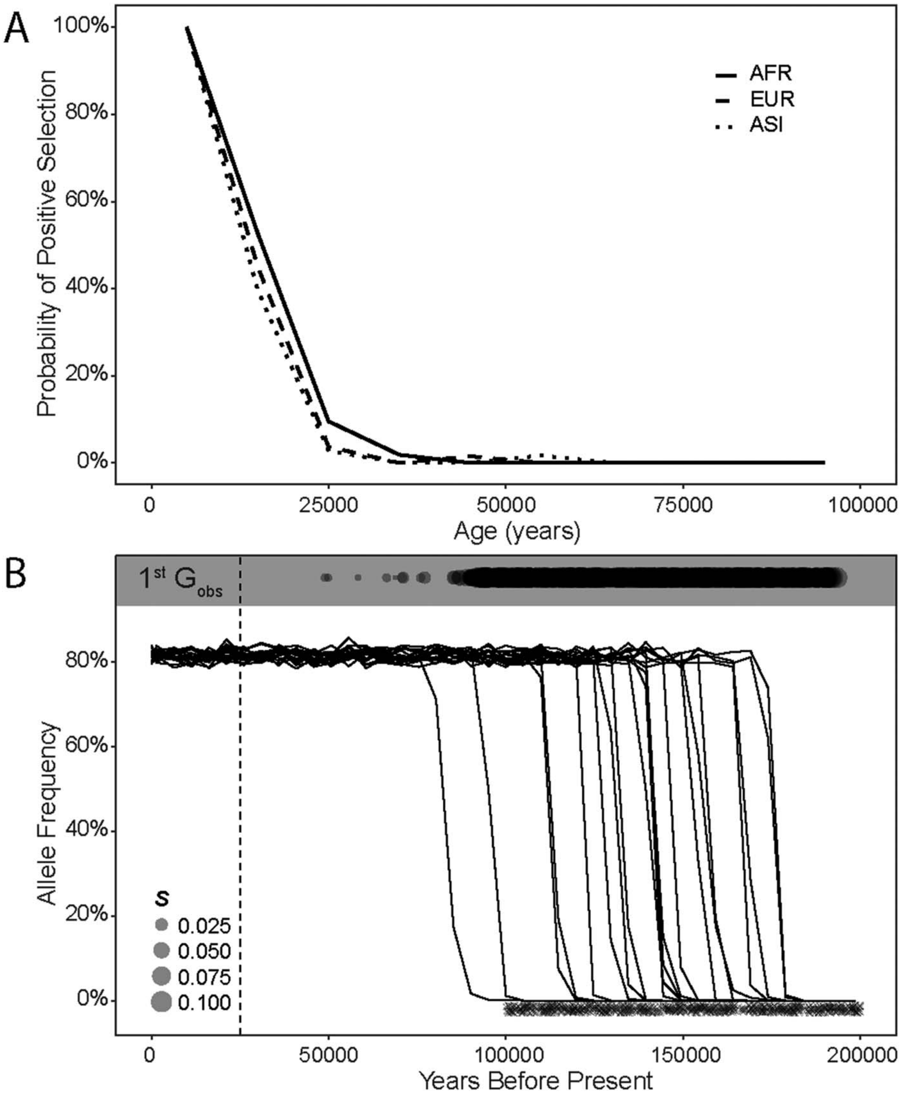

Figure 4. Rejection of the positive selection scenario. (**A**) Probability of an ongoing positive selection scenario assessed by simulation of the model used for our ABC estimations. To compute this probability, we performed 1000 simulations according to the demographic and selection models described for each superpopulation. Simulations were classified as consistent with ongoing positive selection if G_{obs} (0.8) was reached for the first time in the last 5000 years. For each bin of age, the probability of positive selection was quantified as the number of simulations consistent with ongoing positive selection divided by the number of simulations in this bin. (**B**) Assessment of the mode of selection in AFR. One-thousand simulations were performed according to the demographic and selection models described for the AFR superpopulation, restricting the emergence of the selected mutation to the 95% CIs of our ABC estimations. The date of emergence of each simulated mutation (mimicking rs6967330) is represented by an x on the *x*-axis. The date at which the derived allele for each of the 1000 simulated mutations reached Gobs for the first time is indicated by circles in the shaded boxes, with the size of circle reflecting the selection coefficient, s. The derived allele frequency trajectory for 20 randomly drawn simulations is pictured. To formally reject a scenario of ongoing positive selection, the proportion of simulations for which the derived allele of the selected mutation reached Gobs in the last 25 000 years were quantified ($P = 0.001$). (See Supplementary Material, Fig. S8 for similar plots based on EUR and AFR superpopulations.)

reached G_{obs} in a manner consistent with a classic positive selection scenario (Fig. 4A). For example, approximately half of the simulations where the mutation arose between 10 000 and 20 000 years ago reached G_{obs} in relatively recent generations (the last 5000 years). On the contrary, most simulations where the mutation arose >40 000 years ago reached G_{obs} long ago and oscillated around G_{obs} until present under the modeled negative frequency-dependent selection scenario.

In order to formally reject a scenario of classic positive selection acting on the rs6967330 locus, we performed additional simulations restricting the origin of the mutation to the inferred 95% CIs for each super population, and recomputed this probability (Fig. 4B and Supplementary Material, Fig. S8). We found very low support for ongoing positive selection acting on this locus ($P \leq 0.002$). Collectivity, our results show that the frequency of the derived G allele at rs6967330 has likely been maintained by balancing selection in anatomically modern humans.

Discussion

Following its identification as an asthma susceptibility locus (12) and its demonstration as the cellular receptor exploited by RV-C (7), *CDHR3* has become a gene of significant biological interest. We sought to characterize the evolutionary history and the role of natural selection in shaping patterns of diversity at rs6967330, the missense variant in the receptor with demonstrated roles in disease susceptibility. We found genome-wide significant signatures of selection across populations with haplotype-based selection scans (Supplementary Material, Fig. S2) and dated the emergence of the derived mutation to ∼150 000 years ago (Fig. 3 and Supplementary Material, Fig. S6). These haplotype-based selection scans are best at detecting positive selection occurring in the recent past (e.g. ∼30 000 years), and have low power to detect selection occurring on standing variation with intermediate frequency. If the derived allele at rs6967330 had been selected upon prior to divergence of African, Asian and/or European populations and subsequently become no longer advantageous in a given continent, we would not expect to find such high percent ranks of iHS or nS_L in all populations as there would have been sufficient time for mutation and recombination to break up haplotype homozygosity in the region. Thus, we interpret our genome-wide significant MP scores as evidence that selection has acted on all populations simultaneously (e.g. in the last ∼30 000 years). However, we acknowledge that this multipopulation test cannot rule out that the derived allele very recently became neutral in some sub-continental populations. Interestingly, preliminary dating estimates of RV-C point to a recent origin of the virus in the last few thousand years (47), suggesting that an alternative selective agent(s) must have been/be acting on the locus prior to emergence of RV-C (e.g. another virus). These findings well parallel to those observed at the chemokine receptor gene-5 (*CCR-5*), where an unknown historical selective pressure maintained a deletion in *CCR-5* that attenuates infectivity and disease progression of HIV (reviewed in 48), because that mutation clearly predates the emergence of AIDS.

We hypothesized that rs6967330 might be the target of balancing selection as this could reconcile the seemingly contradictory results of significant haplotype-based signatures of selection at the locus (Supplementary Material, Fig. S2), the age of the mutation (Fig. 3 and Supplementary Material, Fig. S6) and the lack of fixation of the advantageous allele in any of the populations examined (Fig. 1). Until recently, evidence for balancing selection in the human genome was limited to a few classical cases such as the heterozygous advantage conferred by the *HbS* sickle cell mutation against malaria (43–45), genes of the major histocompatibility complex/human leukocyte antigen complex (40) and the ABO blood group (49). Balancing selection, however, has recently been recognized as more prevalent than previously thought, particularly in shaping human immune system phenotypes (38,50,51). Our ABC method allowed us to reject a model of ongoing positive selection, favoring a model of balancing selection in shaping the global frequencies of rs6967330 (Fig. 4 and Supplementary Material, Fig. S8).

We modeled negative frequency-dependent selection, an evolutionary process by which the relative fitness of a genotype/phenotype depends on its frequency in the population. We reasoned that negative frequency-dependent selection is a form of balancing selection compatible with typical transmission dynamics of communicable respiratory viruses that might exploit CDHR3. For example, viral spread could be attenuated in populations harboring high frequencies of the protective allele. Another form of balancing selection we believe could be compatible here is one in which the selection coefficient, *s*, varies temporally (e.g. seasonality of virus outbreak or pandemics). However, we chose to model the former as we speculated that it is more likely to result in equivalent equilibrium frequencies across populations. We note that we did not model heterozygote advantage at the locus because in Danish children with severe asthma having even one copy of the risk variant is associated with increased risk of exacerbation and hospitalization (12).

In conclusion, our analyses combined *a priori* knowledge of a genetic variant underlying variable susceptibility to RV-C infections (7,10,14) with population genomics and Bayesian inferences to uncover support that this genetic susceptibility locus has been maintained by balancing selection in humans over a relatively short-term evolutionary history (∼150 000 years ago). This study reinforces the importance of balancing selection in shaping human genetic diversity and provides a well-supported example where frequency-dependent selection potentially has acted on a human disease susceptibility locus.

Materials and Methods

Datasets

1000 genomes project. Individual level phased sequencing data from the 1000 Genomes Project Phase 3 dataset were downloaded from ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/ (19).

Neandertal and Denisovan. Genotypes for the Vindija, Altai and Denisovan genomes generated using snpAD, an ancient DNA damage-aware genotyper, were downloaded from http://cdna.eva.mpg.de/neandertal/Vindija/VCF/ (52).

Great apes. Genotypes of primate sequences were obtained from https://eichlerlab.gs.washington.edu/greatape/data/ and converted to the corresponding human regions with the LiftOver software (30).

Haplotype networks

Indels and multiallelic sites were filtered out with bcftools (53). Variants having a Hardy–Weinberg equilibrium exact test *P*-value below 1×10^{-5} as calculated using the –hwe midp function in PLINK1.9 (54) in any of the 26 populations were removed from all populations. The core haplotype surrounding rs6967330 was identified using biallelic markers within ±50 kb of rs6967330 in Haploview (55) from all 26 populations in the 1000 Genomes Phase 3 release (Supplementary Material, Fig. S1). A large haplotype block was defined on chromosome 7 from 105 657 078 to 105 680 022, and a smaller haplotype of chromosome 7 from 105 657 078 to 105 659 873. Haplotypes within the defined haplotype blocks were extracted from biallelic markers with a minor allele frequency (MAF) > 0.01 with the Pegas package (20,21). Haplotypes occurring at >0.5% (at least 26 individuals) in the total 1000G dataset were constructed into networks. Genotypes from two high quality Neanderthal genomes and a Denisovan genome were similarly extracted and used in network analyses.

Haplotype-based selection scans

We computed iHS and nS_L using sliding windows of 100 kb around each mutation (36). As iHS and nS_L are sensitive to the inferred ancestral/derived state of an allele, we computed these statistics only when the derived state was determined unambiguously. Results were normalization by derived allele frequency (DAF) bins (from 0 to 1, increments of 0.025) (22,36). We also minimized the false-positive discovery by excluding SNPs with a DAF below 0.2, as the power to detect positive selection has been shown to be limited at such low frequencies (22,36). For each statistic, we considered the percent rank at rs6967330 relative to the genome-wide distribution in each population.

Multipopulation (MP) combined statistic

For each SNP and each population, we determined the empirical *P*-value of iHS and nS_L, i.e. the rank of each statistic in the genome-wide distribution divided by the number of SNPs. We then used the combine_pvalues function implementation of Fisher's method (56) of the SciPy python package to compute MP-iHS and MP-nS_L for every SNP (passing the criteria above to compute iHS and nS_L) found in at least two populations.

β test for long-term balancing selection

β scores for each population in the 1000 Genomes Project were obtained from https://github.com/ksiewert/BetaScan.

Human demography

To investigate the age of rs6967330 in different continental populations, we incorporated a previously inferred demographic model (46) in our ABC framework. This three-population isolation-with-migration model inferred using the YRI, CEU and CHB populations has been used to calibrate the genome-wide detection of positive selection in the 1000G populations (26,46). This demographic model incorporates an ancient African expansion, an Out-of-Africa exodus ∼100 000 years ago (assuming a generation time of 29 years (57,58)) followed by a bottleneck and a split of Eurasians into European and Asian populations ∼58 000 years ago and different migration rates between continents with a probability of the order of 10^{-5} per haploid genome per generation. One key feature is the presence of two population bottlenecks in non-African populations, the second bottleneck being stronger in the Asian population (for details see (46)).

Simulating genetic data

To simulate the rs6967330 region according to the demographic scenario described above, we used SLiM v2 (59) with the following parameters: the mutation rate $\mu = 2.3 \times 10^{-8}$ per generation per site (44) and the pedigree-based recombination rate observed in the *CDHR3* region $r = 2.10 \times 10^{-8}$ per generation per site (60). In order to compute iHS and nS_L as was done for the 1000G data, we first simulated genetic data with a selected mutation (to mimic rs6967330) inserted in the middle. We next matched the simulated data with our empirical data by randomly drawing the number of SNPs observed around rs6967330 in accordance with the allele frequency spectrum observed in this region. Because SLiM v2 is a forward-in-time simulator, computation times are large. We thus rescaled effective population sizes, generation times, the recombination rate and the mutation rate according to N/λ, t/λ, $\lambda\mu$ and λr, with $\lambda = 10$(45,61) The selection parameter *s* was multiplied by the same factor λ in order to scale the increase in frequency of the selected allele in each generation, as classically done in the case of positive selection (45,61).

ABC estimation of the age of the rs6967330 mutation

In ABC, estimations are performed by comparing summary statistics of data simulated according to specified prior distributions to those of empirical data (62). To minimize unwanted noise due to limited sample sizes, we merged samples from multiple populations into super populations according to their continent of origin (AFR, EUR, EAS), and excluded populations with significant levels of admixture. For the simulated datasets, the age of the derived mutations were drawn from a flat prior distribution ranging from present until 200 000 years ago. The selection coefficients, *s*, for the derived mutation were drawn from a flat prior distribution ranging from very low selection coefficients (0.001) to very high selection coefficients (0.1). As summary statistics, we used four haplotype-based statistics as previously described: iHS, nS_L, ΔiHH and DIND (45). In addition, we also used the current derived frequency of the selected mutation and the number of segregating sites (θ_S) computed 100 kb around the site, in order to ensure that our estimates were generated from simulations that closely match the genetic diversity observed in the studied populations. We used three different ABC methods implemented in the abc R package (21,63) setting the tolerance parameter ('tol') equal to 0.005. The three posterior distributions obtained were combined into a single posterior distribution as previously described (45). As punctual estimates of the age of the mutation, we cite the posterior mean and the 95% CIs obtained from posterior distributions.

We tested the accuracy of each ABC method used, including the combined one, by treating simulated data as empirical data for which parameter values were known (Supplementary Material, Fig. S7). To this end, we compared the estimated and simulated parameter values, denoted here by $\hat{\theta}_i$ and θ_i, respectively, using classic accuracy indices: the averaged relative error *rError* (i.e. averaged difference between estimated and true values, expressed as a proportion of the true value, $rError = \frac{1}{J}\sum_i^J(\hat{\theta}_i - \theta_i)/\theta_i$, with J simulated datasets), the root of the relative mean square error and the proportion of true values falling between bounds of the 95% CIs computed from posteriors, $95\%CIcov = \frac{1}{J}\sum_1^J 1(q_1 < \theta_i < q_2\ )$ where 1(C) is the indicative function (equal to 1 when C is true, 0 otherwise) and q_1 and $q_{2,}$ the corresponding percentiles of the posterior distributions.

Supplementary Material

Supplementary Material is available at *HMG* online.

Acknowledgements

We wish to thank Andrew Kitchen for his advice on phylogenetic analyses and Lluis Quintana-Murci and Etienne Patin for feedback on the manuscript.

Conflict of Interest statement

None declared.

Funding

National Institutes of Health (1R01AI113287-01A1 to C.S.P.).

References

1. Barreiro, L.B. and Quintana-Murci, L. (2010) From evolutionary genetics to human immunology: how selection shapes host defence genes. *Nat. Rev. Genet.*, **11**, 17–30.
2. Fumagalli, M. and Sironi, M. (2014) Human genome variability, natural selection and infectious diseases. *Curr. Opin. Immunol.*, **30**, 9–16.
3. Karlsson, E.K., Kwiatkowski, D.P. and Sabeti, P.C. (2014) Natural selection and infectious disease in human populations. *Nat. Rev. Genet.*, **15**, 379–393.

4. Siddle, K.J. and Quintana-Murci, L. (2014) The Red Queen's long race: human adaptation to pathogen pressure. *Curr. Opin. Genet. Dev.*, **29**, 31–38.
5. Quach, H. and Quintana-Murci, L. (2017) Living in an adaptive world: genomic dissection of the genus homo and its immune response. *J. Exp. Med.*, **214**, 877–894.
6. Fumagalli, M., Sironi, M., Pozzoli, U., Ferrer-Admettla, A., Pattini, L. and Nielsen, R. (2011) Signatures of environmental genetic adaptation pinpoint pathogens as the main selective pressure through human evolution. *PLoS Genet.*, **7**, e1002355.
7. Bochkov, Y.A., Watters, K., Ashraf, S., Griggs, T.F., Devries, M.K., Jackson, D.J., Palmenberg, A.C. and Gern, J.E. (2015) Cadherin-related family member 3, a childhood asthma susceptibility gene product, mediates rhinovirus C binding and replication. *PNAS*, **112**, 5485–5490.
8. Everman, J.L., Sajuthi, S., Saef, B., Rios, C., Stoner, A.M., Numata, M., Hu, D., Eng, C., Oh, S., Rodriguez-Santana, J. *et al.* (2019) Functional genomics of CDHR3 confirms its role in HRV-C infection and childhood asthma exacerbations. *J. Allergy Clin. Immunol.*
9. Basnet, S., Bochkov, Y.A., Brockman-Schneider, R.A., Kuipers, I., Aesif, S.W., Jackson, D.J., Lemanske, R.F., Jr., Ober, C., Palmenberg, A.C. and Gern, J.E. (2019) CDHR3 asthma-risk genotype affects susceptibility of airway epithelium to rhinovirus C infections. *Am. J. Respir. Cell Mol. Biol.*, **61**, 450–458.
10. Bønnelykke, K., Coleman, A.T., Evans, M.D., Thorsen, J., Waage, J., Vissing, N.H., Carlsson, C.J., Stokholm, J., Chawes, B.L., Jessen, L.E. *et al.* (2017) Cadherin-related family member 3 genetics and rhinovirus C respiratory illnesses. *Am. J. Respir. Crit. Care Med.*, **197**, 589–594.
11. Hammar, K.S., Niespodziana, K., van Hage, M., Kere, J., Valenta, R., Hedlin, G. and Söderhäll, C. (2018) Reduced CDHR3 expression in children wheezing with rhinovirus. *Pediatr. Allergy Immunol.*, **29**, 200–206.
12. Bønnelykke, K., Sleiman, P., Nielsen, K., Kreiner-Møller, E., Mercader, J.M., Belgrave, D., den Dekker, H.T., Husby, A., Sevelsted, A., Faura-Tellez, G. *et al.* (2014) A genome-wide association study identifies CDHR3 as a susceptibility locus for early childhood asthma with severe exacerbations. *Nat. Genet.*, **46**, 51–55.
13. Kanazawa, J., Masuko, H., Yatagai, Y., Sakamoto, T., Yamada, H., Kaneko, Y., Kitazawa, H., Iijima, H., Naito, T., Saito, T. *et al.* (2017) Genetic association of the functional CDHR3 genotype with early-onset adult asthma in Japanese populations. *Allergol. Int.*, **66**, 563–567.
14. Scully, E.J., Basnet, S., Wrangham, R.W., Muller, M.N., Otali, E., Hyeroba, D., Grindle, K.A., Pappas, T.E., Thompson, M.E., Machanda, Z. *et al.* (2018) Lethal respiratory disease associated with human rhinovirus C in wild chimpanzees, Uganda, 2013. *Emerg. Infect. Dis.*, **24**, 267–274.
15. Bryce, J., Boschi-Pinto, C., Shibuya, K. and Black, R.E. (2005) WHO estimates of the causes of death in children. *Lancet*, **365**, 1147–1152.
16. Busse, W.W., Lemanske, R.F. and Gern, J.E. (2010) The role of viral respiratory infections in asthma and asthma exacerbations. *Lancet*, **376**, 826–834.
17. Ferkol, T. and Schraufnagel, D. (2014) The global burden of respiratory disease. *Ann ATS*, **11**, 404–406.
18. Gern, J.E. (2010) The ABCs of rhinoviruses, wheezing, and asthma. *J. Virol.*, **84**, 7418–7426.
19. The 1000 Genomes Project Consortium (2015) A global reference for human genetic variation. *Nature*, **526**, 68–74.
20. Paradis, E. (2010) Pegas: an R package for population genetics with an integrated–modular approach. *Bioinformatics*, **26**, 419–420.
21. R Development Core Team R: A Language and Environment for Statistical Computing *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.
22. Voight, B.F., Kudaravalli, S., Wen, X. and Pritchard, J.K. (2006) A map of recent positive selection in the human genome. *PLoS Biol.*, **4**, e72.
23. Ferrer-Admetlla, A., Liang, M., Korneliussen, T. and Nielsen, R. (2014) On detecting incomplete soft or hard selective sweeps using haplotype structure. *Mol. Biol. Evol.*, **31**, 1275–1291.
24. Sabeti, P.C., Reich, D.E., Higgins, J.M., Levine, H.Z.P., Richter, D.J., Schaffner, S.F., Gabriel, S.B., Platko, J.V., Patterson, N.J., McDonald, G.J. *et al.* (2002) Detecting recent positive selection in the human genome from haplotype structure. *Nature*, **419**, 832–837.
25. Ferrari, S.L., Ahn-Luong, L., Garnero, P., Humphries, S.E. and Greenspan, S.L. (2003) Two promoter polymorphisms regulating Interleukin-6 gene expression are associated with circulating levels of C-reactive protein and markers of bone resorption in postmenopausal women. *J. Clin. Endocrinol. Metab.*, **88**, 255–259.
26. Grossman, S.R., Shylakhter, I., Karlsson, E.K., Byrne, E.H., Morales, S., Frieden, G., Hostetter, E., Angelino, E., Garber, M., Zuk, O. *et al.* (2010) A composite of multiple signals distinguishes causal variants in regions of positive selection. *Science*, **327**, 883–886.
27. Barreiro, L.B., Ben-Ali, M., Quach, H., Laval, G., Patin, E., Pickrell, J.K., Bouchier, C., Tichit, M., Neyrolles, O., Gicquel, B. *et al.* (2009) Evolutionary dynamics of human toll-like receptors and their different contributions to host defense. *PLoS Genet.*, **5**, e1000562.
28. Deschamps, M., Laval, G., Fagny, M., Itan, Y., Abel, L., Casanova, J.-L., Patin, E. and Quintana-Murci, L. (2016) Genomic signatures of selective pressures and introgression from archaic hominins at human innate immunity genes. *Am. J. Hum. Genet.*, **98**, 5–21.
29. Karolchik, D., Hinrichs, A.S., Furey, T.S., Roskin, K.M., Sugnet, C.W., Haussler, D. and Kent, W.J. (2004) The UCSC table browser data retrieval tool. *Nucleic Acids Res.*, **32**, D493–D496.
30. Rosenbloom, K.R., Armstrong, J., Barber, G.P., Casper, J., Clawson, H., Diekhans, M., Dreszer, T.R., Fujita, P.A., Guruvadoo, L., Haeussler, M. *et al.* (2015) The UCSC genome browser database: 2015 update. *Nucleic Acids Res.*, **43**, D670–D681.
31. Cooper, G.M., Stone, E.A., Asimenos, G., Green, E.D., Batzoglou, S. and Sidow, A. (2005) Distribution and intensity of constraint in mammalian genomic sequence. *Genome Res.*, **15**, 901–913.
32. Prado-Martinez, J., Sudmant, P.H., Kidd, J.M., Li, H., Kelley, J.L., Lorente-Galdos, B., Veeramah, K.R., Woerner, A.E., O'Connor, T.D., Santpere, G. *et al.* (2013) Great ape genetic diversity and population history. *Nature*, **499**, 471–475.
33. Fu, Q., Li, H., Moorjani, P., Jay, F., Slepchenko, S.M., Bondarev, A.A., Johnson, P.L.F., Petri, A.A., Prüfer, K., de Filippo, C. *et al.* (2014) The genome sequence of a 45,000-year-old modern human from western Siberia. *Nature*, **514**, 445–449.
34. Mathieson, I., Lazaridis, I., Rohland, N., Mallick, S., Patterson, N., Roodenberg, S.A., Harney, E., Stewardson, K., Fernandes, D., Novak, M. *et al.* (2015) Genome-wide patterns of selection in 230 ancient Eurasians. *Nature*, **528**, 499–503.

35. Sabeti, P.C., Schaffner, S.F., Fry, B., Lohmueller, J., Varilly, P., Shamovsky, O., Palma, A., Mikkelsen, T.S., Altshuler, D. and Lander, E.S. (2006) Positive natural selection in the human lineage. *Science*, **312**, 1614–1620.
36. Fagny, M., Patin, E., Enard, D., Barreiro, L.B., Quintana-Murci, L. and Laval, G. (2014) Exploring the occurrence of classic selective sweeps in humans using whole-genome sequencing data sets. *Mol. Biol. Evol.*, **31**, 1850–1868.
37. Stephan, W. (2016) Signatures of positive selection: from selective sweeps at individual loci to subtle allele frequency changes in polygenic adaptation. *Mol. Ecol.*, **25**, 79–88.
38. Key, F.M., Teixeira, J.C., de Filippo, C. and Andrés, A.M. (2014) Advantageous diversity maintained by balancing selection in humans. *Curr. Opin. Genet. Dev.*, **29**, 45–51.
39. Siewert, K.M. and Voight, B.F. (2017) Detecting long-term balancing selection using allele frequency correlation. *Mol. Biol. Evol.*, **34**, 2996–3005.
40. Leffler, E.M., Gao, Z., Pfeifer, S., Ségurel, L., Auton, A., Venn, O., Bowden, R., Bontrop, R., Wall, J.D., Sella, G. *et al.* (2013) Multiple instances of ancient balancing selection shared between humans and chimpanzees. *Science*, **339**, 1578–1582.
41. Wiuf, C., Zhao, K., Innan, H. and Nordborg, M. (2004) The probability and chromosomal extent of *trans*-specific polymorphism. *Genetics*, **168**, 2363–2372.
42. Teixeira, J.C., de Filippo, C., Weihmann, A., Meneu, J.R., Racimo, F., Dannemann, M., Nickel, B., Fischer, A., Halbwax, M., Andre, C. *et al.* (2015) Long-term balancing selection in LAD1 maintains a missense trans-species polymorphism in humans, chimpanzees and bonobos. *Mol. Biol. Evol.*, **32**, 1186–1196.
43. Allison, A.C. (1954) Protection afforded by sickle-cell trait against subtertian malarial infection. *Br. Med. J.*, **1**, 290–294.
44. Shriner, D. and Rotimi, C.N. (2018) Whole-genome-sequence-based haplotypes reveal single origin of the sickle allele during the holocene wet phase. *Am. J. Hum. Genet.*, **102**, 547–556.
45. Laval, G., Peyrégne, S., Zidane, N., Harmant, C., Renaud, F., Patin, E., Prugnolle, F. and Quintana-Murci, L. (2019) Recent adaptive acquisition by African rainforest hunter-gatherers of the late Pleistocene sickle-cell mutation suggests past differences in malaria exposure. *Am. J. Hum. Genet.*, **104**, 553–561.
46. Grossman, S.R., Andersen, K.G., Shlyakhter, I., Tabrizi, S., Winnicki, S., Yen, A., Park, D.J., Griesemer, D., Karlsson, E.K., Wong, S.H. *et al.* (2013) Identifying recent adaptations in large-scale genomic data. *Cell*, **152**, 703–713.
47. Palmenberg, A.C. (2017) Rhinovirus C, asthma, and cell surface expression of virus receptor, CDHR3. *J. Virology*, 2017 Mar13; 91(7) PMID 28100615; PMCID 5355607; doi: 10.1127/JVI.00072-17.
48. Arenzana-Seisdedos, F. and Parmentier, M. (2006) Genetics of resistance to HIV infection: role of co-receptors and co-receptor ligands. *Semin. Immunol.*, **18**, 387–403.
49. Ségurel, L., Thompson, E.E., Flutre, T., Lovstad, J., Venkat, A., Margulis, S.W., Moyse, J., Ross, S., Gamble, K., Sella, G. *et al.* (2012) The ABO blood group is a trans-species polymorphism in primates. *PNAS*, **109**, 18493–18498.
50. Ferrer-Admetlla, A., Bosch, E., Sikora, M., Marquès-Bonet, T., Ramírez-Soriano, A., Muntasell, A., Navarro, A., Lazarus, R., Calafell, F., Bertranpetit, J. *et al.* (2008) Balancing selection is the main force shaping the evolution of innate immunity genes. *J. Immunol.*, **181**, 1315–1322.
51. Andrés, A.M., Hubisz, M.J., Indap, A., Torgerson, D.G., Degenhardt, J.D., Boyko, A.R., Gutenkunst, R.N., White, T.J., Green, E.D., Bustamante, C.D. *et al.* (2009) Targets of balancing selection in the human genome. *Mol. Biol. Evol.*, **26**, 2755–2764.
52. Prüfer, K., Filippo, C.D., Grote, S., Mafessoni, F., Korlević, P., Hajdinjak, M., Vernot, B., Skov, L., Hsieh, P., Peyrégne, S. *et al.* (2017) A high-coverage Neandertal genome from Vindija Cave in Croatia. *Science*, **358**, 655–658.
53. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G. and Durbin, R. (2009) The sequence alignment/map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.
54. Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A., Bender, D., Maller, J., Sklar, P., de Bakker, P.I., Daly, M.J. *et al.* (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.*, **81**, 559–575.
55. Barrett, J.C., Fry, B., Maller, J. and Daly, M.J. (2005) Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics*, **21**, 263–265.
56. Fisher, R. A. (1992) Statistical methods for research workers. *Breakthroughs in Statistics*, Springer Series in Statistics, Springer, New York, NY, pp. 66–70.
57. Fenner, J.N. (2005) Cross-cultural estimation of the human generation interval for use in genetics-based population divergence studies. *Am. J. Phys. Anthropol.*, **128**, 415–423.
58. Moorjani, P., Sankararaman, S., Fu, Q., Przeworski, M., Patterson, N. and Reich, D. (2016) A genetic method for dating ancient genomes provides a direct estimate of human generation interval in the last 45,000 years. *PNAS*, **113**, 5652–5657.
59. Haller, B.C. and Messer, P.W. (2017) SLiM 2: flexible, interactive forward genetic simulations. *Mol. Biol. Evol.*, **34**, 230–240.
60. Matise, T.C., Chen, F., Chen, W., Vega, F.M.D.L., Hansen, M., He, C., Hyland, F.C.L., Kennedy, G.C., Kong, X., Murray, S.S. *et al.* (2007) A second-generation combined linkage–physical map of the human genome. *Genome Res.*, **17**, 1783–1786.
61. Hoggart, C.J., Chadeau-Hyam, M., Clark, T.G., Lampariello, R., Whittaker, J.C., Iorio, M.D. and Balding, D.J. (2007) Sequence-level population simulations over large genomic regions. *Genetics*, **177**, 1725–1731.
62. Beaumont, M.A., Zhang, W. and Balding, D.J. (2002) Approximate Bayesian computation in population genetics. *Genetics*, **162**, 2025–2035.
63. Csilléry, K., François, O. and Blum, M.G.B. (2012) abc: an R package for approximate Bayesian computation (ABC). *Methods Ecol. Evol.*, **3**, 475–479.
64. Marcus, J.H. and Novembre, J. (2017) Visualizing the geography of genetic variants. *Bioinformatics*, **33**, 594–595.