Evolutionary Genetics of a Disease Susceptibility Locus in *CDHR3*

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

[1]Laboratory of Genetics, University of Wisconsin-Madison, Madison, WI 53706, USA
[2]Department of Medical Microbiology and Immunology, University of Wisconsin-Madison, Madison, WI 53706, USA
[3]Human Evolutionary Genetics Unit, Institut Pasteur, Paris 75015, France; CNRS, URA3012, Paris 75015, France; Center of Bioinformatics, Biostatistics and Integrative Biology, Institut Pasteur, Paris 75015, France.
[4]Institute for Molecular Virology and Department of Biochemistry, University of Wisconsin—Madison, Madison, Wisconsin, USA
[5]Department of Medicine, University of Wisconsin-Madison, Madison, WI 53706, USA

* mary.oneill@wisc.edu
** cspepper@medicine.wisc.edu

Abstract

Selection pressures imposed by pathogens have varied among populations throughout human evolution, leading to inter-population differences at genetic loci mediating susceptibility to infectious diseases. A common polymorphism resulting in a C_{529} versus T_{529} change in the Cadherin-Related Family Member 3 (CDHR3) receptor is associated with severe childhood asthma exacerbations. Transfection of this nonsynonymous variant in lung epithelial cell lines results in increased cell surface expression of the protein and increased rhinovirus-C binding and replication. Given the morbidity and mortality associated with rhinovirus-C-dependent respiratory infections and asthma, we hypothesized that the 'protective' variant has been under positive selection in worldwide populations. Using publicly available phased, whole-genome sequence data for 2504 individuals from 26 human populations, we sought to determine the evolutionary history and role of selection acting on this infectious disease susceptibility locus. The 'risk' allele is the ancestral allele and is found at highest frequency in African populations and lowest frequency in East Asian populations. There is minimal population structure among haplotypes and strong evidence that the 'protective' allele arose in anatomically modern humans prior to their migrations out of Africa. We detect shared signatures of selection across human populations using haplotype based selection scans; however, the patterns observed here are not consistent with a classical selective sweep. We hypothesize that patterns may indicate short term balancing selection and/or polygenic selection.

Main Text

An increasing number of studies have described signatures of natural selection at immunity-related genes (reviewed in [1–5]), supporting the idea that infectious diseases have been important selective forces on human populations.[6] For many candidate loci, the mechanisms and phenotypic effects underlying the observed patterns of variation remain elusive. In the present

study, we investigated the evolutionary history and signatures of selection at a locus for which there are experimental data linking genotype with phenotypes that appear to modulate disease susceptibility. A non-synonymous variant in *CDHR3*, the host receptor exploited by rhinovirus-C (RV-C), results in a 10-fold difference in virus binding and replication.[7] The rs6967330 variant was discovered to affect cell surface expression of CDHR3 in transduced lung-epithelial cells.[8] Rhinoviruses, particularly RV-C, are a primary trigger of asthma exacerbations, and rs6967330 is associated with a form of childhood asthma characterized by recurrent and severe exacerbations.[9] These differences at the cellular level afford a functional explanation for differing susceptibility to RV-C infection mediated by host genetics. As severe asthma (particularly prior to the availability of modern medical interventions) and respiratory infections represent significant threats to human health,[10–13] this locus is a promising candidate target of natural selection.

Examination of the locus in an alignment of 100 vertebrate genomes[14,15] revealed that the *CDHR3* locus is highly conserved, with homologs present in 85 species (Figure 1). Tyrosine is the ancestrally encoded amino acid at the homologous position 529 in the human protein sequence. There appears to have been an A→T mutation introduced in a common ancestor of opossums, Tasmanian devils, and wallabies that results in the encoding of a phenylalanine at the homologous position. Numerous vertebrates (elephant, horse, rabbit, pika, naked mole-rat, rat, mouse, golden hamster, Chinese hamster, prairie vole, and lesser Egyptian jerboa) spread throughout the species tree also encode a different amino acid (histidine) at this position. Whether these mutations are fixed in each of these species requires further sequencing, and the effect of these nonsynonymous substitutions on protein expression and function remain to be explored.

Excluding *Homo sapiens*, sequencing data from the remaining extant species comprising all hominids (great apes) are invariant at the position corresponding to rs6967330,[16] suggesting that this allele arose somewhere in hominin evolution. The genomes of the Vindija and Altai Neanderthals and the genome of a Denisovan carry the ancestral A allele at the position corresponding to rs6967330

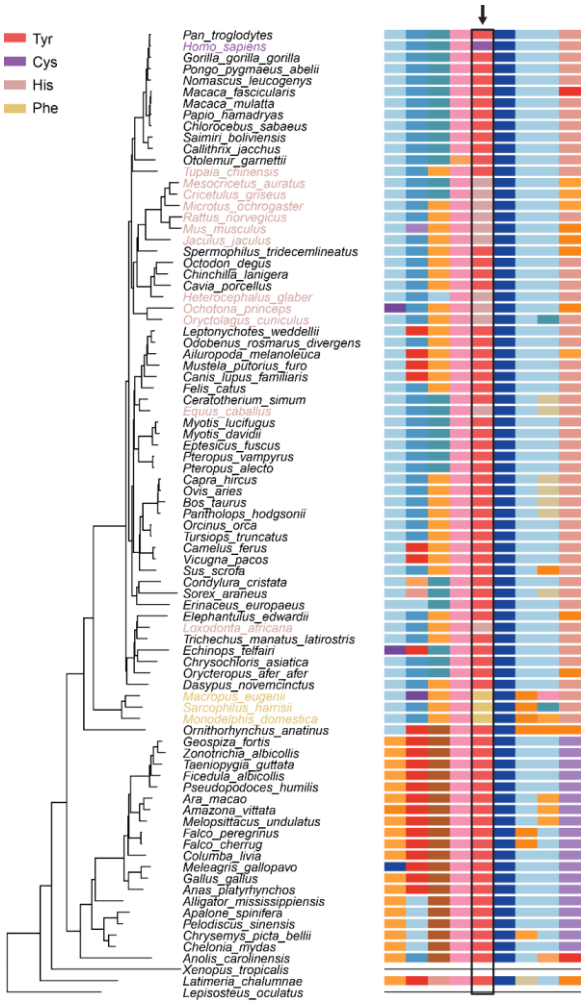

Figure 1. CDHR3 protein. (Left) Species tree for the multi sequence alignment of 85 species in the UCSC multiz alignment. (Right) Multi sequence protein alignment surrounding position 529 of the human CDHR3 protein. Residue 529 is outlined in black and designated with an arrow. Humans carry the Tyr and Cys alleles. Data obtained from the UCSC Genome Browser.[14,15]

in the human genome. Ancient specimens of anatomically modern humans do carry both alleles. In low coverage sequencing data of *H. sapiens* estimated to have lived between 5000-8000 years ago (using sequencing reads from 45 aDNA samples for which we felt confident making diploid genotype calls) we estimate the allele frequencies at rs6967330 to be 34.4% A and 65.5% G (Table S1).[17] Higher coverage aDNA extracted from a 7,000 year old skeleton found in Germany and an 8,000 year old skeleton from the Loschbour rock shelter in Luxembourg reveals heterozygotes at the locus.[18] Finally, a 45,000 year old early *H. sapiens* from western Siberia is a homozygote for the derived allele.[19]

Remarkably, the locus represents a shared polymorphism in contemporary worldwide populations. Based on whole genome sequence data for 2504 individuals from 26 different populations,[20] we find that the derived G allele is most common at the super population level 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%). At the individual population level, allele frequencies of the derived G allele range from 68.8% ("Mende in Sierra Leone", MSL) to 95.3% ("Peruvians from Lima, Peru", PEL) (Figure 2). To help determine the potential significance of observing a segregating site at high derived allele frequency (DAF) in all populations jointly, we quantified the number of biallelic sites with a minor allele frequency (MAF) ≥ 0.01 that have a DAF greater than or equal to that of rs6967330 in MSL (the lowest DAF observed) across five chromosomes (2, 5, 7, 9, and 17). Only 3.7% of SNPs meet these criteria in all 26 populations (Figure S1).

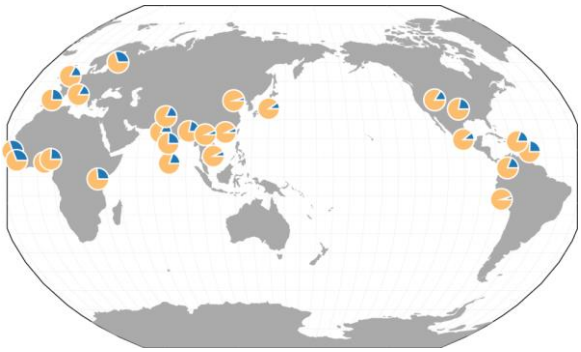

Figure 2. Global distribution of allele frequencies at rs6967330. Pie charts represent the relative allele frequencies of the ancestral A allele (blue) and the derived G allele (yellow) in each of the 26 populations of the 1000 Genomes Project.[20] This image was generated through the *Geography of Genetic Variants* browser.[46]

We explored linkage disequilibrium (LD) patterns around *CDHR3* with Haploview.[21] A tight LD block was identified on chromosome 7 from 105,657,078-105,659,873, with only moderate LD decay extending up to 105,680,022. Considering only biallelic SNPs with a MAF > 0.01, haplotypes within these blocks were identified with the Pegas package in R[22,23]. Relationships among haplotypes occurring at >1% were inferred using network analysis (Figure 3). Eight haplotypes at frequency ≥ 1% were identified from the region of negligible LD, with the majority of individuals (n = 3848) in all populations carrying the same haplotype with the derived G allele; two less frequent haplotypes carrying the derived allele are found in various geographic regions. The ancestral allele is found in the remaining five haplotypes, which vary in their distributions across regions. Twenty-eight haplotypes were found using the same criteria of a MAF > 0.01 and haplotype frequency of > 1% in the larger genomic block containing moderate LD. Phylogenetic reconstruction of the resulting haplotypes in this region resulted in a clear separation of those carrying the ancestral and derived alleles. Interestingly, Neanderthal and

Denisovan haplotypes reside within extant diversity of modern populations of *H. sapiens* (Figure S2).

Given the morbidity and mortality associated with viral respiratory infections[11,13] and severe childhood asthma,[10,12] particularly pre modern medical interventions, we hypothesized that the 'protective' G allele (resulting in a Cys_{529}) at rs6967330 would be under positive selection in human populations. We performed scans for positive selection across the five autosomes containing GWAS hits for severe childhood asthma[9] in the 26 populations of the 1000 Genomes Project.[20] The integrated haplotype score (iHS)[24] and the number of segregating sites by length (nS_L)[25] were calculated and normalized by frequency-bin for all loci with a MAF ≥ 0.01 on chromosomes 2, 5, 7, 9, and 17 in each population using Selscan[26] (Figure S3). Thirteen populations have an extreme iHS score (≥ 95th percentile of the distribution of absolute value of iHS scores in the population) at rs6967330, while 18/26 populations have an extreme nS_L score at the locus. Both iHS and nS_L statistics have been shown to dete ct hard (and with less power, soft), ongoing selective sweeps. *CDHR3* lies in a region of high recombination in the human genome (1.8cM/Mb) and rs6967330 resides in a male specific recombination hotspot.[15,27] This may explain the different patterns observed in nS_L and iHS statistics. (An excess of extreme values of iHS have been observed at regions of low recombination.[25]) Interestingly, we detected a local dip in Tajima's D (calculated in sliding-windows with vcftools[28]) immediately downstream of rs6967330 in several populations (Figure S4), indicating a skew in the site frequency spectrum towards rare alleles. This can be indicative of past selection, where recent mutation introduces low frequency variants onto a homogeneous selected background.[29]

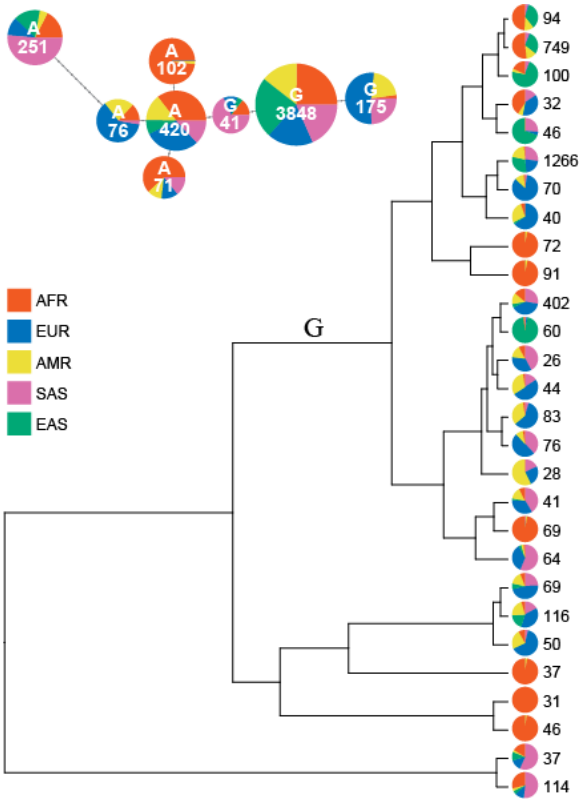

Figure 3. Haplotype structure of anatomically modern humans from the 1000 Genomes Project. (Top) Haplotype network from chromosome 7 between 105,657,078 and 105,659,873. Haplotype network was constructed from phased genome sequences of 2504 individuals with variation at 16 sites in the 2,795bp region. Only haplotypes occurring at > 1% are shown. 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 individuals were derived for a larger haplotype block. Again, only haplotypes occurring at > 1% are shown and colors reflect super population designation of individuals. The tree is based on 83 SNPs (68 informative) from the 22,945bp region.

To investigate whether selection acted on all populations simultaneously, we implemented a multi-population (MP) combined statistic approach for iHS and nS_L. We defined the MP-iHS and MP-nS_L as the mean value of the respective statistics across all populations, and did this for every SNP found in all 26 populations at a MAF ≥ 0.01 across the 5 autosomes examined (1,226,480 loci). We obtained a MP-iHS score of -1.85 (*p*-value 0.009) and a MPSC-

nS_L score of -2.27 (*p*-value 0.004), suggestive that non-neutral processes have acted on this locus globally.

Haplotype-based statistics are designed to capture rapid increases in selected variants and linked variation, with extreme values supporting positive selection occurring in the past ~25,000 years.[24,30,31] Theoretically, such methods should not detect ancient selective events. Peculiarly, several lines of evidence point to the derived allele at rs6967330 arising before human dispersals of Africa,[32] including a lack of haplotype structure (Figure 3) and the presence of the allele in aDNA from anatomically modern humans.[17–19] That the 'protective' variant has not fixed in any of the contemporary populations examined, despite having had adequate time to do so, raises the question of what evolutionary forces have shaped variation at this locus. We venture that the patterns observed cannot be explained by a classic sweep from selective pressures imposed by RV-C infection (alone), as originally hypothesized. In fact, preliminary dating estimates of this particular virus species point to a relatively recent origin, and thus if RV-C has exerted a selective pressure, it has only been in the last few thousand years.[8] Our findings suggest that an alternative selective pressure may have been/be acting on the *CDHR3* locus prior to the emergence of RV-C. These findings are reminiscent of the suggestion that an unknown historical selective pressure maintained a deletion in the chemokine receptor gene-5 (*CCR-5*) that attenuates infectivity and disease progression of HIV (reviewed in [33]), because that mutation also clearly predates the emergence of AIDS.

One possible scenario to explain the observed patterns of variation at our locus is balancing selection. 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,[34] and genes of the major histocompatibility complex/human leukocyte antigen complex, [35] and ABO blood group.[36] Balancing selection, however, has recently been recognized as more prevalent than previously thought, particularly in shaping human immune system phenotypes.[37–39] Because shared polymorphisms between human populations are expected under neutrality, most scans for balancing selection have focused on detecting long term selection resulting in trans species polymorphisms.[35,40,41] Short term balancing selection (i.e. balancing selection within a single species) clearly plays a role in shaping human diversity (e.g. *HbS* sickle cell mutation), although these legacies are much more difficult to discern from genomic signatures alone. Given that the derived variant at rs6967330 was present at high frequencies in ancient human specimens and in the homozygous state in the case of a 45,000 year old fossil, we posit that this allele started to increase in frequency as a result of balancing selection prior to human migrations out of Africa, and it reached its current equilibrium frequency more recently and independently, in worldwide populations. Such a scenario could explain the signals we detected with haplotype-based statistics, as genomic signatures of short term balancing selection are predicted to be indistinguishable from incomplete sweeps of positive selection.[24,39] As a reminder, we found rs6967330 to be an outlier in haplotype-based methods of detecting positive selection both at the global and individual population level, even when compared with other SNPs with high DAF (Figures S1 and S2). Furthermore, population differentiation as measured with F_{ST} (calculated with PLINK1.9[24]) at this locus is low relative to other SNPs with the same frequency (Figure S5), a finding which is

also expected under balancing selection.[35,37,39] However, some aspects of a short term balancing selection scenario are not yet clear. Frequencies of the derived allele range from 68.8-95.3% in modern human populations, suggesting that the equilibrium frequency of the derived mutation is high. Balancing selection can maintain functional diversity through frequency-dependent selection, heterozygote advantage, pleiotropy, and fluctuating selection,[37] and we cannot at present determine which of these scenarios may be at play here.

Polygenic selection is another possible explanation for our findings, and putatively, one that could have led to the high derived allele frequency at rs6967330 observed in contemporary human populations. Selection acts at the level of phenotypes and thus selection acting on a polygenic trait is predicted to lead to modest allele frequency shifts at many loci simultaneously (selection on standing variation) and possibly, would not be detected by commonly implemented methods of inferring positive selection.[42] Indeed, in examining haplotype-based selection statistics at the top five SNPs identified in the GWAS for severe childhood asthma, we did not detect strong signatures of selection at all loci (Figure S3).[9] (Although we did detect extreme positive values of nS_L in some EAS populations at rs928413, indicating possible positive selection for the ancestral variant at this locus. The same was not true, however, for iHS at this locus.) To better address the role of polygenic adaption, we also examined changes in allele frequencies between ancient European populations and the EUR populations of the 1000 Genomes Project, but at present, do not feel confident drawing strong conclusions from those small sample sizes (Table S1).

In total, our analyses combined *a priori* knowledge of a genetic variant underlying susceptibility to asthma[9] dependent upon RV-C infections[7] with population genetic analyses of whole genome sequence data to investigate the evolutionary history of the locus in *CDHR3*. The conservation of the protein, combined with its complex evolutionary history, exemplifies the biological importance of CDHR3, which may or may not be ultimately relevant to its function as the cellular surface receptor for RV-C. We detected a worldwide signature of selection at this locus, but also found that patterns of variation do not conform to the classic selective sweep model. Instead, we posit the possibility of short term balancing selection operating at this locus which warrants more investigation into genotypic and biochemical effects of the variant (e.g. whether there is any phenotypic benefit to being heterozygous). In Danish children with severe asthma, having even one copy of the risk variant was associated with increased risk of exacerbation and hospitalization.[9] An understanding of the function of this protein is of interest to the asthma, viral, and population genetics communities. An alternative (and not mutually exclusive) scenario to explain the patterns detected is that of polygenic adaptation. We believe further exploration of these alternative models of selection are warranted as additional data and methods are developed.

Methods

Datasets. *1000 Genomes*. Individual level phased sequencing data for Chromosomes 2, 5, 7, 9, and 17 from the 1000 Genomes Project Phase 3 dataset were downloaded from ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/.[20] Indels and multi-allelic sites were filtered out with bcftools.[43] Variants having a Hardy-Weinberg equilibrium exact test p-value

below $1x10^{-5}$ as calculated using the –hwe midp function in PLINK1.9[44] in any of the 26 populations were removed from all populations.

Mathieson et al ancient. BAM files for each of the 230 individuals included in the study by Mathieson *etal* [17] were downloaded and converted to MPILEUP format using samtools.[43] Low coverage sequencing data at asthma susceptibility loci were manually inspected for each individuals for which at least one read had mapped to the site. Manual diploid genotyping calls were made for aDNA samples for which we felt confident making diploid genotype calls (Table S1).

Neandertal and Denisovan. Geontypes 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/.

Great Apes. Genotypes of primates sequences were obtained from https://eichlerlab.gs.washington.edu/greatape/data/ and converted to the corresponding human regions with the LiftOver software.[15]

Nucleotide Diversity, Tajima's D, and Fay & Wu's H. Sliding-window estimates of Tajima's D were calculated with vcftools using 10KB and 50KB bins.[28]

Haplotype Networks. The core haplotype surrounding rs6967330 was identified using biallelic markers within 100kb of rs6967330 in Haploview[21] from all 26 populations in the 1000 Genomes Phase 3 release. 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 [22,23]. Haplotypes occurring at >1% (at least 26 individuals) in the total 1000 genomes 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 Selection Statistics. Selscan[26] was used to calculate iHS and nS_L statistics, with a MAF threshold of 0.01. All biallelic SNPs on Chromosomes 2, 5, 7, 9, and 17 were used for normalization across allele frequency bins.

Population Differentiation. Plink 1.9[44] was used to calculate the Weir and Cockerham estimate of F_{ST}[45] between each population and the YRI and LWK populations for all biallelic SNPs on Chromosomes 2, 5, 7, 9, and 17.

Acknowledgements

We wish to thank Andrew Kitchen for his advice on phylogenetic analyses. This work was supported by the National Science Foundation Graduate Research Fellowship Program (DGE-1255259) to MBO. ACP RV-C work is supported by the National Institute of Health (U19-AI070503). CSP is supported by the National Institutes of Health (R01AI113287).

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. Siddle, K.J., and Quintana-Murci, L. (2014). The Red Queen's long race: human adaptation to pathogen pressure. Current Opinion in Genetics & Development *29*, 31–38.

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. Fumagalli, M., and Sironi, M. (2014). Human genome variability, natural selection and infectious diseases. Current Opinion in Immunology *30*, 9–16.

5. Quach, H., and Quintana-Murci, L. (2017). Living in an adaptive world: Genomic dissection of the genus Homo and its immune response. Journal of Experimental Medicine *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 Genetics *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. Palmenberg, A.C. (2017). Rhinovirus C, Asthma, and Cell Surface Expression of Virus Receptor, CDHR3. J. Virol. JVI.00072-17.

9. 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.

10. Gern, J.E. (2010). The ABCs of Rhinoviruses, Wheezing, and Asthma. J. Virol. *84*, 7418–7426.

11. Bryce, J., Boschi-Pinto, C., Shibuya, K., and Black, R.E. (2005). WHO estimates of the causes of death in children. The Lancet *365*, 1147–1152.

12. 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.

13. Ferkol, T., and Schraufnagel, D. (2014). The Global Burden of Respiratory Disease. Annals ATS *11*, 404–406.

14. 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*, D493-496.

15. 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.

16. 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.

17. 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 *advance online publication*,.

18. Lazaridis, I., Patterson, N., Mittnik, A., Renaud, G., Mallick, S., Kirsanow, K., Sudmant, P.H., Schraiber, J.G., Castellano, S., Lipson, M., et al. (2014). Ancient human genomes suggest three ancestral populations for present-day Europeans. Nature *513*, 409–413.

19. 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.

20. The 1000 Genomes Project Consortium (2015). A global reference for human genetic variation. Nature *526*, 68–74.

21. 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.

22. Paradis, E. (2010). pegas: an R package for population genetics with an integrated–modular approach. Bioinformatics *26*, 419–420.

23. R Development Core Team R: A Language and Environment for Statistical Computing (Vienna, Austria: R Foundation for Statistical Computing).

24. 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.

25. 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 msu077.

26. Szpiech, Z.A., and Hernandez, R.D. (2014). selscan: An Efficient Multithreaded Program to Perform EHH-Based Scans for Positive Selection. Mol Biol Evol *31*, 2824–2827.

27. Kong, A., Thorleifsson, G., Gudbjartsson, D.F., Masson, G., Sigurdsson, A., Jonasdottir, A., Walters, G.B., Jonasdottir, A., Gylfason, A., Kristinsson, K.T., et al. (2010). Fine-scale recombination rate differences between sexes, populations and individuals. Nature *467*, 1099–1103.

28. Danecek, P., Auton, A., Abecasis, G., Albers, C.A., Banks, E., DePristo, M.A., Handsaker, R.E., Lunter, G., Marth, G.T., Sherry, S.T., et al. (2011). The variant call format and VCFtools. Bioinformatics *27*, 2156–2158.

29. Tajima, F. (1989). Statistical Method for Testing the Neutral Mutation Hypothesis by DNA Polymorphism. Genetics *123*, 585–595.

30. 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.

31. 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.

32. Nielsen, R., Akey, J.M., Jakobsson, M., Pritchard, J.K., Tishkoff, S., and Willerslev, E. (2017). Tracing the peopling of the world through genomics. Nature *541*, 302–310.

33. Arenzana-Seisdedos, F., and Parmentier, M. (2006). Genetics of resistance to HIV infection: Role of co-receptors and co-receptor ligands. Seminars in Immunology *18*, 387–403.

34. Allison, A.C. (1954). Protection Afforded by Sickle-cell Trait Against Subtertian Malarial Infection. Br Med J *1*, 290–294.

35. 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.

36. 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.

37. Key, F.M., Teixeira, J.C., de Filippo, C., and Andrés, A.M. (2014). Advantageous diversity maintained by balancing selection in humans. Current Opinion in Genetics & Development *29*, 45–51.

38. 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. The Journal of Immunology *181*, 1315–1322.

39. 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.

40. Wiuf, C., Zhao, K., Innan, H., and Nordborg, M. (2004). The Probability and Chromosomal Extent of *trans*-specific Polymorphism. Genetics *168*, 2363–2372.

41. 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.

42. Pritchard, J.K., Pickrell, J.K., and Coop, G. (2010). The Genetics of Human Adaptation: Hard Sweeps, Soft Sweeps, and Polygenic Adaptation. Curr Biol *20*, R208–R215.

43. 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.

44. 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.

45. Weir, B.S., and Cockerham, C.C. (1984). Estimating F-Statistics for the Analysis of Population Structure. Evolution *38*, 1358–1370.

46. Marcus, J.H., and Novembre, J. (2017). Visualizing the geography of genetic variants. Bioinformatics *33*, 594–595.