THE INTERPLAY BETWEEN DISEASES AND ADAPTATION IN THE HUMAN GENOME

by

Chenlu Di

A Dissertation Submitted to the Faculty of the

DEPARTMENT OF ECOLOGY AND EVOLUTIONARY BIOLOGY

In Partial Fulfillment of the Requirements

For the Degree of

DOCTOR OF PHILOSOPHY

In the Graduate College

THE UNIVERSITY OF ARIZONA

2023

THE UNIVERSITY OF ARIZONA
GRADUATE COLLEGE

As members of the Dissertation Committee, we certify that we have read the dissertation prepared by: Chenlu Di, titled:

The interplay between diseases and adaptation in the human genome

and recommend that it be accepted as fulfilling the dissertation requirement for the Degree of Doctor of Philosophy.

David Enard — Date: Dec 14, 2022

Michael S. Barker — Date: Dec 14, 2022

Ryan N. Gutenkunst — Date: Dec 14, 2022

Joanna Masel
Joanna Masel — Date: Dec 14, 2022

Final approval and acceptance of this dissertation is contingent upon the candidate's submission of the final copies of the dissertation to the Graduate College.

I hereby certify that I have read this dissertation prepared under my direction and recommend that it be accepted as fulfilling the dissertation requirement.

Date: Dec 14, 2022

David Enard
Dissertation Committee Chair
Department of Ecology & Evolutionary Biology

Acknowledgement

I sincerely thank my advisor, David Enard, for all the discussions about research, the happy discussions on exciting findings and the tough discussions to sort out puzzles. His intelligence and deep understanding of the evolutionary process has greatly helped me learn and think better about scientific questions. David is a supportive mentor who tells me “it’s not the end of the world” when problems occurred. The fun journey of making troubles and solving problems with David has trained me to be a much better scientist. The philosophy of work-life balance and great sympathy of students has also made David a great mentor, especially in the tough time during the pandemic.

I’d like to thank my committee members, Michael Barker, Ryan Gutenkunst and Joanna Masel for the massive pool of knowledge, great suggestions on my research and support for my career development. Specially for each member, I’d like to thank Mike for always being so supportive and encouraging of my research ideas; Ryan, for lending me my currently favorite population genetic book that I have read multiple times, sincerely; Joanna, for joining my committee at the last moment and helping me head out.

I’d like to thank Michael Worobey and Michael Sanderson who were in my committee for years. Thank Mike W for providing me with great freedom in exploring the questions that I am enthusiastic with. Thank Mike S for the kindness and patience in listening to my naive research ideas as well as teaching the best phylogenetic course! I would also like to thank Peter Chesson for his generous time and encouragement that helped me develop my first naive proposal draft of this dissertation. I’d also like to thank some other scientists who have contributed time and ideas to improve the research in this dissertation, especially, Elise Lauterbur and Diego Salazar Tortosa. I’d like to thank the faculty and staff of the Ecology and Evolutionary Biology department for creating a supportive environment to work, especially Pennie Rabago and Lilian Schwartz.

I want to thank my wonderful friends in and outside the EEB for spending the best time together to chat about research, hike around Tucson, try all the new and exciting activities and share happiness. Lastly, I must thank my parents and sister who provide me with tremendous support and always be there for me.

Table of Contents

Abstract

Human health is largely influenced by genetic architecture and living environments. Evolutionary processes, especially past adaptation to changing environments, shaped the genetic architecture and might deeply influence current disease risks. Advances in genomic sequencing dramatically improved our understanding of the genetic basis of diseases in the past ten years. Thousands of genes have been found to be associated with non-infectious and infectious diseases. However, the adaptation experienced by disease-associated genes is not well characterized, let alone the potential causal relationships between disease and genomic adaptation. Here, we use human genomic data to characterize the interplay between adaptation and human non-infectious diseases: what disease gene attributes may influence adaptation, and conversely how past adaptation may have shaped the landscape of disease variants.

In the first chapter, I study an important prerequisite: accounting for confounders when studying adaptation in groups of genes, for example, disease genes, relative to the rest of the genome. I show how the lack of accounting for confounding factors other than the biological categories of interest can cause spurious results in the framework of Gene Set Enrichment Analysis (GSEA) of past adaptation. I propose a pipeline that specifically addresses the methodological problems of GSEA applied to recent adaptation in the form of selective sweeps.

In the second chapter, I use the GSEA approach established in the first chapter to study the relationship between human non-infectious disease and recent adaptation. I specifically try to clarify the dominant causal direction of this relationship. Adaptation might increase the risk of diseases. For example, deleterious mutations may increase in frequency by hitchhiking with advantageous mutations and thus genes carrying deleterious variants may experience more recent adaptation compared to non-disease genes. Alternatively, pre-existing disease status associated with disease genes might affect the occurrence of selective sweeps at disease genes through the specific attributes that differentiate disease genes from non-disease genes. We find a deficit of selective sweeps in Mendelian non-infectious disease genes compared to non-disease genes in the human genome. This deficit is due to linked disease variants substantially slowing down adaptation at disease gene loci. This highlights a dominant causal relationship direction, without however excluding the possibility that selective sweeps have also increased the frequency of linked disease variants, albeit not at a sufficiently large number of genes to create a visible selective sweep enrichment to counteract the observed deficit, caused by the more predominant opposite action of disease variants slowing down linked adaptive variants. Thus, the picture that emerges from

these results is that predominantly, some pre-existing specific attributes of disease genes have limited recent adaptation at their corresponding loci.

Taking a step back to the definition of disease, disease is a phenotype that largely deviated from the optimum. What processes might increase the risk of having a largely deviated phenotype? Past strong adaptations, including those that took place a long evolutionary time ago, may have taken the associated phenotypes further from the current optimum compared to the hypothetical situation where these adaptations had not occurred. For example, for a protein whose optimal abundance is high in the current and most historical environments, past adaptation to one particular environment that lowers the abundance to the edge of the disease-causing value may increase the risk of association with diseases. Any mutation that slightly further decreases the abundance may push the abundance of the protein below the critical disease level. In this respect, past strong and rapid adaptation, as opposed to weak and slow adaptation, should have been particularly prone to cause pronounced shifts away from phenotypic optima.

An important difficulty then is to first identify past strong adaptations in the human genome. This challenge presented an opportunity for me to connect my work on non-infectious diseases and adaptation to the work done by the rest of the lab on virus-driven adaptation. As mentioned, past strong adaptation should have been more prone to distance phenotypes away from the current optima. We happen to know that viruses drove strong adaptation in human host genomes during ancient viral epidemics, in genes that interact physically with viruses (VIPs for Virus-Interacting Proteins). This strong adaptation notably likely involved adaptive changes in gene expression and abundance, a phenotype that has been shown many times to be connected to genetic diseases. Although we do not have access to past changes in protein abundance directly, we can infer past changes in protein stability, the protein property that affects abundance of folded, functional proteins.

In the third chapter, I therefore study host protein adaptations in response to viruses that were driven by changes in protein stability of VIPs. We find that past strong adaptation in VIPs mostly involved large stability changes. This result indicates that host VIP protein stability and thus protein abundance is a phenotype that was strongly selected during ancient viral epidemics. However, the optimal protein stability during past epidemics may be deviated from the current optimum after the selective pressure is weak or gone. In fact, we find compensatory evolution that keeps protein stability stable following viral epidemics in proviral VIPs which have broadly conserved non-immune host native functions.

At the same time, specifically, many genetic diseases are known to carry disease variants that decrease thermodynamic stabilities. It is possible that strong past adaptation to viral infections that largely changed protein stability in VIPs increases the risk for following mutations to be deleterious. However, further research is needed to connect these virus-driven adaptive changes in VIP stability to the present occurrence of non-infectious disease variants at VIPs. This connection represents a logical further avenue of research to continue to characterize the relationship between non-infectious disease genes and adaptation.

Chapter 1: Introduction

1.1 Why study the interplay between genetic diseases and adaptive evolution?

1.1.1 Genetic basis of diseases

The classification of diseases is a complicated biological and ethical problem. Whether a phenotype is considered as a disease varies in different cultural contexts and at different historical times. For example, homosexuality was viewed as a mental disease during the first half of the twentieth century in some contries and was finally deleted from the official listing of psychiatric disorders in 1974 (Bayer and Spitzer, 1982). However, in some other cultural contexts and historical times, homosexuality was considered as a personal preference. On the contrary, osteoporosis, which causes bones to become weak and brittle, was not officially recognized as a disease by the WHO until 1994 (World Health Organization 1994). Regardless of all these varied cognitions of diseases, the main element in identifying diseases is the biological function which contributes to survival and reproduction (Scully 2004; Lewens and McMillan 2004). The most notable definition of diseases based on biological theory is from Christopher Boorse – "Apart from universal environmental injuries, diseases are internal states that depress a functional ability below species-typical levels" (Boorse 1975; 1977). Although the biological definition of diseases is not perfectly applicable in all circumstances, it is still largely adopted in practice and fits most common sense. Here, we relied on expert-curated biomedical databases to get human non-infectious diseases and their associated genes (Piñero et al. 2015; Landrum et al. 2014; Buniello et al. 2019; The UniProt Consortium 2014). These phenotypes may not all perfectly fit the biological definition, however, a majority of these phenotypes are likely to decrease fitness in humans.

In the past ten years, Genome-Wide Association Studies or GWAS have dramatically advanced the understanding of the genetic basis of many diseases (Rappaport et al. 2013). Thousands of human genes have now been associated with Mendelian diseases (Piñero et al. 2015; 2020) and tens of thousands of variants are found to be associated with complex diseases (Claussnitzer et al. 2020). It seems that all genes have the potential to carry pathogenic mutations and may be associated with diseases, especially with complex diseases. More and more pathogenic variants are being found for both common diseases and rare diseases. With the increasing data, we know that not all genes have the same risk of having pathogenic variants, but it is not clear why specific genes are at higher risk of carrying disease variants. Moreover, we need to be aware that we are not likely to observe all the possible mutations let alone their multiple possible combinations. The pathogenic variants that we observe are the result of evolutionary history, attributes of the organism, environments and behaviors. This complexity of diseases highlights the importance of studying the origin of diseases and evaluating the fitness effects of mutations.

1.1.2 The basis of adaptation and methods to detect adaptation

Adaptation is a key evolutionary process driven by the tendency that a favored character can increase in frequency under positive selection (Darwin and Wallace 1858). Deciphering the process and detecting signals of adaptation have been central goals in evolutionary biology. In Haldane's deterministic model, a favored character can increase in frequency and may finally replace all the other alternative characters (Haldane 1924). However, not all the favored variants will be fixed in the population. For example, favored variants may fail to fix because of environmental changes and the variant is no longer favored, or the interference between the beneficial variants and deleterious variants stopped it from further increasing in frequency. In empirical analyses, we may also observe an emerging beneficial variant on the way of increasing frequency.

Fortunately, positive selection that didn't drive the beneficial variant all the way to fixation is still detectable. A beneficial variant that increases fast to a high frequency will also take the linked variants to a high frequency. These linked variants can increase to a high frequency in a short time compared to slowly increasing neutral variants that are not linked to beneficial variants. Thus, during this shorter time, fewer mutations and recombinations happen in the linked region. Therefore, a lower genetic diversity (Smith and Haigh 1974; Nielsen et al. 2005) and a long-range linkage-disequilibrium (Sabeti et al. 2002) will be observed around the selected variant. This decrease of standing variation in regions linked to a recently selected beneficial mutation is known as a "selective sweep". We are able to detect positive selection through detecting the signals of selective sweeps. However, the signals of selective sweeps decay over time. Accumulated recombination gradually breaks the linkage between loci and new mutations increase the genetic diversity. Current methods such as iHS (Voight et al., 2006) and nSL (Ferrer-Admetlla et al., 2014) , which are used in this work, detect positive selection that happened at most 50,000 years ago (Sabeti et al. 2006). More recent methods may be able to extend the time to 5000 generations which is about 140,000 years (Lauterbur, Munch, and Enard 2022). This time limitation is however informative in telling us when positive selection happened and thus, we can quantify recent adaptation by the signals of selective sweeps. In proteins, older positive selection can be detected by identifying an excess of fixed non-synonymous mutations. For example, using chimpanzees as an outgroup, approaches such as the McDonald-Kreitman test (MK test) (McDonald and Kreitman 1991) capture the cumulative signals of positive selection since humans and chimpanzees had a common ancestor, likely more than 6 million years ago (Sarich and Wilson 1967; Langergraber et al. 2012; Katoh et al. 2016). Other methods based on codon-substitution models (Goldman and Yang 1994) are also widely used to detect old adaptation, such as the branch-site test, which is a likelihood ratio test that detects positive selection along prespecified lineages of a phylogeny, by identifying accelerated nonsynonymous substitutions in aligned codons (Goldman and Yang 1994; Yang 1998; Yang and Nielsen 2002; Yang and dos Reis 2011).

The power to detect the signal of selective sweeps can be influenced by many factors, for example, recombination rate. In addition, other evolutionary processes can also confound the signal of selective sweeps. For example, negative selection against deleterious variants can also decrease the genetic diversity of linked locus (background selection) (Charlesworth, Morgan, and Charlesworth 1993). The demographic history, such as population expansion and reduction can also confound the signals of positive selection (Nielsen 2001; Macpherson et al. 2008; Pickrell et al. 2009; Torres, Szpiech, and Hernandez 2018; Cuadros-Espinoza et al. 2022). Thus, we should be careful about these confounding factors when studying positive selection. In addition to controlling confounding factors between test and control genes, the comparative approach I used systematically, where I compare large groups of genes in the same genome, solves a number of issues such as confounding by demography (the compared groups have experienced the same demography on average).

The ability to detect recent and old positive selection strongly supports the prevalence of adaptation in humans and other species. These discoveries about genomic adaptation and the development of methods for detecting positive selection are fundamental for us to further explore the interplay between adaptation and human diseases.

1.1.3 Evolutionary processes influence the risk of diseases

The risk of diseases is influenced by the genomic architecture, biological and non-biological environments, and behaviors (Benton et al. 2021). Mendelian genetic diseases are usually due to variations in a single gene. Instead, complex diseases, such as diabetes, are associated with genetic and environmental factors. Current genetic variants interact together with environments to determine disease risks in human populations. Over evolutionary time scales, the current composition of human genomes and allele frequencies has been shaped by past evolutionary events, such as migration out from Africa (Groucutt et al. 2015; Beyer et al. 2021), adaptation to agriculture (Mathieson and Mathieson 2018), past pandemics (Klunk et al. 2022) and other unknown events. These evolutionary processes reflect the interactions between humans and environments in the past and are thus informative to understand the biological function and environmental factors involved in different diseases. Therefore, the emerging field of evolutionary medicine, which integrates evolutionary perspectives into clinical studies, seeks to identify disease-associated variants, find underlying mechanisms, and promote clinical treatments.

Although we cannot completely decipher the comprehensive connections between all evolutionary events and human diseases, recent studies have found evolutionary perspectives helpful in understanding the mechanisms of diseases. Adaptation, as an evolutionary process that closely connects organisms and environments, provides a framework for understanding the origin of diseases. For example, environmental shifts can increase the risk of diseases. The variants that are neutral or beneficial in the past environment might not fit the current environment. This mismatch is happening to the human population whose living environments have dramatically changed in very recent hundreds of years. It has been found that the variants that are metabolically beneficial in times of starvation, however, increased the risk for obesity in many current human populations due to the easy access to higher-calorie foods (Minster et al. 2016; Corbett et al. 2018). In this case, obesity associated with these variants can be probably treated by adjusting behaviors. In my work, I show that the adaptation process is deeply connected to the risk of diseases and studying this connection is also helpful to understand biological functions involved in different diseases.

Understanding how past evolution influences the current diseases and the frequencies of disease variants help us better understand the origin of diseases. We can thus better predict the reproducibility of diseases association results in different populations which influence the transferability with diseases studies between populations or even species. Studying the causal reasons of diseases from the evolutionary perspective provides clues on treatment and serves as a conceptual foundation for precision medicine.

1.2 Explanation of dissertation format

The next chapter is a summary of three first-author papers focusing on addressing the connection between adaptation and human diseases. The first one improved a comparative method, and the other two papers used the method to study the interplay between adaptation and non-infectious diseases and infectious diseases specifically. In the first paper, we discuss and solve issues in gene set enrichment tests, especially for comparing recent genomic adaptation, by controlling for confounding factors and running the test in combination with block-randomized genomes. The second points out that the interference between deleterious and beneficial variants can impede adaptation and thus cause a deficit of recent adaptation in human disease-associated genes. The last draft addresses that a large portion of adaptation, especially strong adaptation, against past viral infection happened through changing host protein stability.

The first paper needs to be resubmitted, the second paper has been published and the third paper will soon be a preprint. My contribution to each of the papers will be explained at the end of each summary section.

Chapter 2: Present Study

The methods, results, and conclusions of this study are presented in the papers appended to this dissertation. The following is a summary of the most important findings of these papers. The figures, tables and references refer to the corresponding items in the attached manuscripts.

2.1 Gene set enrichment analysis of genome scans for positive selection

Gene Set Enrichment Analysis (GSEA) is frequently used to extract biological information from a set of genes of interest (Mootha et al., 2003) through comparison with a set of control genes. This comparison finds enriched biological features in genes of interest and provides an interpretation for the experiment that identifies the gene set of interest. Biological features used to conduct a GSEA are very diverse, ranging from discrete functional annotations such as those provided by the Gene Ontology Consortium (Gene Ontology, 2015), to continuous, quantitative annotations such as the gene expression level or evolutionary conservation level of genes. Here, focusing on selection signals, we argue that the power of GSEA is largely limited by the outlier approach that has been widely used to identify candidate selected loci. We further detail a number of other issues with the way GSEA of recent adaptation has typically been conducted, especially a general neglect of confounding factors. To solve these issues, we provide a solution that matches confounding factors in controls and uses GSEA in combination with block-randomized genomes. These improvements of GSEA makes it a more effective tool to interpret human genomic data. We also suggest that there is ample room for improvement to make GSEA an important tool to interpret the vast amount of ecological genomics data. In this paper we further show how to use bootstrap test together with block randomized genomes in GSEA, using the comparison of selective sweeps in human proteins that physically interact with viruses (VIPs for Virus-Interacting Proteins) and proteins not known to physically interact with viruses (non-VIPs).

2.1.1 Problems of the outlier approach

GSEA emerged ~20 years ago and is still being improved up to this day. At that time, population geneticists rapidly adopted GSEA to try to make biological interpretations of candidate genes that are identified for natural selection by the first human genome-wide genetic variation datasets (Gibbs et al. 2003). In order to identify a set of candidate selected genes, population geneticists used, and still use to this day, the outlier approach. This approach identifies genes whose values of summary statistics sensitive to positive selection are unexpected under the demographic history of the studied population alone. Using this approach, early study found enriched selection in immune genes since humans split from chimpanzees (Bustamante et al. 2005). However, older scans for more recent selection, in the form of selective sweeps, often failed to detect these enrichments, likely due to lower statistical power. The lower power may result from a combination of reasons, such as, less versatile and less powerful sweep detection statistics. On top of this, a main issue of the outlier approach has the potential to severely decrease the power of GSEA in selective sweeps. The statistics of weaker or older sweeps may not fall outside of the neutral range and thus only leaving the opportunity for very strong, unusual sweeps to be detected by the outlier approach. Fortunately, new statistical developments such as powerful sweep detection by machine learning (Schrider and Kern 2017; Sugden et al. 2018), and a greater focus on the whole distribution of summary statistics in entire groups of genes (Josephine T. Daub et al. 2015; J. T. Daub et al. 2017; Gouy, Daub, and Excoffier 2017), instead of a focus on single candidate genes, have greatly reduced the problematic reliance on the outlier approach.

2.1.2 Accounting for confounding factors

Persistent issues with the way that GSEA of recent selection scans conducted today still exist, and results in a risk of false biological interpretations. For example, both in the specific case of genome scans for selection and more broadly, GSEA have been conducted with very little consideration for potential confounding factors. More often than not, gene sets of interest and control genes often differ in many other factors than just belonging or not to that particular biological function. These factors may also influence positive selection and will confound the enrichment test. For example, gene length, one of the few confounding factors often but not always taken into account, can bias enrichment tests due to a greater preponderance of long genes to overlap with selective sweep regions in the genome (Pavlidis et al. 2012). We can alleviate the problem of confounding factors by using control sets that match the tested gene set with regard to confounding factors. The effects of confounding factors on selective sweeps are canceled out after matching the genes of interest with control genes with the same values of confounding factors. The idea is to progressively add control genes to a growing control set until it has the same size as the gene set of interest, while checking that the growing control set matches the set of interest for the desired confounding factors (Fig. 2).

2.1.3 Clustering of genes in selective sweeps

Selective sweeps, and in particular large selective sweeps, can overlap not just one but multiple neighboring genes, as is for example the case of the lactase sweep (Bersaglieri et al. 2004; Tishkoff et al.

2007). Correlated sweep summary statistics are thus expected between neighboring genes. Therefore, applying a simple permutation test or the Fisher's Exact test with GSEA to find a significant sweep enrichment is potentially inconclusive for two reasons. First, the power of the test is limited because a large sweep can include both test and control genes. This limitation is even worse when the test gene set is way smaller than the control gene set. Second, the variance of the expected null distribution, that is the distribution of the number of genes expected in a sweep under no enrichment, is increased because of gene clusters in the same selective sweep.

In order to avoid too many control genes overlapping with the same sweeps as the genes of interest, and thus decreasing the power of the GSEA, we sample control genes that are far enough from the genes of interest. However, this process, in addition to the process of matching confounding factors, limits the number of control genes that can make the enrichment test too liberal or too conservative. Fortunately, block-randomized genomes is a simple solution to all of these issues. Block-randomized genomes are genomes of which the order of genes has been shuffled randomly. However, instead of simply shuffling genes individually, large blocks of contiguous genes are shuffled. Thus, the actual order of genes are preserved within each of these large blocks (Fig. 5). For example, we split human genes into 100 blocks (each with the same number of genes) based on their order on chromosomes, and then randomly shuffled these 100 blocks. Because the expected number of genes in sweeps, even in the largest sweep, is much smaller than the number of genes in each block, the block-randomized genomes largely preserve the original clustering structure of genes in sweeps observed in the real genome. Using a large number of block randomized genomes, it is then possible to get a null distribution of the expected number of genes in sweeps given the same clustering structure. By counting the numbers of genes in sweeps at the original locations of both genes in the set of interest and the control genes, we can get the null distribution of sweep enrichment (Fig. 6). The number of genes in the set of interest and the number of control genes are exactly the same as in the real genome, and thus we can estimate the null expected enrichment distribution while fully accounting for the extra variance observed in the bootstrap test due to the effect of the sizes of control sets. Therefore, contrary to using the bootstrap test p-values alone, the block randomized genomes allow to estimate an actual, unbiased false positive risk that takes both clustering and the size of the bootstrap test control sets into account. In summary, block randomized genomes account for clustering, biases of the bootstrap test, and avoid relying on the outlier approach in the context of GSEA.

2.1.4 Contributions to the work presented in Appendix A

The inspiration for the work presented here and in Appendix A was from David Enard. The script used in this work is modified from David Enard's previous work. All data analysis was performed by me and the first draft of the paper was written by me. Substantial editing was done by David Enard.

2.2 Decreased recent adaptation at human Mendelian disease genes as a possible consequence of interference between advantageous and deleterious variants

In the past ten years, Genome-Wide Association Studies or GWAS have dramatically advanced the understanding of the genetic basis of many diseases (Rappaport et al. 2013). Thousands of human genes have now been associated with different diseases (Piñero et al. 2015; 2020). However, despite this fast development and despite the fact that multiple evolutionary processes might connect disease and genomic adaptation at the gene level, the potential causal relationships between disease and genomic adaptation are currently unclear, especially in the case of recent genomic adaptation. A gene may have a higher risk of being associated with disease after recent adaptation because deleterious mutations may hitchhike together with advantageous mutations. If this process was sufficiently widespread, disease genes might exhibit more recent adaptation than non-disease genes. Alternatively, disease genes might adapt more slowly compared to non-disease genes. For example, variants that were neutral in a stable environment might be deleterious in a new environment and thus cause diseases. It takes time for new adaptive mutations to happen and achieve a new phenotypic optimum. If this process is the dominant reason of associating with diseases, we will observe less recent adaptation in disease genes compared with non-disease genes. However, genes may also be more likely to associate with diseases and evolve more slowly because of being more constrained. We have known that disease genes are more constrained (Blekhman et al. 2008) but this fact represents a consequence of varying constraint between genes and says little specific about disease genes. Thus, in order to find relationships between adaptation and disease beyond the simple effect of constraint, we compare disease genes with non-disease genes in the same level of constraint.

2.2.1 Less recent adaptation in disease genes

We compare the rate of strong recent adaptation in the form of selective sweeps between Mendelian, non-infectious disease genes and non-disease genes across distinct human populations from the 1000 Genomes Project. We measure recent adaptation around human protein coding genes using the integrated Haplotype Score iHS (Voight et al. 2006) and the number of Segregating sites by Length nSL (Ferrer-Admetlla et al. 2014) which detect sweeps correspond to a time window of at most 50,000 years (Sabeti et al. 2006). Using the Gene Set Enrichment Analysis (GSEA) described above, we find a strong significant depletion in sweep signals at disease genes, especially in Africa (False Positive Risk =$3*10^{-4}$, Figure 3A). We find that Mendelian disease genes have experienced far fewer selective sweeps compared to non-disease genes especially in Africa.

2.2.2 Disease genes do not experience less long-term adaptation

What will be the possible causes of the sweep deficit at disease genes? Disease genes may have a constitutive tendency to experience fewer adaptive mutations because of a higher pleiotropy (Otto 2004), and/or because the new mutations in Mendelian with large effect (Quintana-Murci 2016) tend to overshoot the fitness optimum, and thus unlikely to be beneficial. Regardless of the underlying processes, this tendency predicts not only less recent adaptation but also a deficit of long-term adaptation during evolution.

To test whether Mendelian disease genes experienced less long-term adaptation, we use the McDonald-Kreitman test based method ABC-MK (Castellano et al. 2019) and GRAPES (Galtier 2016) to capture the cumulative signals of adaptive events since humans and chimpanzees had a common ancestor, likely more than 6 million years ago. We compare the proportion of adaptive non-synonymous substitutions in Mendelian disease genes with non-disease controls and find no significant difference between disease and control non-disease genes (Figure 5A,B,C,D,E). This result indicates that Mendelian disease genes do not have constitutively fewer adaptive mutations. Processes that are stable over evolutionary time such as a higher pleiotropy, or overshooting the fitness optimum, may not result in the sweep deficit at disease genes. Moreover, based on this result, we can also exclude that purifying selection in constrained genes alone can explain the sweep deficit at disease genes, because purifying selection would then also have decreased long-term adaptation. These results suggest that more transient evolutionary processes impede disease genes from adapting as fast as the rest of the genome.

2.2.3 A possible role of interference of deleterious mutations

An obvious difference between Mendelian disease versus non-disease genes is that we have found segregating disease variants in disease genes and a majority of disease variants are recessive (Amberger et al. 2019; Balick et al. 2015). Recessive deleterious mutations have been shown to be able to slow down the linked advantageous mutations from increasing in frequency (Assaf, Petrov, and Blundell 2015) . Thus, interference between deleterious mutations and advantageous mutations may decrease the recent adaptations in disease genes. This interference is strongest in regions with many deleterious variants and low recombination rate. Since the number of deleterious segregating variants at a given locus and recombination hotspot are likely to vary over evolutionary time, this process may result in a deficit in recent adaptation but not in long-term adaptation.

If the number of known segregating disease variants correlates well enough with the known disease variants in Mendelian disease genes, we will expect the sweep deficit to be particularly strong at disease genes with both many disease variants and lower recombination rates. As expected, we find a strong deficit at disease genes with both low recombination rates and high numbers of associated disease variants (Figure 6, FPR=$8*10^{-4}$), but almost no deficit at disease genes with higher recombination rates or lower numbers of associated disease variants (Figure 6, FPR=0.74). We also find that deleterious variants or recombination rates alone is not enough to explain the sweep deficit in disease genes. These observations together suggest that adaptation in disease genes has been slowed down by the interference of recessive deleterious variants, especially in low recombination regions.

2.2.4 Contributions to the work presented in Appendix B

The inspiration for the work presented here and in Appendix B was from Chenlu Di and David Enard. The first draft was written by Chenlu Di, substantial editing was done by David Enard and all the authors contributed to writing, reviewing and editing for the final published version. Most of the analyses except for long-term adaptation was done by Chenlu Di. Confounding factors table was prepared by Diego F Salazar-Tortosa, and analyses of long-term adaptation were done by Jesus Murga Moreno and David Enard during the revision.

2.3 Stability evolution as a major mechanism of human protein adaptation in response to viruses

Viruses are a major driver of adaptation in humans and other mammals. Best studied examples are genes, such as protein kinase R (PKR) and TRIM5, involved in immune functions that can adapt fast against viral infections (Elde et al. 2009; Rothenburg et al. 2009; Sawyer et al. 2005; Águeda-Pinto et al. 2019). Case studies also have frequently found positive selection targets on the interacting surface between host proteins and viral proteins. For example, multiple adaptive changes on the surface of PKR were found to defeat viral infections (Elde et al. 2009). However, adaptation against viral infections is not limited to fast evolving immune genes but is more generally found in virus-interacting proteins (VIPs) which are host proteins interacting with viral proteins (Enard et al. 2016; Castellano et al. 2019). Most (80%) of the VIPs are not known to have antiviral or broader immune functions (Enard et al. 2016). Moreover, in VIPs, only a few percent of amino acid residues are on the interacting surfaces. For example, only ~6% amino acid residues are at the interacting surface of human proteins interacting with SARS-CoV-2 (Wierbowski et al. 2021). However, the proportion of adaptive substitutions is around 50% in the proteins that interact with coronaviruses (Figure 4). This gap indicates that the widespread positive selection in VIPs is not restricted to contact interfaces.

What other protein evolution mechanisms then may drive the widespread positive selection in VIPs? Mutations may change the host native molecular functions or affect protein conformation. However, VIPs are conserved proteins whose structures may not change notably across mammals and beyond (Castellano et al. 2019). Repeated changes in host native molecular activities and protein conformation might thus be less likely to happen and explain frequent adaptation in VIPs (Lee, Redfern, and Orengo 2007; Radivojac et al. 2013; Konaté et al. 2019). Instead, evolution in protein stability, that is the abundance of the folded, functional form of proteins, may be more likely to happen to conserved VIPs because of the following two reasons. First, many amino acid changes in many parts of a protein can change stability, including parts that outside of evolutionarily conserved active sites of VIPs (Goldenzweig and Fleishman 2018; Stein et al. 2019; Li et al. 2020). This large pool matches with the large amount of adaptation found in these conserved VIPs. Second, multiple lines of evidence have shown that changes in protein stability can influence biological processes and the host fitness (Araya et al. 2012; Martelli et al. 2016; Gerasimavicius, Liu, and Marsh 2020; Cagiada et al. 2021; Høie et al. 2022).

Here, we hypothesize that adaptation against past viral infections might happen by changing protein stability in virus-interacting proteins (VIPs). To test if protein stability was the dominant driver of adaptation in VIPs, we ask if substitutions that significantly changed stability have (i) experienced more positive selection in VIPs compared to control non-VIPs, and (ii) experienced more positive selection compared to substitutions that altered stability to a lesser extent.

2.3.1 Abundant stability-changing, adaptive substitutions in VIPs

We use a recent version of the McDonald-Kreitman test (McDonald and Kreitman 1991) called ABC-MK (Urrichio et al. 2019) to estimate the proportion of adaptive amino acid substitutions based on coding sequence substitutions that occurred specifically in the human branch since divergence with chimpanzees (Methods), and variants from the 1,000 Genome projects groups located in Africa (Auton et al. 2015; Urrichio et al. 2019). We also need to know the protein stability changes caused by substitutions and variants. However, it is currently impossible to measure experimentally protein stability changes in the scale of tens of thousands of proteins and millions of amino acids mutations. Thus, we estimate stability changes caused by amino acid substitutions and variants (Figure 1A,B) using the computational method ThermoNet, one of the best performing machine-learning methods for estimating the protein stability changes by extracting protein properties from protein structures. The protein structures are from Alpha Fold and we only focus on orthologs across mammals with high-confidence structures (Jumper et al. 2021; Varadi et al. 2022, 2) which are 2900 out of ~5500 VIPs and 5700 non-VIPs (Table S1: https://www.biorxiv.org/content/biorxiv/early/2022/12/01/2022.12.01.518739/DC2/embed/media-2.xls?download=true).

We find that the proportion of adaptive amino acids in VIPs is elevated in mutations that cause large stability changes compared to mutations that change stability to a lesser extent. We call the mutations which cause stability changes above the median the Large Stability Changes group (LSC), and the group below the Small Stability Changes group (SSC). In VIPs, the proportion of adaptive non-synonymous substitutions in LSC is 34% versus only 15% in SSC. In addition, in LSC, the adaptation rate for VIPs is significantly higher than control genes (p-value=3.01×10^{-4}) which are shuffled VIPs and control non-VIPs (Figure 2). Although the adaptation rate for VIPs in SSC is still higher than control genes (p-value=4.38×10^{-2}), the p-value is two orders of magnitude larger than the p-value for LSCs. Notably, the elevated adaptation rate in LSC is mostly driven by strongly beneficial mutations. The proportion of adaptive non-synonymous substitution contributed by strongly beneficial mutations is 27% compared with a total of 34%. We further calculate the number of strong adaptive substitutions driven by viruses and largely changing the protein stability. We calculate the number by minus strong adaptive substitutions in matched non-VIPs from VIPs and we get 177 in LSC and 42 in SSC. Thus, 81% (=177/(177+42)) strong adaptive substitutions are driven by viruses and largely change the protein stability.

2.3.2 Stability explains increased VIP adaptation at buried residues

We note that mutations happening at less exposed sites which have lower relative solvent accessibility (RSA) are more likely to cause large stability changes (Tokuriki et al. 2007). This correlation might confound the results and the stability might be a by-stander factor of RSA in terms of contributing to adaptation in VIPs. As expected, we find that VIPs have experienced more adaptation in the buried (39%) than exposed parts (19%) of their protein structure (Figure 3A, B). If stability is a by-stander, we will expect to see no difference in LSCs and SSCs in buried parts. However, the elevated adaptation in the buried parts of VIPs is strongly dependent on protein stability. The LSCs in the buried parts have strongly elevated adaptation (40%) in VIPs compared to control genes (Figures 3D and FigureS2:

https://www.biorxiv.org/content/biorxiv/early/2022/12/01/2022.12.01.518739/DC1/embed/media-1.pdf?download=true), while all SSCs in the buried parts only have 10% adaptive substitutions, not different from control genes (FDR=0.31) (Figure 3F). We also observe a slightly higher adaptation in LSCs (28%) at the exposed sites compared to SSCs (20%). Together, these results suggest that protein stability changes is the major driver of increased adaptation in VIPs, especially of strong adaptation, in more buried parts. This also narrows down the possible mechanisms explaining the elevated adaptation in VIPs, since more buried parts are also less likely to actively interact with other molecules.

2.3.3 Increased stability-changing adaptive substitutions response to more RNA than DNA viruses

We further confirm the results by VIPs interacting with different viruses and participating in different biological functions. Previous work found particularly strong and abundant selection during human evolution in VIPs that interact with RNA viruses whose genome is coded by RNA but not in VIPs that interact with DNA viruses (Enard and Petrov 2018; 2020; Souilmi et al. 2021). If protein stability changes is the major mechanism of adaptation to past viral infections, we will expect to observe elevated adaptation in LSCs particularly for VIPs interacting with RNA viruses. In agreement with this prediction, we find significantly increased adaptive LSCs in their specific VIPs (compared to control genes, Table 1) in seven out of nine tested RNA viruses and only one of the six tested DNA viruses, Kaposi's sarcoma Herpesvirus KSHV (Table 1) which has strongly elevated adaptation rate in its specific VIPs (Enard et al. 2016). Moreover, we observe directional selection of protein stability in immune antiviral VIPs whose optimum may frequently shift and stabilizing selection in proviral VIPs which have many broadly conserved non-immune host native functions.

In conclusion, we find that protein stability evolution may have been an important mechanism of human adaptation in response to viruses. Our results show that the studying positive selection of quantitative patterns in VIPs as a whole group can provide new insights on molecular mechanisms that hosts applied to adapt repeatedly to viral infections.

2.3.4 Contributions to the work presented in Appendix C

The inspiration for the work presented here and in Appendix C was from David Enard. The attached draft was written by Chenlu Di and David Enard and figures are made by Chenlu Di. Most of the analyses were done by Chenlu Di. The analysis about compensatory adaptation was done by David Enard. Jesus Murga Moreno helped with using ABC-MK.

2.4 Conclusion

Adaptation happened in the further past and recent times deeply influenced the risk of diseases in current human populations. In my dissertation, I find a dominant relationship between recent adaptation and human non-infectious diseases. Focusing on Mendelian non-infectious disease genes, I find a strong deficit of recent adaptation in disease genes compared to the rest of the genome. This deficit might be due to the interference from recessive deleterious variants. In low recombination regions, linked deleterious variants impede the emerging beneficial variants from increasing in frequency. These genes thus cannot adapt fast to changing environments and are therefore associated with diseases.

I also find that disease genes do not adapt more slowly than other genes in the long term. Insead of never being able to adapt, disease genes may adapt more slowly in a transient stage because of pre-existing attributes of being more vulnerable to mutations. This high risk of having deleterious variants can possibly be driven by strong past adaptation that largely changed phenotypes. Focusing on human proteins interacting with viruses, which strongly adapted in response to past viral infections, I find that strong past adaptation in VIPs mostly happened through large changes to protein stability. The changes of protein stability influences the abundance of folded functional proteins and may thus affect host fitness. Although I haven't directly connected the past strong adaptation in VIPs with the risks of diseases, I find evidence showing that VIPs may be at the edge of or fall out from its optimum stability after strong adaptation. For proviral proteins, compensatory substitutions may help VIPs to stay in stable stability after epidemics. This process usually needs multiple substitutions and thus during this period of time, proteins may be more sensitive to mutations that change the stability in an unfavored direction. However, further studies about VIPs that are associated with diseases are needed to test this hypothesis.

In addition, the dominant relationship we found between adaptation and diseases and the main protein property that drives adaptation in response to past viral infections are not mutually exclusive from other mechanisms. Instead, I present a way to investigate the origin of diseases by disentangling the complicated evolutionary processes that interact in a temporal (past recent and now) and "spatial" (different genes in the whole genome) manner. This work points out a possible connection between adaptation and human disease and highlights insights from evolutionary biology to human health and disease.

Appendix A -- Gene set enrichment analysis of genome scans for positive selection

Authors: Chenlu Di[1] and David Enard[1]

Authors affiliation: 1. Department of Ecology and Evolutionary Biology, University of Arizona, Tucson, AZ, USA

Abstract

Gene Set Enrichment Analysis (GSEA) is often used to attempt making biological sense of genome scans for recent adaptation in humans and a growing number of other species with available genome-wide variation information. This approach has had several issues in the past, among which a substantial risk of storytelling and limited statistical power. Here, we provide an updated perspective on the past limitations of GSEA of recent genomic adaptation in the light of the most recent progress made in the area of genome scans for natural selection. We argue that the past limitations of GSEA of recent adaptation were largely a reflection of the limitations of the outlier approach that has been widely used to identify candidate loci. We further detail a number of other issues with the way GSEA of recent adaptation has typically been conducted, including a general neglect of confounding factors. We provide evidence that using GSEA in combination with block-randomized genomes solves many of these issues. Even though GSEA has had limited success in the past, we suggest that there is ample room for improvement to make GSEA an important tool to make sense of the vast amount of ecological genomics data that is already available.

Introduction

In genomics, Gene Set Enrichment Analysis (GSEA) is one of the most frequently used methods to extract biological information from a set of genes of interest (Mootha et al., 2003). GSEA aims at providing a biological interpretation for an experiment that identified a gene set of interest, by identifying biological features that are significantly enriched or depleted within that gene set, compared to a set of control genes. Often the set of control genes is the rest of the genome not included in the gene set of interest. The biological features used to conduct a GSEA are very diverse and can range from discrete functional annotations such as those provided by the Gene Ontology Consortium (Gene Ontology, 2015), to continuous, quantitative annotations such as the length or evolutionary conservation level of genes. GSEA originated ~20 years ago in response to the fast-growing need to make sense of sets of candidate genes identified by the then popular microarray studies, and is still being improved up to this day

(Al-Shahrour et al., 2006; Alonso et al., 2015; Curtis et al., 2005; Hukku et al., 2020; Hung et al., 2012; Kim and Volsky, 2005; Mootha et al., 2003). At that time, GSEA was most typically used to provide biological interpretations for sets of genes that had been found to be differentially expressed by microarray expression comparisons between two experimental conditions (for example, treatment vs. placebo, K.O. vs. wild-type mice, etc.).

The introduction of GSEA happened to coincide closely with the release of the first publicly available human genome-wide genetic variation datasets (International HapMap, 2003). Seizing the opportunity, population geneticists rapidly adopted GSEA to try to make biological sense of the candidate genes they had identified by conducting the first genome-wide scans for natural selection in the human genome. For example, in 2005 Bustamante et al. (Bustamante et al., 2005) used GSEA to identify which biological functions were enriched among genes with elevated amounts of positive or negative selection in their coding sequences. Among all the significantly enriched biological functions found by Bustamante et al., gametogenesis (the production of gametes for reproduction) and the immune response were particularly expected, given the large body of evidence on pathogen and reproductive selective pressures during evolution. This provided one of the first validations of the GSEA approach used to make biological sense of patterns of natural selection in the human genome. While Bustamante et al. had been focusing on natural selection in human coding sequences since the split of human and chimpanzee lineages from their common ancestor, multiple other studies were published that focused on finding enrichments for recent positive selection in the form of selective sweeps (a reduction of linked neutral diversity near positively selected mutations) (Barreiro et al., 2008; Carlson et al., 2005; Sabeti et al., 2006; Sabeti et al., 2007; Tang et al., 2007; Voight et al., 2006; Williamson et al., 2007). Although the different scans for selective sweeps did not always identify gametogenesis genes and immune genes as enriched in recent positive selection, more recent and powerful scans have unambiguously shown that these two functions are indeed strongly enriched for selective sweeps signals (Enard and Petrov, 2020; Schrider and Kern, 2017). Older scans for selective sweeps often failed to detect these enrichments likely due to lower statistical power, resulting from a combination of smaller samples of genotyped individual genomes, high levels of ascertainment bias (Clark et al., 2005), less versatile and powerful sweep detection statistics, and very incomplete Gene Ontology functional annotations at the time (Skunca et al., 2012; The Gene Ontology, 2019).

The statistical power of GSEA of early scans for selective sweeps was also likely affected by the widespread use of the so-called outlier approach (Kelley et al., 2006; Teshima et al., 2006; Thornton and Jensen, 2007). In order to identify a set of candidate selected genes, population geneticists used -and still use to this day- the outlier approach, which consists in identifying genes in the genome whose values of summary statistics sensitive to positive selection cannot be expected under the demographic history of the studied population alone. The expected null distribution of the selection-sensitive summary statistic in the absence of natural selection is usually determined by running population simulations under demographic models that are supposed to match the actual demographic history of the studied population as closely as possible. The two main issues with this approach are that (i) there is no guarantee that the demographic models used to determine the values of a statistic under neutral expectations are accurate, and the fact that (ii), there is no reason to expect that weaker selective sweeps, or older selective sweeps, should fall outside of the neutral expected range of summary statistics. The latter issue has the potential to severely limit the power of the outlier approach, especially when the expected neutral range overlaps with a large range of possible selective sweeps, leaving only the opportunity for very strong, unusual sweeps to be detected.

In the context of GSEA, the limitations of the outlier approach were self-imposed and could have largely been avoided. In fact, because false positive sweeps due for example to past demographic events are expected to occur at random in the genome and are certainly not expected to preferentially target specific biological functions, GSEA would have still identified genuine functional enrichments even when using sets of sweep candidates with some level of false positives. In this respect, the GSEA of recent positive selection in the human genome would have benefited from balancing the widespread concern for false positives, with an equal concern for false negatives generated by the low statistical power resulting from the outlier approach. Gouy et al. and Daub et al. clearly identified this power issue, and proposed a number of approaches to better detect moderate selection signals in the context of GSEA (Daub et al., 2015; Daub et al., 2017; Gouy et al., 2017).

Since the early examples of the use of GSEA to interpret genome-wide scans for selected genes, genome-wide genetic variation datasets have become much larger and much more common in a large number of species, well beyond the few evolutionary model species such as humans or Drosophila. At the same time, new statistical developments such as powerful sweep detection by machine learning (Schrider and Kern, 2017; Sugden et al., 2018), and a greater focus on the whole distribution of summary statistics in entire groups of genes (Daub et al., 2015; Daub et al., 2017; Gouy et al., 2017), rather than a focus on single candidate genes, have greatly reduced the problematic reliance on the outlier approach. The rapid increase in the availability of whole genome datasets has made genome-wide scans for natural selection very popular, especially in the context of domestication genomics and in the context of ecological genomics, where natural positive selection plays a central role. More specifically, many new genome scans for selection in diverse species seek to identify genes that have experienced recent positive selection in the form of selective sweeps, and then attempt to interpret the biological relevance of candidate sweeps thanks to GSEA. There are however persistent issues with the way that GSEA of recent selection scans are still conducted today, with as a result a risk of false biological interpretations. Both in the specific case of genome scans for selection and more broadly, GSEA have for example been conducted with very little consideration for potential confounding factors. Moreover, in the specific context of genome scans for selective sweeps, additional issues arise that are related to the clustering of multiple genes in large selective sweeps.

Here, we assess the current problematic GSEA practices when conducted in the context of genome-wide scans for selective sweeps. We then highlight solutions, and we more specifically suggest that block-randomized genomes represent a simple solution to many of the issues discussed. Throughout the paper we further provide a detailed example of the use of a bootstrap test together with block randomized genomes when comparing sweeps in human genes that code for proteins that interact physically with viruses (VIPs for Virus-Interacting Proteins), with genes that code for proteins not known to interact physically with viruses (non-VIPs).

Main issues of GSEA of genome scans for selective sweeps: confounding factors and the clustering of genes in the same sweeps

Confounding factors

One problematic GSEA practice derives from the fact that genes within, and control genes outside of the biological function of interest being tested, often differ in many other ways than just belonging or not to that particular biological function. For example, the thousands of human genes known to interact physically with viruses (Virus-Interacting Proteins or VIPs), differ from genes not know to interact with any virus (non-VIPs) in many other different ways, including higher expression levels or higher levels of selective constraint and conservation (Enard et al., 2016; Enard and Petrov, 2018). It is then easy to imagine a hypothetical case where an enrichment in selective sweeps at genes that interact with viruses would result not from causal interactions with viruses, but from the correlated fact that highly expressed genes are in general enriched for sweeps, and genes that interact with viruses just happen to be highly expressed genes.

Genes in a biological function and the control genes used as comparison to measure an enrichment can thus differ in many other ways, that can act as confounding factors when interpreting the enrichment in positive selection of specific biological functions. Unfortunately, very little attention has usually been paid to address the potential confounding factors that can differ between the genes in the biological function of interest and the control genes. For example, given the impact of recombination on selective sweeps signals (O'Reilly et al., 2008), no scan for selective sweeps should ever report biological functions enriched for sweeps, without verifying first that genes in those functions have different or similar recombination rates. Indeed, in that case there is a serious risk of confusion between a lower average recombination rate and a genuine excess of sweep signals (O'Reilly et al., 2008).

Confounding factors can result in spurious interpretations in diverse other ways. For example, gene length, one of the few confounding factors often but not always taken into account, can bias enrichment tests due to a greater preponderance of long genes to overlap with selective sweep regions in the genome (Pavlidis et al., 2012). Although some biological functions may only have a few confounding factors such as gene length that are easily corrected for, many biological functions are likely to exhibit many simultaneous confounding factors that may be much more difficult to address. For example, human genes that interact physically with viruses differ from other genes in the genome for many potential confounding factors (Enard et al., 2016; Enard and Petrov, 2018). Perhaps more problematic in this specific example is the fact that genes that interact with viruses, compared to other genes, are also enriched in many host Gene Ontology biological functions that could drive selective sweeps rather than interactions with viruses (Enard and Petrov, 2020). These host GO functions notably include the immune response, apoptosis or the mitotic cell cycle. All these factors clearly need to be considered before concluding that viruses increased the number of sweeps at virus-interacting genes (Enard and Petrov, 2020). This large number of potential confounding factors is however problematic, because there might be very few or no control genes at all that match the genes of interest for all those factors, and that could be used to set a fair comparison required to estimate an unbiased enrichment genuinely due to interactions with viruses.

After we detail a number of other issues with the GSEA applied to positive selection, we describe in the second part of this paper (Solving GSEA problems) a bootstrap test to measure positive selection enrichments while simultaneously accounting for multiple confounding factors all at once.

Clustering of genes in selective sweeps

While confounding factors likely influence any GSEA, GSEA of scans for selective sweeps have the additional issue of clustering. Selective sweeps, and in particular large selective sweeps, can overlap not just one but multiple neighboring genes, as is for example the case of the lactase sweep (Bersaglieri et al., 2004; Tishkoff et al., 2007). Correlated sweep summary statistics are thus expected between neighboring genes. This makes neighboring genes not independent from each other when it comes to the evidence for selective sweeps (Fig. 1). This non-independence makes applying GSEA with a simple permutation test or the Fisher's Exact test to find a significant sweep enrichment potentially inconclusive for two reasons. First, genes from a specific biological function that is genuinely enriched in sweeps are close to other genes not involved in that function that are typically used as controls. This means that especially large sweeps increase the number of genes found in a sweep in the gene set of interest, but also in the set of control genes, which is likely to limit the statistical power to detect significant sweep enrichments. This power limitation may be especially severe when testing biological functions with many genes spread out across the genome. Indeed, in that case many "control" genes may actually be close to a gene with the tested biological function, and will then be included as sweep candidate genes. This is expected to render the enrichment test severely underpowered. For example, there are currently 5,291 human genes known to interact with viruses (as of July 2020, latest update by D. Enard), and 6,638 other genes in the genome are found at a distance less than 100kb form these virus-interacting genes. Although there is a simple solution that consists in selecting control genes far enough from the genes of interest, this has only been done in very few cases of GSEA applied to scans for selective sweeps (Enard and Petrov, 2020).

Second, the clustering of multiple genes in the same selective sweep makes simple permutations or the Fisher's Exact test inconclusive for GSEA, because the clustering of sweep candidate genes increases the variance of the expected null distribution of the number of genes expected in a sweep under no enrichment. To understand why, one can imagine the extreme hypothetical case where a specific biological function has 100 genes in the genome, and all the 100 genes are found packed together within a single cluster. In this extreme case, all the 100 genes happen to overlap a strong, highly significant, single sweep signal. In the entire genome, 1% of genes in total are found in strong sweep signals. In this case, a simple permutation test or a Fisher's exact test will conclude that 100% of genes with the tested biological function in a sweep represents a very strong and significant enrichment compared to the 1% expected by chance with no enrichment. In this example, an extreme level of clustering is confused with what would be an extreme enrichment only if all the 100 genes were independent from each other and spread out in different genomic locations. Although this is an extreme example, even more moderate levels of genes clustering in the same selective sweeps can also lead to erroneously conclude for a sweep enrichment when there is none. In the second part of the paper, we describe how block-randomized genomes make it possible to run GSEA while taking clustering into account.

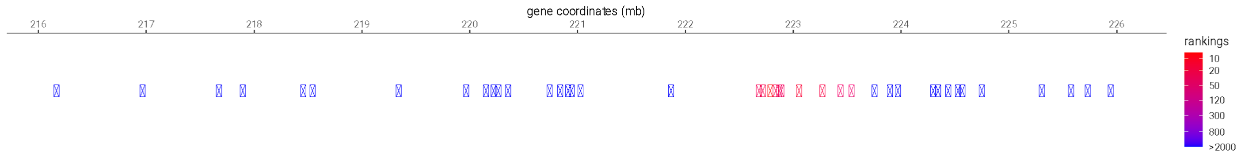


Figure 1. Clustering of multiple genes in the same candidate sweep.
Protein coding genes are represented by a dot located at the center of their genomic coordinates on human chromosome 1. Genes were ranked according to the iHS summary statistic in the 1000 Genomes Project ACB population. Multiple genes are clearly clustered in the same candidate sweep around 223 Mb.

Solving sweep GSEA limitations combining a confounding factors-aware bootstrap test and block-randomized genomes

Accounting for confounding factors with a bootstrap test

As discussed above, there are multiple issues with running a naïve GSEA of scans for selective sweeps, including confounding factors and clustering. We now seek to provide possible solutions to these issues. More often than not, gene sets of interest differ from the control genes used in a GSEA by more than just the biological difference of interest that is actually being tested by contrasting a specific gene set with control genes. For example, when comparing sweeps at genes that interact with viruses with genes that don't, the biological difference of interest is the presence/absence of interactions with viruses. But VIPs also happen for example to have many more interactions in the human Protein-Protein Interaction (PPI) network, compared to non-VIPs (Enard et al., 2016). Interestingly, Luisi et al. (Luisi et al., 2015) and Schrider & Kern (Schrider and Kern, 2017) found that human genes with more PPI in the human interaction network overlap with more selective sweeps than genes with less PPI. In this case, we are not able to tell whether VIPs are enriched in selective sweeps due to interactions with viruses or due to interactions with other human proteins, by just running a naïve GSEA. PPIs are thus a potential confounding factor that can result in a spurious interpretation of the GSEA for selective sweeps between VIPs and non-VIPs. Other factors such as recombination rate are also correlated with sweep signals, or the ability to detect sweeps, and can also be confounding factors. The interpretation of the results from a GSEA is ambiguous without taking confounding factors into consideration.

We can alleviate the problem of confounding factors by using control sets that match the tested gene set with regard to confounding factors. For example, we don't expect to observe any difference in selective sweeps between two sets of genes with the same number of PPI if PPI number is the only factor that influences the number of sweeps. In other words, if PPI are the only confounding factor, any observed difference in the number of sweeps after matching PPI due to the tested biological feature, must be due to the latter. The effects of confounding factors on selective sweeps are canceled out after matching the genes of interest with control genes with the same values of confounding factors. Then the question is: how can we build sets of control genes with confounding factors matching the set of interest? For example, if VIPs have three times more PPIs than non-VIPs, how do we build control sets of non-VIPs also with three times more PPIs from the whole pool of non-VIPs? And importantly, can we do this for not just one factor, but for multiple confounding factors simultaneously? We can build control sets as already described in Castellano et al. (Castellano, 2019) and Enard & Petrov (Enard and Petrov, 2020). The idea is to progressively add control genes to a growing control set until it has the same size as the gene set of interest, while checking that the growing control set matches the set of interest for the desired confounding factors (Fig. 2). This is however much easier said than done.

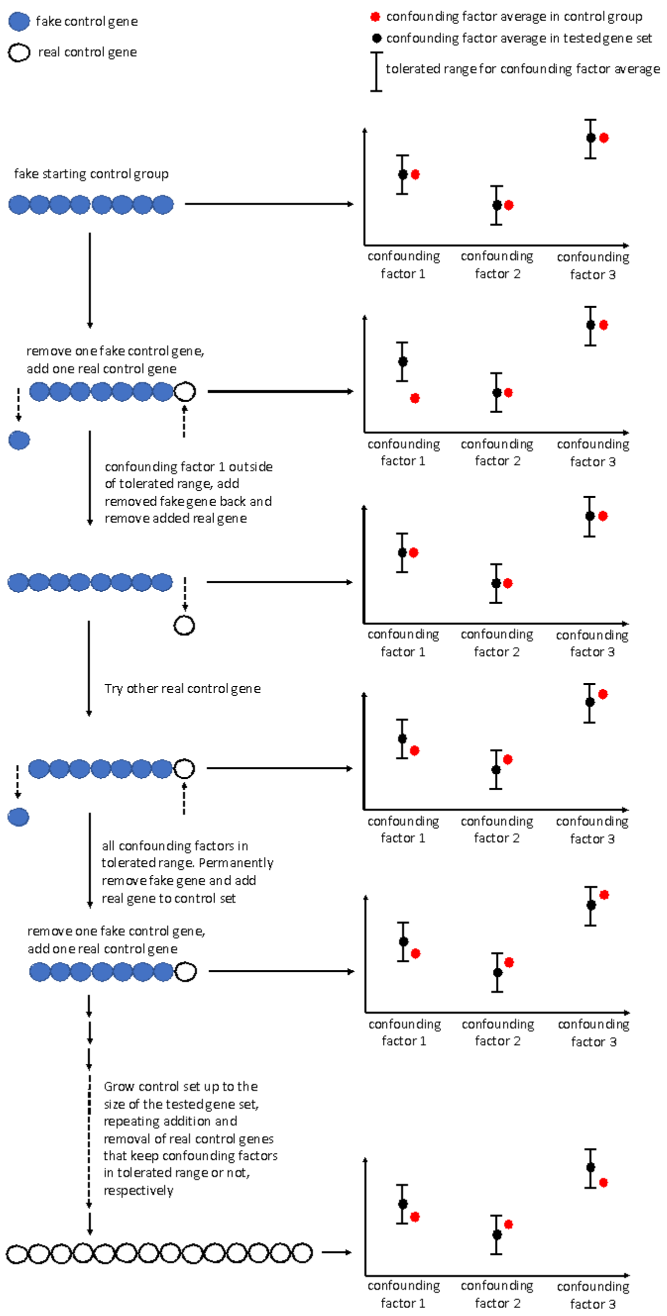

fake control gene
real control gene
confounding factor average in control group
confounding factor average in tested gene set
tolerated range for confounding factor average
fake starting control group
remove one fake control gene, add one real control gene
confounding factor 1 outside of tolerated range, add removed fake gene back and remove added real gene
Try other real control gene
all confounding factors in tolerated range. Permanently remove fake gene and add real gene to control set
remove one fake control gene, add one real control gene
Grow control set up to the size of the tested gene set, repeating addition and removal of real control genes that keep confounding factors in tolerated range or not, respectively
confounding factor 1
confounding factor 2
confounding factor 3

Figure 2. Building a control set with matching confounding factors

In practice, a candidate control gene is picked up randomly from all potential control genes (Fig. 2). The candidate control gene is then added to the control set if the confounding factors in the growing control set still match the set of interest (Fig. 2). To decide whether the growing control set matches the set of interest, different matching criteria can be used. For example, we can just check that the average values of confounding factors in the growing control set are close enough to their average value in the set of interest, within a margin of error. For example, close enough may be defined as an average for the growing control set that falls within an interval between 95% and 105% (plus or minus 5%) of the average for the set of interest. Or, to match not only averages but more precisely the whole distributions of each confounding factor, we can check that multiple quantiles of the distribution stay close to the same quantiles in the set of interest. In this matching process, any newly added gene to the control set that pushes it outside of the defined "close enough" range is discarded, and a new candidate control gene can be randomly sampled until one matches (Fig. 2). The process can be iterated until the control set has reached the size of the tested set of interest. In the sampling process, a potential control gene may be resampled multiple times and included in the control set, thus making it a bootstrap. Using the same control gene multiple times is required if the number of control genes is limited, and if using control genes only one time is insufficient to reach the size of the tested set. Note that using the same control genes multiple times in a bootstrap test can make the latter non-nominal, i.e. either too liberal or too conservative due to a small gene sample size from which to pick up controls. However, we show in the next part how using block-randomized genomes solve this issue by enabling the recalculation of nominal p-values that represent true false positive risks. Resampling the same control genes in a bootstrap process is necessary when the number of available control genes is limited. A scarcity of control genes can happen for multiple reasons, including (i) the necessity of sampling control genes far enough from the genes in the set of interest, (ii) testing a gene set of interest with one or multiple confounding factors with values that differ greatly from other genes, and (iii) attempting to match a large number of confounding factors.

Because sweeps can extend over multiple genes, it is crucial to sample control genes that are far enough from the genes of interest, in order to avoid too many control genes overlapping with the same sweeps as the genes of interest, and thus decreasing the power of the GSEA. Far enough here ideally means further than the size of the largest sweeps. However, excluding nearby genes will especially limit the pool of available control genes when testing a large set of interest, because then there are fewer places left in the genome that are far from genes in this set. For example, there are 5,291 known VIPs in the human genome, and 14,967 non-VIPs. However, of these 14,967 non-VIPs, only 8,329 are more than 100kb away from the nearest VIP, and only 2,424 are more than 500kb away, which is closer to the size of large sweeps in the human genome (Enard and Petrov, 2020; Voight et al., 2006; Williamson et al., 2007). There are therefore half the number of appropriate non-VIP controls as there are VIPs, which clearly illustrates the need for a bootstrap with repeated sampling of the same control genes. Note that to our knowledge, the proximity of control genes to the genes in the set of interest has almost never been taken into account in past GSEA of selective sweep genome scans, which might have limited their statistical power to find sweep enrichments in biological functions represented by many genes.

Another difficulty when building control sets is that the values of confounding factors in the gene set of interest may be very different from the values in the set of potential control genes. This is expected to make the matching of confounding factors more difficult, because only a subset of potential control genes will actually preserve the matching of confounding factors when progressively building the matched control sets. In the same vein, only a limited subset of potential controls will actually preserve the matching of confounding factors when attempting to match a large number of confounding factors.

The scarcity of appropriate, matching control genes can make the building of control sets difficult, and this is especially true at the starting phase when there are no or only a few newly added genes in the growing control set. This is because the first potential control genes cannot be added to the growing control set unless they match closely the tolerance ranges of all the confounding factors. In practice, very few genes may fulfill these requirements, especially when there are many confounding factors included in the matching. In order to successfully run the bootstrap while avoiding using always the same control genes that individually closely match the tolerated range, we start the building of control sets by using a "fake" starting group of control genes (Fig. 2). Each individual gene in the fake starting group matches the average of all the confounding factors in the genes of interest perfectly. This group then acts as a buffer, making it less likely that any newly added, real candidate control gene will push the growing control set out of the tolerated range (due to the small weight of one gene in the average calculated including the "fake" group; Fig. 2). Each time a new actual control gene is added, a fake control gene can be removed until the fake control group is completely removed (Fig. 2), at which point the actual control set has grown sufficiently to buffer on its own the confounding factor deviations of newly added genes. The size of the initial fake group can be adjusted, but in practice we have found that groups of 50 fake starting genes work well to jump-start the building of control sets. Importantly, this procedure increases the range of genes that can be used as controls, by allowing more genes with diverse combinations of confounding factors to be added to the growing control set without falling outside of the tolerated range. In summary, there are many confounding factors that can affect a GSEA of recent selection scans that have rarely been considered. The bootstrap approach we just described to take these factors into account is implemented in a pipeline available at https://github.com/DavidPierreEnard/Gene_Set_Enrichment_Pipeline.

Applying the bootstrap test to compare VIPs and non-VIPs.

Here we further illustrate the usefulness of the bootstrap test to match confounding factors when comparing VIPs and non-VIPs. Altogether, there currently are 5,291 human VIPs known to interact with different viruses that infect humans (Table S1). We previously showed using a slightly smaller set of VIPs that they are substantially enriched for large selective sweeps compared to non-VIPs (Enard and Petrov, 2020). This reference also provides more details on how VIPs are identified. Comparing VIPs and non-VIPs is interesting for our purpose, because the large number of VIPs and the differences in potential confounding factors between VIPs and non-VIPs make it challenging to run the bootstrap test. It is therefore a good example of what to do in a rather extreme case. Running the bootstrap test is difficult, first because the large number of VIPs means that there are not that many potential control non-VIPs far enough from VIPs. To run the bootstrap test we decided to use control non-VIPs more than 500kb from VIPs, because large sweeps in the human genome can extend this far or even further. This leaves us with 2,424 potential control non-VIPs. Second, running the bootstrap test is difficult in that case because there is a significant number of potential confounding factors to match between VIPs and control non-VIPs. This is expected to further limit the number of non-VIPs that can be used to build the control sets. Table 1 lists 13 potential confounding factors considered, all of which are significantly different between VIPs and non-VIPs and might affect the occurrence of sweeps (two-sided Wilcoxon test P<0.05).

factor	Wilcoxon test p-value
GTEx v8 average expression	<10^-16
GTEX v8 lymphocyte expression	<10^-16
GTEX v8 testis expression	<10^-16
# of gene neighbors	0.02863
deCode recombination	0.001058
GC content	4.41E-15
CDS density	3.03E-07
PhastCons conserved density	<10E-16
DNase I density	<10E-16
gene length	2.02E-09
# of protein-protein interactions	<10E-16
# of bacteria-interacting proteins	<10E-16
# of immune genes	<10E-16

Table 1. List of potential confounding factors when comparing VIPs and non-VIPs
Expression data is from GTEx release 8 (Consortium, 2013). Because we look specifically for large sweeps, we measured the following factors in 500kb genomic windows: # of gene neighbors, deCode 2019 recombination map (Halldorsson et al., 2019), GC content, CDS density, PhastCons conserved elements density,and DNAse I density. DNAse I density provides a measure of regulatory sequence density. The number of bacteria-interacting proteins is from the Intact database (Orchard et al., 2014).

Running the bootstrap test trying to match all these 13 confounding factors fails, due to an insufficient number of adequate control non-VIPs (the bootstrap fails after a number of unsuccessful sampling of control genes to add to the control set; see https://github.com/DavidPierreEnard/Gene_Set_Enrichment_Pipeline). There is however no need to match all the factors, because some, even if they differ between VIPs and non-VIPs, may either (i) not affect the sweep enrichment at all, or (ii) affect it in the conservative direction; that is, not matching these specific factors results in a smaller and thus conservative estimate of the enrichment. To identify which factors need to be taken into account, we need to first estimate separately the impact of each factor on the sweep enrichment at VIPs.

To estimate sweep enrichments, we use the iHS summary statistic measured on the 1,000 Genomes Project phase 3 genomes (Genomes Project et al., 2015). We define a one Mb window centered at the genomic center of each Ensembl protein-coding gene in the human genome, and measure the average iHS value across SNPs in this entire window. We use large windows because we previously specifically found an enrichment in strong, large sweeps at VIPs (Enard and Petrov, 2020). We then rank protein-coding genes according to their corresponding average iHS value in a specific human population. For more details, please refer to (Enard and Petrov, 2020). Importantly, we then do not use just one set of outlier iHS sweeps, and instead explore the enrichment in iHS sweeps across all the 26 1,000 Genomes Project populations at VIPs using a whole range of top ranks, from the top 2,000 iHS sweep candidates, to the top 10 iHS sweep candidates (Fig. 3A,B and C). Indeed, as already explained in the introduction, there is no valid reason to limit GSEA to an arbitrary set of outliers defined using the outlier approach, because false positive sweeps happen at random in the genome. Thus, instead of estimating the sweep enrichment for one arbitrarily defined set of outlier candidates, we estimate a more agnostic enrichment curve that better captures the effect of a whole range of possible sweep signals, from weaker, or older fading sweeps (top 2,000), to stronger, very recent sweeps (top 10). If there is no sweep enrichment, the enrichment curve is expected to be flat.

Using this enrichment curve, we find that seven of the 13 factors have either a negligible impact on the VIP sweep enrichment (the enrichment curve stays largely the same), or impact the estimated sweep enrichment in the conservative direction, i.e, not matching these factors between VIPs and the control non-VIPs results in smaller enrichment estimates.

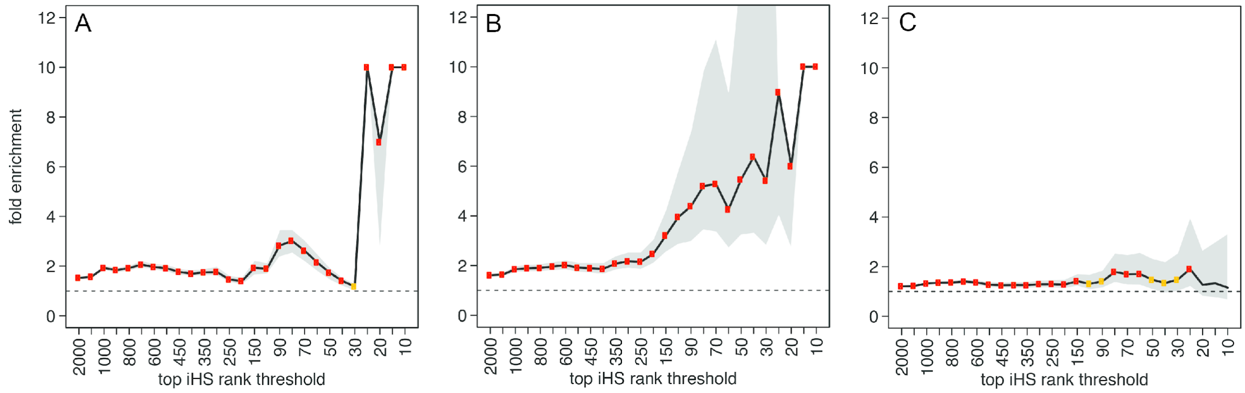


Figure 3. Enrichment of iHS sweeps at VIPs compared to non-VIPs
Fold enrichment (y-axis) is the number of VIPs in candidate sweeps divided by the average number of control-non-VIPs in candidate sweeps. VIPs and non-VIPs in candidate sweeps are counted if they belong to top x iHS genes (x-axis), where x is a rank threshold that slides from top 2,000 to top 10, taking in total 27 values (2,000; 1,500; 1,000; 900; 800; 700; 600; 500; 450; 400; 350; 300; 250; 200;150; 100 ;90 ; 80 ;70 ;60 ;50 ;40 ;30 ;25 ;20; 15**;** 10). A fold enrichment of y=k at top x=50 means that there are k times more VIPs in the top 50 iHS genes than control non-VIPs on average (10, 000 control sets of non-VIPs). The black line shows observed fold enrichment at VIPs. The grey area shows the 95% confidence interval of the fold enrichment. Fold enrichments that exceed ten are represented at ten. In this case, the confidence interval is not represented. However, the lowest edge of the confidence intervals not represented are all above one. The highest or lowest edge can be infinite when none of non-VIPs are in candidate sweeps. When this happens, the infinite edge is not shown. Orange dots means bootstrap test P<0.05 and red dots indicate bootstrap test P<0.001. Dashed line indicates fold enrichment of one, that is, no enrichment. (A) Enrichment curve when the minimum distance between VIPs and non-VIPs is 500kb and matching confounding factors. (B) Enrichment curve when the minimum distance between VIPs and non-VIPs is 500kb and confounding factors are not matched. (C) Enrichment curve when minimum distance is 100kb and confounding factors are matched.

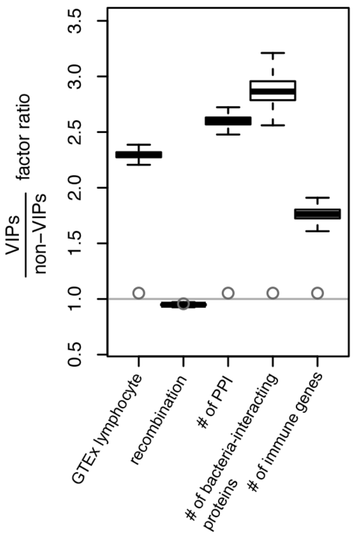


Figure 4. Confounding factors before and after matching
The five factors that need to be controlled for are GTEx expression in lymphocytes, recombination, the number of protein-protein interactions (PPI), the number of bacteria-interacting proteins, and the number of immune genes. The whiskers are the default R boxplot whiskers, and were estimated using 1,000 random sets of non-VIPs the same size as the set of VIPs (5,291). For each random non-VIP set, we measured the ratio of the average of a confounding factor in VIPs, over the average of the confounding factor in the random set of non-VIPs. The circles represent the same ratio, but for 1,000 non-VIP control sets generated by the bootstrap test. All the ratios are close to one, as expected if the bootstrap test matches the factors between VIPs and the control non-VIPs, and can thus be represented by just one point (gray circles).

Matching the five remaining factors (Fig. 4), the bootstrap test no longer fails, and finds a highly significant iHS sweep enrichment at VIPs when counting iHS candidate sweeps across all the 26 populations included in the 1,000 Genomes Project (Fig. 3A; see (Enard and Petrov, 2020) for more details). Although substantial, the enrichment found when matching confounding factors is not as strong as when not matching confounding factors (Fig. 3B), thus showing the importance of taking confounders into account to avoid spurious conclusions. We further show in figure 3C the enrichment when using control non-VIPs at least 100kb away from VIPs, instead of 500kb as in figure 3A. The enrichment is clearly much lower, which confirms that selecting controls far enough form the set of interest is critical. Note that we already previously found that the sweep enrichment at VIPs is not due to a confounding effect of the host biological functions enriched in VIPs, such as the cell cycle or DNA repair (Enard and Petrov, 2020). As described below, we can now estimate an unbiased false positive risk for the whole top 2,000 to top 10 enrichment curve shown in figure 3A, by using block-randomized genomes (see below). Using block-randomized genomes indeed enables the estimation of unbiased false positive risks (the other name for False Discovery Rate estimates but used not in the context of multiple testing) that fully account for the small sample size of the non-VIP controls that we used, and likely make the bootstrap test non-nominal.

Estimating false positive risks and accounting for clustering and the non-independence of biological functions all at once with block-randomized genomes.

Although the bootstrap test allows to control for multiple GSEA confounding factors simultaneously, it suffers from two main limitations that require additional steps in order to verify a sweep enrichment. First, the matching process can result in a limited number of control genes that can make the bootstrap test too liberal or too conservative, as already explained. Second, the sampling of control genes during the bootstrap test does not take clustering into account, since control genes are not selected to reproduce the clustering observed for the genes in the set of interest. In an extreme hypothetical case, we can imagine that all the genes in the set of interest are clustered together in one single locus, while the sampled control genes are not clustered but instead scattered between distant genome locations. Such a configuration could result in an overly liberal bootstrap test, if just one sweep happens to overlap the whole cluster of genes of interest.

Fortunately, randomized genomes represent a simple solution to all of these issues. More specifically, block-randomized genomes solve both the non-nominal nature and the clustering issues of the bootstrap test. Block-randomized genomes are genomes where the order of genes has been shuffled randomly. However, instead of simply shuffling genes individually, large blocks of contiguous genes are shuffled, with the actual order of genes being preserved within each block (Fig. 5). For example, human genes in the human genome can be split into 100 blocks (each with the same number of genes) based on their order on chromosomes, and these 100 blocks can then be randomly shuffled. Because the number of genes in each block is much larger than the expected number of genes even in the largest sweeps, the block-randomized genomes largely preserve the original clustering structure of genes in sweeps observed in the real genome. By using a large number of block randomized genomes, it is then possible to get a null distribution of the expected sweep enrichment given the same clustering structure, by counting the new numbers of genes in sweeps at the original locations of both the genes in the set of interest and the control genes (Fig. 6). The number of genes in the set of interest and the number of control genes are exactly the same as in the real genome, and thus the null expected enrichment distribution fully takes into account the effect of the size of the control set of genes on the variance of the enrichment observed with the bootstrap test. Contrary to the bootstrap test p-values alone, the block randomized genomes therefore allow to estimate an actual, unbiased false positive risk that takes both clustering and the size of the bootstrap test control sets into account.

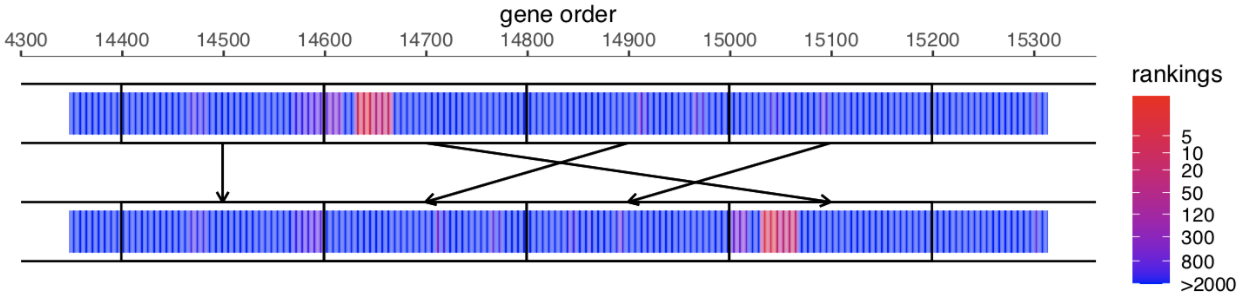


Figure 5. Random shuffling of genomic blocks
Legend as in figure 1. Genes are represented as vertical lines, and colored as in figure 1 according to the strength of the iHS signal. Genes are initially ordered (x-axis) as they are in the genome.

Importantly, the false positive risks estimated by the block-randomized genomes to unbias the bootstrap test are completely equivalent to False Discovery Rates (Colquhoun, 2014), which is the name that they are given in the specific context of multiple testing. Because they preserve the clustering and interdependency of gene attributes, block randomized genomes can thus be used to estimate false discovery rates when testing a large number of gene sets of interest, for example when running GSEA for a large number of Gene Ontology annotations. In this specific example, block-randomized genomes preserve the parent/daughter structure of the Gene Ontology, which is for instance not taken into account by the frequently used Fisher's Exact test.

Because block randomized genomes provide unbiased false positive risks, using the classic outlier approach to designate sweep candidates does not make much sense. Indeed, the false positive risk can be assessed for any rule to decide which genes are candidate sweeps, and which are not. We could for example decide to run a GSEA using the top 2,000, or top 1,000, or top 100 gene candidates, with no unjustified preconception about the expected number or strength of sweeps, and no preconception about whether or not they should fall outside of an often questionably defined neutral distribution. One can even use more complex statistics to estimate the false positive risk, such as the sum of enrichments measured over many sweep rank thresholds (for example, top 2,000 to top 100 with increments of 100), in order to detect a range of situations going from an enrichment in many weak or older sweep signals (top 2,000) to an enrichment in a few strong and recent sweeps signals (top 10).

In summary, block randomized genomes account for clustering, biases of the bootstrap test, and eliminate the need to rely on the outlier approach in the context of GSEA. Using block-randomized genomes to estimate false positive risks can however be computationally costly, as they require running the entire GSEA pipeline again and again. A pipeline to run both the bootstrap test and false positive risk estimations with block randomized genomes is available at https://github.com/DavidPierreEnard/Gene_Set_Enrichment_Pipeline.

Applying block randomized genomes to VIPs and non-VIPs

As shown in figure 3A, the bootstrap test detects a strong sweep enrichment at VIPs compared to control non-VIPs. The small pool of control VIPs compared to the number of VIPs however likely makes the bootstrap test non-nominal, meaning that the p-values reported by the bootstrap test (Fig. 3A) likely do not represent the real false positive risk (being either lower of higher, depending on the direction of the noise due to the small size of the controls sample). The bootstrap test also does not take clustering into account. To solve all these issues, we estimate an actual false positive risk for the enrichment curve presented in figure 3A by using block randomized genomes. The enrichment curve represents the relative enrichment at different top iHS sweep thresholds from top 2,000 to top 10. The relative enrichment is the ratio of the observed number of VIPs in sweep candidates (given a specific threshold), divided by the average expected number according to the matched controls from the bootstrap test. To measure the false positive risk, we use a slightly different metric, the sum of the difference (instead of the ratio) between the observed and the expected numbers of VIPs in candidate sweeps across all the top iHS thresholds. We use the difference between observed and expected instead of the ratio, because there is a distinction between a ratio of two reflecting two VIPs in sweeps instead of one, and a ratio reflecting 200 VIPs in sweeps instead of 100. The ratio is just easier for visualization. We compute the false positive risk by comparing this sum with the same sum calculated 5,000 times for 5,000 block randomized genomes. The estimated false positive risk for the VIP sweep enrichment curve is 0.0002, thus confirming a strong excess of sweeps at VIPs not due to the restricted number of non-VIP controls, nor due to the clustering of genes in the same sweeps.

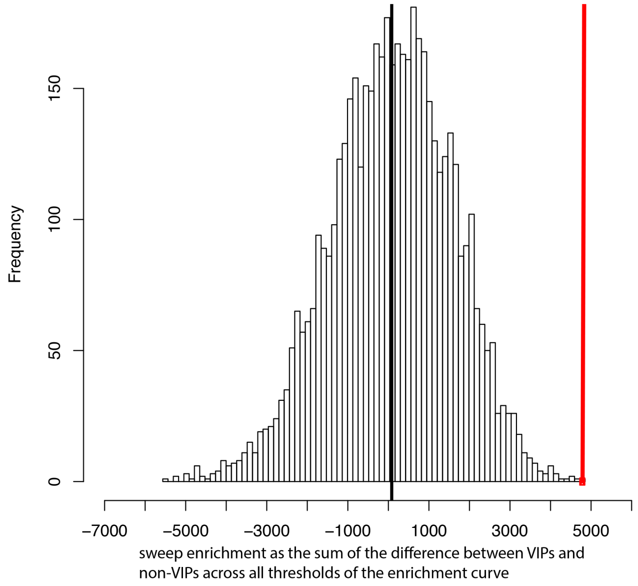


Figure 6. Estimation of the false positive risk.
Red vertical line: real sum of the difference between the number of sweeps at VIPs and control non-VIPs. The distribution shows the same sum but for 5,000 block-randomized genomes. Black vertical line: average over the 5,000 block-randomized genomes.

Discussion

Here, we described the many potential problems with naïve GSEA of positive selection, and we provide a number of potential solutions. In the future, more and more robust GSEA pipelines will hopefully be developed to gain better and better functional genomics insights based on sweep signals.

Since population geneticists started running GSEA of selection scans, GSEA have often been derided as that "one analysis you did and put at the end of a paper, just because you could and it was easy enough to run". This negative view of GSEA stems both from (i) the issues with the quality of early biological functional gene annotations that have since been greatly improved, and (ii) from a serious risk of excessive story-telling. It is noteworthy that the randomized genomes-based false positive risk analysis that we describe makes GSEA robust against story-telling, because unbiased false positive risks evaluate whether a biological function is enriched for selection just by chance out of thousands of biological functions tested (there always are outliers even in the absence of selection), or genuinely enriched for selection. The legitimate concerns around GSEA may have however devalued it enough that they took the focus away from improving GSEA in the context of population genetics.

These concerns must however not distract from the great potential of GSEA to help population geneticists better understand genomic evolution. For instance, it has become clear that the inference of demographic processes is biased by selection, and that the inference of selection in return is biased by demographic processes (Schrider et al., 2016). These confounding effects of different evolutionary processes have fueled a lengthy and sometimes frustrating debate on the role of natural selection versus neutral processes in molecular evolution. It can be argued that population geneticists have been "running in circles" fighting over arguments strictly rooted in population genetics, at a time when GSEA can provide critical answers. Indeed, a good example of GSEA helping the debate is the finding by Schrider & Kern that sweep candidate loci in the human genome are very strongly enriched in VIPs (Schrider and Kern, 2017). Finding sweeps where there is strong prior biological knowledge to expect them is very strong evidence of their reality and importance in the human genome. Whether or not Schrider and Kern's approach to detect sweeps is, for example, sensitive to demography or not then no longer really matters that much. Indeed, they may detect false positive sweeps, but not so many that there are not enough real sweeps among their candidates that a biologically meaningful enrichment cannot be identified. Note that although Schrider & Kern did not control for confounding factors, we have confirmed their result here and previously (Enard and Petrov, 2020). Such strong transversal biological evidence should be decisive in the selection versus neutrality debate (Kern and Hahn, 2018), but has far too often been ignored. Our hope is to have shown that GSEA has great potential, is far from being trivial, and worth more efforts for improvement at a time when ecological genomics are about to explode.

References

Al-Shahrour, F., Minguez, P., Tarraga, J., Montaner, D., Alloza, E., Vaquerizas, J.M., Conde, L., Blaschke, C., Vera, J., and Dopazo, J. (2006). BABELOMICS: a systems biology perspective in the functional annotation of genome-scale experiments. Nucleic Acids Res 34, W472-476.

Alonso, R., Salavert, F., Garcia-Garcia, F., Carbonell-Caballero, J., Bleda, M., Garcia-Alonso, L., Sanchis-Juan, A., Perez-Gil, D., Marin-Garcia, P., Sanchez, R., et al. (2015). Babelomics 5.0: functional interpretation for new generations of genomic data. Nucleic Acids Res 43, W117-121.

Barreiro, L.B., Laval, G., Quach, H., Patin, E., and Quintana-Murci, L. (2008). Natural selection has driven population differentiation in modern humans. Nat Genet 40, 340-345.

Bersaglieri, T., Sabeti, P.C., Patterson, N., Vanderploeg, T., Schaffner, S.F., Drake, J.A., Rhodes, M., Reich, D.E., and Hirschhorn, J.N. (2004). Genetic signatures of strong recent positive selection at the lactase gene. Am J Hum Genet 74, 1111-1120.

Bustamante, C.D., Fledel-Alon, A., Williamson, S., Nielsen, R., Hubisz, M.T., Glanowski, S., Tanenbaum, D.M., White, T.J., Sninsky, J.J., Hernandez, R.D., et al. (2005). Natural selection on protein-coding genes in the human genome. Nature 437, 1153-1157.

Carlson, C.S., Thomas, D.J., Eberle, M.A., Swanson, J.E., Livingston, R.J., Rieder, M.J., and Nickerson, D.A. (2005). Genomic regions exhibiting positive selection identified from dense genotype data. Genome Res 15, 1553-1565.

Castellano, D.U., L.H.; Munch, K.; Enard, D. (2019). Viruses rule over adaptation in conservd human proteins. bioRxiv.

Clark, A.G., Hubisz, M.J., Bustamante, C.D., Williamson, S.H., and Nielsen, R. (2005). Ascertainment bias in studies of human genome-wide polymorphism. Genome Res 15, 1496-1502.

Colquhoun, D. (2014). An investigation of the false discovery rate and the misinterpretation of p-values. R Soc Open Sci 1, 140216.

Consortium, G.T. (2013). The Genotype-Tissue Expression (GTEx) project. Nat Genet 45, 580-585.

Curtis, R.K., Oresic, M., and Vidal-Puig, A. (2005). Pathways to the analysis of microarray data. Trends Biotechnol 23, 429-435.

Daub, J.T., Dupanloup, I., Robinson-Rechavi, M., and Excoffier, L. (2015). Inference of Evolutionary Forces Acting on Human Biological Pathways. Genome Biol Evol 7, 1546-1558.

Daub, J.T., Moretti, S., Davydov, II, Excoffier, L., and Robinson-Rechavi, M. (2017). Detection of Pathways Affected by Positive Selection in Primate Lineages Ancestral to Humans. Mol Biol Evol 34, 1391-1402.

Enard, D., Cai, L., Gwennap, C., and Petrov, D.A. (2016). Viruses are a dominant driver of protein adaptation in mammals. Elife 5.

Enard, D., and Petrov, D.A. (2018). Evidence that RNA Viruses Drove Adaptive Introgression between Neanderthals and Modern Humans. Cell 175, 360-371 e313.

Enard, D., and Petrov, D.A. (2020). Ancient RNA virus epidemics through the lens of recent adaptation in human genomes. bioRxiv.

Gene Ontology, C. (2015). Gene Ontology Consortium: going forward. Nucleic Acids Res 43, D1049-1056.

Genomes Project, C., Auton, A., Brooks, L.D., Durbin, R.M., Garrison, E.P., Kang, H.M., Korbel, J.O., Marchini, J.L., McCarthy, S., McVean, G.A., et al. (2015). A global reference for human genetic variation. Nature 526, 68-74.

Gouy, A., Daub, J.T., and Excoffier, L. (2017). Detecting gene subnetworks under selection in biological pathways. Nucleic Acids Res 45, e149.

Halldorsson, B.V., Palsson, G., Stefansson, O.A., Jonsson, H., Hardarson, M.T., Eggertsson, H.P., Gunnarsson, B., Oddsson, A., Halldorsson, G.H., Zink, F., et al. (2019). Characterizing mutagenic effects of recombination through a sequence-level genetic map. Science 363.

Hukku, A., Quick, C., Luca, F., Pique-Regi, R., and Wen, X. (2020). BAGSE: a Bayesian hierarchical model approach for gene set enrichment analysis. Bioinformatics 36, 1689-1695.

Hung, J.H., Yang, T.H., Hu, Z., Weng, Z., and DeLisi, C. (2012). Gene set enrichment analysis: performance evaluation and usage guidelines. Brief Bioinform 13, 281-291.

International HapMap, C. (2003). The International HapMap Project. Nature 426, 789-796.

Kelley, J.L., Madeoy, J., Calhoun, J.C., Swanson, W., and Akey, J.M. (2006). Genomic signatures of positive selection in humans and the limits of outlier approaches. Genome Res 16, 980-989.

Kern, A.D., and Hahn, M.W. (2018). The Neutral Theory in Light of Natural Selection. Mol Biol Evol 35, 1366-1371.

Kim, S.Y., and Volsky, D.J. (2005). PAGE: parametric analysis of gene set enrichment. BMC Bioinformatics 6, 144.

Luisi, P., Alvarez-Ponce, D., Pybus, M., Fares, M.A., Bertranpetit, J., and Laayouni, H. (2015). Recent positive selection has acted on genes encoding proteins with more interactions within the whole human interactome. Genome Biol Evol 7, 1141-1154.

Mootha, V.K., Lindgren, C.M., Eriksson, K.F., Subramanian, A., Sihag, S., Lehar, J., Puigserver, P., Carlsson, E., Ridderstrale, M., Laurila, E., et al. (2003). PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. Nat Genet 34, 267-273.
O'Reilly, P.F., Birney, E., and Balding, D.J. (2008). Confounding between recombination and selection, and the Ped/Pop method for detecting selection. Genome Res 18, 1304-1313.

Orchard, S., Ammari, M., Aranda, B., Breuza, L., Briganti, L., Broackes-Carter, F., Campbell, N.H., Chavali, G., Chen, C., del-Toro, N., et al. (2014). The MIntAct project--IntAct as a common curation platform for 11 molecular interaction databases. Nucleic Acids Res 42, D358-363.

Pavlidis, P., Jensen, J.D., Stephan, W., and Stamatakis, A. (2012). A critical assessment of storytelling: gene ontology categories and the importance of validating genomic scans. Mol Biol Evol 29, 3237-3248.
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.

Sabeti, P.C., Varilly, P., Fry, B., Lohmueller, J., Hostetter, E., Cotsapas, C., Xie, X., Byrne, E.H., McCarroll, S.A., Gaudet, R., et al. (2007). Genome-wide detection and characterization of positive selection in human populations. Nature 449, 913-918.

Schrider, D.R., and Kern, A.D. (2017). Soft Sweeps Are the Dominant Mode of Adaptation in the Human Genome. Mol Biol Evol 34, 1863-1877.

Schrider, D.R., Shanku, A.G., and Kern, A.D. (2016). Effects of Linked Selective Sweeps on Demographic Inference and Model Selection. Genetics 204, 1207-1223.

Skunca, N., Altenhoff, A., and Dessimoz, C. (2012). Quality of computationally inferred gene ontology annotations. PLoS Comput Biol 8, e1002533.

Sugden, L.A., Atkinson, E.G., Fischer, A.P., Rong, S., Henn, B.M., and Ramachandran, S. (2018). Localization of adaptive variants in human genomes using averaged one-dependence estimation. Nat Commun 9, 703.

Tang, K., Thornton, K.R., and Stoneking, M. (2007). A new approach for using genome scans to detect recent positive selection in the human genome. PLoS Biol 5, e171.

Teshima, K.M., Coop, G., and Przeworski, M. (2006). How reliable are empirical genomic scans for selective sweeps? Genome Res 16, 702-712.

The Gene Ontology, C. (2019). The Gene Ontology Resource: 20 years and still GOing strong. Nucleic Acids Res 47, D330-D338.

Thornton, K.R., and Jensen, J.D. (2007). Controlling the false-positive rate in multilocus genome scans for selection. Genetics 175, 737-750.

Tishkoff, S.A., Reed, F.A., Ranciaro, A., Voight, B.F., Babbitt, C.C., Silverman, J.S., Powell, K., Mortensen, H.M., Hirbo, J.B., Osman, M., et al. (2007). Convergent adaptation of human lactase persistence in Africa and Europe. Nat Genet 39, 31-40.

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.

Williamson, S.H., Hubisz, M.J., Clark, A.G., Payseur, B.A., Bustamante, C.D., and Nielsen, R. (2007). Localizing recent adaptive evolution in the human genome. PLoS Genet 3, e90.

Appendix B -- Decreased recent adaptation at human Mendelian disease genes as a possible consequence of interference between advantageous and deleterious variants

eLife RESEARCH ARTICLE

Decreased recent adaptation at human mendelian disease genes as a possible consequence of interference between advantageous and deleterious variants

Chenlu Di[1], Jesus Murga Moreno[2], Diego F Salazar-Tortosa[1], M Elise Lauterbur[1], David Enard[1]*

[1]University of Arizona Department of Ecology and Evolutionary Biology, Tucson, United States; [2]Institut de Biotecnologia i de Biomedicina and Departament de Genètica i de Microbiologia, Universitat Autònoma de Barcelona, Barcelona, Spain

Abstract Advances in genome sequencing have improved our understanding of the genetic basis of human diseases, and thousands of human genes have been associated with different diseases. Recent genomic adaptation at disease genes has not been well characterized. Here, we compare the rate of strong recent adaptation in the form of selective sweeps between mendelian, non-infectious disease genes and non-disease genes across distinct human populations from the 1000 Genomes Project. We find that mendelian disease genes have experienced far less selective sweeps compared to non-disease genes especially in Africa. Investigating further the possible causes of the sweep deficit at disease genes, we find that this deficit is very strong at disease genes with both low recombination rates and with high numbers of associated disease variants, but is almost non-existent at disease genes with higher recombination rates or lower numbers of associated disease variants. Because segregating recessive deleterious variants have the ability to interfere with adaptive ones, these observations strongly suggest that adaptation has been slowed down by the presence of interfering recessive deleterious variants at disease genes. These results suggest that disease genes suffer from a transient inability to adapt as fast as the rest of the genome.

Introduction

Advances in genome sequencing have dramatically improved our understanding of the genetic basis of human diseases, and thousands of human genes have been associated with different diseases (***Amberger et al., 2019***; ***Piñero et al., 2020***). Despite our expanding knowledge of mendelian disease gene associations, and despite the fact that multiple evolutionary processes might connect disease and genomic adaptation at the gene level, these connections are yet to be studied more thoroughly, especially in the case of recent genomic adaptation. Different evolutionary processes have the potential to make the occurrence of mendelian disease genes and adaptation not independent from each other in the human genome. For instance, hitchhiking of deleterious mutations linked to advantageous mutations might increase the risk of mendelian disease-causing variants at genes subjected to past directional adaptation. Disease genes might then appear to have experienced more adaptation than non-disease genes if this specific process was sufficiently widespread. Conversely, higher evolutionary constraint, and higher pleiotropy might reduce adaptation at disease genes compared to genes not involved in diseases (***Otto, 2004***). There is currently considerable uncertainty about how any of these non-exclusive evolutionary processes, or other processes, might have influenced adaptation at disease genes. It is even not well-known whether human non-

***For correspondence:** denard@email.arizona.edu

Competing interests: The authors declare that no competing interests exist.

Funding: See page 24

Preprinted: 31 March 2021
Received: 01 April 2021
Accepted: 02 October 2021
Published: 12 October 2021

Reviewing editor: Luis Barreiro, University of Chicago, United States

infectious disease genes have similar, higher or lower levels of adaptation in human populations compared to genes not involved in diseases. Comparing levels of adaptation between disease genes and non-disease genes is a first important step toward better understanding the evolutionary relationship between non-infectious diseases and genomic adaptation.

Multiple recent studies comparing evolutionary patterns between human mendelian disease and non-disease genes have found that mendelian disease genes are more constrained and evolve more slowly (***Blekhman et al., 2008***; ***Quintana-Murci, 2016***; ***Spataro et al., 2017***; ***Torgerson et al., 2009***). An older comparison by ***Smith and Eyre-Walker, 2003*** found that disease genes evolve faster than non-disease genes, but we note that the sample of disease genes used at the time was very limited.

The significant increase of the number of known disease genes since these studies were completed makes it important to update the comparison of evolutionary patterns at disease and non-disease genes. More critically however, past studies all have in common an important limitation that justifies comparing disease genes and non-disease genes again. Disease and non-disease genes may differ by more than just the fact that they have been associated with disease or not. Disease and non-disease genes may also differ in many other factors other than their disease status. Such factors can be a problem when comparing adaptation in disease genes and non-disease genes, because they, instead of the disease status itself, could explain differences in adaptation. For example,

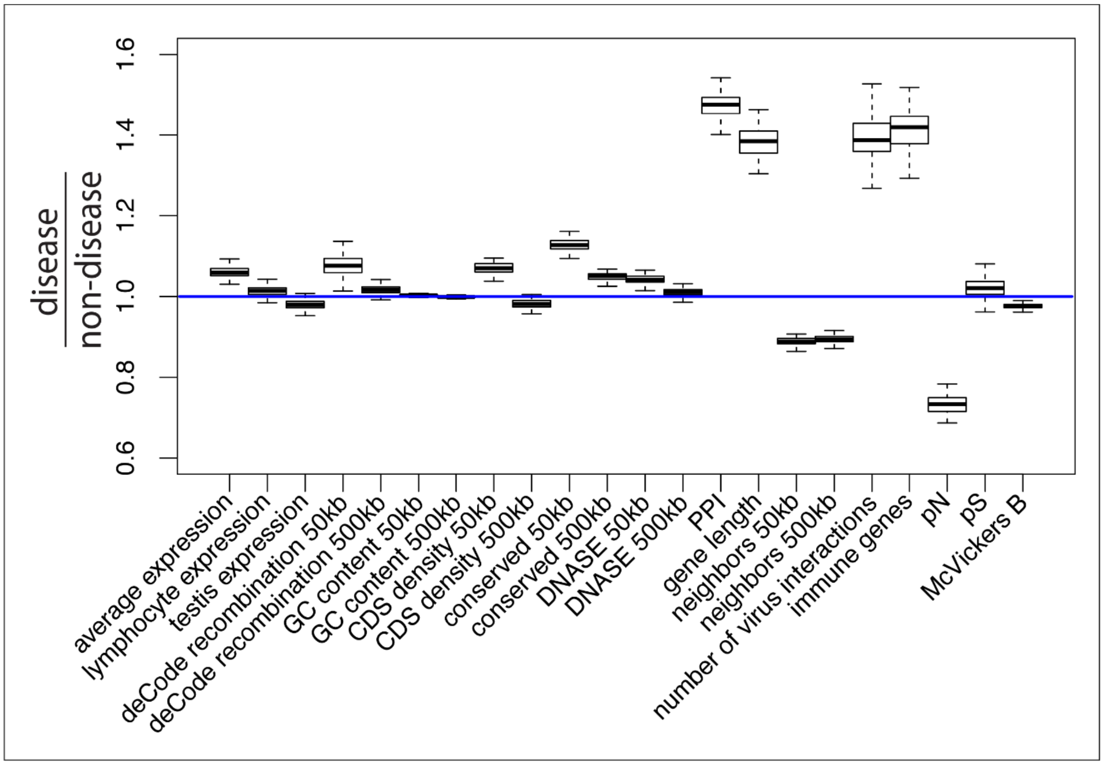


Figure 1. Potential confounding factors in disease versus non-disease genes. Each potential confounding factor is detailed in the Materials and methods. For each confounding factor, the boxplot shows on the y-axis the ratio of the average factor value for disease genes, divided by the average factor value for non-disease genes. The boxplot error bars are obtained by calculating the ratio 1000 times, each time by randomly sampling as many non-disease genes as there are disease genes.

disease genes tend to be more highly expressed than non-disease genes (**Spataro et al., 2017**; **Figure 1**). If higher expression happens to be associated with more adaptation in general, one might detect more adaptation in disease genes in a way that has nothing to do with disease, and just reflects their higher levels of expression. Many other factors may also be important. For example, immune genes, which often adapt in response to infectious pathogens, may further complicate comparisons if they are represented in unequal proportions between non-infectious disease and non-disease genes. Comparing genomic adaptation in disease and non-disease genes thus requires careful consideration of confounding factors.

Among possible confounding factors, it is particularly important to take into account evolutionary constraint, that is the level of purifying selection experienced by different genes. A common intuition is that mendelian disease genes may exhibit less adaptation because they are more constrained (**Blekhman et al., 2008**; **Spataro et al., 2017**; **Torgerson et al., 2009**), leaving less mutational space for adaptation to happen in the first place. Less adaptation at mendelian disease genes might thus represent a trivial consequence of varying constraint between genes (**Kim et al., 2007**), which says little about a specific connection between disease and adaptation. In the same vein, one might expect disease genes to be associated with higher mutation rates, and more frequent adaptation to follow as a trivial consequence of elevated mutation rates. Whether disease genes experience higher mutation rates is however still an open question (**Eyre-Walker and Eyre-Walker, 2014**; **Osada et al., 2009**). In any case, focusing specifically on disease and adaptation requires controlling for confounders such as constraint/purifying selection and mutation rate (see Materials and methods, Results and **Figure 1** for a complete list of confounders accounted for in this analysis).

A specific evolutionary relationship may exist between adaptation and disease beyond the simple effect of constraint, mutation rate or other confounders. In an evolutionary context, once constraint and other confounding factors have been accounted for, we can imagine three potential scenarios for the comparison of adaptation between disease and non-disease genes. Under scenario 1, any potential difference in adaptation between disease and non-disease genes is entirely due to differences in constraint and other confounding factors. Under this scenario, there is no further evolutionary process linking disease and adaptation together. Therefore, there is no difference in adaptation between disease and non-disease genes once confounding factors have been accounted for.

Under scenario 2, always once selective constraint and other confounding factors have been accounted for, disease genes have more adaptation than non-disease genes. For example, as already mentioned above, deleterious mutations can hitchhike together with adaptive mutations to high frequencies in human populations (**Barreiro and Quintana-Murci, 2010**; **Birky and Walsh, 1988**; **Chun and Fay, 2011**; **Quintana-Murci and Barreiro, 2010**). Other, less well established, cases can be imagined where past adaptation decreased the robustness of a specific gene, and subsequent mutations become more likely to be associated with diseases (**Xu and Zhang, 2014**). Scenario 2 thus favors a relationship between adaptation and disease, where past adaptation precedes and influences the likelihood of a gene being associated with disease.

Under scenario 3, disease genes have less adaptation than non-disease genes even after accounting for confounding factors such as evolutionary constraint. Such a scenario might occur for example if disease genes happen to be genes that can be sensitive to changes in the environment, with a fitness optimum that can change over time, but where adaptation has not occurred yet to catch up with the new optimum. Such an adaptation lag (or lag load, to reuse the terminology introduced by **Maynard Smith, 1976**) may occur for example if higher pleiotropy at disease genes (**Ittisoponpisan et al., 2017**) makes it less likely for new mutations to be advantageous (**Otto, 2004**) (in addition to increasing the level of constraint already accounted for as a confounding factor). An adaptation lag may also occur if deleterious mutations interfere with and slow down adaptation at disease genes more than at non-disease genes (**Assaf et al., 2015**; **Hill and Robertson, 1966**). Alternatively, disease genes could have constitutively less adaption, because of pleiotropy or because new mutations tend to be large effect mutations that often overshoot the fitness optimum, which would prevent them from being advantageous. In the latter scenario of a constitutive deficit of adaptation, disease genes should have not only a deficit of recent adaptation, but also a deficit of older, long-term adaptation, that can be estimated with approaches such as the McDonald-Kreitman test (**McDonald and Kreitman, 1991**).

Even though uncovering the underlying evolutionary processes that govern the relationship between disease and adaptation will take a lot more work, it is important to find first which scenario is the most likely to be true, that is whether disease genes have as much, more, or less adaptation than non-disease genes. Finding out which out of the three possible scenarios is true may give a preliminary basis to further hypothesize which evolutionary processes are more likely to dominate the relationship between disease and adaptation genome-wide.

Here, we compare recent adaptation in mendelian disease and non-disease genes in order to disentangle the connections between adaptation and disease. We specifically compare the abundance of recent selective sweeps signals, where hitchhiking has raised haplotypes that carry an advantageous variant to higher frequencies (***Smith and Haigh, 1974***). Note that this means that we can only compare adaptation at specific loci between disease and non-disease genes that was strong enough to induce hitchhiking, hence we do not take into account polygenic adaptation distributed across a large number of loci that did not leave any hitchhiking signals (see Discussion). As mentioned above, confounding factors may affect the comparison between disease and non-disease genes. In contrast with previous studies, we systematically control for a large number of confounding factors when comparing recent adaptation in human mendelian disease and non-disease genes, including evolutionary constraint, mutation rate, recombination rate, the proportion of immune or virus-interacting genes, etc. (please refer to Materials and methods for a full list of the confounding factors included). In addition to controlling for a large number of confounding factors, we estimate false positive risks (FPR) (***Colquhoun, 2019***) for our comparison pipeline that fully take into account the implications of controlling for many confounding factors (see Materials and methods and Results).

As a list of mendelian disease genes to test, we curate human mendelian non-infectious disease genes based on annotations in the DisgeNet (***Piñero et al., 2020***) and OMIM (***Amberger et al., 2019***) databases (Materials and methods). We focus on non-infectious mendelian disease genes rather than all disease genes including complex disease associations, because different evolutionary patterns can be expected between mendelian and complex disease genes based on previous studies (***Blekhman et al., 2008***; ***Quintana-Murci, 2016***; ***Spataro et al., 2017***; ***Torgerson et al., 2009***). In total, we compare 4215 mendelian disease genes with non-disease genes in the human genome. In agreement with scenario 3, we find a strong deficit of selective sweeps at disease genes compared to non-disease genes in Africa, and weaker deficits in East Asia and Europe. We further test multiple potential explanations for this deficit pattern across human populations, and find that it strongly depends on recombination and the number of known disease variants at given mendelian disease genes. This suggests that segregating deleterious mutations at disease genes might interfere with, and slow down genetically linked adaptive variants enough to produce the observed lack of sweeps at disease genes. We further support this possible explanation with forward population simulations (***Haller and Messer, 2019***) that show that stronger interference in Africa compared to East Asia or Europe is expected. We also show that alternative explanations implying an evolutionarily stable, constitutive deficit of advantageous mutations, rather than transient interference, are made unlikely by the fact that although disease genes have experienced less recent adaptation in the form of sweeps, they have not experienced less long-term adaptation during the millions of years of human evolution.

Results

Controlling for confounding factors with a bootstrap test

To compare mendelian disease and non-disease genes, we first ask which potential confounding factors differ between the two groups of genes. As expected, multiple measures of selective constraint are significantly higher in mendelian disease compared to non-disease genes. As a measure of long-term constraint, the density of conserved elements across mammals is slightly higher at disease genes compared to non-disease genes (***Figure 1***: conserved 50 kb, conserved 500 kb; Materials and methods).

As a measure of more recent constraint, we contrast pS, the average proportion of variable synonymous sites, with pN, the average proportion of variable nonsynonymous sites (***Figure 1***; Materials and methods). If the coding sequences of disease genes are more constrained, we expect a drop of pN at disease genes, but no such drop of pS at neutral synonymous sites. Accordingly, pN is lower at disease compared to non-disease genes, while pS is very similar between the two categories of genes (*Figure 1*). Therefore, selective constraint was stronger in the coding sequences of disease genes during recent human evolution.

As another measure of recent constraint, we also use McVicker's B estimator of background selection (*McVicker et al., 2009*). The amount of background selection at a locus can be used as a proxy for recent constraint, since it depends on the number of deleterious mutations that were recently removed at this locus. The lower B, the more background selection there is at a specific locus. In line with higher recent constraint at disease genes, B is slightly, but significantly lower at disease genes (*Figure 1*; Materials and methods). Overall, we find evidence of higher constraint at disease genes.

These differences between disease and non-disease genes highlight the need to compare disease genes with control non-disease genes with similar levels of selective constraint. To do this and compare sweeps in mendelian disease genes and non-disease genes that are similar in ways other than being associated with mendelian disease (as described in the Results below, Less sweeps at mendelian disease genes in Africa), we use sets of control non-disease genes that are built by a bootstrap test to match the disease genes in terms of confounding factors (Materials and methods), including the confounding factors that represent measures of selective constraint/purifying selection (density of conserved elements, pN and pS, and McVicker's B; see Materials and methods). To verify that the measures of selective constraint included indeed control for purifying selection when comparing disease and matched control non-disease genes, we ran a maximum likelihood version of the McDonald-Kreitman test called GRAPES (*Galtier, 2016*), to compare the average selection coefficient of deleterious mutations at disease gene coding sequences, compared to the control non-disease gene coding sequences (Materials and methods). We find that disease genes and their non-disease control genes have undistinguishable average strengths of deleterious variants, suggesting that our controls for selective constraint are sufficient, at least to account for constraint at the coding sequence level (*Figure 2*; comparison test p=0.37). Further down in the Results (Verification of purifying selection controls), we also show that we properly control for purifying selection not just in coding sequences but in the whole genomic regions that we analyze by using GERP (*Davydov et al., 2010*).

In addition to constraint, mutation rate could represent an important confounder. The proportion of variable neutral synonymous sites pS can be used to compare mutation rates, since the number of variable synonymous sites is proportional to the mutation rate under neutrality. As mentioned

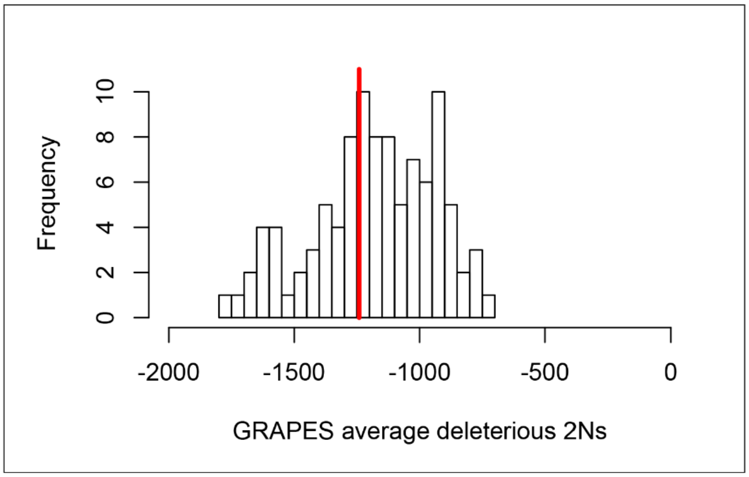


Figure 2. Average strength of deleterious nonsynonymous variants in disease vs control genes. The average strength of deleterious nonsynonymous variants was measured using GRAPES with the DisplGamma distribution of fitness effects, which gave the best fit to disease and control sets. The histogram represents 100 control sets. The red line represents the average strength of deleterious nonsynonymous variants in mendelian disease genes (2Ns=-1241).

already, pS is very similar at disease and non-disease genes (***Figure 1***), suggesting that mutation rates are similar at disease and non-disease genes. This is further supported by the fact that multiple factors that could affect the mutation rate such as GC content or recombination are also similar or very slightly different at disease and non-disease genes (***Figure 1***; Materials and methods). Aside from mutation rate and constraint, multiple other factors that could affect adaptation differ between disease and non-disease genes, notably including the proportion of genes that interact with viruses, the proportion of immune genes, or the number of protein-protein interactions (PPIs) in the human PPIs network. All these factors have been shown to, or could in principle affect adaptation (Materials and methods), further showing the necessity to control for confounding factors when comparing adaptation at disease and non-disease genes. The fact that previous studies comparing adaptation at disease versus non-disease genes did not control for confounding factors, makes it unclear if their conclusions reflect properties tied to genes being associated with disease or not, or tied to other confounding factors not accounted for.

Less sweeps at mendelian disease genes in Africa

For our comparison of disease and non-disease genes, we measure recent adaptation around human protein coding genes (Materials and methods) using the integrated Haplotype Score iHS (***Voight et al., 2006***) and the number of Segregating sites by Length nSL (***Ferrer-Admetlla et al., 2014***) in 26 populations rom the 1,000 Genomes Project (***Auton et al., 2015***) (Materials and methods). The iHS and nS_L statistics are both sensitive to recent incomplete sweeps, and have the advantage over other sweep statistics of being insensitive to the confounding effect of background selection (***Enard et al., 2014***; ***Schrider, 2020***). To evaluate the prevalence of sweeps at disease genes relative to non-disease genes, we do not use the classic outlier approach, and instead use a previously described, more versatile approach based on block-randomized genomes to estimate

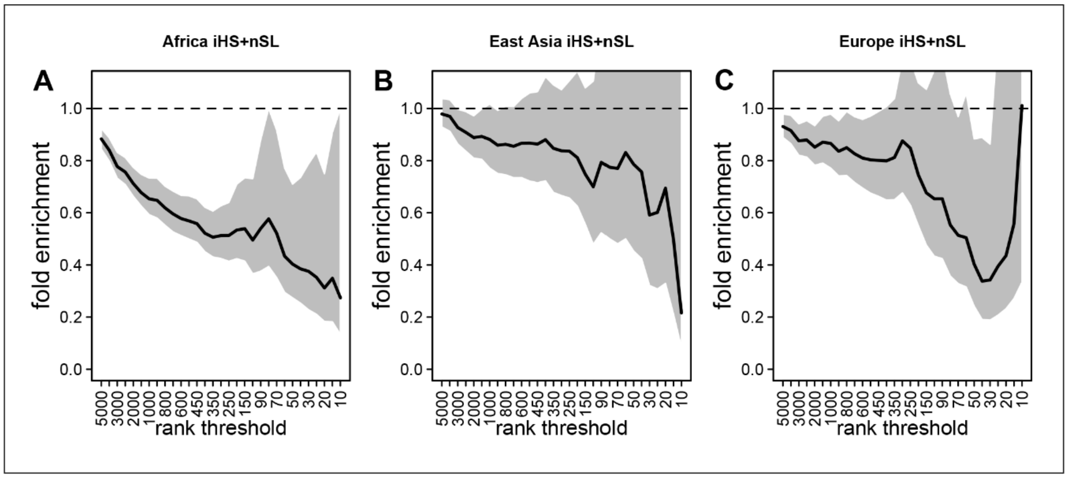


Figure 3. Deficit of iHS and nS_L sweep signals at mendelian disease genes. The figure shows the averaged whole enrichment curves and their averaged confidence intervals from the bootstrap test, averaged over both iHS and nS_L sweep ranks, and over all the populations from each continent (Materials and methods). The y-axis represents the relative sweep enrichment at disease genes, calculated as the number of disease genes in putative sweeps, divided by the number of control non-disease genes in putative sweeps. The gray areas are the 95% confidence interval for this ratio. The number of genes in putative sweeps is measured for varying sweep rank thresholds. For example, at the top 100 rank threshold, the relative enrichment is the number of disease genes within the top 100 genes with the strongest sweep signals (either according to iHS or nS_L), divided by the number of control non-disease genes within the top 100 genes with the strongest sweep signals. We use genes ranked by iHS or nS_L using 200 kb windows, since 200 kb is the intermediate size of all the window sizes we use (50 kb, for the smallest, 1000 kb for the largest; see Materials and methods). (**A**) Africa, average over the ESN, GWD, LWK, MSL, and YRI populations from the 1000 Genomes Project. (**B**) East Asia, average over the CDX, CHB, CHS, JPT, and KHV populations. (**C**) Europe, average over the CEU, FIN, GBR, IBS, and TSI populations.

unbiased false positive risks (FPR) for whole enrichment curves (***Figure 3***; ***Enard and Petrov, 2020***). We first rank genes based on the average iHS or nS_L in genomic windows centered on genes (Materials and methods), from the top-ranking genes with the strongest sweep signals to the genes with the weakest signals. We then slide a rank threshold from a high rank value to a low rank value (from top 5000 to top 10, x-axis on ***Figure 3***). For each rank threshold, we estimate the sweep enrichment (or deficit) at disease relative to non-disease genes (***Figure 3***, y-axis). For example, for rank threshold 200, the relative enrichment (or deficit) is the number of disease genes in the top 200 ranking genes, divided by the number of control non-disease genes in the top 200. By sliding the rank threshold, we estimate a whole enrichment curve that is not only sensitive to the strongest sweeps but also to weaker sweeps signals (for example using the top 5000 threshold; ***Figure 3***). Using block-randomized genomes (Materials and methods), we can then estimate an unbiased false positive risk (FPR) for the whole enrichment curve. In brief, we estimate FPRs by re-running the entire enrichment analysis pipeline many times, but each time on a block-randomized genome instead of the real genome. Block randomized genomes are genomes where gene coordinates have been randomly shuffled in a way that preserves the original genes/sweeps clustering structure. The FPR can then be calculated by comparing the whole sweep deficit/enrichment curve in the real genome with the many observed whole sweep deficit/enrichment curves observed with the block-randomized genomes (see Materials and methods for more details). This strategy makes less assumptions on the expected strength of selective sweeps. The approach also makes it possible to estimate a single false positive risk based on the cumulated enrichment (or deficit) over multiple whole enrichment curves (Materials and methods). Here, we estimate a single false positive risk for both iHS and nS_L curves considered together, and also for multiple window sizes to measure average iHS and nS_L (from 50kb to 1Mb, Materials and methods).

To control for confounding factors (***Figure 1***), we compare sweep signals at disease genes with control sets of non-disease genes that were chosen by a bootstrap test (***Enard and Petrov, 2020***) because they match disease genes in terms of confounding factor values (Materials and methods). Furthermore, control non-disease genes are chosen far from disease genes (>300 kb; Materials and methods). We do this to avoid choosing as controls non-disease genes that are too close to disease genes and thus likely to have the same sweep profile (especially in the case of large sweeps potentially overlapping both neighboring disease and non-disease genes). This, together with the large number of confounding factors that we match, tends to limit the pool of possible control genes (Materials and methods). The statistical impact of a limited control pool is however fully taken into account by the estimation of a FPR with block-randomized genomes (Materials and methods).

Because they have experienced different demographic histories, we test different human populations from distinct continents separately. Specifically, we test African populations, East Asian populations and European populations from the 1000 Genomes Project phase 3 (***Auton et al., 2015***). At this stage, we must consider the fact that most gene-disease associations in our dataset were likely discovered in European cohorts. Because disease genes in Europe may not always be disease genes in other populations, we cannot exclude the possibility that a sweep enrichment or a sweep deficit might be more pronounced in Europe, unless the evolutionary processes that make a gene more likely to be a disease gene predated the split of different human populations.

Using both iHS and nS_L sweep signals, we find a strong depletion in sweep signals at disease genes, especially in Africa (***Figure 3A***) compared to East Asia or Europe (***Figure 3B, C***, respectively). ***Figure 3A, B, C*** show the sweep deficit curves at disease genes compared to control non-disease genes in Africa, East Asia, and Europe, respectively. The corresponding false positive risks that quantify how unexpected the downward or upward skew of these curves are (Materials and methods), show that the sweep deficit is strongly significant in Africa, marginally so in Europe, and not significant at all in East Asia (FPR=3.10 ^{4} in Africa vs. 0.18 in East Asia and 0.05 in Europe, ***Figure 4A, B C*** respectively; Materials and methods). Note that this FPR takes the clustering of multiple genes in the same sweeps into account (***Enard and Petrov, 2020***). A stronger depletion in Africa suggests that the evolutionary processes linking disease and adaptation at the gene level predate the split of African and European populations, given that most gene-disease associations studies involved European cohorts. As we show below, the stronger sweep depletion in Africa can be explained in the evolutionary context of genetic interference between advantageous and deleterious variants at mendelian disease genes.

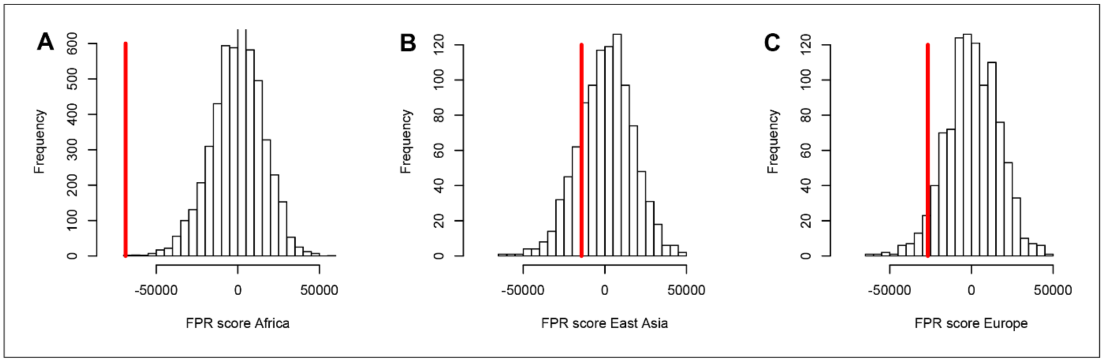


Figure 4. A stronger sweep deficit at disease genes in Africa than in East Asia and Europe. The figure shows the observed sweep enrichment/deficit score used to measure the false positive risk (FPR) in the real genome (red line), compared to the expected null distribution of the score estimated with block-randomized genomes (5000 block-randomized genomes in Africa, 1000 in East Asia and Europe; Materials and methods). The FPR score is based on summing the difference between the number of genes in sweeps at disease genes and the number of genes in sweeps in control genes, over both iHS and nS_L, and different window sizes (Materials and methods). (**A**) FPR score in Africa, estimated summing over the ESN, GWD, LWK, MSL, and YRI populations from the 1000 Genomes Project. (**B**) FPR score in East Asia, estimated summing over the CDX, CHB, CHS, JPT, and KHV populations. (**C**) FPR score in Europe, summing over the CEU, FIN, GBR, IBS, and TSI populations.
The online version of this article includes the following figure supplement(s) for figure 4:

Figure supplement 1. FPR with or without controlling for GERP.

Notably, the stronger depletion observed in Africa likely excludes the possibility that it could be mostly due to a technical artifact, where sweeps themselves might make it harder to identify disease genes in the first place. Sweeps increase linkage disequilibrium (LD) in a way that could make it more difficult to assign a disease to a single gene in regions of the genome with high LD and multiple genes genetically linked to a disease variant. This could result in a depletion of sweeps at monogenic disease genes, simply because disease genes are less well annotated in regions of high LD. However, if this was the case, because most disease gene were identified in Europe, we would expect such an artifact to deplete sweeps at disease genes primarily in Europe, not in Africa. This artifact is also very unlikely due to the fact that recombination rates are only very slightly different between disease and non-disease genes (***Figure 3***). Overall, these results support the third scenario where evolutionary processes decrease recent adaptation at mendelian disease genes. That said, it is important to note that we only detect a deficit of recent adaptation strong enough to leave hitchhiking signals. Our results do not imply that the same is true for adaptation that is too polygenic to leave signals detectable with iHS or nS_L. Note that the sweep deficit at disease genes in Africa is robust to differences in gene functions between disease and non-disease genes according to a Gene Ontology analysis (Materials and methods) (***Gene Ontology Consortium and Gene Ontology, 2021***).

Verification of purifying selection controls

To further verify that constraint/purifying selection is properly controlled for when comparing mendelian disease and control non-disease genes, we also add the GERP score, as well as the density of both coding and non-coding conserved elements identified by GERP (***Davydov et al., 2010***) to the list of matched confounding factors (Materials and methods). The average GERP score in a genomic window estimates the amount of substitutions that never happened during long-term evolution because the said mutations were removed by purifying selection (both in coding and non-coding sequences). The sweep deficit in Africa at disease genes compared to controls is completely unchanged when using GERP or not (***Figure 4—figure supplement 1***). This shows that the measures of selective constraint already included (Materials and methods) are sufficient to control for selective constraint/purifying selection. For this reason, we do not use GERP further (as explained in the

Materials and methods, the larger the number of confounding factors that we match, the lower the power of our approach to detect a sweep enrichment or deficit).

Disease genes do not experience constitutively less long-term adaptive mutations

A deficit of strong recent adaptation (strong enough to affect iHS or nS_L) raises the question of what creates the sweep deficit at disease genes. As already discussed, purifying selection and other confounding factors are matched between disease genes and their controls, which excludes that these factors alone could possibly explain the sweep deficit. Purifying selection alone in particular cannot explain this result, since we find evidence that it is well matched between disease and control genes (***Figure 4*** and ***Figure 4—figure supplement 1***). Furthermore, we find that the 1000 genes in the genome with the highest density of conserved elements do not exhibit any sweep deficit (bootstrap test + block-randomized genomes FPR=0.18; Materials and methods). Association with mendelian diseases, rather than a generally elevated level of selective constraint, is therefore what matters to observe a sweep deficit. What then might explain the sweep deficit at disease genes?

As mentioned in the introduction, it could be that mendelian disease genes experience constitutively less adaptive mutations. This could be the case for example because mendelian disease genes tend to be more pleiotropic (***Otto, 2004***), and/or because new mutations in mendelian are large effect mutations (***Quintana-Murci, 2016***) that tend to often overshoot the fitness optimum, and cannot be positively selected as a result. Regardless of the underlying processes, a constitutive tendency to experience less adaptive mutations predicts not only a deficit of recent adaptation, but also a deficit of more long-term adaptation during evolution. The iHS and nSL signals of recent adaptation we use to detect sweeps correspond to a time window of at most 50,000 years, since these statistics have very little statistical power to detect older adaptation (***Sabeti et al., 2006***). In contrast, approaches such as the McDonald-Kreitman test (MK test) (***McDonald and Kreitman, 1991***) capture the cumulative signals of adaptative events since humans and chimpanzee had a common ancestor, likely more than 6 million years ago.

To test whether mendelian disease genes have also experienced less long-term adaptation, in addition to less recent adaptation, we use the MK tests ABC-MK (***Uricchio et al., 2019***) and GRAPES (***Galtier, 2016***) to compare the rate of protein adaptation (advantageous amino acid changes) in mendelian disease gene coding sequences, compared to confounding factors-matched non-disease controls (Materials and methods). We find that overall, disease and control non-disease genes have experienced similar rates of protein adaptation during millions of years of human evolution, as shown by very similar estimated proportions of amino acid changes that were adaptive (***Figure 5A, B,C,D,E***). This result suggests that disease genes do not have constitutively less adaptive mutations. This implies that processes that are stable over evolutionary time such as pleiotropy, or a tendency to overshoot the fitness optimum, are unlikely to explain the sweep deficit at disease genes. If disease genes have not experienced less adaptive mutations during long-term evolution then the process at work during more recent human evolution has to be transient, and has to has to have limited only recent adaptation. It is also noteworthy that both disease genes and their controls have experienced more coding adaptation than genes in the human genome overall (***Figure 5A***), especially more strong adaptation according to ABC-MK (***Figure 5C***). The fact that the baseline long-term coding adaptation is lower genome-wide, but similarly higher in disease and their control genes, also shows that the matched controls do play their intended role of accounting for confounding factors likely to affect adaptation. The fact that long-term protein adaptation is not lower at disease genes also excludes that purifying selection alone can explain the sweep deficit at disease genes, because purifying selection would then also have decreased long-term adaptation. A more transient evolutionary process is thus more likely to explain our results.

A possible role of interference of deleterious mutations

The underlying evolutionary process at mendelian disease genes must explain the sweep deficit, while simultaneously not implying a long-term deficit of adaptation. A possible explanation is that adaptation may be limited at disease genes due to currently segregating deleterious mutations interfering with, and slowing down advantageous variants. This process may in principle satisfy the condition of decreasing recent adaptation without decreasing long-term adaptation, since the

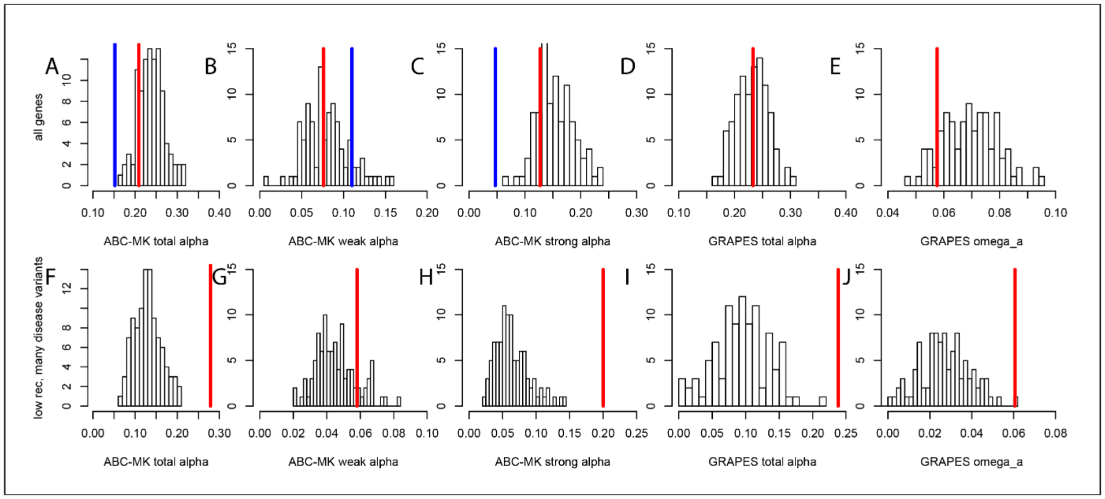


Figure 5. Nonsynonymous coding adaptation in disease vs. control genes. Histograms represent the long-term coding adaptation values in 100 control sets. Red lines represent the long-term coding adaptation value in disease genes. Blue lines represent the long-term adaptation value across the whole coding genome. (**A** to **E**) All disease genes compared to controls. (**F** to **J**) Disease genes with many disease variants vs. controls, in low recombination regions of the genome. (**A** and **F**) Total alpha from ABC-MK. (**B** and **G**) Alpha for weak adaptation according to ABC-MK. (**C** and **H**) Alpha for strong adaptation according to ABC-MK. (**D** and **I**) Total alpha according to GRAPES. (**E** and **J**) Omega_a, the ratio of the rate of advantageous amino acid changes over the rate of synonymous changes, according to GRAPES.

number of deleterious segregating variants at a given locus is likely to vary significantly over evolutionary time due to genetic drift. This explanation, where the sweep deficit is specifically due to segregating deleterious variants, is particularly plausible given our results so far. Indeed, even though more selectively constrained genes, including disease genes, may be arguably more prone to harbor deleterious segregating variants because they are more constrained, we have already gathered evidence that purifying selection alone does not explain the sweep deficit at disease genes; it reflects the amount of deleterious variants that were removed, not the amount of currently segregating ones. Furthermore, we always compare disease and control non-disease genes with matched purifying selection. Lastly, we have shown that genes with a high level of purifying selection (high proportion of conserved elements) do not have any sweep deficit (see above).

The process of interference between deleterious and advantageous variants has been mostly studied in haploid species (***Jain, 2019***; ***Johnson and Barton, 2002***; ***Peck, 1994***). In diploid species including humans, recessive deleterious mutations specifically have been shown to have the ability to slow down, or even stop the frequency increase of advantageous mutations that they are linked with (***Assaf et al., 2015***). Dominant variants do not have the same interfering ability, because they do not increase in frequency in linkage with advantageous variants as much as recessive deleterious do, before the latter can be 'seen' by purifying selection when enough homozygous individuals emerge in a population (***Assaf et al., 2015***). (***Uricchio et al., 2019***) also found evidence of decreased protein adaptation in the regions of the human genome with strong background selection and low recombination. The majority of disease variants at mendelian disease genes are recessive (***Amberger et al., 2019***; ***Balick et al., 2015***). Thus, if segregating recessive deleterious mutations are more common at disease genes, starting with the known disease variants themselves, then their interference could in theory explain the sweep deficit that we observe. This is true even despite the fact that we matched disease and control non-disease genes for multiple measures of selective constraint. Indeed, we use measures of selective constraint such as the density of conserved elements or the proportion of variable non-synonymous sites pN (Materials and methods), that are indicative of the amount of deleterious mutations that were ultimately removed, and not indicative of the number of currently segregating deleterious variants. Disease genes and control non-disease genes may have very similar densities of conserved elements and similar pN, and still very different numbers of currently segregating recessive deleterious variants. Although directly comparing the actual total numbers of recessive deleterious mutations at disease and non-disease genes is difficult notably because estimating dominance coefficients in the human genome is a notoriously hard problem (***Huber et al., 2018***), we can still use indirect comparison strategies. First, if an interference of recessive deleterious mutations is involved then this interference is expected to be stronger in low recombination regions of the genome, where more deleterious mutations are likely to be genetically linked to an advantageous mutation. Therefore, we predict that the sweep deficit should be more pronounced when comparing disease and non-disease genes only in low recombination regions of the genome, where the linkage between deleterious and advantageous variants is higher. Conversely, the sweep deficit should be less pronounced in high recombination regions of the genome. Second, if the number of known segregating disease variants at a given disease gene correlates well enough with the total number of segregating recessive deleterious mutations at this disease gene then we should observe a stronger sweep deficit at disease genes with many known disease variants, compared to disease genes with few known segregating disease variants. Based on these two predictions, the sweep deficit should be particularly strong at disease genes with both many disease variants AND lower recombination. As the number of disease variants for each disease gene, we use the number of disease variants as curated by OMIM/UNIPROT (Materials and methods).

For these comparisons, we focus solely on African populations for which we found the strongest sweep deficit (***Figure 4***). We first compare disease and control non-disease genes both from only regions of the genome with recombination rates lower than the median recombination rate (1.137 cM/Mb). In agreement with recombination being involved, we find that the sweep deficit at low recombination disease genes is much more pronounced than the overall sweep deficit found when considering all disease and control non-disease genes regardless of recombination (***Figure 6***, FPR=2.10^{-4}). Conversely, the sweep deficit at disease genes compared to non-disease genes is much less pronounced when restricting the comparison to genes with recombination rates higher than the median recombination rate (1.137 cM/Mb), and remains only marginally significant (***Figure 6***, FPR=0.029). This provides evidence that genetic linkage may indeed be involved. Low

recombination rate	# of disease variants	deficit
≥0 cM/Mb	≥ 1	22.98
≥ 1.14 cM/Mb	≥ 1	9.92
< 1.14 cM/Mb	≥ 1	37.76
≥0 cM/Mb	< 5	6.97
≥0 cM/Mb	≥ 5	40.18
≥ 1.14 cM/Mb	< 5	2.16
≥ 1.14 cM/Mb	≥ 5	15.36
< 1.14 cM/Mb	< 5	12.91
< 1.14 cM/Mb	≥ 5	66.8

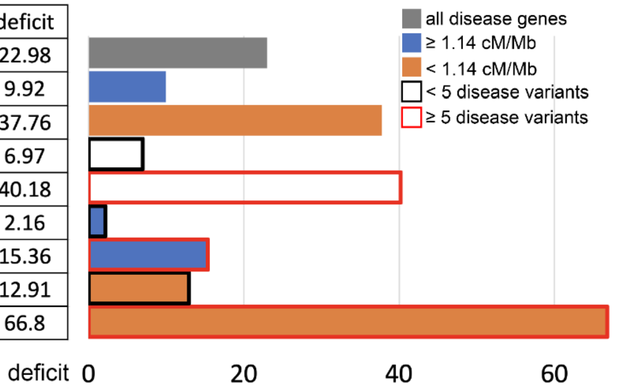


Figure 6. Sweep deficit as a function of recombination and disease variants number. The sweep deficit is measured as the FPR score per gene (to make all tested groups comparable) over all window sizes, and nS_L and iHS, as in ***Figure 2*** (Materials and methods). The different groups are separated according to recombination and numbers of disease variants so that they have approximately the same size (a half or a fourth of the disease genes). All deficits are measured using only African populations.

The online version of this article includes the following source data for figure 6:

Source data 1. Confounding factors differences between low and high recombination disease genes.

recombination is however not sufficient on its own to create a sweep deficit, and we further test if the sweep deficit also depends on the number of disease variants at each disease gene. In our dataset, approximately half of all the disease genes have five or more disease variants, and the other half have four or less disease variants (Materials and methods). In further agreement with possible interference of recessive deleterious variants, the sweep deficit is much more pronounced at disease genes with five or more disease variants (***Figure 6***, FPR=8.10 ^{4}). The sweep deficit at disease genes with four or less disease variants is barely significant compared to control non-disease genes (***Figure 6***, FPR=0.032). In addition, disease genes with five or more disease variants, but with recombination higher than the median recombination rate, do not have a strong sweep deficit either (***Figure 6***, FPR=0.026). A higher number of disease variants alone is thus not enough to explain the sweep deficit. In a similar vein, disease genes with a recombination rate less than the median recombination rate, and with four or less disease variants, do not exhibit a strong sweep deficit (***Figure 6***, FPR=0.021). This confirms that low recombination alone is not enough to explain the sweep deficit at disease genes. Accordingly, disease genes with both low recombination AND five or more disease variants show the strongest sweep deficit (***Figure 6***, FPR=2.10 ^{4}). Disease genes with both high recombination AND less than five disease variants show no sweep deficit at all, with a sweep prevalence undistinguishable from control non-disease genes (***Figure 6***, FPR=0.74). The latter result is important, because it suggests that interference of recessive deleterious variants may be sufficient on its own to explain the whole sweep deficit at disease genes. Both higher linkage and more disease variants seem to be needed to explain the sweep deficit at disease genes. Note that these results are not due to introducing a bias in the overall number of variants by using the number of disease variants, because we always match the level of neutral genetic variation between disease genes and control non-disease genes with pS. The overall level of genetic variation is further matched thanks to pN and thanks to McVicker's B, whose value is directly dependent on the level of genetic variation at a given locus (***McVicker et al., 2009***). Further note that only moderate differences in confounding factors between low and high recombination mendelian disease genes are unlikely to explain the sweep deficit difference (***Figure 6—source data 1***).

These results further imply that the alternative explanation of constitutively less common adaptation at disease genes is less likely than interference. With constitutively less adaptation at disease genes, the stronger sweep deficit observed at disease genes in low, compared to high recombination regions, could only reflect the fact that there is more statistical power to detect sweeps in low recombination regions (***Booker et al., 2020***; ***O'Reilly et al., 2008***), and therefore more statistical power to distinguish a sweep deficit. A higher statistical power to detect sweeps in low recombination regions however does not explain the very strong sweep deficit in low recombination regions with many disease variants, and the marginal sweep deficit in low recombination regions with few disease variants.

We further find that the difference in sweep deficits between high and low recombination regions is not affected when using only nSL as a sweep statistic (Materials and methods). The nSL statistic was initially designed to be more robust to recombination than iHS (***Ferrer-Admetlla et al., 2014***), and to have more similar power in low and high recombination regions, and here we confirm this greater robustness. The two distributions of nSL sweep ranks, one for the lower recombination half and one for the higher recombination half of the genes, are much more similar than the two corresponding distributions of iHS sweep ranks (***Figure 7A,B***). Low recombination regions only have a slight excess of top-ranking nSL signals compared to high recombination regions. Such a small difference is unlikely to generate the substantial discrepancy in power needed to explain the much stronger sweep deficit in low recombination regions. The sweep deficit is substantial when using only nSL on all the disease genes and their controls (***Figure 7C***; FPR<5.10 ^{4}). The nSL-only sweep deficit is only marginally significant in high recombination regions (FPR=0.043, deficit score=-9,227.4), but strongly significant and about four times more pronounced in low recombination regions (FPR<5.10 ^{4}, deficit score=-33,177.2), the same relative difference observed when using both iHS and nSL (***Figure 6***).

More importantly, the fact that constitutively less adaptation at disease genes combined to more power to detect sweeps in low recombination regions does not explain our results, is made even clearer by the fact that disease genes in low recombination regions and with many disease variants have in fact experienced more, not less long-term adaptation according to an MK analysis using both ABC-MK and GRAPES (***Figure 5F,G,H,I,J***). ABC-MK in particular finds that there is a significant

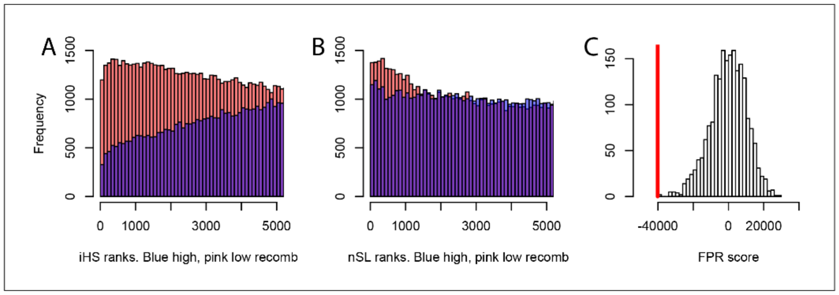


Figure 7. Different sweep detection power response of iHS and nSL to varying recombination rates. (**A**) iHS sweep ranks, shown from 1 to 5000 across all window sizes (50 kb to 1000 kb) in Africa, in low recombination (pink) or high recombination regions (blue). (**B**) Same as A. but for nSL. (**C**) Observed sweep deficit at disease genes (red line) compared to the distribution of the sweep deficit in 2000 block-randomized genomes. Same as ***Figure 2A*** but with only nSL.

excess of long-term strong adaptation (***Figure 5H***, P<0.01) in disease genes with low recombination and with many disease variants, compared to controls, but similar amounts of weak adaptation (***Figure 5G***, P=0.16). It might be that disease genes with many disease variants are genes with more mutations with stronger effects that can generate stronger positive selection. The potentially higher supply of strongly advantageous variants at these disease genes makes it all the more notable that they have a very strong sweep deficit in recent evolutionary times. This further strengthens the evidence in favor of interference during recent human adaptation: the limiting factor does not seem to be the supply of strongly advantageous variants, but instead the ability of these variants to have generated sweeps recently by rising fast enough in frequency.

Decreased interference of recessive deleterious mutations during a bottleneck may explain the weaker sweep deficit in East Asia and Europe

An important observation in our analysis, that any potential explanation needs to account for, is the much weaker sweep deficit at disease genes in Europe and especially in East Asia, compared to Africa. If interference of recessive deleterious variants explains the sweep deficit at disease genes then it should also account for the weaker sweep deficit out of Africa. Previous results suggest that it might be the case. (***Balick et al., 2015***) showed that during a bottleneck of the magnitude of the Out of Africa bottleneck, there should be a sharp decrease of the segregating recessive deleterious variants load, because of all the low-frequency recessive deleterious variants that are removed when the bottleneck occurs. This is especially true for strongly deleterious variants that tend to segregate at lower frequencies. The magnitude of the bottleneck investigated by Balick et al. (a 10-fold decrease in population size) has since been confirmed for the Out of Africa bottleneck by the most recent Ancestral Recombination Graph approaches (***Speidel et al., 2019***). Populations in East Asia in particular went from an ancestral effective population size of ~10,000 to a post-Out of Africa effective population size of ~1000 for extended amounts of time before the very recent explosive human population expansions (***Speidel et al., 2019***). Balick et al. also found (i) evidence of an overall increased burden of recessive deleterious variants at disease genes compared to other genes and (ii) also found that this recessive burden had decreased in Europe, following the bottleneck out of Africa.

Here, we hypothesize that the bottleneck out of Africa decreased the recessive burden enough to cause a possible decrease of interference of recessive segregating variants at mendelian disease genes, and that this decrease of interference might explain the smaller sweep deficit observed at disease genes in Europe and especially in East Asia (***Figure 4***). We test this hypothesis using forward population simulations of loci with concentrations of deleterious variants meant to resemble a number of genic regions (Materials and methods). We find that, as expected given the results of Balick et al., there is much less interference of recessive deleterious variants after a bottleneck similar to the Out of Africa bottleneck (*Table 1*). In *Table 1*, we provide both the fixation probabilities and the time to fixations of advantageous mutations of different strengths, under different simulated demographies matching either past demography in or out of Africa, and including deleterious mutations or not for comparing fixation parameters with or without interference (Materials and methods). In the presence of recessive deleterious variants, the time to fixation of advantageous variants in particular is only slightly increased after an Out of Africa-like bottleneck, compared to the strong fixation time increase when no bottleneck has taken place (*Table 1*). This interference effect is specific to recessive deleterious mutations and not observed with dominant deleterious mutations, as expected (*Assaf et al., 2015*). The effect on fixation time alone is likely sufficient to explain the sharp difference in sweep deficit observed, especially when comparing Africa and East Asia, the latter being the most bottlenecked population investigated here. Indeed, a sharp increase in fixation time is expected to result in substantially weaker sweep signals. This is because a slower increase in frequency of an advantageous mutation will leave more time to a larger number of recombination events to occur, and thus narrow down the breadth of the sweep signature around that advantageous mutation (*Assaf et al., 2015*). A reduction of the segregating recessive burden as observed by Balick et al. at mendelian disease genes, and therefore interference, can thus explain the

Table 1. Decreased interference during a bottleneck.

The table provides the proportion of advantageous mutations that go to fixation (% fixed), and the time to fixation under multiple conditions simulated with SLiM (Materials and methods). For example, s=0.005, 40% constrained, recessive means that we simulate advantageous mutations with s=0.005, surrounded by a genomic region where 40% of sites experience recessive deleterious mutations according to a specific distribution of fitness effets (Materials and methods). The fix. time increase column provides the relative increase in fixation time (ratio of times) in the presence compared to in the absence of deleterious mutations. The time to fixation is in number of generations. The Methods provide more details on the simulations.

	Demography	**Deleterious mutations?**	**Time to fixation**	**% fixed**	**Fix. time increase**	**Fix. prob decrease**
S=0.005, 10% constrained, recessive	**East Asia: 10000->1000**	**No**	**2265**	**0.0050**		
		Yes	**2547**	**0.0033**	**1.12**	**0.66**
	Africa: 10000->10000	**No**	**4204**	**0.0051**		
		Yes	**6530**	**0.0038**	**1.55**	**0.74**
s=0.005, 20% constrained, recessive	demography	deleterious mutations?	time to fixation	% fixed	fix. time increase	fix. prob decrease
	East Asia: 10000->1000	no	2265	0.0050	1.16	0.66
		yes	2617	0.0033		
	Africa: 10000->10000	no	4204	0.0051	1.69	0.62
		yes	7113	0.0032		
s=0.005, 40% constrained, recessive	demography	deleterious mutations?	time to fixation	% fixed	fix. time increase	fix. prob decrease
	East Asia: 10000->1000	no	2265	0.0050	1.17	0.65
		yes	2642	0.0033		
	Africa: 10000->10000	no	4204	0.0051	2.10	0.56
		yes	8809	0.0028		
s=0.01, 40% constrained, recessive	demography	deleterious mutations?	time to fixation	% fixed	fix. time increase	fix. prob decrease
	East Asia: 10000->1000	no	1530	0.0092	1.37	0.78
		yes	2090	0.0072		
	Africa: 10000->10000	no	2546	0.0098	2.05	0.82
		yes	5209	0.0080		
s=0.005, 40% constrained, dominant	demography	deleterious mutations?	time to fixation	% fixed	fix. time increase	fix. prob decrease
	East Asia: 10000->1000	no	2265	0.0050	0.96	0.93
		yes	2169	0.0046		
	Africa: 10000->10000	no	4204	0.0051	0.99	0.92
		yes	4169	0.0047		

observed patterns in Africa versus East Asia and Europe. This reduction might be due to a number of concurrent processes: a reduction in the number of recessive segregating variants as observed by Balick et al., but hypothetically, might also be due to a decrease in the effective deleteriousness of the remaining segregating variants due to the lower effective population size. This further supports the idea that interference with recessive deleterious variants may explain our observation of a strong sweep deficit at disease genes in Africa, and of weaker sweep deficits out of Africa.

Similar levels of sweep depletion in mendelian disease genes across MeSH disease classes

Because we find an overall sweep depletion at mendelian disease genes, we further ask if genes associated with different diseases might show different patterns of depletion (always in African populations). We classify disease genes into different classes according to the Medical Subject Headings (MeSH) annotation for diseases in DisGeNet (***Piñero et al., 2020***). The MeSH annotations organize the disease genes into broad disease categories that overlap with distinct organs or large physiological systems (for example the endocrine system). We find significant (FPR<0.05) sweep depletions in Africa for all but one disease MeSH classes (FPR<0.05; ***Figure 8***). The sweep deficit is comparable across MeSH disease classes (***Figure 8***), suggesting that the evolutionary process at the origin of the sweep deficit is not disease-specific. This is compatible with a non-disease specific explanation such as recessive deleterious variants interfering with adaptive variants, irrespective of the specific disease type. The only non-significant deficit is for the MeSH term immune system diseases. Interestingly, there is evidence that past adaptation at disease genes in response to diverse pathogens has resulted in increased prevalence of specific auto-immune diseases (***Barreiro and Quintana-Murci, 2010***), and we can speculate that this might be why we do not see a sweep deficit at those genes.

Discussion

We found a depletion of the number of genes in recent sweeps at human non-infectious, mendelian disease genes compared to non-disease genes. Although more work is needed, the lack of sweeps

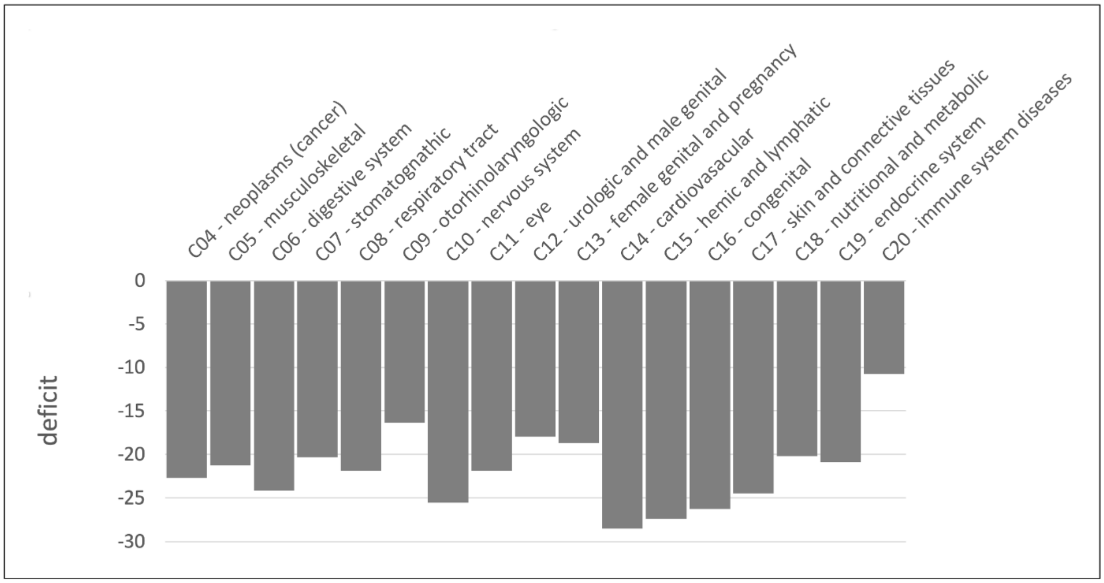


Figure 8. Sweep deficit per MeSH disease classes. The sweep deficit is measured as the overall FPR score per gene (Materials and methods), to make all MeSH classes comparable even if they include different numbers of genes.

at disease genes already favors specific evolutionary processes over others. For example, it makes it unlikely that past adaptations increasing the occurrence of disease variants through hitchhiking would be the dominant process linking disease and adaptation at the gene level. The lack of sweeps at mendelian disease genes especially in Africa also seems to be unrelated to any difference in mutation accumulation between disease and non-disease genes, since we find no sign of a difference in mutation rates between the two categories of genes in the first place, and since we match metrics accounting for mutation rate in our comparisons (e.g., GC content and pS). Instead, a lack of sweeps, once selective constraint has been controlled for, seems to favor a relationship involving a decrease of recent adaptation at disease genes, beyond simple constraint (measured by the amount and strength of deleterious mutations that are removed).

Multiple mechanisms might explain such a lack of recent adaptation. A first possible hypothesis is that disease genes are genes that can be sensitive to the environment and whose fitness optimum can change during evolution when the environment changes. However, when this happens, adaptation may be constitutively less common at disease genes. Although higher pleiotropy is a tempting hypothesis to explain such a lag (**Otto, 2004**), disease genes have not experienced less long-term protein adaptation. Since gene pleiotropy is a stable property over evolutionary time, it is difficult to see how it would generate a recent adaptation deficit without also generation a deficit of long-term adaptation. This also likely excludes that mendelian disease genes just happen to be genes where mutations are rarely advantageous because they have strong effects that tend to overshoot, and thus miss the fitness optimum. On the contrary, we find that mendelian disease genes (and their controls) have experienced more long-term protein adaptation than the genomic baseline (**Figure 5**). Given our results on genetic interference, disease genes may experience as much, or even more long-term adaptive substitutions while also showing a deficit of strong recent adaptation because the list of current human disease genes defined by the presence of segregating disease variants varies over evolutionary time. The list of genes that are mendelian disease genes may evolve over evolutionary time based on where the transiently segregating, recessive deleterious variants that define them, are found in the genome as a result of the interplay between gene constraint and genetic drift. In this case, the millions of years of human evolution would be more than enough to see a substantial turnover of genes with segregating, recessive deleterious variants in the genome.

A plausible explanation for all our observations is indeed genetic interference, where selective sweeps are impeded at disease genes due to the interference of genetically linked recessive deleterious variants. The deleterious effects of these variants can reveal themselves when they hitchhike together with an advantageous variant that is just starting to increase in frequency (**Assaf et al., 2015**). Accordingly, we find a marked sweep depletion in Africa when restricting the comparison to disease and non-disease genes in low recombination regions of the genome and with higher numbers of disease variants (**Figure 6**). We also show through simulations that a stronger sweep deficit in Africa is expected if genetic interference indeed explains our results. All these comparisons are however indirect; we do not quantify directly the effect of recessive deleterious mutations at disease or non-disease genes. That said, the majority of mendelian disease variants are known to be recessive (**Balick et al., 2015**), and using the number of disease variants, as done in the present study, should be a good proxy of the actual number of segregating recessive deleterious mutations. Estimating dominance may prove challenging, however, since it is difficult to distinguish selection coefficient changes from dominance coefficient changes (**Huber et al., 2018**). Again, our results provide preliminary evidence to further test in the future.

In addition to suggesting possible explanatory evolutionary scenarios, our results highlight a number of potential limitations and biases that also need to be further explored. First, the lack of sweeps at disease genes suggests the possibility of a technical bias against the annotation of disease genes in sweep regions with high LD, as described in the Results. This bias is unlikely to be the dominant explanation for our results, because then we would expect a stronger sweep deficit at disease genes in Europe than in Africa, given that most disease genes were annotated in Europe. The recombination rate at disease genes is also only slightly different from the recombination rate at non-disease genes (**Figure 1**), and we match the recombination rate between disease genes and controls. The increase of the sweep deficit when comparing disease and non-disease genes only in low recombination regions (**Figure 6**), where disease annotation would then be more difficult regardless of overlapping a sweep or not, also suggests that this bias is unlikely.

Further work is now required regarding the connection between the sweep deficit and polygenic adaptation not leaving hitchhiking signals. Our results could also be explained by a different balance between sweeps and polygenic adaptation at mendelian disease genes, with less sweeps but more polygenic adaptation that would be less affected by interference with deleterious variants. That said, we do find that mendelian disease genes have experienced more long-term adaptive protein evolution than the genomic baseline (***Figure 5***), suggesting mutations that were advantageous enough to go all the way to fixation. It may be possible to use recent polygenic adaptation quantification tools such as PALM (***Stern et al., 2021***) to compare its prevalence between mendelian disease and non-disease genes.

Finally, there are multiple directions to further analyze the sweep deficit at disease genes that we have not explored in this manuscript. For instance, analyzing the sweep deficit as a function of the time of onset of diseases (early or late in life), might further provide clues to why the sweep deficit exists in the first place. Preliminary comparison of the sweep deficit at specific MeSH disease classes (***Figure 8***) with known early (congenital diseases) or mostly late onsets (cancer, cardiovascular), however, suggests that the average onset time of diseases might not make much of a difference.

In conclusion, although our analysis reveals a strong deficit of selective sweeps at human disease genes in Africa that seems to be due to genetic interference, it also suggests that more work is needed to better understand the evolutionary processes at work, and the biases that may have skewed our interpretations. Despite these limitations, our comparison already suggests that specific evolutionary relationships between disease genes and adaptation might be more prevalent than others, especially interference between segregating recessive deleterious and advantageous variants. As an important follow-up question, it may now be important to ask how the sweep deficit at disease genes might have hidden interesting adaptive patterns in previous functional enrichment analyses, especially in gene functions that are often annotated based on disease evidence in the first place. For example, metabolic genes are believed to be of particular interest for adaptation to climate change. But metabolic genes are often found due to their role in metabolic disorders, and a strong representation of disease genes among all metabolic genes could then in theory mask any sweep enrichment. A sweep enrichment at metabolic genes might only become visible once controlling for the proportion of disease genes, in addition to the list of controls that we already use in the present analysis. Our results thus highlight the great complexity of studying functional patterns of adaptation in the human genome.

Materials and methods

Disease gene lists

We consider genes that are known to be associated (mendelian type of association) with diseases as mendelian disease genes. We focus on protein-coding genes associated with human mendelian non-infectious diseases. By non-infectious, we mean that we excluded genes with known infectious disease-associated variants. This does not exclude most virus-interacting genes since most of them are not associated at the genetic variant level with infectious diseases. It is important to note that the effect of virus interactions is accounted for by matching the number of interacting viruses between mendelian disease genes and controls (see below). Complex diseases are associated with several loci and environmental factors. Patterns of positive selection at complex disease and mendelian disease genes may differ (***Blekhman et al., 2008***; ***Quintana-Murci, 2016***; ***Torgerson et al., 2009***), which is why we restrict our analysis to mendelian disease genes. We also restrict our analyses to non-infectious disease genes, since disease, genetic associations with pathogens are an entirely different problem. We nevertheless control for the proportion of genes that are immune genes or interact with viruses (see below), since it has been shown that immune genes and interactions with viruses drive a large proportion of genomic adaptation in humans (***Enard and Petrov, 2020***). Therefore, different proportions of immune and virus-interacting genes between disease and non-disease genes might confound their comparison. Moreover, although diseases can be associated with non-coding genes, we only use protein-coding genes. We curate disease genes defined as genes associated with diseases in a mendelian fashion according to both DisGeNet (***Piñero et al., 2020***) and OMIM (***Amberger et al., 2019***), to ensure that we focus on high-confidence mendelian disease genes. DisGeNet is a comprehensive database including gene-disease associations (GDAs) from many sources. In order to get disease genes with high confidence, we further only use GDAs curated by UniProt. These gene-disease associations are extracted and carefully curated from the scientific literature and the OMIM (Online Mendelian Inheritance in Man) database, which reports phenotypes either mendelian or possibly mendelian (**Amberger et al., 2019**). We also exclude all genes associated with infectious diseases according to MeSH annotation (disease class C01). In the end, we curate 4215 non-infectious mendelian disease genes from DisGeNet also curated by OMIM and Uniprot. Although we rely on GDAs from Uniprot to curate high-quality disease genes, we also include GDAs of DisGeNet from other sources when classifying disease genes into different MeSH classes and measuring pleiotropy, as long as a disease gene has at least one GDA curated by OMIM and Uniprot. We completely exclude GDAs that are only reported by CTD (Comparative Toxicogenomics Database) (**Davis et al., 2021**) in this study. This is because CTD includes a broad range of chemical-induced diseases that might only happen when people are exposed to these chemicals, especially some inorganic chemicals that may not be present in natural environments (**Davis et al., 2021**).

In order to study different types of diseases, we also divide disease genes into different classes according to the annotated MeSH classes in DisGeNet (**Piñero et al., 2020**). Those diseases without MeSH class are annotated as 'unclassfied'. Genes belonging to more than one MeSH class are counted in each MeSH class where they are present. MeSH classes including less than 50 genes are not considered in this study. We classify all the non-infectious disease genes into 17 MeSH classes including Neoplasms (C04), Musculoskeletal Diseases (C05), Digestive System Diseases (C06), Stomatognathic Diseases (C07), Respiratory Tract Diseases (C08), Otorhinolaryngologic Diseases (C09), Nervous System Diseases (C10), Eye Diseases (C11), Male Urogenital Disease (C12), Female Urogenital Diseases and Pregnancy Complications (C13), Cardiovascular Diseases (C14), Hemic and Lymphatic (C15), Congenital, Hereditary, and Neonatal Diseases and Abnormalities (C16), Skin and Connective Tissue Diseases (C17), Nutritional and Metabolic Diseases (C18), Endocrine System Diseases (C19), Immune System Diseases (C20), and 'unclassified'.

Detecting recent selection signals at human genes

All the analyses were conducted human genome version hg19. We use two different methods to detect selective sweeps in human populations: iHS (**Voight et al., 2006**) and nSL (**Ferrer-Admetlla et al., 2014**). Both approaches are haplotype-based statistics calculated with polymorphism data. We use human genome data from the 1,000 Genomes Project phase 3, which includes 2504 individuals from 26 populations (**Auton et al., 2015**).

We measure iHS and nSL in windows centered on human coding genes (i.e. windows whose center is located half-way between the most upstream transcript start site and most downstream transcript stop site of protein coding genes). We use windows of sizes ranging from 50 kb to 1000 kb (50kb, 100kb, 200kb, 500kb and 1000kb) since we do not want to presuppose of the size of sweeps, and since the size of the selective sweeps may vary between different genes. Moreover, to avoid any preconception related to the expected strength or number of sweep signals, we use a moving rank threshold strategy to measure the enrichment or deficit in sweeps at disease genes. For example, we select the top 500 genes with the stronger sweep signals according to a specific statistic (iHS or nSL). We then compare the number of diseases and non-disease genes within the top 500 genes with the strongest iHS or nS_L signals. This was repeated for different top thresholds and the corresponding ranks from top 5,000 to top 10 (5000,4000,3000,2500,2000,1500,1000,900,800,700,600,500,450,400,350,300,250,200,150,100,90,80,70,60,50,40,30,25,20,15,10). Using a range of rank thresholds makes less assumptions and provides more flexibility than the classic outlier approach, even though we still have to arbitrarily determine a list of rank thresholds to include. This is because we can get a significant result not only due to an enrichment of only the top, absolutely strongest sweeps, but also due for example to a large excess of weak or moderate sweeps, that would for example increase the expected numbers in the top 5000 or top 2000, without increasing the number of sweeps in the top 100 or top 50. Therefore, our approach is sensitive to a more diverse range of sweeps than the classic outlier approach, that makes a very restrictive assumption that sweeps have to be necessarily be strong. Genes are ranked based on the average iHS or nSL in their gene centered windows. Both iHS and nSL measure, individually for each SNP in the genome, how much larger haplotypes linked to the derived SNP allele are compared to haplotypes linked to the ancestral allele. For each window, we measure the average of the absolute value of iHS or nS_L over all the SNPs in that window with an iHS or nS_L value. The average iHS or nSL values in a window provide high power to detect recent select sweeps (**Enard and Petrov, 2020**).

Comparing recent adaptation between disease and non-disease genes

We use a previously developed gene-set enrichment analysis pipeline to compare recent adaptation between disease and non-disease genes (**Enard and Petrov, 2020**) available at https://github.com/DavidPierreEnard/Gene_Set_Enrichment_Pipeline. This pipeline includes two parts. The first part is a bootstrap test that estimates the whole sweep enrichment or depletion curve at genes of interest (mendelian disease genes in our case) while controlling for confounding factors. The second part is a false positive risk (also known as false discovery rate in the context of multiple testing) that estimates the statistical significance of the whole sweep enrichment curve using block-randomized genomes (**Enard and Petrov, 2020**).

To compare disease and non-disease genes, we first need to select control non-disease genes that are sufficiently far away from disease genes. In that way, we avoid using as controls non-disease genes that overlap the same sweeps as neighboring disease genes, thus resulting in an underpowered comparison. The question is then how far do we need to choose non-disease control genes? Ideally, we would choose non-disease control genes as far as possible from disease genes in the human genome, further than the size of the largest known sweeps (e.g. the lactase sweep), which would be on the order of a megabase. However, because there are many disease genes in our dataset (4215), there are very few non-disease genes in the human genome that are more than one megabase away from the closest disease gene. This is a problem, because the available number of potential control non-disease genes is an important parameter that can affect both the type I error, false positive rate, and type II error, false negative rate of the disease vs. non-disease genes comparison. Indeed, the smaller the control set, the more likely it is to deviate from being representative of the true null expectation at non-disease genes. The noise associated with a small sample could go either way. Either the small control sample happens by chance to have less sweeps, and the bootstrap test we use to compare disease and non-disease genes will become too liberal to detect sweep enrichments, and to conservative to detect sweep deficits. Or the small control sample happens by chance to have more sweeps than a larger control sample would, and the bootstrap test becomes too conservative to detect sweep enrichments, and too liberal to detect sweep deficits.

After trying distances between disease genes and control disease genes of 100 kb, 200 kb, 300 kb, 400 kb, and 500 kb, we find that the sweep deficit observed at disease genes increases steadily from 100 kb to 300 kb (**Table 2**), showing that 100 kb or 200 kb are likely insufficient distances. Further than 300 kb at 400 kb, we do not observe much stronger sweep deficits than at 300 kb, while at the same time the risks of type I and type II errors keep increasing due to shrinking non-disease genes control sets. This would translate in a decreased power to possibly exclude the null hypothesis of no sweep enrichment or deficit in the second part of the pipeline, when estimating the actual pipeline FPR. Because of this, we set the required distance of potential control non-disease genes from disease genes at 300 kb. This is also the distance where there are still approximately as many control genes (3455) as there are disease genes that we can use for the comparison (3030; those genes out of the 4215 disease genes with sweep data and data for all the confounding factors).

Another important aspect of the bootstrap test (first part of the pipeline), aside from setting up the minimal distance of the control non-disease genes, is the matching of potential confounding

Table 2. Sweep deficit as a function of the minimal distance of control non-disease genes.
The sweep deficit is measured by the FPR score, that is the cumulative difference between the number of genes in sweeps at disease and control non-disease genes, across window sizes, sweep summary statistics, and African populations (see the rest of the Materials and methods).

Minimal distance	Sweep deficit
100 kb	−20889
200 kb	−35009
300 kb	−68928
400 kb	−88546

factors likely to influence sweep occurrence. We choose non-disease control genes that have the same confounding factors characteristics as disease genes (for example, control non-disease genes that have the same gene expression level across tissues as disease genes). The precise matching algorithm is detailed in **Enard and Petrov, 2020**. In brief, the bootstrap test builds sets of control genes that have the same overall average values for confounding factors as disease genes. For example, the bootstrap test can build 100 control sets, with each set having the same overall average GC content as disease genes. Note that this means that disease genes are not individually matched one by one with one control gene that happens to have the same GC content. Matching genes individually, instead of matching the overall gene sets averages, would indeed limit the pool of potential control genes too drastically. For more details on this, please refer to **Enard and Petrov, 2020**.

When comparing disease and non-disease genes with the bootstrap test, we control for the following potential confounding factors that could influence the occurrence of sweeps at genes:

- Average overall expression in 53 GTEx v7 tissues (**GTEx Consortium, 2020**) (https://www.gtexportal.org/home/). We used the log (in base 2) of TPM (Transcripts Per Million).
- Expression (log base 2 of TPM) in GTEx lymphocytes. Expression in immune tissues may impact the rate of sweeps.
- Expression (log base 2 of TPM) in GTEx testis. Expression in testis might also impact the rate of sweeps.
- deCode recombination rates 50 kb and 500 kb: recombination is expected to have a strong impact on iHS and nSL values, with larger, easier to detect sweeps in low recombination regions but also more false positive sweeps signals. The average recombination rates in the gene-centered windows are calculated using the most recent deCode recombination map (**Halldorsson et al., 2019**). We use both 50 kb and 500 kb window estimates to account for the effect of varying window sizes on the estimation of this confounding factor (same logic for other factors where we also use both 50 kb and 500 kb windows).
- GC content is calculated as a percentage per window in 50 kb and 500 kb windows. It is obtained from the USCS Genome Browser.
- The density of coding sequences in 50 kb and 500 kb windows centered on genes. The density is calculated as the proportion of coding bases respect to the whole length of the window. Coding sequences are Ensembl v99 coding sequences.
- The density of mammalian phastCons conserved elements (**Siepel et al., 2005**) (in 50 kb and 500 k windows), downloaded from the UCSC Genome Browser. We used a threshold considering 10% of the genome as conserved, as it is unlikely that more than 10% of the whole genome is constrained according to previous evidence (**Siepel et al., 2005**). Given that each conserved segment had a score, we considered those segments above the 10% threshold as conserved.
- The density of regulatory elements, as measured by the density of DNASE1 hypersensitive sites (in 50 kb and 500 kb windows) also from the UCSC Genome Browser.
- The number of protein-protein interactions (PPIs) in the human protein interaction network (**Luisi et al., 2015**). The number of PPIs has been shown to influence the rate of sweeps (**Luisi et al., 2015**). We use the log (base 2) of the number of PPIs.
- The gene genomic length, that is the distance between the most upstream and the most downstream transcription start sites.
- The number of gene neighbors in a 50 kb window, and the same number in 500 kb window centered on the focal genes: it is the number of coding genes within 25 kb or within 250 kb.
- The number of viruses that interact with a specific gene (**Enard and Petrov, 2020**).
- The proportion of immune genes. The matched control sets have the same proportion of immune genes as disease genes, immune genes being genes annotated with the Gene Ontology terms GO:0002376 (immune system process), GO:0006952 (defense response) and/or GO:0006955 (immune response) as of May 2020 (**Gene Ontology Consortium and Gene Ontology, 2021**).
- The average non-synonymous polymorphism rate pN in African populations, and the the synonymous rate pS. We matched pN to build control sets of non-disease genes with the same average amount of strong purifying selection as disease genes. Also, pS can be a proxy for mutation rate and we can build control sets of non-disease genes with similar level of mutation rates.
- McVicker's B value which can be used to account for more recent selective constraint (**McVicker et al., 2009**).

Similar to the selection of control genes far enough from disease genes, the matching of many confounding factors decreases the number of non-disease genes that can effectively be used as controls. This further increases the risk of type I and type II errors of the bootstrap test, as previously described (**Enard and Petrov, 2020**). In addition, the bootstrap test only provides a p-value for each tested sweep rank threshold separately, in the whole enrichment (or deficit) curve (**Figure 3**). It does not provide any estimate of the significance of the whole curve, which is needed to estimate the significance of a sweep enrichment or deficit without making too many assumptions on how many sweeps are expected or how strong they are.

To address the increased type I and type II error risks of the bootstrap test, as well to get an unbiased significance estimate for whole enrichment curves, the second part of our pipeline conducts a false positive risk (FPR) analysis based on block-randomized genomes (**Enard and Petrov, 2020**). Briefly, we re-estimate many whole enrichment curves reusing the same mendelian disease and control non-disease genes used in the first part of the pipeline by the bootstrap test, but after having randomly shuffled the locations of genes or clusters of neighboring genes in sweeps at those disease and control non-disease genes. To do this, we order the disease and control non-disease genes as they appear in the genome. We then define blocks of neighboring genes, whose limits do not interrupt clusters of genes in the same putative sweep. Then, we randomly shuffle the order of these blocks. Because we do not cut any cluster of genes that might be in the same sweep, the resulting block-randomized genomes preserve the same clustering of the genes in the same putative sweeps as in the real genome. With this approach, we look at the exact same set of disease and control non-disease genes and just shuffle sweep locations between them. Thus, by using many block-randomized genomes, we can estimate the null expected range of whole enrichment curves while fully accounting for the extra variance expected from having a limited sample of control non-disease genes. We can then estimate a false positive risk (FPR) for the whole enrichment or deficit curve by comparing the real observed one with the distribution of random curves generated with block-randomized genomes.

To measure the FPR for a curve, we need to define a metric to compare the real curve with the randomly generated ones. In **Figure 3**, we show relative enrichments at each sweep rank threshold, the number of disease genes in sweeps divided by the number of control non-disease genes in sweeps. As a summary metric for the curve, we could then use the sum of the relative enrichments over all thresholds. However, the issue with this approach is that a relative enrichment is the same whether we have two disease genes in sweeps and one control non-disease gene in sweeps, or we have 200 disease genes in sweeps and 100 control non-disease genes in sweeps. Thus, although relative enrichments are convenient for visualization on a figure, they are not adequate to measure the FPR. Instead of the relative enrichment, we use as a score for estimating the FPR (FPR score) the difference between disease and non-disease genes, that is, the number of disease genes in sweeps, minus the average number of control non-disease genes across control sets built by the bootstrap test. We then get this score for a whole curve the sum of differences over all the rank thresholds. We use this sum of differences to estimate the enrichment or deficit curve FPR, as the proportion of block-randomized genomes where the FPR score (the sum of differences) exceeds the observed sum of differences for an enrichment (one minus this proportion for a deficit).

We can write this FPR score as follows. With t being the number t threshold belonging to T, the set of rank threshold numbers, $\mathrm{D_t}$ the number of disease genes in sweeps at threshold number t, and $\mathrm{C_t}$ the number of control genes in sweeps at threshold number t then:

$$\mathrm{FPRscore} = \sum_{\mathrm{t} \in \mathrm{T}} (\mathrm{D_t} - \mathrm{C_t})$$

Importantly, although so far we have described the case where we measure the FPR for one enrichment curve, nothing prevents us from calculating a single sum of differences over an entire group of enrichment or deficit curves. This way, we can measure a single FPR for any number of curves considered together. In our analysis, we measure a single FPR adding iHS and nS_L curves together, and also adding together the curves for 50kb, 100kb, 200kb, 500kb, and 1000kb windows (10 curves in total, 2 statistics*5 window sizes). Then, with W the set of window sizes, M the set of summary statistics used for detecting sweeps, and P the set of populations, we have

$$\mathrm{FPRscore} = \sum_{m\in M}\sum_{w\in W}\sum_{p\in P}\sum_{t\in T}\left(D_{t,m,w,p} - C_{t,m,w,p}\right)$$

In this FPR score, it is important to note that the strongest sweeps signals in the top ranks weight more than the weaker ranks. For example, if the top rank threshold used is 10 for the top 10 genes with the strongest sweep signals, then a disease gene in the top 10 is also in the top 20, or top 50, or top 100 or any other less restrictive defined rank threshold. Such a disease gene thus contributes to D_1, D_2, D_3 (….) up to D_n, where n is the number of rank thresholds used. It follows that the genes in the top rank threshold are weighted by a factor of n, and that the genes in the second top rank threshold, but not in the top rank threshold, are weighted by a factor of n-1, and so on up to the genes in the last nth rank threshold being weighted by a factor of 1. We can thus define d_t as the number of genes with ranks lower than rank threshold number t, but higher than rank threshold number t-1. Then, with c_t being the equivalent of d_t but for control genes:

$$\mathrm{FPRscore} = \sum_{m\in M}\sum_{w\in W}\sum_{p\in P}\sum_{t\in T}\left(d_{t,m,w,p} - c_{t,m,w,p}\right)*(n+1-t)$$

This weighting scheme is justified, as it makes sense to give more weight to stronger, and therefore higher confidence sweep signals.

Additional GERP confounding factors

To test if the confounding factors enumerated above properly account for purifying selection/selective constraint when comparing disease and control genes, we add several GERP-based metrics (***Davydov et al., 2010***) to the list of matched confounding factors, and check whether it makes a difference or not when estimating the sweep deficit in disease genes. Indeed, GERP provides a quantification of purifying selection in a given genomic window. So if adding GERP data to the confounding factors makes a difference for the sweep deficit, then it means that purifying selection was not already accounted for by the existing confounding factors. On the contrary, if adding GERP data to the confounding factors makes no difference for the sweep deficit, then it shows that the already included confounding factors, already meant to account for purifying selection such as the density of phastCons conserved elements or McVicker's B statistic (among others), already control well for purifying selection. As additional confounding factors, we consider the average GERP score in 50 kb and 500 kb windows centered on genes, as well as the density of GERP conserved elements also in 50 kb and 500 kb windows (four additional confounding factors in total) downloaded from the Sidow lab website for human genome assembly hg19 (http://mendel.stanford.edu/SidowLab/downloads/gerp/).

McDonald-Kreitman analysis of long-term coding adaptation

We use ABC-MK (https://github.com/jmurga/Analytical.jl, ***Moreno, 2021***) and GRAPES (https://github.com/BioPP/grapes, ***Dutheil, 2021***) to estimate the long-term rate of protein adaptation (both with the alpha and the omega_a estimates, see ***Figure 5***). As divergence data to get DN (number of non-synonymous fixed substitutions) and DS (number of synonymous fixed substitutions), we count the number of human-specific fixed substitutions (***Uricchio et al., 2019***). As non-synonymous and synonymous genetic variation data, we use the variants from the 1000 Genomes phase 3 for the 661 African individuals included (***Uricchio et al., 2019***). We use GRAPES to estimate the average strength of deleterious non-synonymous variants with the DisplGamma distribution (***Galtier, 2016***).

Sweep deficit at high and low recombination disease genes, and at high and low disease variant number disease genes

To generate ***Figure 6***, we separate disease genes in groups of approximately the same size based on their recombination rate and numbers of disease variants annotated in OMIM/Uniprot. We separate the disease genes into two groups of equal size, those with recombination lower than 1.137 cM/Mb, and those with recombination higher than this value. To count the disease variants at each disease gene, we count not only the OMIM/Uniprot disease variants for that gene, but also all the other OMIM/Uniprot disease variants that occur in a 500 kb window centered on that gene. We do this because the recessive deleterious variants form other nearby disease genes may also interfere with adaptation. Half of disease genes have less than five OMIM/Uniprot disease variants, and half have five or more.

Selection of the 1000 genes with the highest density of conserved elements

To compare highly constrained genes with other genes in the genome to check if purifying selection alone can create a sweep deficit, we first identify the 1000 genes with both complete confounding factors and complete sweep data. We then ask if these 1000 genes have a sweep deficit compared to other genes in the genome far enough from them (>300kb as for the disease vs. control genes). For this comparison, we cannot match the following confounding factors because they are themselves measures of, or related to purifying selection: the density of conserved elements, coding and regulatory densities, pN, pS, the number of gene neighbors and genes's genomic length. We still match all other confounding factors.

Population simulations of interference with and without a bottleneck

To investigate if reduced interference of recessive deleterious variants could explain the waker sweep deficit observed, we use SLiM (***Haller and Messer, 2019***) to simulate advantageous mutations in the presence of recessive deleterious mutations. To estimate the average fixation times and probabilities included in ***Table 1***, we get the average of these values over 1000 simulations each time. We simulate a one megabase genomic region. Across this entire region, we set the percentage of conserved sites that can experience deleterious mutations at 8%, believed to be the average proportions of sites that are conserved in the human genome (***Davydov et al., 2010***; ***Siepel et al., 2005***). At the center of the one megabase region, we include a 100 kb region where the proportion of sites that can experience deleterious mutations is higher, from 10%, to 20%, to 40%. This is meant to simulate the fact that we center our analysis on genes. The 10% proportion of sites that can experience deleterious mutations represents approximately the median GERP density of conserved elements in 100 kb windows centered on genes (***Figure 9***). The 20% and 40% proportions represent moderately elevated, and very strongly elevated values, respectively (***Figure 9***). We simulate a distribution of deleterious fitness effects (DFE) with a relatively flat profile across orders of magnitude for s, as found by recent human DFE estimates (***Kim et al., 2017***), with four negative selection coefficients, from weakly to strongly deleterious (s=-0.002,–0.02,−0.1,–0.5). The population size is initially set to 10,000 individuals and stays that way to simulate African populations (***Speidel et al., 2019***). After a burn-in period of 20,000 generations or 2N (deleterious mutations reach equilibrium faster than neutral ones), we set the population size to 1000 to simulate the demography of non-African populations (***Balick et al., 2015***; ***Speidel et al., 2019***). The advantageous mutation is then

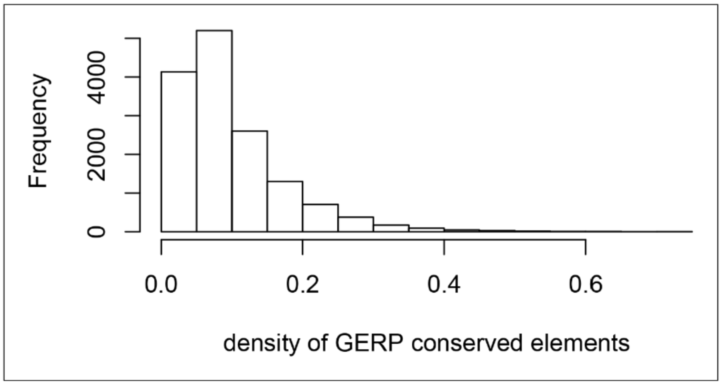


Figure 9. Density of GERP conserved elements around genes. The histogram represents the density of GERP conserved elements in 100kb windows centered on Ensembl protein-coding genes.

introduced 500 generations later (~15,000 years later in human evolution) Counting an Out of Africa bottleneck around 60,000 years ago, and if we count that in addition a sweep will take a few more tens of thousands of years to reach frequencies in the iHS or nSL sensitivity range, this brings us within the time window where iHS and nSL still have power (sweeps not older than 30,000 years). We rewind the simulation back to 20,500 generations as many times as necessary until one advantageous mutation goes to fixation. The average number of times we need to rewind the simulations gives us the probability of fixation. We simulate co-dominant advantageous mutations. Each simulation configuration is repeated 1000 times to get the estimates provided in ***Table 1***.

The recombination rate is set for the entire one megabase simulated locus at a low value of 0.1 cM/Mb, or ~10% of the average human recombination rate, to maximize the interference effect even during the simulated bottleneck, to see how much a bottleneck can decrease even strong interference.

Impact of functional differences between disease and non-disease genes on the sweep deficit

The sweep deficit at disease genes could be due to a different representation of gene functions at disease genes compared to control non-disease genes. In this case, disease genes would have less adaptation not because they are disease genes, but because the gene functions that are enriched among disease genes compared to non-disease happen to experience less adaptation. We can test this possibility using Gene Ontology (GO) (***Gene Ontology Consortium and Gene Ontology, 2021***) functional annotations as follows. If GO gene functions that are enriched in disease genes experience less adaptation independently of the disease status of genes, then we can predict that non-disease genes with these functions should also experience less adaptation than non-disease genes that do not have these GO functions. In total, we find that 3,097 GO annotations are enriched in disease genes compared to confounding factors-matched controls (bootstrap test $p \leq 0.01$). In our dataset, half of non-disease genes have 20 or more of these GO annotations, and half have less than 20 (very few have none). We find no difference in the sweep prevalence between the two groups (20 or more annotations vs. less than 20 annotations at least 300 kb away; FPR=0.15). The sweep deficit at disease genes is therefore unlikely to be due to the gene functions that are more represented in disease genes compared to controls. In addition, such a scenario would not explain the lack of sweep deficit observed at disease genes with high recombination rates and low numbers of disease variants (***Figure 6***).

Acknowledgements

We wish to thank Dan Shrider for helpful comments on the results presented in the manuscript.

Additional information

Funding

Funder	Grant reference number	Author
University of Arizona	startup to David Enard	David Enard

The funders had no role in study design, data collection and interpretation, or the decision to submit the work for publication.

Author contributions

Chenlu Di, Conceptualization, Data curation, Formal analysis, Validation, Investigation, Visualization, Writing - original draft, Writing - review and editing; Jesus Murga Moreno, Formal analysis, Methodology; Diego F Salazar-Tortosa, Writing - original draft, Writing - review and editing; M Elise Lauterbur, Writing - review and editing; David Enard, Conceptualization, Data curation, Software, Formal analysis, Supervision, Funding acquisition, Validation, Investigation, Visualization, Methodology, Writing - original draft, Project administration, Writing - review and editing, Interpretation of Results

Author ORCIDs

Diego F Salazar-Tortosa http://orcid.org/0000-0003-4289-7963
David Enard https://orcid.org/0000-0003-2634-8016

Decision letter and Author response

Decision letter https://doi.org/10.7554/eLife.69026.sa1
Author response https://doi.org/10.7554/eLife.69026.sa2

Additional files

Supplementary files

- Transparent reporting form

Data availability

The entire article is based on publicly available disease genes and genomic data. The disease genes used and sweep data and the sweep enrichment analysis pipeline (bootstrap test and False Positive risk estimation) with the required input files including the confounding factors are available at https://github.com/DavidPierreEnard/Gene_Set_Enrichment_Pipeline (copy archived at https://archive.softwareheritage.org/swh:1:rev:7b755c0c23dd4d7c3f54c4b53e74366e4041ac8f).

The following previously published datasets were used:

Author(s)	Year	Dataset title	Dataset URL	Database and Identifier
Auton A	2020	DisGeNET	https://www.disgenet.org/	DisGeNET, disgenet.org/
Auton A	2015	1000 Genomes	ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/phase3/data	1000 Genomes Phase 3, 1000genomes.ebi.ac.uk/vol1/ftp/phase3/data

References

Amberger JS, Bocchini CA, Scott AF, Hamosh A. 2019. Omim.org: leveraging knowledge across phenotype-gene relationships. *Nucleic Acids Research* **47**:D1038–D1043. DOI: https://doi.org/10.1093/nar/gky1151, PMID: 30445645

Assaf ZJ, Petrov DA, Blundell JR. 2015. Obstruction of adaptation in diploids by recessive, strongly deleterious alleles. *PNAS* **112**:2658–2666. DOI: https://doi.org/10.1073/pnas.1424949112, PMID: 25941393

Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, Korbel JO, Marchini JL, McCarthy S, McVean GA, Abecasis GR. 2015. A global reference for human genetic variation. *Nature* **526**:68–74. DOI: https://doi.org/10.1038/nature15393, PMID: 26432245

Balick DJ, Do R, Cassa CA, Reich D, Sunyaev SR. 2015. Dominance of deleterious alleles controls the response to a population bottleneck. *PLOS Genetics* **11**:e1005436. DOI: https://doi.org/10.1371/journal.pgen.1005436, PMID: 26317225

Barreiro LB, Quintana-Murci L. 2010. From evolutionary genetics to human immunology: how selection shapes host defence genes. *Nature Reviews. Genetics* **11**:17–30. DOI: https://doi.org/10.1038/nrg2698, PMID: 19953080

Birky CW, Walsh JB. 1988. Effects of linkage on rates of molecular evolution. *PNAS* **85**:6414–6418. DOI: https://doi.org/10.1073/pnas.85.17.6414, PMID: 3413105

Blekhman R, Man O, Herrmann L, Boyko AR, Indap A, Kosiol C, Bustamante CD, Teshima KM, Przeworski M. 2008. Natural selection on genes that underlie human disease susceptibility. *Current Biology : CB* **18**:883–889. DOI: https://doi.org/10.1016/j.cub.2008.04.074, PMID: 18571414

Booker TR, Yeaman S, Whitlock MC. 2020. Variation in recombination rate affects detection of outliers in genome scans under neutrality. *Molecular Ecology* **29**:4274–4279. DOI: https://doi.org/10.1111/mec.15501, PMID: 32535981

Chun S, Fay JC. 2011. Evidence for hitchhiking of deleterious mutations within the human genome. *PLOS Genetics* **7**:e1002240. DOI: https://doi.org/10.1371/journal.pgen.1002240, PMID: 21901107

Colquhoun D. 2019. The false positive risk: a proposal concerning what to do about P-values. *The American Statistician* **73**:192–201. DOI: https://doi.org/10.1080/00031305.2018.1529622

Davis AP, Grondin CJ, Johnson RJ, Sciaky D, Wiegers J, Wiegers TC, Mattingly CJ. 2021. Comparative toxicogenomics database (CTD): update 2021. *Nucleic Acids Research* **49**:D1138–D1143. DOI: https://doi.org/10.1093/nar/gkaa891, PMID: 33068428

Davydov EV, Goode DL, Sirota M, Cooper GM, Sidow A, Batzoglou S. 2010. Identifying a high fraction of the human genome to be under selective constraint using GERP++. *PLOS Computational Biology* **6**:e1001025. DOI: https://doi.org/10.1371/journal.pcbi.1001025, PMID: 21152010

Dutheil JY. 2021. grapes. *GitHub*. 3.0. https://github.com/BioPP/grapes

Enard D, Messer PW, Petrov DA. 2014. Genome-wide signals of positive selection in human evolution. *Genome Research* **24**:885–895. DOI: https://doi.org/10.1101/gr.164822.113, PMID: 24619126

Enard D, Petrov DA. 2020. Ancient RNA virus epidemics through the lens of recent adaptation in human genomes. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences* **375**: 20190575. DOI: https://doi.org/10.1098/rstb.2019.0575, PMID: 33012231

Eyre-Walker YC, Eyre-Walker A. 2014. The role of mutation rate variation and genetic diversity in the architecture of human disease. *PLOS ONE* **9**:e90166. DOI: https://doi.org/10.1371/journal.pone.0090166, PMID: 24587257

Ferrer-Admetlla A, Liang M, Korneliussen T, Nielsen R. 2014. On detecting incomplete soft or hard selective sweeps using haplotype structure. *Molecular Biology and Evolution* **31**:1275–1291. DOI: https://doi.org/10.1093/molbev/msu077, PMID: 24554778

Galtier N. 2016. Adaptive protein evolution in animals and the effective population size hypothesis. *PLOS Genetics* **12**:e1005774. DOI: https://doi.org/10.1371/journal.pgen.1005774, PMID: 26752180

Gene Ontology Consortium, Gene Ontology C. 2021. The gene ontology resource: enriching a GOld mine. *Nucleic Acids Research* **49**:D325–D334. DOI: https://doi.org/10.1093/nar/gkaa1113, PMID: 33290552

GTEx Consortium. 2020. The GTEx consortium atlas of genetic regulatory effects across human tissues. *Science* **369**:1318–1330. DOI: https://doi.org/10.1126/science.aaz1776, PMID: 32913098

Halldorsson BV, Palsson G, Stefansson OA, Jonsson H, Hardarson MT, Eggertsson HP, Gunnarsson B, Oddsson A, Halldorsson GH, Zink F, Gudjonsson SA, Frigge ML, Thorleifsson G, Sigurdsson A, Stacey SN, Sulem P, Masson G, Helgason A, Gudbjartsson DF, Thorsteinsdottir U, et al. 2019. Characterizing mutagenic effects of recombination through a sequence-level genetic map. *Science* **363**:eaau1043. DOI: https://doi.org/10.1126/science.aau1043, PMID: 30679340

Haller BC, Messer PW. 2019. SLiM 3: forward genetic simulations beyond the Wright-Fisher model. *Molecular Biology and Evolution* **36**:632–637. DOI: https://doi.org/10.1093/molbev/msy228, PMID: 30517680

Hill WG, Robertson A. 1966. The effect of linkage on limits to artificial selection. *Genetical Research* **8**:269–294. DOI: https://doi.org/10.1017/S0016672300010156, PMID: 5980116

Huber CD, Durvasula A, Hancock AM, Lohmueller KE. 2018. Gene expression drives the evolution of dominance. *Nature Communications* **9**:2750. DOI: https://doi.org/10.1038/s41467-018-05281-7, PMID: 30013096

Ittisoponpisan S, Alhuzimi E, Sternberg MJ, David A. 2017. Landscape of pleiotropic proteins causing human disease: structural and system biology insights. *Human Mutation* **38**:289–296. DOI: https://doi.org/10.1002/humu.23155, PMID: 27957775

Jain K. 2019. Interference effects of deleterious and beneficial mutations in large asexual populations. *Genetics* **211**:1357–1369. DOI: https://doi.org/10.1534/genetics.119.301960, PMID: 30700529

Johnson T, Barton NH. 2002. The effect of deleterious alleles on adaptation in asexual populations. *Genetics* **162**:395–411. DOI: https://doi.org/10.1093/genetics/162.1.395, PMID: 12242249

Kim PM, Korbel JO, Gerstein MB. 2007. Positive selection at the protein network periphery: evaluation in terms of structural constraints and cellular context. *PNAS* **104**:20274–20279. DOI: https://doi.org/10.1073/pnas.0710183104, PMID: 18077332

Kim BY, Huber CD, Lohmueller KE. 2017. Inference of the distribution of selection coefficients for new nonsynonymous mutations using large samples. *Genetics* **206**:345–361. DOI: https://doi.org/10.1534/genetics.116.197145, PMID: 28249985

Luisi P, Alvarez-Ponce D, Pybus M, Fares MA, Bertranpetit J, Laayouni H. 2015. Recent positive selection has acted on genes encoding proteins with more interactions within the whole human interactome. *Genome Biology and Evolution* **7**:1141–1154. DOI: https://doi.org/10.1093/gbe/evv055, PMID: 25840415

Maynard Smith J. 1976. What determines the rate of evolution? *The American Naturalist* **110**:973.

McDonald JH, Kreitman M. 1991. Adaptive protein evolution at the adh locus in *Drosophila*. *Nature* **351**:652–654. DOI: https://doi.org/10.1038/351652a0, PMID: 1904993

McVicker G, Gordon D, Davis C, Green P. 2009. Widespread genomic signatures of natural selection in hominid evolution. *PLOS Genetics* **5**:e1000471. DOI: https://doi.org/10.1371/journal.pgen.1000471, PMID: 19424416

Moreno JM. 2021. ABC-MK. *GitHub*. 1.6. https://github.com/jmurga/Analytical.jl

O'Reilly PF, Birney E, Balding DJ. 2008. Confounding between recombination and selection, and the ped/Pop method for detecting selection. *Genome Research* **18**:1304–1313. DOI: https://doi.org/10.1101/gr.067181.107, PMID: 18617692

Osada N, Mano S, Gojobori J. 2009. Quantifying dominance and deleterious effect on human disease genes. *PNAS* **106**:841–846. DOI: https://doi.org/10.1073/pnas.0810433106, PMID: 19139396

Otto SP. 2004. Two steps forward, one step back: the pleiotropic effects of favoured alleles. *Proceedings. Biological Sciences* **271**:705–714. DOI: https://doi.org/10.1098/rspb.2003.2635, PMID: 15209104

Peck JR. 1994. A ruby in the rubbish: beneficial mutations, deleterious mutations and the evolution of sex. *Genetics* **137**:597–606. DOI: https://doi.org/10.1093/genetics/137.2.597, PMID: 8070669

Piñero J, Ramírez-Anguita JM, Saüch-Pitarch J, Ronzano F, Centeno E, Sanz F, Furlong LI. 2020. The DisGeNET knowledge platform for disease genomics: 2019 update. *Nucleic Acids Research* **48**:D845–D855. DOI: https://doi.org/10.1093/nar/gkz1021, PMID: 31680165

Quintana-Murci L. 2016. Understanding rare and common diseases in the context of human evolution. *Genome Biology* **17**:225. DOI: https://doi.org/10.1186/s13059-016-1093-y, PMID: 27821149

Quintana-Murci L, Barreiro LB. 2010. The role played by natural selection on mendelian traits in humans. *Annals of the New York Academy of Sciences* **1214**:1–17. DOI: https://doi.org/10.1111/j.1749-6632.2010.05856.x, PMID: 21175682

Sabeti PC, Schaffner SF, Fry B, Lohmueller J, Varilly P, Shamovsky O, Palma A, Mikkelsen TS, Altshuler D, Lander ES. 2006. Positive natural selection in the human lineage. *Science* **312**:1614–1620. DOI: https://doi.org/10.1126/science.1124309, PMID: 16778047

Schrider DR. 2020. Background selection does not mimic the patterns of genetic diversity produced by selective sweeps. *Genetics* **216**:499–519. DOI: https://doi.org/10.1534/genetics.120.303469, PMID: 32847814

Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, Rosenbloom K, Clawson H, Spieth J, Hillier LW, Richards S, Weinstock GM, Wilson RK, Gibbs RA, Kent WJ, Miller W, Haussler D. 2005. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Research* **15**:1034–1050. DOI: https://doi.org/10.1101/gr.3715005, PMID: 16024819

Smith NG, Eyre-Walker A. 2003. Human disease genes: patterns and predictions. *Gene* **318**:169–175. DOI: https://doi.org/10.1016/s0378-1119(03)00772-8, PMID: 14585509

Smith JM, Haigh J. 1974. The hitch-hiking effect of a favourable gene. *Genetical Research* **23**:23–35. DOI: https://doi.org/10.1017/S0016672300014634, PMID: 4407212

Spataro N, Rodríguez JA, Navarro A, Bosch E. 2017. Properties of human disease genes and the role of genes linked to mendelian disorders in complex disease aetiology. *Human Molecular Genetics* **26**:489–500. DOI: https://doi.org/10.1093/hmg/ddw405, PMID: 28053046

Speidel L, Forest M, Shi S, Myers SR. 2019. A method for genome-wide genealogy estimation for thousands of samples. *Nature Genetics* **51**:1321–1329. DOI: https://doi.org/10.1038/s41588-019-0484-x, PMID: 31477933

Stern AJ, Speidel L, Zaitlen NA, Nielsen R. 2021. Disentangling selection on genetically correlated polygenic traits via whole-genome genealogies. *American Journal of Human Genetics* **108**:219–239. DOI: https://doi.org/10.1016/j.ajhg.2020.12.005, PMID: 33440170

Torgerson DG, Boyko AR, Hernandez RD, Indap A, Hu X, White TJ, Sninsky JJ, Cargill M, Adams MD, Bustamante CD, Clark AG. 2009. Evolutionary processes acting on candidate cis-regulatory regions in humans inferred from patterns of polymorphism and divergence. *PLOS Genetics* **5**:e1000592. DOI: https://doi.org/10.1371/journal.pgen.1000592, PMID: 19662163

Uricchio LH, Petrov DA, Enard D. 2019. Exploiting selection at linked sites to infer the rate and strength of adaptation. *Nature Ecology & Evolution* **3**:977–984. DOI: https://doi.org/10.1038/s41559-019-0890-6, PMID: 31061475

Voight BF, Kudaravalli S, Wen X, Pritchard JK. 2006. A map of recent positive selection in the human genome. *PLOS Biology* **4**:e72. DOI: https://doi.org/10.1371/journal.pbio.0040072, PMID: 16494531

Xu J, Zhang J. 2014. Why human disease-associated residues appear as the wild-type in other species: genome-scale structural evidence for the compensation hypothesis. *Molecular Biology and Evolution* **31**:1787–1792. DOI: https://doi.org/10.1093/molbev/msu130, PMID: 24723421

Appendix C -- Stability evolution as a major mechanism of human protein adaptation in response to viruses

Authors: Chenlu Di[1] and David Enard[1]

Affiliations: 1. Department of Ecology and Evolutionary Biology, University of Arizona, Tucson AZ 85719, USA.

Corresponding author: David Enard denard@arizona.edu

Abstract

Pathogens were a major driver of genetic adaptation during human evolution. Viruses in particular were a dominant driver of adaptation in the thousands of proteins that physically interact with viruses (VIPs for Virus-Interacting Proteins). This however poses a conundrum. The best understood cases of virus-driven adaptation in specialized immune antiviral factors or in host viral receptors are numerically vastly insufficient to explain abundant adaptations in VIPs. What adaptive mechanisms can then at least partly close this gap? VIPs tend to be broadly conserved proteins with conserved host native molecular functions. Because many amino acid changes in a protein can alter its stability –the balance between the folded and unfolded forms of a protein-- without destroying conserved native activities, here we ask if stability evolution was an important mechanism of virus-driven human protein adaptation. Using predictions of protein stability changes based on Alphafold 2 structures and validated by multiple lines of evidence, we find that amino acid changes that altered stability experienced highly elevated adaptative evolution in VIPs, compared to changes with a weaker impact on stability. We further find that RNA viruses, rather DNA viruses, predominantly drove strong adaptation through stability changes in VIPs. We also observe that stability in immune antiviral VIPs preferentially evolved under directional selection. Conversely, stability in proviral VIPs needed by viruses evolved under compensatory evolution following viral epidemics. Together, these results suggest that stability evolution, and thus functional host protein abundance evolution, was a prominent mechanism of host protein adaptation during viral epidemics.

Introduction

Virus-driven adaptation includes some of the most compelling and best understood examples of protein adaptation on a mechanistic level in human and other mammals. Host proteins with well understood adaptation in response to viruses notably include prominent immune antiviral proteins such as TRIM5 (Sawyer et al., 2005), PKR (Elde et al., 2009) and APOBEC3G (Compton et al., 2012; Sawyer et al., 2004; Yang et al., 2020) that directly attack viral molecules and trigger downstream molecular processes meant to destroy them. Other well understood cases include adaptive amino acid changes at the host-virus contact interface that decrease the binding affinity of viruses for their host cell surface receptors. The filovirus receptor NPC1 in bats is such an example, with one specific amino acid position in the contact interface explaining most of bat species-specific susceptibility (Ng et al., 2015). Interestingly however, viral receptors can also harbor many positively selected amino acid changes well outside of the contact interface with no identified mechanism to date to explain their adaptive nature (Enard et al., 2016; Uricchio et al., 2019). This is the case of ANPEP, a coronavirus receptor, where many selected sites well outside of the contact interface make it one of the most positively selected proteins in mammals (Enard *et al.*, 2016).

More generally, recent evidence shows that, far from being restricted to specialized immune antiviral factors or viral receptors, adaptation has been abundant among thousands of currently known VIPs, including proviral ones, with a more than two-fold increase in adaptation compared to proteins that do no interact with viruses (noted non-VIPs) (Enard *et al.*, 2016; Uricchio *et al.*, 2019). VIPs in particular experienced highly increased rates of strong adaptation, where the viral selective pressure was likely intense and led to rapid fixation of the advantageous amino acid changes (Enard and Petrov, 2020; Uricchio *et al.*, 2019). In VIPs, the surface contact interfaces with viruses only represent a very small proportion of all the amino acids, typically a few percent (Wierbowski et al., 2021). For example, the contact interface of ACE2 with SARS-CoV-2 represents only 21 amino acids, or 2.6% of ACE2 (Yan et al., 2020).

What protein evolution mechanisms then may explain widespread positive selection in VIPs not restricted to contact interfaces? What protein attributes changed that provided a selective advantage in response to viruses, while preserving the host native functions of VIPs? We can sort protein mutations into four non-exclusive categories that may be selected during viral epidemics: (i) mutations that affect protein conformation, (ii) mutations that alter the host-virus contact interface, (iii) mutations that change the host native molecular functions, and (iv) mutations that change the stability, i.e the abundance of the folded, functional form of VIPs (Dasmeh et al., 2013). Even though we do not exclude that any of these different mechanisms occurred in VIPs, it is important to consider that VIPs tend to be broadly conserved across mammals and beyond (Castellano, 2019). This is true irrespective of whether VIPs were discovered through low-throughput hypothesis-driven virology studies (75% with orthologs across mammals; Methods), or through high throughput mass-spectrometry assays that are blind to previous knowledge (74% with orthologs) (Batra et al., 2018; Jager et al., 2011; Shah et al., 2018; Watanabe et al., 2014). This excludes any difference in research attention and corresponding publications artificially biasing VIPs towards more conserved genes. This greater conservation might make repeated protein conformation changes less likely, because recurrent, notable conformational changes may be incompatible with the broad conservation of VIPs and their host native molecular functions (Konate et al., 2019; Lee et al., 2007; Radivojac et al., 2013). The latter might also limit adaptation through amino acid changes that modify the host native molecular activities of VIPs. Mutations at catalytic residues are very likely to disrupt these activities (Firnberg et al., 2014). In addition, similarly to contact interfaces, catalytic residues responsible for these activities only represent a small percentage of a protein (less than 5%), mainly restricted to protein surfaces (Nelson et al., 2013).

Conversely, protein stability changes may represent a good candidate mechanism for virus-driven adaptation. Protein stability can be defined as the thermodynamic balance between the folded functional form of a protein, and the unfolded, non-functional form that is typically targeted for degradation (Clausen et al., 2019). Protein stability changes are quantified as ΔΔG, defined as the change in the thermodynamic quantity ΔG, the Gibbs free energy (in kcal/mol) that determines the stability of a protein. Amino acid changes with positive ΔΔG are destabilizing, negative ones are stabilizing. A change of ΔΔG of one kcal/mol roughly corresponds to a fivefold change in folded protein abundance at body temperature (Dasmeh *et al.*, 2013; Serohijos et al., 2013). Protein stability is a major determinant of the abundance of the functional form of a protein in cells, and many known disease-causing variants destabilize and decrease the abundance of essential proteins (Nielsen et al., 2017; Scheller et al., 2019; Stein et al., 2019). We can then imagine how VIP stability changes may be advantageous during a viral epidemic. For example, lower stability and thus lower abundance of a proviral factor required by a virus to replicate may be advantageous for the host.

Experimental data, notably from the disease variants literature (Stein *et al.*, 2019) and from protein design studies (Goldenzweig and Fleishman, 2018), shows that many amino acid changes in many parts of a protein can change stability (Li et al., 2020; Serohijos and Shakhnovich, 2014). The large number of possible stability-altering amino acid changes may thus in theory match the large number of adaptive amino acid changes observed in VIPs. Furthermore, stability evolution may be easier in otherwise conserved proteins (Dasmeh *et al.*, 2013), including VIPs; a large pool of possible stability-altering amino acid changes might (i) happen outside of evolutionarily conserved active sites of VIPs responsible for conserved host native functions, and (ii) a large pool might make compensatory evolution easier in the case where a host native function is no longer optimal after a change in folded protein abundance (see below, Compensatory stability evolution in proviral VIPs). Finally, stability changes are particularly likely for amino-acid changes that occur in the buried parts of protein structures, and below we observe that many adaptive amino acid changes in VIPs occurred within the buried part of VIPs (Results below).

Here, we focus on protein stability changes as a possible mechanism of adaptation in response to viral epidemics. We find that amino acid substitutions in human evolution that altered protein stability substantially were much more likely to be adaptive in VIPs, compared to amino acid changes that changed stability to a lesser extent. We further find that VIPs have experienced more adaptation in the buried than in the outer parts of their protein structure. This is in agreement with the fact that changes of buried residues are more likely to affect protein stability, and we confirm indeed that the elevated adaptation in the buried part of VIPs is driven by those amino acid changes that modify protein stability. We further observe that antiviral and proviral VIPs have experienced stability-driven adaptation, but in different ways. Immune antiviral VIPs whose main function is to impede viruses have overall experienced directional stability evolution, likely as a result of changing optima depending on shifting pathogen pressures over evolutionary time. Conversely, most proviral VIPs do not have immune functions but have many broadly conserved non-immune host native functions under stabilizing selection. As expected, likely due to these conserved native functions, non-immune proviral VIPs experienced predominantly compensatory stability evolution. Together, our result suggest that protein stability evolution may have been an important mechanism of host adaptation in response to viruses. Our results further suggest a model where further compensatory evolution may occur after viral epidemics, to bring proviral VIPs back to their initial, and perhaps more optimal stability in the absence of viral selective pressure.

Results

To test if protein stability was a determinant of adaptation in VIPs, we use a recent version of the McDonald-Kreitman test (McDonald and Kreitman, 1991) called ABC-MK (Uricchio *et al.*, 2019) to estimate the proportion of amino acid substitutions that were adaptive among those that significantly altered stability, compared to those that did not during human evolution since divergence with chimpanzees. McDonald-Kreitman approaches estimate the percentage of nonsynonymous substitutions that were adaptive by contrasting the total observed number of nonsynonymous substitutions with what this number would be under neutrality, if adaptation had not occurred. This neutral expectation can be derived from the Site-Frequency-Spectrum of present non-synonymous variants, while at the same time controlling for past fluctuations of the mutation rate by contrasting nonsynonymous and synonymous substitutions and variants from the same coding sequences (Uricchio *et al.*, 2019). We use ABC-MK with coding sequence substitutions that occurred specifically in the human branch since divergence with chimpanzees (Methods), and variants from the 1,000 Genome projects groups located in Africa (Genomes Project et al., 2015), as described in (Uricchio *et al.*, 2019). Notably, ABC-MK has the ability to distinguish between weak and strong adaptation (Uricchio *et al.*, 2019) (Methods), and here we use this functionality as it was previously shown that viruses drive particularly strong adaptation (Enard and Petrov, 2020; Uricchio *et al.*, 2019).

We estimate stability changes caused by amino acid substitutions and variants (Figure 1A,B) with the computational method Thermonet (Li *et al.*, 2020) used on high confidence Alphafold 2 protein structures (Jumper et al., 2021) (available at https://alphafold.ebi.ac.uk/; Methods). Thermonet uses a deep convolutional network trained on experimental stability data to predict the stability changes caused by amino acid substitutions or variants within a given protein structure. Experimental measures of stability changes are not currently available beyond a limited number of human proteins (Pancotti et al., 2022).

Although Thermonet provides computational estimates, it was recently shown to have good, balanced performance when benchmarked with new experimental stability data that was not used for its convolutional network training (Pancotti *et al.*, 2022). We further validate Thermonet estimates in multiple ways. First, Thermonet correctly identifies a known destabilizing genetic variant (R105G) in antiviral VIP APOBEC3H as strongly destabilizing (stability change $\Delta\Delta G = 0.71$ kcal/mol) (Chesarino and Emerman, 2020). Second and most importantly, we find strong, highly significant evidence of compensatory evolution of protein stability in non-immune proviral VIPs with multiple amino acid changes, where it is the most expected (see below). It would not be possible to observe such compensatory evolution if Thermonet estimates were not sufficiently correlated with the actual stability changes that occurred during human protein evolution.

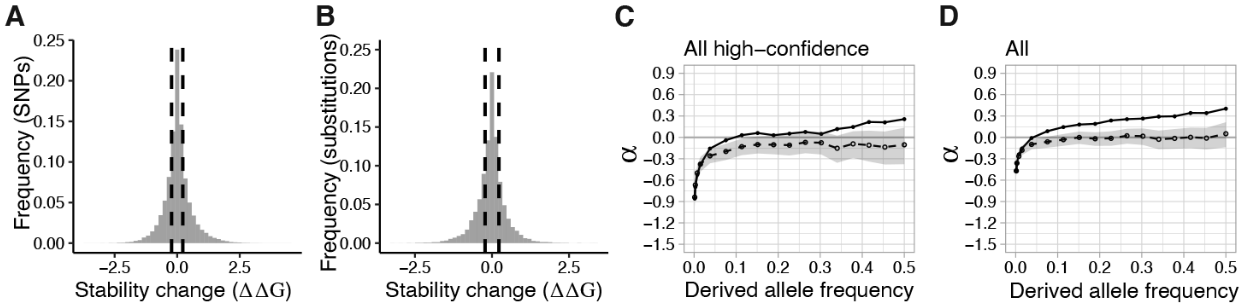


Figure 1. Distributions of stability changes and adaptation in VIPs and control non-VIPs.
A) Distribution of Thermonet estimated ΔΔG values for human nonsynonymous variants (Methods). The dashed lines represent the ΔΔG=-0.225 and ΔΔG=0.225 limits. B) Distribution of Thermonet estimated ΔΔG values for human nonsynonymous fixed substitutions (Methods). C) α-curves used by ABC-MK to estimate the proportion α of adaptive nonsynonymous substitutions (y-axis; Methods) using only variants and substitutions from high confidence Alphafold 2 residues (pLDDT≥70) with stability predictions (for the nonsynonymous ones; Methods) in VIPs (continuous curve) and control non-VIPs (dashed line for the average and grey area for the 95% confidence interval). D) Same as C) but using all variants and substitutions known at VIPs and control non-VIPs, not just the ones restricted at high confidence Alphafold 2 residues.

Out of ~5,500 VIPs known to date (Table S1), 2,900 have high confidence Alphafold 2 structures and can be compared with 5,700 non-VIPs also with high confidence structures (Table S1; Methods). These include only VIPs and non-VIPs with orthologs across mammals (Methods) since we previously showed that viruses increase adaptation more specifically in VIPs that are conserved across mammals and beyond (Castellano, 2019). This also limits the risk of confounding due to gene age (Moutinho et al., 2022). We compare VIPs with non-VIPs to highlight the evolutionary patterns that are specific to VIPs (Figure 1C,D).

Here, we specifically ask if stability-altering substitutions have (i) experienced more positive selection in VIPs than in non-VIPs, and (ii) more positive selection than substitutions that altered stability to a lesser extent. We do not compare VIPs with any non-VIPs, but match VIPs with control non-VIPs that look like VIPs in ways other than interacting with viruses, that could affect adaptation and confound the comparison (Methods). The matching is done with a previously described bootstrap procedure (Di et al., 2021; Enard and Petrov, 2020) and includes many potential confounders (Methods). We then estimate how significant increased adaptation is in VIPs by repeating the measurements of adaptation in sets of the same size as the VIPs set, but made of randomly sampled VIPs and non-VIPs. This effectively estimates an unbiased false discovery rate by generating null distributions of estimates of adaptation expected if there was no impact of interactions with viruses (FDR; Methods).

Considering only substitutions and variants with predicted stability changes (Methods), the 2,900 VIPs with good Alphafold 2 structures have experienced 3.6 times more adaptation than control non-VIPs, with 27.5% of all amino acid substitutions estimated to be adaptive vs. only 7.4% for control non-VIPs (Figure 1C; the proportion of adaptive substitutions is noted α). This includes 19% of strongly selected substitutions in VIPs, compared to only 3% in control non-VIPs. Considering all substitutions and variants including those with no stability change prediction (Methods), the 2,900 VIPs have experienced 3.5 times more adaptation than control non-VIPs, with 35% of all amino acid substitutions estimated to be adaptive vs. only 10% for control non-VIPs (Figure 1D). This includes 29% of strongly selected substitutions in VIPs compared to only 4% in control non-VIPs.

Abundant stability-changing, adaptive substitutions in VIPs

In order to study the impact of protein stability on adaptation, we split nonsynonymous variants and fixed substitutions in the human branch according to stability changes (given by their $|\Delta\Delta G|$ absolute value in kcal/mol) into two groups: those below the variants' absolute stability change median ($|\Delta\Delta G|<0.225$), and those above ($|\Delta\Delta G|>0.225$). This also happens to roughly correspond to a transition point in the leptokurtic distribution of $\Delta\Delta G$ among variants or substitutions (Figure 1A,B), between a large concentration of SNPs around $\Delta\Delta G=0$ and more spread out, symmetric tails on both sides. We do not try to distinguish between destabilizing ($\Delta\Delta G>0$) and stabilizing ($\Delta\Delta G<0$) variants that decrease or increase protein stability, respectively, but instead we use the absolute $|\Delta\Delta G|$. Indeed, VIPs can be proviral or antiviral, making it tempting to assign an expectation of what stability change direction should be adaptive. One might expect that it is advantageous to increase the stability and therefore increase the abundance of an antiviral VIP indefinitely. This may however have adverse effects, notably in the case of antiviral VIPs related to inflammation (Yong et al., 2022). More importantly, we found through a large, multi-year effort of manual curation of 4,477 virology publications that antiviral VIPs for a virus are very frequently subverted and made proviral by the same or other viruses (Table S1, Methods), thus making it difficult to assign a specific expected direction of adaptive $\Delta\Delta G$. Retasking by viruses to accomplish proviral steps was also recently noticed to occur even with very prominent antiviral factors (King and Mehle, 2022; Tran et al., 2020). Finally, assigning an adaptive $\Delta\Delta G$ direction might also be made difficult by the fact that there might be compensatory evolution of stability, especially in proviral VIPs that have important, conserved non-immune functions in the host (immune antiviral VIPs are more specialized in attacking viruses and likely less impeded by other native host functions, see below).

Thus, we estimate the rate of adaptive substitutions with ABC-MK for just two categories, all substitutions with absolute stability changes below, and all substitutions above the variants' absolute median for |ΔΔG|, respectively (Figure 1A,B). We cut the data this way to compare two groups of equal sizes, and thus similarly powered comparisons between VIPs and non-VIPs. We call the group above the median the Large Stability Changes group (LSC), and the group below the Small Stability Changes group (SSC). We only include variants and substitutions at amino acids with high Alphafold 2 structural prediction confidence (Alphafold 2 accuracy score pLDDT≥70; Methods).

We find much higher rates of adaptation for the LSC group than for the SSC group. ABC-MK estimates that 34% or 250 of the 737 LSC substitutions were advantageous in VIPs, versus only 15% or 125 of the 832 SSC substitutions (Figure 2A,B). Thus, a substantial majority (67% or 250/375) of advantageous nonsynonymous substitutions with ΔΔG predictions in VIPs altered protein stability. The estimated 34% of adaptive LSCs in VIPs is also much higher than the estimated 10% in sets built randomly sampling both VIPs and control non-VIPs to estimate a false discovery rate (FDR=$3*10^{-4}$; Methods). In contrast, VIPs only have a marginally higher adaptive percentage of SSCs compared to random sets (15% vs. 7% respectively, FDR=0.04).

By comparing adaptive proportions in VIPs and control non-VIPs, we can also estimate the amount of adaptation that can be attributed to interactions with viruses. In control non-VIP sets, 8% LSC substitutions were advantageous on average, more than four times less than in VIPs. A total of 34% minus 8% (26%), or 192 of the 737 LSC substitutions in VIPs, can thus be attributed to virus-driven adaptation. SSC substitutions were also 8% advantageous in control non-VIPs, implying 58 (15%-8%) of the 832 SSC substitutions in VIPs can be attributed to viruses. Thus, 77% (192/(192+58)) of the adaptative substitutions attributable to viruses in VIPs changed stability above the |ΔΔG| median.

ABC-MK also has the ability to distinguish between strong and weak past protein adaptation (Methods). We find that in VIPs, it is in particular the rate of strong adaptation that is increased among LSC substitutions, with an estimated 27% of substitutions being strongly advantageous, vs. only 5% in random sets and 3% in control non-VIPs (Figure 2A; FDR=$4*10^{-4}$). In VIPs SSC substitutions are 9% strongly advantageous compared to 4% in control non-VIPs. Using the same logic as above, this means that 177 LSC and 42 SSC substitutions can be attributed to strong, virus-driven adaptation, respectively, with 81% of strong virus-driven adaptation then involving stability changes above the |ΔΔG| median. This corresponds to the expected evolutionary pattern if adaptative evolution with LSCs in VIPs was indeed driven by viruses, since we previously showed that virus-driven adaptation was disproportionately strong adaptation (Enard and Petrov, 2020; Uricchio *et al.*, 2019).

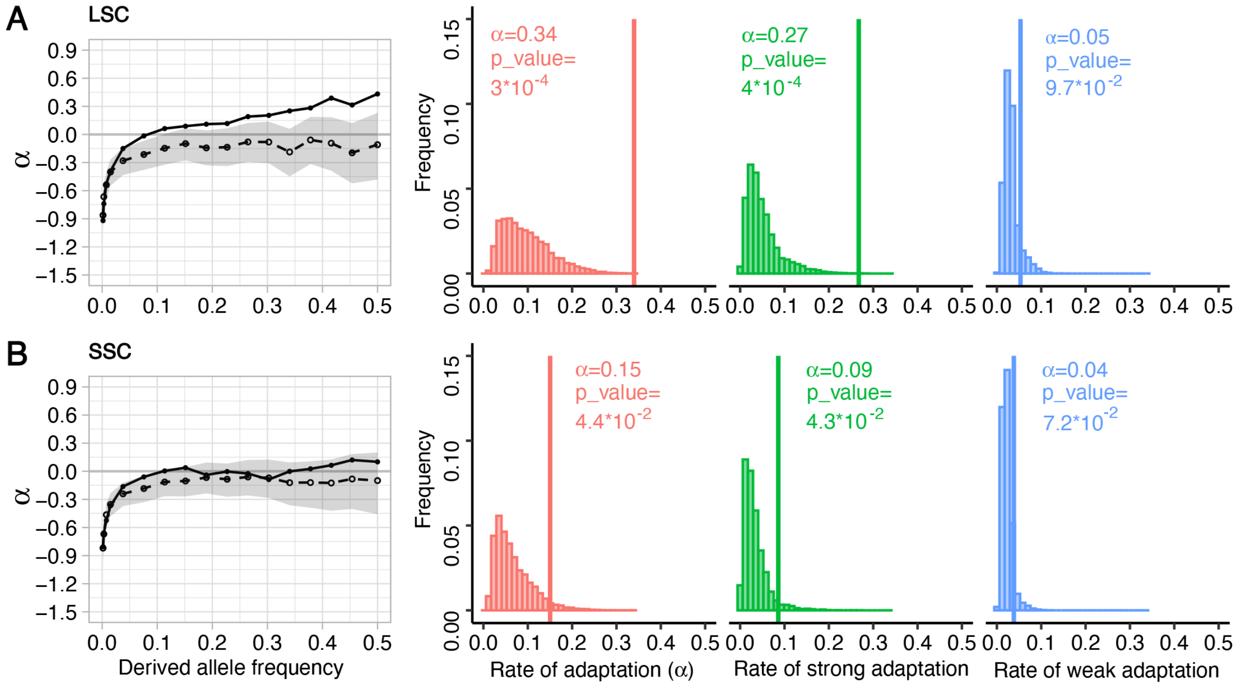


Figure 2. Stronger adaptation through large stability changes in VIPs.
We compared adaptation among LSC substitutions (A) with adaptation among SSC substitutions (B). The α-curves used by ABC-MK to estimate α are as in Figure 1C. The red graph shows the total α in VIPs (vertical line) vs. the expected null distribution when randomly sampling VIPs and control non-VIPs. Green graph: same as red but only for strong adaptation as measured by ABC-MK. Blue graph: same as red but only for weak adaptation as measured by ABC-MK.

Stability, rather than other protein attributes explains increased VIP adaptation at buried residues

We further assess potential confounding protein attributes that might be the primary selected attributes, instead of stability. Specifically, the effect of a given residue on stability is known to depend strongly on the position of the residue in the protein structure. More hydrophobic residues closer to the buried core of proteins are known to contribute disproportionately to protein stability (Nick Pace et al., 2014). In agreement with this, LSCs tend to be more buried, further from the structural surface of VIPs than SSCs. Using Relative Solvent Accessibility (RSA) measured by DSSP (Kabsch and Sander, 1983) (Methods) as a measure of how buried or exposed to the surface a given protein residue is, we observe that LSCs have an overall median RSA of 0.23, vs. 0.4 for SSCs (see also Figure S1). This however raises the possibility that buried residues in VIPs have elevated adaptation regardless of their impact on stability. VIPs might experience more adaptation at buried residues because changes in protein conformation and/or changes in allostery, where functional signals are transmitted through protein structure (Nelson *et al.*, 2013; Swain and Gierasch, 2006), were the primary adaptive protein attributes. The increased rate of adaptation of LSCs might then only be a secondary, bystander effect of the fact that substitutions in buried parts of proteins tend to also affect protein stability to a greater extent. Accordingly, we find that VIPs experienced more adaptive substitutions below than above the overall median RSA (RSA=0.32) in their structures, with 39% below or 262 adaptive substitutions, vs. 19% above or 173 adaptive substitutions estimated by ABC-MK, respectively (Figure 3A,B). The increased adaptation below the median RSA strongly sets VIPs apart from control non-VIPs (random sets FDR=2*10^{-4}; Figure 3B), and this is even more the case for strong adaptation (Figure 3B). Notably, LSCs were less adaptive in high RSA (Figures 3C and S2) than in small RSA (Figures 3D and S2) parts of VIPs, which excludes that adaptive LSCs overall can represent a bystander effect of adaptation at contact interfaces or molecular activities located at the protein surface.

We further observe that the increased adaptation in the low RSA parts of VIPs is strongly dependent on protein stability. While LSCs below their RSA median or the SSCs' RSA median (0.23 and 0.4, respectively) have strongly elevated adaptation compared to random sets (Figures 3D and S2), all SSCs in VIPs below a 0.4 RSA only have 10% adaptive substitutions, a percentage that is not different from random FDR sets (FDR=0.31) and lower than the 20% adaptive SSCs in high RSA parts of VIPs (Figure 3E,F). Together, these results suggest that protein stability changes are the primary driver of increased adaptation, especially strong adaptation, in more buried, lower RSA parts of VIPs, rather than changes of protein conformation and/or allostery. This further narrows down the possible mechanistic explanations to protein stability, since more buried parts of VIPs are also less likely to include contact interfaces or active catalytic pockets typically found at the surface of proteins.

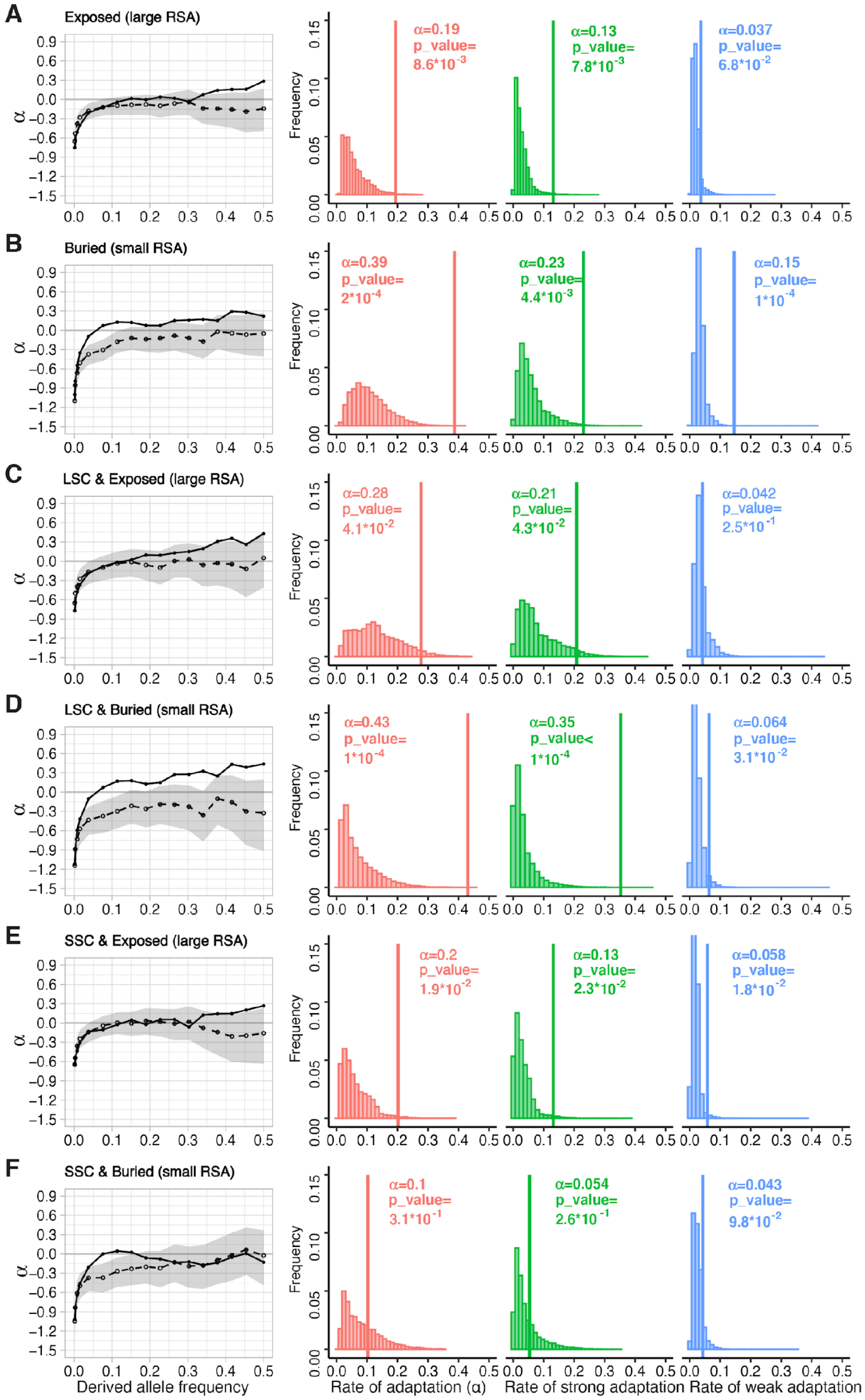
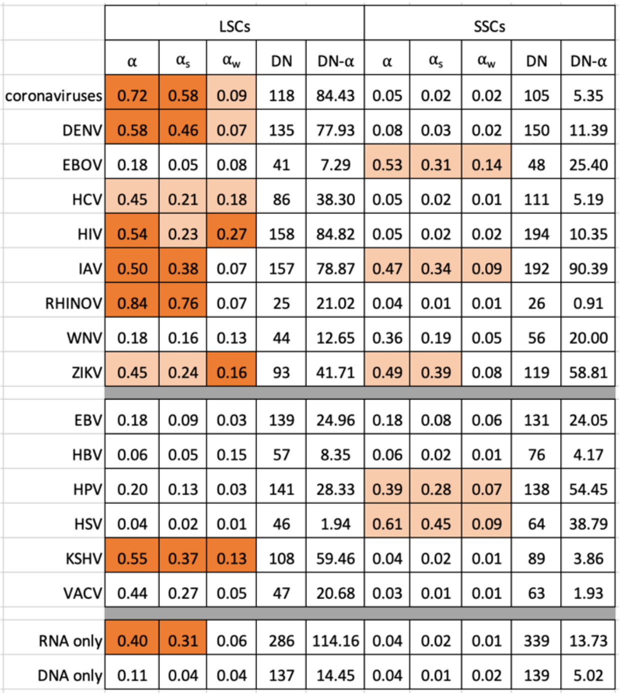


Figure 3. Adaptation as a function of position in the protein structure.
Legend same as Figure 2. A) adaptation (LSC+SSC) in more exposed parts of VIPs and control non-VIPs with RSA≥0.32, the median for LSC and SSC combined. B) same as A) but for more buried parts with RSA<0.32. C) LSC adaptation above the LSCs' RSA median of 0.23. D) LSC adaptation below the LSCs' RSA median of 0.23. E) SSC adaptation above the SSCs' RSA median of 0.4. F) SSC adaptation below the SSCs' RSA median of 0.4. All four subgroups represented in C, D, E and F have separating

RSA medians such that the four subgroups have similar sizes (and thus variance and power to test hypotheses).

Increased adaptation through large stability changes in more RNA than DNA viruses

Having found broad patterns across VIPs, we then estimate which viruses have a particularly increased associated percentage of adaptive LSC substitutions. We previously found that RNA viruses whose genomes are coded by RNA, rather than DNA viruses, have driven particularly strong and abundant selection at their respective VIPs during human evolution (Enard and Petrov, 2018; 2020; Souilmi et al., 2021). If abundant adaptive evolution of LSCs in VIPs is indeed a hallmark of virus-driven adaptation, we should then be able to observe that this is particularly the case for VIPs of RNA viruses. In agreement with this prediction, we find seven RNA viruses out of nine tested with significantly increased adaptive LSCs in their specific VIPs (compared to random FDR sets, Table 1), compared to only one of the six tested DNA viruses, Kaposi's sarcoma Herpesvirus KSHV (Table 1). Five of the seven significant RNA viruses have VIPs with very strongly elevated percentages of adaptive LSCs at or above 50%, including coronaviruses (72%, Figure 4A,B), rhinovirus RV-B14 (84%, Figure 4C,D), Dengue Virus (58%, Figure S3A), Human Immunodeficiency Virus HIV (54%, Figure S3B) and Influenza A Virus (50%, Figure S3C). In the VIPs of coronaviruses, dengue virus, Hepatitis C virus, HIV, rhinovirus RV-B14 and KSHV, the difference in adaptation between LSCs and SSCs is extreme (Table 1 and Figures 4 and S3). All VIPs that interact only with RNA viruses had a significant elevation of adaptive LSCs (40%, Table 1), while all VIPs that interact only with DNA viruses did not (10%, Table 1). Together, these results show that RNA viruses were the predominant drivers of strong adaptation through large stability changes in VIPs.



	LSCs					SSCs				
	α	α_s	α_w	DN	DN-α	α	α_s	α_w	DN	DN-α
coronaviruses	0.72	0.58	0.09	118	84.43	0.05	0.02	0.02	105	5.35
DENV	0.58	0.46	0.07	135	77.93	0.08	0.03	0.02	150	11.39
EBOV	0.18	0.05	0.08	41	7.29	0.53	0.31	0.14	48	25.40
HCV	0.45	0.21	0.18	86	38.30	0.05	0.02	0.01	111	5.19
HIV	0.54	0.23	0.27	158	84.82	0.05	0.02	0.02	194	10.35
IAV	0.50	0.38	0.07	157	78.87	0.47	0.34	0.09	192	90.39
RHINOV	0.84	0.76	0.07	25	21.02	0.04	0.01	0.01	26	0.91
WNV	0.18	0.16	0.13	44	12.65	0.36	0.19	0.05	56	20.00
ZIKV	0.45	0.24	0.16	93	41.71	0.49	0.39	0.08	119	58.81
EBV	0.18	0.09	0.03	139	24.96	0.18	0.08	0.06	131	24.05
HBV	0.06	0.05	0.15	57	8.35	0.06	0.02	0.01	76	4.17
HPV	0.20	0.13	0.03	141	28.33	0.39	0.28	0.07	138	54.45
HSV	0.04	0.02	0.01	46	1.94	0.61	0.45	0.09	64	38.79
KSHV	0.55	0.37	0.13	108	59.46	0.04	0.02	0.01	89	3.86
VACV	0.44	0.27	0.05	47	20.68	0.03	0.01	0.01	63	1.93
RNA only	0.40	0.31	0.06	286	114.16	0.04	0.02	0.01	339	13.73
DNA only	0.11	0.04	0.04	137	14.45	0.04	0.01	0.02	139	5.02

Table 1. LSC and SSC adaptation for the VIPs and control non-VIPs of 15 different viruses. α is the total alpha. α_s: strong adaptation. α_w: weak adaptation. DN: number of nonsynonymous substitutions (LSC left or SSC right) in the VIPs of each virus. DN-α: corresponding estimated number of adaptive nonsynonymous substitutions. Upper rows: nine RNA viruses. Middle rows: six DNA viruses. Lower rows: VIPs that interact only with RNA viruses and VIPs that interact only with DNA viruses. Light orange: random shuffling of VIPs and control non-VIPs FDR<0.05. Dark orange: FDR<0.001.

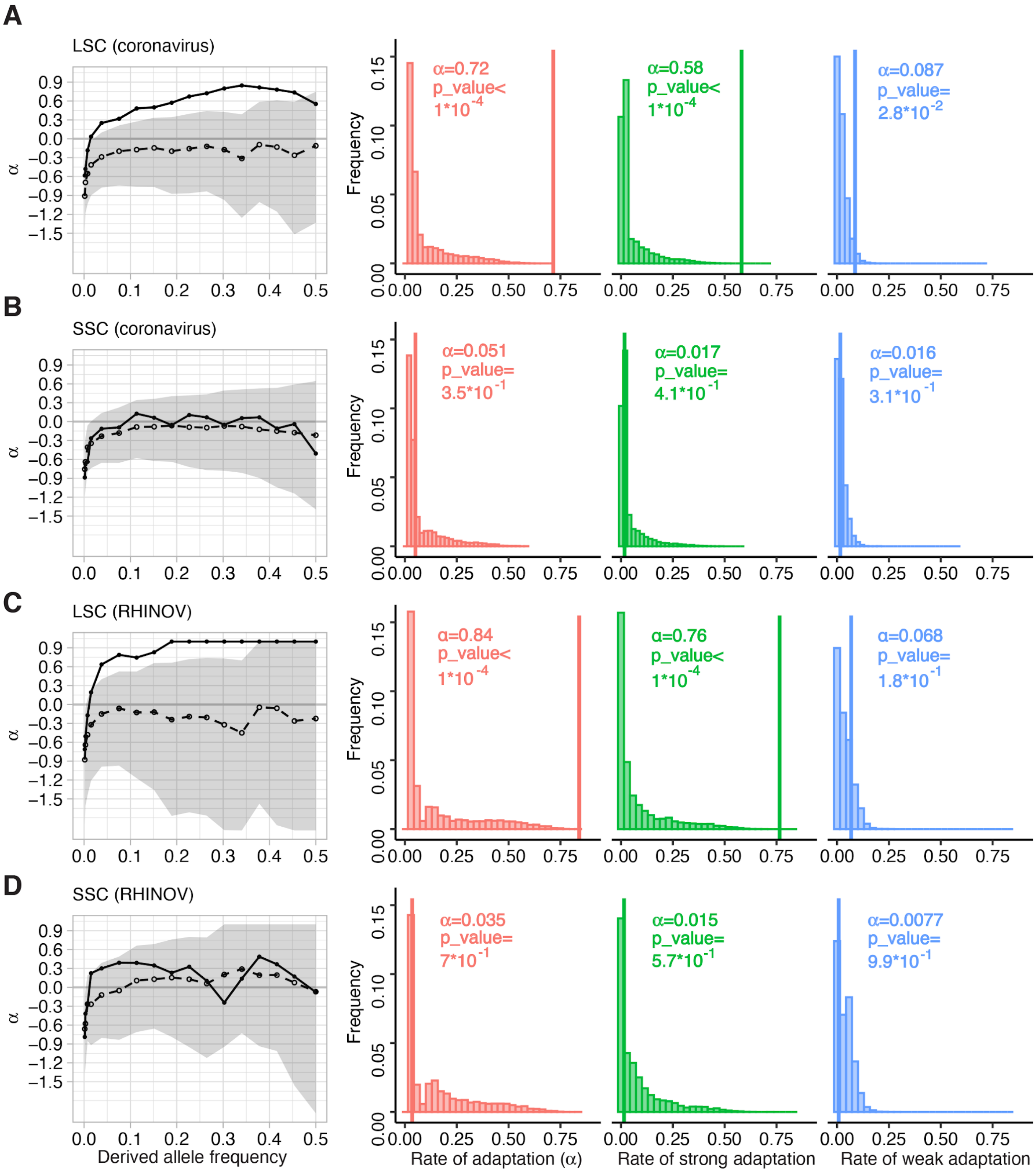


Figure 4. Adaptation in large or small stability changes for the VIPs of two RNA viruses with high LSC adaptation. Legend same as Figure 2. A) ABC-MK results for 448 coronavirus VIPs and control non-VIPs for LSCs. B) Same but for SSCs. C) ABC-MK results for 164 rhinovirus RV-B14 VIPs for LSCs. D) Same but for SSCs. The numbers of VIPs for each virus type correspond to the number of VIPs with enough high confidence Alphafold 2 residues (Methods) and with orthologs across mammals (Methods).

Directional evolution in immune antiviral VIPs, and compensatory evolution of stability in non-immune proviral VIPs

Because we use computational estimations from Thermonet, we seek to further validate these estimations by testing predictions about the expected evolution of protein stability of different functional classes of VIPs, that can only be verified if Thermonet's estimations are of sufficient quality. Different VIPs have very diverse native functions in the host (Enard and Petrov, 2018; 2020), but can still be sorted in a number of categories depending on their functional effects on viruses. In particular, many VIPs can be classified as immune antiviral VIPs that actively engage viral molecules to directly degrade, or target them for degradation, or activate degradation pathways (Table S1). On the other end of the functional spectrum, many VIPs do not have any immune function and are in fact proviral factors that assist viruses when the latter hijack their molecular functions to complete multiple steps of their replication cycle (Table S1). For example, viruses subvert host transcription regulators to activate the expression of their own viral genes during infection (Chau et al., 2008; Scoggin et al., 2001; Shen et al., 2022). We predict that immune antiviral and non-immune proviral VIPs should have experienced different patterns of protein stability evolution. In addition to uncovering new evolutionary patterns of host adaptation to viruses, verifying these predictions would also provide strong additional validation of the quality of the ΔΔG estimated by Thermonet.

We predict that immune antiviral VIPs should have experienced predominantly directional stability evolution, while non-immune proviral VIPs should have experienced predominantly compensatory evolution of stability. Indeed, the main, often highly specialized function of immune antiviral VIPs is to engage with viruses, with limited pleiotropic constraints in their way to often evolve very rapidly (Elde *et al.*, 2009; Sawyer *et al.*, 2004; Sawyer *et al.*, 2005). Their incessant arms race with viruses suggests that stability as an adaptive protein attribute, might have then evolved under directional selection, with stability optima changing over time together with the ever-changing landscape of pathogenic viruses. Conversely, non-immune proviral VIPs have non-immune, conserved native functions in the host. Thus, in any given non-immune proviral VIP with stability itself under stabilizing selection, adaptation during viral epidemics might take away protein stability from its usual optimum to complete native host functions, when there is no viral selective pressure to temporarily shift that optimum. This predicts that compensatory evolution of protein stability should then occur in non-immune proviral VIPs after viral epidemics, in the form of amino acid changes that tend to bring stability back closer to its pre-epidemic level (Chaurasia and Dutheil, 2022). This prediction has the advantage that it can be tested in a straightforward way, provided that Thermonet estimations are of sufficient quality. If compensatory stability evolution occurred in a non-immune proviral VIP, then we expect that the sum of stability changes caused by each amino acid change in this VIP should be closer to zero than expected by chance (regardless of the chronological order of these changes, which is unknown). Stability changes in this VIP should indeed compensate each other, i.e. should tend to cancel each other out. Conversely, we might observe the opposite for an immune antiviral VIP under directional stability evolution, with the sum of stability changes caused by each amino acid substitution being further from zero than expected by chance, if the direction of selection was preferentially toward increasing, or toward decreasing stability.

This is straightforward to look at because we can compare the sum of stability changes associated with each substitution in a VIP with the null expected distribution under neither stabilizing compensatory nor directional selection. We get the null expected distribution by randomly shuffling estimates of protein stability changes between genes included in the analysis (Methods), while also preserving the absolute average stability change per substitution in each VIP with shuffled stability changes (Methods). Crucially, the comparison with null expectations also provides a test of the quality of Thermonet ΔΔG predictions; we do not expect any departure from random expectations if the predictions are too far from the actual effects of the substitutions on protein stability.

The difficulty is then to identify immune antiviral, and non-immune proviral VIPs. To know which VIPs fall into these two categories, we conducted an extensive manual curation of 4,477 virology articles on VIPs, looking for the reported functional effects of VIPs on the replication cycle of a wide range of viruses (Methods). Through this effort we were able to identify 772 immune antiviral VIPs and 1,434 non-immune proviral VIPs as of October 2022 (Table S1; Methods).

We must further consider a few important potential limitations. In the millions of years since divergence with chimpanzees (a short amount of time for protein divergence), most VIPs with amino acid substitutions have accumulated only one such substitution and are therefore not appropriate for testing directional or compensatory evolution that imply multiple substitutions. Thus, we restrict this analysis to 67 immune antiviral VIPs and 43 non-immune proviral VIPs with two or more amino acid substitutions in the human branch (Table S1). We further consider that it might take more than one additional substitution to complete an episode of compensatory evolution after an epidemic, or more than two overall substitutions to start seeing a clear unidirectional pattern of directional selection. Thus, in addition to the test with VIPs with two or more nonsynonymous substitutions, we also further restrict our analysis to VIPs with three or more substitutions, in particular with the expectation that the signatures of compensatory evolution should be more visible in this subset of VIPs. We also restrict the test to VIPs with a minimal number of increasingly large absolute stability changes. Indeed, VIPs with no pronounced stability change may be entirely intolerant to such changes, or may not require compensatory evolution at all in the first place.

Using this specific, increasingly restrictive testing design, we first find that the 67 immune antiviral VIPs with at least two substitutions (Table S1) have predominantly experienced directional stability evolution. On average across these VIPs, the sum of stability changes is 1.17 fold further from zero than expected by chance (stability shuffling test P=2.5*10^{-3}; Methods). As predicted, this departure tends to become more pronounced when focusing on immune antiviral VIPs with a larger number of increasing high stability changes (Figure 5A,B). Individual antiviral VIPs with a strong signature of directional stability evolution notably include the APOBEC3G antiviral factor, albeit with marginal statistical significance (Table S1; sum of stability changes two times higher than expected, stability shuffling test P=0.07). Although immune antiviral VIPs have an overall collective trend toward directional selection, we notice a few notable exceptions. TRIM5 has a signature of compensatory, not directional stability evolution (Table S1). The sum of the six stability changes in TRIM5 is 10.7 times closer to zero than expected by chance (stability shuffling test P=0.049). Similarly, the prominent antiviral factor OAS1 has three stability changes with a signature of compensatory evolution (sum 20.1 times closer to zero than expected by chance, stability shuffling test P=0.028).

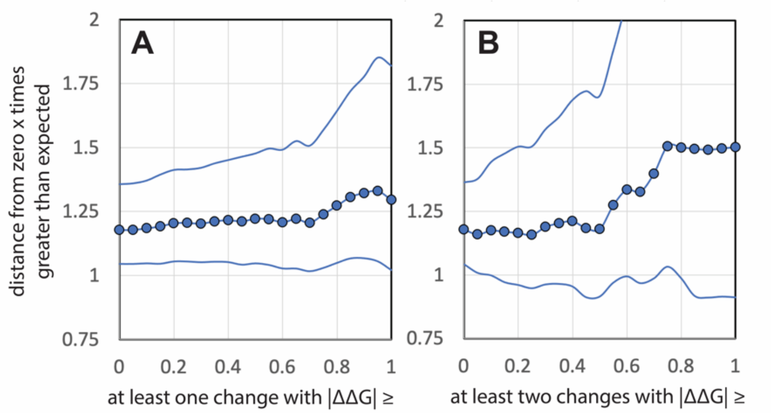


Figure 5. Directional evolution in immune antiviral VIPs.
The y-axis represents how many times further from zero the average of the per-VIP |sum| of stability changes is compared to random expectations. Central curve: compared to average random expectations over 1,000,000 random iterations (Methods). Lower curve: 95% confidence interval lower boundary. Upper curve: 95% confidence interval upper boundary. The x-axis represents the |ΔΔG| threshold to restrict the shuffling test only to groups of immune antiviral VIPs each with (A) at least two substitutions in total including at least one at or above the |ΔΔG| threshold, or (B) at least two substitutions in total including at least two at or above the |ΔΔG| threshold.

But compensatory stability evolution is most visible in non-immune proviral VIPs. For these VIPs we observe an overall trend of the sum of stability changes being on average closer to zero than expected by chance (Figure 6, y-axis). Overall, the sum is 1.2 times closer to zero than expected by chance in the 43 non-immune proviral VIPs with at least two nonsynonymous substitutions (stability shuffling test P=0.02; Methods). This trend becomes much more pronounced when restricting the test to non-immune proviral VIPs with three or more nonsynomymous substitutions, and when restricting the test to VIPs with an increasing number of more pronounced stability changes (Figure 6A,B,C). For example, the sum of stability changes is on average 2.5 times closer to zero than expected in the 16 non-immune proviral VIPs with at least three nonsynonymous substitutions including two with |ΔΔG|≥0.2 (Figure 6C and Table S1, stability shuffling test P=1.6*10^{-5}). This strong increase of the compensatory evolution signature compared to when including any nonsynonymous substitutions makes sense, given that many non-immune VIPs with weaker |ΔΔG| substitutions likely do not need compensatory evolution in the first place.

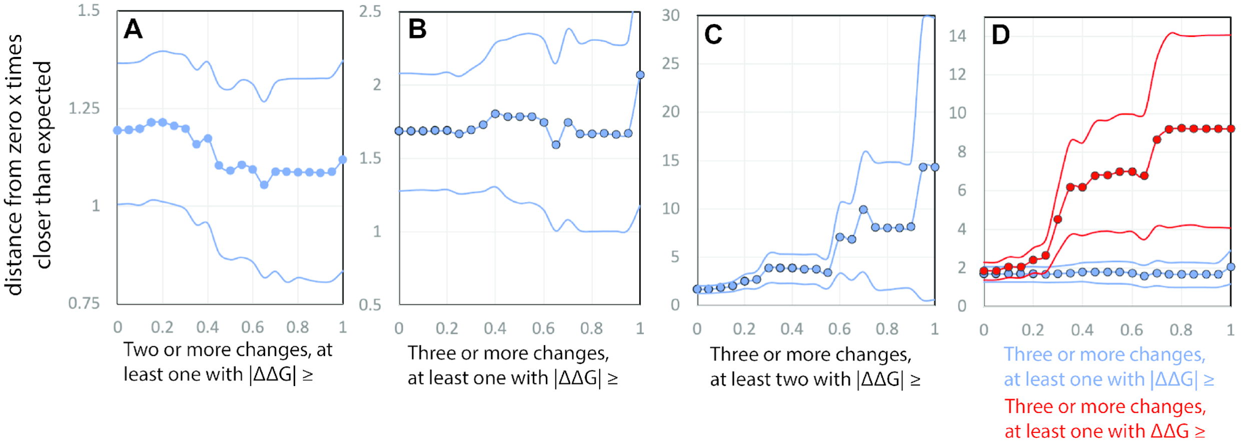


Figure 6. Compensatory evolution in non-immune proviral VIPs.
The y-axis represents how many times closer to zero the average of the per-VIP |sum| of stability changes is compared to random expectations. Central curve: compared to average random expectations over 1,000,000 random iterations (Methods). Lower curve: 95% confidence interval lower boundary. Upper curve: 95% confidence interval upper boundary. The x-axis represents the |ΔΔG| threshold to restrict the shuffling test only to groups of non-immune proviral VIPs each with (A) at least two substitutions in total including at least one at or above the |ΔΔG| threshold, or (B) at least three substitutions in total including at least one at or above the |ΔΔG| threshold, or (C) at least three substitutions in total including at least two at or above the |ΔΔG| threshold. The blue curves in (D) are the same as in (B). The red curves in (D) represent how much closer the average of the per-VIP |sum| of stability changes is compared to random expectations in VIPs with at least three substitutions in total, including at least one at or above the destabilizing ΔΔG threshold on the x-axis. Note the very different ranges of the y-axis for A,B,C and D.

Having clarified the possible impact of compensatory evolution in non-immune proviral VIPs, we further test one last prediction. Because proviral VIPs benefit the viruses that subvert them, strongly destabilizing substitutions that strongly decrease the abundance of non-immune proviral VIPs should be particularly advantageous during viral epidemics, and may also require similarly strong compensatory evolution afterwards. Thus, we expect that compensatory evolution should be particularly visible in non-immune proviral VIPs with at least one strongly destabilizing substitution, compared to non-immune proviral VIPs with at least one substitution affecting stability similarly but in either direction, stabilizing or destabilizing. When we compare these two situations, we find a much stronger signal of compensatory evolution in non-immune proviral VIPs specifically with at least one strongly destabilizing substitution (Figure 6D). This signal increases when restricting the test to VIPs with at least one increasingly destabilizing substitution (Figure 6D). This supports the predicted model where strongly destabilizing substitutions made proviral VIPs less available to viruses during epidemics, followed by compensatory evolution with stabilizing substitutions.

Together, these results highlight important differences in the stability evolution of two functional types of VIPs. The predominant signature of directional selection in immune antiviral VIPs, and most importantly the strong signature of compensatory evolution in non-immune proviral VIPs, cannot be expected if the Thermonet predictions were not of good, sufficient quality. Indeed, it is difficult to see how the sum of stability change predictions would be so much closer to zero than expected in non-immune proviral VIPs, if those predictions were far from the actual, real effect of the corresponding amino acid changes on protein stability. It would also be hard to explain how the compensatory evolution signature could be so much stronger in non-immune proviral VIPs with strongly destabilizing substitutions. This further provides strong support for the Thermonet ΔΔG predictions.

Discussion

Using Thermonet predictions of protein stability changes, we found that such changes were likely an important mechanism of virus-driven host adaptation. Although the computational nature of these predictions may raise doubt about their quality, it is important to reiterate that poor predictions not much better than random would not have allowed us to observe a large adaptation difference between LSCs and SSCs in VIPs, or the marked differences between immune antiviral and non-immune proviral VIPs (directional vs. compensatory evolution). Specifically, our results with non-immune proviral VIPs suggests an evolutionary model with positive selection during and after viral epidemics, with changes in stability during epidemics that are later compensated back to pre-epidemic stability levels that are likely more optimal for native host functions. This evolutionary model however requires further investigation. In particular, we do not have access to the chronological order of substitutions that would allow to further test it.

Together with previous studies, our results further highlight a common theme of host adaptation through changes in RNA and/or protein abundance of genes that interact functionally with pathogens. We previously found that strong selection during an ancient coronavirus or related virus epidemic in East Asia predominantly occurred at or in close linkage to eQTLs of coronavirus VIPs (Souilmi *et al.*, 2021). Similarly, Klunk et al. recently found that the most strongly selected variants during the Black Death are associated with changes in the expression level of immune genes (Klunk et al., 2022). There is also evidence that adaptive introgression in response to viruses, from Neanderthals to Eurasian modern human ancestors, involved Neanderthal variants that affect gene expression (Enard and Petrov, 2018; Nedelec et al., 2016; Quach et al., 2016).

Importantly however, our results do not exclude that mechanisms other than VIP stability changes also have played an important role in virus-driven protein host adaptation. Adaptive evolution in viral receptors might cut epidemics short by blocking viruses from infecting cells altogether. Similarly, adaptation in prominent antiviral factors such as TRIM5 or APOBECs may have had an oversized impact compared to VIPs not specialized in attacking viruses (Sawyer *et al.*, 2004; Sawyer *et al.*, 2005). Quantifying the relative quantitative contributions of different mechanisms of host adaptation is however complicated by the fact that the location of the vast majority of physical contact interfaces between viruses and VIPs are currently unknown. Further understanding of virus-driven adaptation will likely require a better knowledge of these interfaces. That said, we still do not expect that adaptation at contact interfaces may be able to fully explain the abundant adaptation we observe for large stability changes, in a scenario where these would only be a secondary bystander effect. Indeed, we found strong stability-dependent adaptation in more buried parts of the structures of VIPs. This however does not exclude the possibility of adaptive changes in binding between viruses and VIPs also due to allosteric, distance effects of buried adaptive amino-acid substitutions, through conformational and/or structural flexibility changes. It is also important to note that we only focused on parts of proteins that are well-structured with high Alphafold 2 confidence scores (Methods). Intrinsically disordered protein segments are known to result in poor structure confidence scores, and were thus completely excluded from our analysis. Whether intrinsically disordered proteins or protein segments also participate substantially to virus-driven adaptation remains an open question (Lou et al., 2016).

Our results suggest that in the future, identifying the biological mechanisms involved in virus-driven adaptation might enable more discriminant detection of ancient viral epidemics, where the involvement of a specific virus in past epidemics may be recognized not only through an overall increase in adaptation in its specific VIPs, but more specifically through an increase in adaptative evolution of specific mechanistic attributes such as protein stability. In that respect, we find that RNA viruses clearly stand out during human evolution compared to DNA viruses. Taken together, our results show that in addition to the detailed functional study of specific gene candidates by evolutionary virologists, the study of quantitative patterns of adaptive evolution in VIPs as a whole group can provide new insights on the functional evolutionary changes that gave hosts a fighting chance against repeated viral epidemics.

Methods

Identifying human protein coding genes with orthologs across mammals

We previously found that viruses increase adaptation specifically in VIPs with orthologs across mammals (Castellano, 2019). We therefore restrict all analyses to VIPs and non-VIPs with orthologs across mammals. We use an updated list of Ensembl v99 human genes (Cunningham et al., 2019) with orthologs found in at least 251 out of 261 mammalian genome assemblies. These 261 assemblies were extracted from NCBI Genome (https://www.ncbi.nlm.nih.gov/genome/) and are the deposited assemblies that had a N50 contig size of at least 30kb as of July 2021, in order to limit the number of truncated genes. Ensembl human protein coding genes are selected as mammals-wide orthologs if they have best-reciprocal hits using the largest number of identical nucleotide hits from Blat alignments (Kent, 2002), with at least 251 out of the 261 genome assemblies (to account for the fact that orthologs in some species may have not been sequenced, and located in assembly gaps). This process finds 13,495 such human Ensembl v99 genes with orthologs across mammals (Table S1). The list of mammalian species and the corresponding assembly versions are provided in Table S2.

Thermonet predictions with Alphafold structures

We use the Thermonet software to predict the ΔΔG caused by specific amino acid changes in VIPs and non-VIPs. Thermonet uses a convolutional neural network to make ΔΔG predictions. Thermonet's neural network is trained using images of the biophysical properties of the close three-dimensional environment of the amino-acid change location. The neural network was trained first on experimental datasets of ΔΔG measurements. Because Thermonet uses the three-dimensional local environment, it requires a protein structure as input. To run Thermonet we chose to use public Alphafold v2 structures (from https://alphafold.ebi.ac.uk/) rather than experimental structures because using only available human experimental structures would have strongly limited the number of VIPs and the statistical power of the analysis. Alphafold 2 however generates structures very close to the experimental ones when the latter are available to use by Alphafold as input (Jumper *et al.*, 2021). This means that for those proteins with experimental structures, we expect very little difference when using the Alphafold structures. The advantage of Alphafold is then also that it provides good quality predictions for a substantially larger number of human proteins than are experimentally available (Jumper *et al.*, 2021). This is because Alphafold accurately predicts structures made of local folds that are well represented in its input database (Jumper *et al.*, 2021). Nevertheless, Alphafold still fails to properly predict a subset of proteins or parts of proteins that then have a mixture of well and poorly predicted local structure regions. Note that this is particularly true for proteins or parts of proteins that are intrinsically disordered and thus do not have one single structure to predict in the first place.

Fortunately, Alphafold provides a site-by-site confidence score, noted pLDDT (Jumper *et al.*, 2021). The pLDDT score has been shown by comparison with experimental structures to strongly predict the per-residue structure accuracy. The pLDDT score varies from zero to 100, with 100 indicating the most accurate per-residue structural prediction possible. A pLDDT score above 70 is usually indicative of a highly accurate structure prediction. For this reason, in our analysis we only use Thermonet ΔΔG predictions at sites with a pLDDT equal to or greater than 70. We also only use Alphafold structures with 50% or more sites with pLDDT≥70. In total 76% of the VIPs and non-VIPs that we compare have such high quality Alphafold 2 structures (same % for VIPs and non-VIPs), for a total of 2,909 and 5707 non-VIPs. In addition, to avoid any confounding effect of discrepancies between the accuracy of Alphafold structures between VIPs and non-VIPs, we match VIPs with control non-VIPs with similar average per-residue pLDDT and percentage of sites with pLDDT≥70 (see below, VIPs and control non-VIPs with matching confounding factors).

Using the filters described above, we use 86,244 and 10,337 Thermonet ΔΔG predictions for coding variants and substitutions, respectively. Figure 1 represents the distribution of the predicted ΔΔG for variants (Figure 1A) and substitutions (Figure 1B). We run Thermonet using the amino acid change from the ancestral to the derived amino-acid, from the ancestral to the fixed human amino acid for substitutions, and from the ancestral to the derived allele for variants. Finally, it is also important to note that Alphafold only provides publicly the structures of the canonical coding sequence of each protein coding gene, but not for their other isoforms. Here we thus use only the corresponding Ensembl v100 canonical coding sequences with an Alphafold structure, which excludes variants and substitutions in other isoforms that do not overlap with the canonical isoform.

Variants and substitutions data for ABC-MK

We use ABC-MK as previously extensively described in (Uricchio *et al.*, 2019), using the human coding variants and substitutions dataset also described in the same publication. The main difference is that we split nonsynonymous variants and substitutions into two groups, noted LSCs and SSCs in the main text, according to their predicted ΔΔG that we then run ABC-MK on separately. We separate the two groups according to the variants absolute median of |ΔΔG| (0.225) in the 2,909 VIPs included in the analysis. We do this so that the groups have similar amounts of information, and thus the same variance of ABC-MK estimates of the proportion of selected substitutions. This proportion is usually noted α, calculated as α=1-(PN*DS)/(PS*DN), where PN and PS are the numbers of nonsynonymous and synonymous variants, respectively, and DN and DS are the numbers of nonsynonymous and synonymous fixed substitutions, respectively. The classic MK test simply uses this calculation of α. ABC-MK uses a more complex approach by computing the α-curve, which is the curve of α measured specifically for bins of derived allele frequencies, as for example in Figure 1C. This is required to account for segregating non-synonymous deleterious variants among other things (Uricchio *et al.*, 2019). ABC-MK uses Approximate Bayesian Computation to match the observed α-curve with the best-fitting ones among many analytically predicted α-curves. Each analytical α-curve is generated for millions of combinations of varying distributions and amounts of deleterious, weakly advantageous, and strongly advantageous mutations. The difference between weakly and strongly advantageous variants that is exploited by ABC-MK is that weakly advantageous mutations do not go to fixation so fast that their contribution to nonsynonymous polymorphism is negligeable as it is for strongly advantageous variants (Uricchio *et al.*, 2019). This affects the shape of the α-curve, especially at higher derived allele frequencies where weakly advantageous variants tend to segregate before eventually reaching fixation (selective sweeps tend to have a long pre-fixation phase after the faster exponential one). This translates into a downward trend of the α-curve at higher frequencies that can be detected by ABC-MK. It is important to note that another possible cause of a downward trend of the α-curve at higher frequencies is mispolarization, where low frequency derived nonsynonymous alleles may be mistaken for high frequency derived ones. This can happen when a nucleotide site experienced a substitution in the human branch, but subsequently experienced a mutation back to the initial ancestral nucleotide. This new derived allele will then be mistakenly annotated as the ancestral (Hernandez et al., 2007). Hernandez et al. have shown that in the human genome this issue affects derived alleles with a frequency greater than 0.7. We therefore run ABC-MK using nonsynonymous and synonymous variants with a derived allele frequency less than 0.7. It is also important to mention that ABC-MK uses the shape of the α-curve at low derived allele frequencies to estimate the distribution of deleterious fitness effects (Urrichio et al.). An important last detail about how we run ABC-MK is that because as described above, we only use amino acids with a pLDDT score at or above 70, we only use the synonymous variants and substitutions in the corresponding codons. We use the same set of synonymous variants and substitutions as a neutral reference when we run ABC-MK either with LSCs or SSCs, since focusing on smaller subsets of nonsynonymous changes (either LSCs or SSCs) is readily accounted for by the PN over DN ratio in the equation $\alpha=1-(PN*\mathrm{DS})/(DN*\mathrm{PS})$. ABC-MK is available at https://github.com/jmurga/MKtest.jl.

Controlling for confounding factors when matching VIPs and non-VIPs

To highlight evolutionary patterns of adaptation that are specific to VIPs, we compare them with sets of control non-VIPs that have been matched with VIPs so that the former have the same average values of confounding factors as the latter. Here confounding factors are factors other than physical interaction with viruses that in principle might affect adaptation, and thus explain differences between VIPs and non-VIPs instead of physical interactions with viruses. We know for example that VIPs tend to be much more highly expressed at the mRNA level than non-VIPs. Higher expression might hypothetically be associated with increased adaptation, and thus explain the increased adaptation in VIPs rather than interactions with viruses. We match control non-VIPs with VIPs using a bootstrap procedure that was already extensively previously described (Di *et al.*, 2021; Enard and Petrov, 2020). In total we match 17 potential confounding factors between VIPs and control non-VIPs:

-Ensembl canonical coding sequence length, since they correspond to the isoform used by Alphafold.

- the average GC content for each coding sequence.

- the average GC1 content at the first codon nucleotide position for each coding sequence.

- the average GC2 content at the second codon nucleotide position for each coding sequence.

- the average GC3 content at the third codon nucleotide position for each coding sequence. GC1, GC2, and GC3 control for possible differences in GC content between nonsynonymous and synonymous sites that might distort the α-curve.

- average GTEx v8 (Consortium, 2020) TPM (Transcripts Per Million) mRNA expression across 53 tissues (in log base 2).

- average GTEx v8 TPM mRNA expression in lymphocytes (in log base 2).

- average GTEx v8 TPM mRNA expression in testis (in log base 2).

- the number of protein-protein interactions (in log base 2) in the human protein interaction network (Luisi et al., 2015).

- the proportion of immune genes as annotated with the Gene Ontology terms GO:0002376 (immune system process), GO:0006952 (defense response) and/or GO:0006955 (immune response) as of May 2020 (Gene Ontology, 2015).

- the recombination rate (Halldorsson et al., 2019) in 500kb windows centered on genes, to account for potential mutational biases related to recombination such as biased gene conversion that could differentially affect synonymous and nonsynonymous sites with different GC content. We use large 500kb windows because they better represent the long-term recombination rate in a given genomic window compared to smaller windows.

- McVicker's B value (McVicker et al., 2009), a measure of background selection that we used to account for the recent prevalence of segregating deleterious variants in the genomic environment surrounding a coding sequence and that could affect adaptation (Di *et al.*, 2021).

- the density of GERP conserved elements (Davydov et al., 2010) in 50 kb and 500 kb windows centered on genes, to further account for the possible prevalence of segregating deleterious variants in the genomic environment surrounding a coding sequence.

- the proportion of amino acids in the Alphafold v2 structure with a pLDDT score of 50 or above.

- the proportion of amino acids in the Alphafold v2 structure with a pLDDT score of 70 or above.

- the average pLDDT score in the entire Alphafold structure.

Estimation of Relative Solvent Accessibilty

Relative Solvent Accessibility (RSA) provides a measure of how exposed at the surface or buried close to the protein core specific amino acids are in the Alphafold structures. We measure RSA using DSSP (Kabsch and Sander, 1983).

Functional annotation of VIPs

As part of an ongoing effort to annotate the functional impacts of VIPs on viruses, we have to date manually curated 4,477 virology articles from Pubmed to collect their proviral and/or antiviral effects (Table S1), as reported by the virology experiments described in those virology articles. We also annotated if the proviral and antiviral effects occurred through an immune function of the VIPs, either directly reported by the virology articles or because of the fact that the corresponding VIPs are annotated with the Gene Ontology terms GO:0002376 (immune system process), GO:0006952 (defense response) and/or GO:0006955 (immune response) as of May 2020. Through this annotation we identified 772 VIPs with an immune antiviral impact, and 1,434 VIPs with a non-immune, proviral impact (Table S1). Interestingly, we also found that 321 of the 772 antiviral immune VIPs for at least one virus, also have proviral effects for the same or different viruses. A detailed inspection of such cases shows that it happens for example when an expressed antiviral immune VIP has a molecular activity involved in the immune response that is subverted by viruses. For example, the antiviral VIP CREBBP is an important transcription activator involved in interferon beta production (Qu et al., 2021), that is exploited by the Human T-cell Leukemia virus to activate the transcription of its own viral genes (Scoggin *et al.*, 2001). Antiviral immune VIPs expressed during infection are likely good targets for proviral repurposing by viruses, due to their broad availability at the precise time of need of their molecular activities by viruses.

Directional and compensatory evolution of VIP stability

To detect directional or compensatory evolution of stability in VIPs, we design a random shuffling test based on the sum of stability changes in individual VIPs, then averaged over VIPs with multiple amino acid changes tested together. For example, in non-immune proviral VIPs we expect each individual VIP with multiple amino acid changes to have stability changes that tend to cancel out each other, thus resulting in a sum of stability changes closer to zero than expected by chance. As a statistic we use the average across VIPs of the absolute value of the sum of stability changes in each VIP. We then compare this average with its random expectation. This random expectation is generated as follows: for each VIP with a number x of stability changes, we randomly sample x stability changes from the entire pool of all predicted stability changes. We iterate this random sampling for each given VIP, until the randomly sampled stability changes have an average |ΔΔG| that matches the observed average |ΔΔG| for this VIP (plus or minus 2%). We do this to account for the fact that different VIPs may have different spreads of their distribution of possible stability changes, which could affect the null random expected sum of stability changes for each VIP. For the whole set of tested VIPs, we iterate this process 1,000,000 times to determine the average random expectation and the statistical significance of any observed departure of the real average sum of stability changes from this expectation. The sets of VIPs tested together must fulfill a number of pre-requisites, such as a minimal total number of stability changes, and a minimum number of those stability changes having their individual |ΔΔG| above a fixed threshold. For example, for non-immune proviral VIPs, we use these pre-requisites with increasingly stringent thresholds supposed to restrict the test to a number of VIPs where signatures of compensatory evolution are expected to be more visible. Indeed, compensatory evolution is more likely to have occurred in VIPs with a larger number of stability changes (compensatory changes had the chance to occur in the first place), and with a minimum number of large |ΔΔG| changes (compensatory evolution is likely required only when large |ΔΔG| changes occurred in the first place). We represent in Figure 5 the ratio of the observed average absolute value of the sum of stability changes, divided by the average random expectation (and the inverse for Figure 6) and its 95% confidence intervals upper and lower values generated by the 1,000,000 random samplings.

Acknowledgements

We thank Mike Barker, Joanna Masel and Ryan Gutenkunst for their comments on the manuscript. David Enard is funded by NIH NIGMS MIRA R35GM142677.

References

Batra, J., Hultquist, J.F., Liu, D., Shtanko, O., Von Dollen, J., Satkamp, L., Jang, G.M., Luthra, P., Schwarz, T.M., Small, G.I., et al. (2018). Protein Interaction Mapping Identifies RBBP6 as a Negative Regulator of Ebola Virus Replication. Cell 175, 1917-1930 e1913. 10.1016/j.cell.2018.08.044.

Castellano, D.U., L.H.; Munch, K.; Enard, D. (2019). Viruses rule over adaptation in conservd human proteins. bioRxiv. 10.1101/555060.

Chau, C.M., Deng, Z., Kang, H., and Lieberman, P.M. (2008). Cell cycle association of the retinoblastoma protein Rb and the histone demethylase LSD1 with the Epstein-Barr virus latency promoter Cp. J Virol 82, 3428-3437. 10.1128/JVI.01412-07.

Chaurasia, S., and Dutheil, J.Y. (2022). The Structural Determinants of Intra-Protein Compensatory Substitutions. Mol Biol Evol 39. 10.1093/molbev/msac063.

Chesarino, N.M., and Emerman, M. (2020). Polymorphisms in Human APOBEC3H Differentially Regulate Ubiquitination and Antiviral Activity. Viruses 12. 10.3390/v12040378.

Clausen, L., Abildgaard, A.B., Gersing, S.K., Stein, A., Lindorff-Larsen, K., and Hartmann-Petersen, R. (2019). Protein stability and degradation in health and disease. Adv Protein Chem Struct Biol 114, 61-83. 10.1016/bs.apcsb.2018.09.002.

Compton, A.A., Hirsch, V.M., and Emerman, M. (2012). The host restriction factor APOBEC3G and retroviral Vif protein coevolve due to ongoing genetic conflict. Cell Host Microbe 11, 91-98. 10.1016/j.chom.2011.11.010.

Consortium, G.T. (2020). The GTEx Consortium atlas of genetic regulatory effects across human tissues. Science 369, 1318-1330. 10.1126/science.aaz1776.

Cunningham, F., Achuthan, P., Akanni, W., Allen, J., Amode, M.R., Armean, I.M., Bennett, R., Bhai, J., Billis, K., Boddu, S., et al. (2019). Ensembl 2019. Nucleic Acids Res 47, D745-D751. 10.1093/nar/gky1113.

Dasmeh, P., Serohijos, A.W., Kepp, K.P., and Shakhnovich, E.I. (2013). Positively selected sites in cetacean myoglobins contribute to protein stability. PLoS Comput Biol 9, e1002929. 10.1371/journal.pcbi.1002929.

Davydov, E.V., Goode, D.L., Sirota, M., Cooper, G.M., Sidow, A., and Batzoglou, S. (2010). Identifying a high fraction of the human genome to be under selective constraint using GERP++. PLoS Comput Biol 6, e1001025. 10.1371/journal.pcbi.1001025.

Di, C., Murga Moreno, J., Salazar-Tortosa, D.F., Lauterbur, M.E., and Enard, D. (2021). Decreased recent adaptation at human mendelian disease genes as a possible consequence of interference between advantageous and deleterious variants. Elife 10. 10.7554/eLife.69026.

Elde, N.C., Child, S.J., Geballe, A.P., and Malik, H.S. (2009). Protein kinase R reveals an evolutionary model for defeating viral mimicry. Nature 457, 485-489. 10.1038/nature07529.

Enard, D., Cai, L., Gwennap, C., and Petrov, D.A. (2016). Viruses are a dominant driver of protein adaptation in mammals. Elife 5. 10.7554/eLife.12469.

Enard, D., and Petrov, D.A. (2018). Evidence that RNA Viruses Drove Adaptive Introgression between Neanderthals and Modern Humans. Cell 175, 360-371 e313. 10.1016/j.cell.2018.08.034.

Enard, D., and Petrov, D.A. (2020). Ancient RNA virus epidemics through the lens of recent adaptation in human genomes. Philos Trans R Soc Lond B Biol Sci 375, 20190575. 10.1098/rstb.2019.0575.

Firnberg, E., Labonte, J.W., Gray, J.J., and Ostermeier, M. (2014). A comprehensive, high-resolution map of a gene's fitness landscape. Mol Biol Evol 31, 1581-1592. 10.1093/molbev/msu081.

Gene Ontology, C. (2015). Gene Ontology Consortium: going forward. Nucleic Acids Res 43, D1049-1056. 10.1093/nar/gku1179.

Genomes Project, C., Auton, A., Brooks, L.D., Durbin, R.M., Garrison, E.P., Kang, H.M., Korbel, J.O., Marchini, J.L., McCarthy, S., McVean, G.A., and Abecasis, G.R. (2015). A global reference for human genetic variation. Nature 526, 68-74. 10.1038/nature15393.

Goldenzweig, A., and Fleishman, S.J. (2018). Principles of Protein Stability and Their Application in Computational Design. Annu Rev Biochem 87, 105-129. 10.1146/annurev-biochem-062917-012102.

Halldorsson, B.V., Palsson, G., Stefansson, O.A., Jonsson, H., Hardarson, M.T., Eggertsson, H.P., Gunnarsson, B., Oddsson, A., Halldorsson, G.H., Zink, F., et al. (2019). Characterizing mutagenic effects of recombination through a sequence-level genetic map. Science 363. 10.1126/science.aau1043.

Hernandez, R.D., Williamson, S.H., and Bustamante, C.D. (2007). Context dependence, ancestral misidentification, and spurious signatures of natural selection. Mol Biol Evol 24, 1792-1800. 10.1093/molbev/msm108.

Jager, S., Cimermancic, P., Gulbahce, N., Johnson, J.R., McGovern, K.E., Clarke, S.C., Shales, M., Mercenne, G., Pache, L., Li, K., et al. (2011). Global landscape of HIV-human protein complexes. Nature 481, 365-370. 10.1038/nature10719.

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Zidek, A., Potapenko, A., et al. (2021). Highly accurate protein structure prediction with AlphaFold. Nature 596, 583-589. 10.1038/s41586-021-03819-2.

Kabsch, W., and Sander, C. (1983). Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 22, 2577-2637. 10.1002/bip.360221211.

Kent, W.J. (2002). BLAT--the BLAST-like alignment tool. Genome Res 12, 656-664. 10.1101/gr.229202.

King, C.R., and Mehle, A. (2022). Retasking of canonical antiviral factors into proviral effectors. Curr Opin Virol 56, 101271. 10.1016/j.coviro.2022.101271.

Klunk, J., Vilgalys, T.P., Demeure, C.E., Cheng, X., Shiratori, M., Madej, J., Beau, R., Elli, D., Patino, M.I., Redfern, R., et al. (2022). Evolution of immune genes is associated with the Black Death. Nature 611, 312-319. 10.1038/s41586-022-05349-x.

Konate, M.M., Plata, G., Park, J., Usmanova, D.R., Wang, H., and Vitkup, D. (2019). Molecular function limits divergent protein evolution on planetary timescales. Elife 8. 10.7554/eLife.39705.

Lee, D., Redfern, O., and Orengo, C. (2007). Predicting protein function from sequence and structure. Nat Rev Mol Cell Biol 8, 995-1005. 10.1038/nrm2281.

Li, B., Yang, Y.T., Capra, J.A., and Gerstein, M.B. (2020). Predicting changes in protein thermodynamic stability upon point mutation with deep 3D convolutional neural networks. PLoS Comput Biol 16, e1008291. 10.1371/journal.pcbi.1008291.

Lou, D.I., Kim, E.T., Meyerson, N.R., Pancholi, N.J., Mohni, K.N., Enard, D., Petrov, D.A., Weller, S.K., Weitzman, M.D., and Sawyer, S.L. (2016). An Intrinsically Disordered Region of the DNA Repair Protein Nbs1 Is a Species-Specific Barrier to Herpes Simplex Virus 1 in Primates. Cell Host Microbe 20, 178-188. 10.1016/j.chom.2016.07.003.

Luisi, P., Alvarez-Ponce, D., Pybus, M., Fares, M.A., Bertranpetit, J., and Laayouni, H. (2015). Recent positive selection has acted on genes encoding proteins with more interactions within the whole human interactome. Genome Biol Evol 7, 1141-1154. 10.1093/gbe/evv055.

McDonald, J.H., and Kreitman, M. (1991). Adaptive protein evolution at the Adh locus in Drosophila. Nature 351, 652-654. 10.1038/351652a0.

McVicker, G., Gordon, D., Davis, C., and Green, P. (2009). Widespread genomic signatures of natural selection in hominid evolution. PLoS Genet 5, e1000471. 10.1371/journal.pgen.1000471.

Moutinho, A.F., Eyre-Walker, A., and Dutheil, J.Y. (2022). Strong evidence for the adaptive walk model of gene evolution in Drosophila and Arabidopsis. PLoS Biol 20, e3001775. 10.1371/journal.pbio.3001775.

Nedelec, Y., Sanz, J., Baharian, G., Szpiech, Z.A., Pacis, A., Dumaine, A., Grenier, J.C., Freiman, A., Sams, A.J., Hebert, S., et al. (2016). Genetic Ancestry and Natural Selection Drive Population Differences in Immune Responses to Pathogens. Cell 167, 657-669 e621. 10.1016/j.cell.2016.09.025.

Nelson, D.L., Lehninger, A.L., and Cox, M.M. (2013). Lehninger principles of biochemistry, 6th Edition (W.H. Freeman).

Ng, M., Ndungo, E., Kaczmarek, M.E., Herbert, A.S., Binger, T., Kuehne, A.I., Jangra, R.K., Hawkins, J.A., Gifford, R.J., Biswas, R., et al. (2015). Filovirus receptor NPC1 contributes to species-specific patterns of ebolavirus susceptibility in bats. Elife 4. 10.7554/eLife.11785.

Nick Pace, C., Scholtz, J.M., and Grimsley, G.R. (2014). Forces stabilizing proteins. FEBS Lett 588, 2177-2184. 10.1016/j.febslet.2014.05.006.

Nielsen, S.V., Stein, A., Dinitzen, A.B., Papaleo, E., Tatham, M.H., Poulsen, E.G., Kassem, M.M., Rasmussen, L.J., Lindorff-Larsen, K., and Hartmann-Petersen, R. (2017). Predicting the impact of Lynch syndrome-causing missense mutations from structural calculations. PLoS Genet 13, e1006739. 10.1371/journal.pgen.1006739.

Pancotti, C., Benevenuta, S., Birolo, G., Alberini, V., Repetto, V., Sanavia, T., Capriotti, E., and Fariselli, P. (2022). Predicting protein stability changes upon single-point mutation: a thorough comparison of the available tools on a new dataset. Brief Bioinform 23. 10.1093/bib/bbab555.

Qu, Z., Meng, F., Shi, J., Deng, G., Zeng, X., Ge, J., Li, Y., Liu, L., Chen, P., Jiang, Y., et al. (2021). A Novel Intronic Circular RNA Antagonizes Influenza Virus by Absorbing a microRNA That Degrades CREBBP and Accelerating IFN-beta Production. mBio 12, e0101721. 10.1128/mBio.01017-21.

Quach, H., Rotival, M., Pothlichet, J., Loh, Y.E., Dannemann, M., Zidane, N., Laval, G., Patin, E., Harmant, C., Lopez, M., et al. (2016). Genetic Adaptation and Neandertal Admixture Shaped the Immune System of Human Populations. Cell 167, 643-656 e617. 10.1016/j.cell.2016.09.024.

Radivojac, P., Clark, W.T., Oron, T.R., Schnoes, A.M., Wittkop, T., Sokolov, A., Graim, K., Funk, C., Verspoor, K., Ben-Hur, A., et al. (2013). A large-scale evaluation of computational protein function prediction. Nat Methods 10, 221-227. 10.1038/nmeth.2340.

Sawyer, S.L., Emerman, M., and Malik, H.S. (2004). Ancient adaptive evolution of the primate antiviral DNA-editing enzyme APOBEC3G. PLoS Biol 2, E275. 10.1371/journal.pbio.0020275.

Sawyer, S.L., Wu, L.I., Emerman, M., and Malik, H.S. (2005). Positive selection of primate TRIM5alpha identifies a critical species-specific retroviral restriction domain. Proc Natl Acad Sci U S A 102, 2832-2837. 10.1073/pnas.0409853102.

Scheller, R., Stein, A., Nielsen, S.V., Marin, F.I., Gerdes, A.M., Di Marco, M., Papaleo, E., Lindorff-Larsen, K., and Hartmann-Petersen, R. (2019). Toward mechanistic models for genotype-phenotype correlations in phenylketonuria using protein stability calculations. Hum Mutat 40, 444-457. 10.1002/humu.23707.

Scoggin, K.E., Ulloa, A., and Nyborg, J.K. (2001). The oncoprotein Tax binds the SRC-1-interacting domain of CBP/p300 to mediate transcriptional activation. Mol Cell Biol 21, 5520-5530. 10.1128/MCB.21.16.5520-5530.2001.

Serohijos, A.W., Lee, S.Y., and Shakhnovich, E.I. (2013). Highly abundant proteins favor more stable 3D structures in yeast. Biophys J 104, L1-3. 10.1016/j.bpj.2012.11.3838.

Serohijos, A.W., and Shakhnovich, E.I. (2014). Contribution of selection for protein folding stability in shaping the patterns of polymorphisms in coding regions. Mol Biol Evol 31, 165-176. 10.1093/molbev/mst189.

Shah, P.S., Link, N., Jang, G.M., Sharp, P.P., Zhu, T., Swaney, D.L., Johnson, J.R., Von Dollen, J., Ramage, H.R., Satkamp, L., et al. (2018). Comparative Flavivirus-Host Protein Interaction Mapping Reveals Mechanisms of Dengue and Zika Virus Pathogenesis. Cell 175, 1931-1945 e1918. 10.1016/j.cell.2018.11.028.

Shen, Z., Wu, J., Gao, Z., Zhang, S., Chen, J., He, J., Guo, Y., Deng, Q., Xie, Y., Liu, J., and Zhang, J. (2022). High mobility group AT-hook 1 (HMGA1) is an important positive regulator of hepatitis B virus (HBV) that is reciprocally upregulated by HBV X protein. Nucleic Acids Res 50, 2157-2171. 10.1093/nar/gkac070.

Souilmi, Y., Lauterbur, M.E., Tobler, R., Huber, C.D., Johar, A.S., Moradi, S.V., Johnston, W.A., Krogan, N.J., Alexandrov, K., and Enard, D. (2021). An ancient viral epidemic involving host coronavirus interacting genes more than 20,000 years ago in East Asia. Curr Biol 31, 3504-3514 e3509. 10.1016/j.cub.2021.05.067.

Stein, A., Fowler, D.M., Hartmann-Petersen, R., and Lindorff-Larsen, K. (2019). Biophysical and Mechanistic Models for Disease-Causing Protein Variants. Trends Biochem Sci 44, 575-588. 10.1016/j.tibs.2019.01.003.

Swain, J.F., and Gierasch, L.M. (2006). The changing landscape of protein allostery. Curr Opin Struct Biol 16, 102-108. 10.1016/j.sbi.2006.01.003.

Tran, V., Ledwith, M.P., Thamamongood, T., Higgins, C.A., Tripathi, S., Chang, M.W., Benner, C., Garcia-Sastre, A., Schwemmle, M., Boon, A.C.M., et al. (2020). Influenza virus repurposes the antiviral protein IFIT2 to promote translation of viral mRNAs. Nat Microbiol 5, 1490-1503. 10.1038/s41564-020-0778-x.

Uricchio, L.H., Petrov, D.A., and Enard, D. (2019). Exploiting selection at linked sites to infer the rate and strength of adaptation. Nat Ecol Evol 3, 977-984. 10.1038/s41559-019-0890-6.
Watanabe, T., Kawakami, E., Shoemaker, J.E., Lopes, T.J., Matsuoka, Y., Tomita, Y., Kozuka-Hata, H.,

Gorai, T., Kuwahara, T., Takeda, E., et al. (2014). Influenza virus-host interactome screen as a platform for antiviral drug development. Cell Host Microbe 16, 795-805. 10.1016/j.chom.2014.11.002.

Wierbowski, S.D., Liang, S., Liu, Y., Chen, Y., Gupta, S., Andre, N.M., Lipkin, S.M., Whittaker, G.R., and Yu, H. (2021). A 3D structural SARS-CoV-2-human interactome to explore genetic and drug perturbations. Nat Methods 18, 1477-1488. 10.1038/s41592-021-01318-w.

Yan, R., Zhang, Y., Li, Y., Xia, L., Guo, Y., and Zhou, Q. (2020). Structural basis for the recognition of SARS-CoV-2 by full-length human ACE2. Science 367, 1444-1448. 10.1126/science.abb2762.

Yang, L., Emerman, M., Malik, H.S., and McLaughlin, R.N.J. (2020). Retrocopying expands the functional repertoire of APOBEC3 antiviral proteins in primates. Elife 9. 10.7554/eLife.58436.

Yong, Y.Y., Zhang, L., Hu, Y.J., Wu, J.M., Yan, L., Pan, Y.R., Tang, Y., Yu, L., Law, B.Y., Yu, C.L., et al. (2022). Targeting autophagy regulation in NLRP3 inflammasome-mediated lung inflammation in COVID-19. Clin Immunol 244, 109093. 10.1016/j.clim.2022.109093.

Supplements

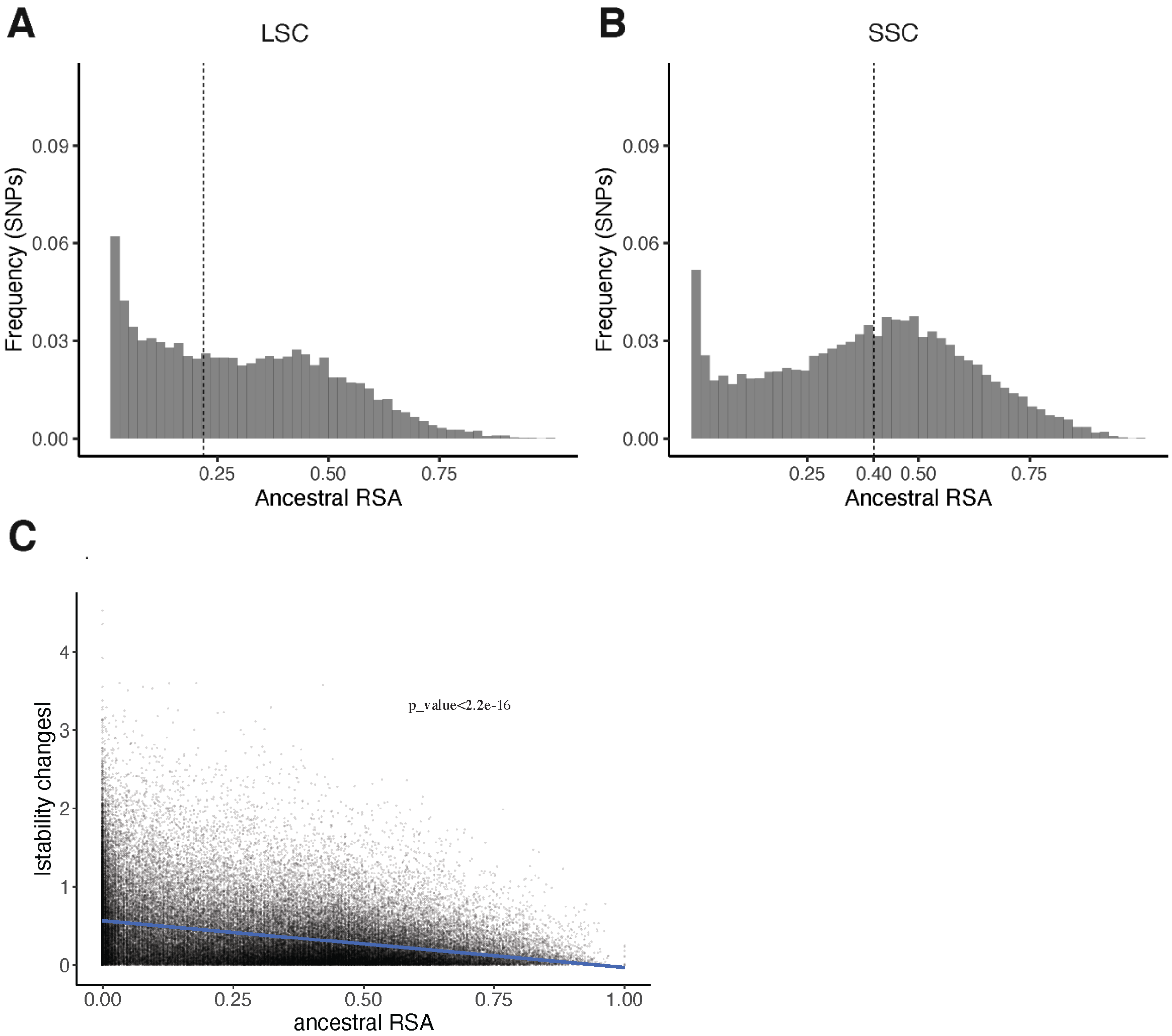


Figure S1. Distributions of ancestral RSA of LSC and SSC and correlation between stability changes and RSA.

Distributions of ancestral RSA of A) LSC and B) SSC variants. The dashed line is the median, the median RSA in LSC is 0.23 smaller than the median RSA in SSC (0.4). C) The correlation between the absolute value of stability changes and ancestral RSA.

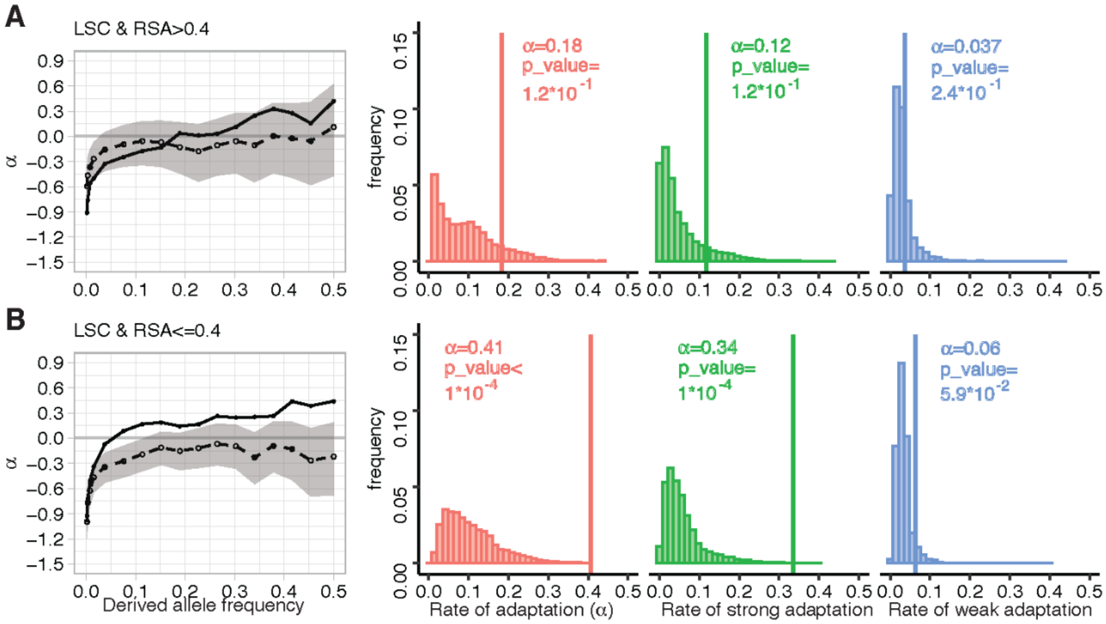
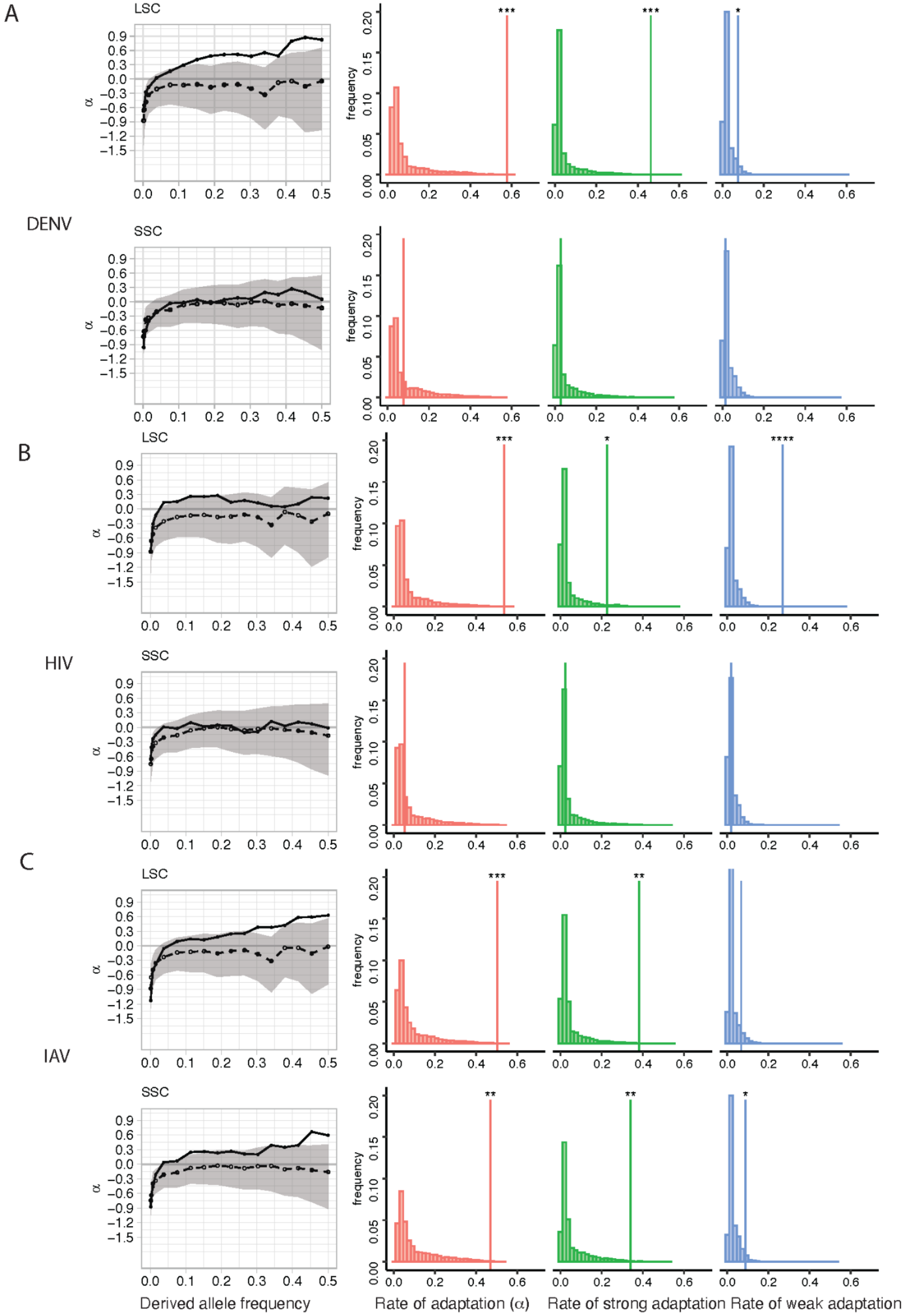
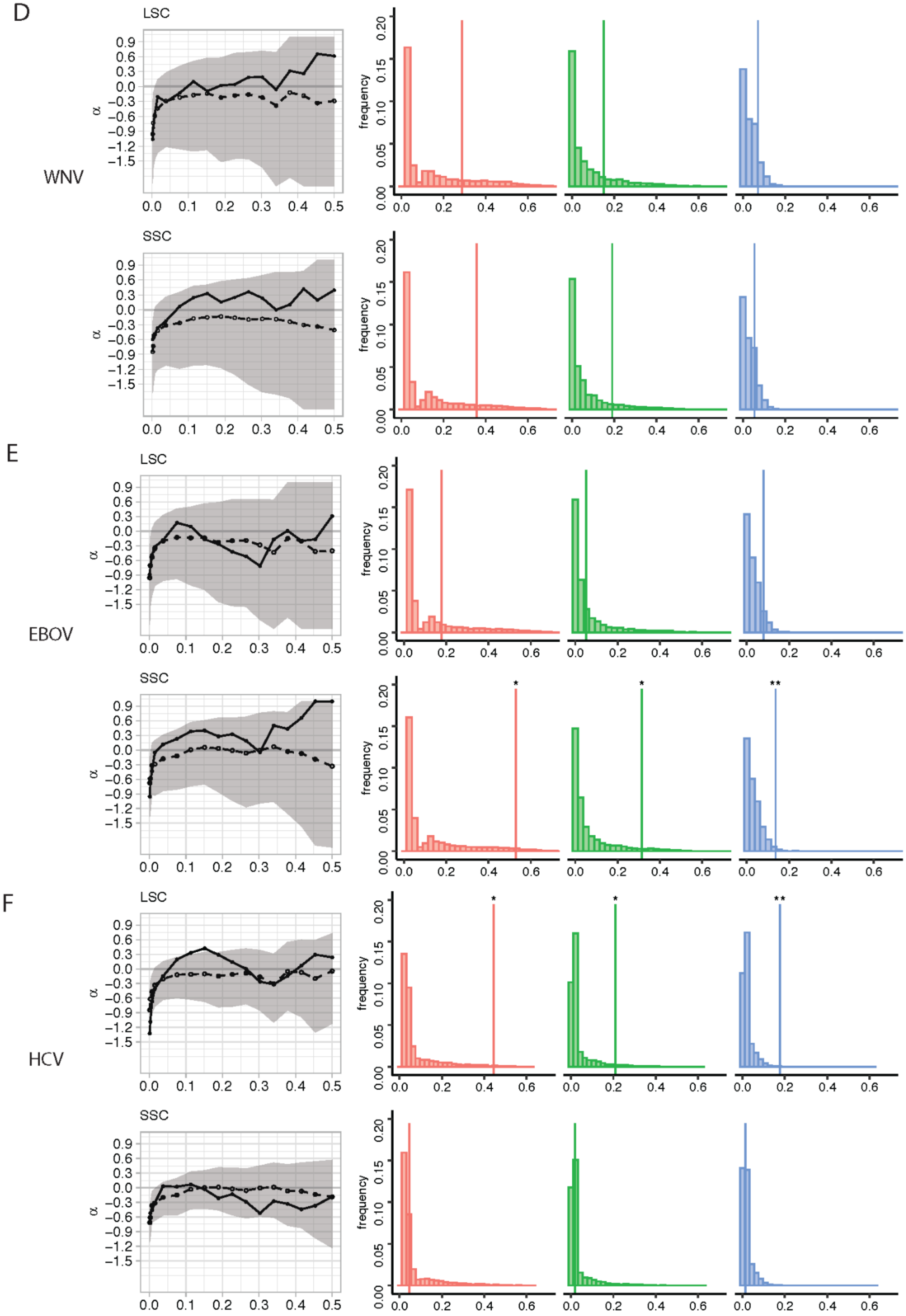
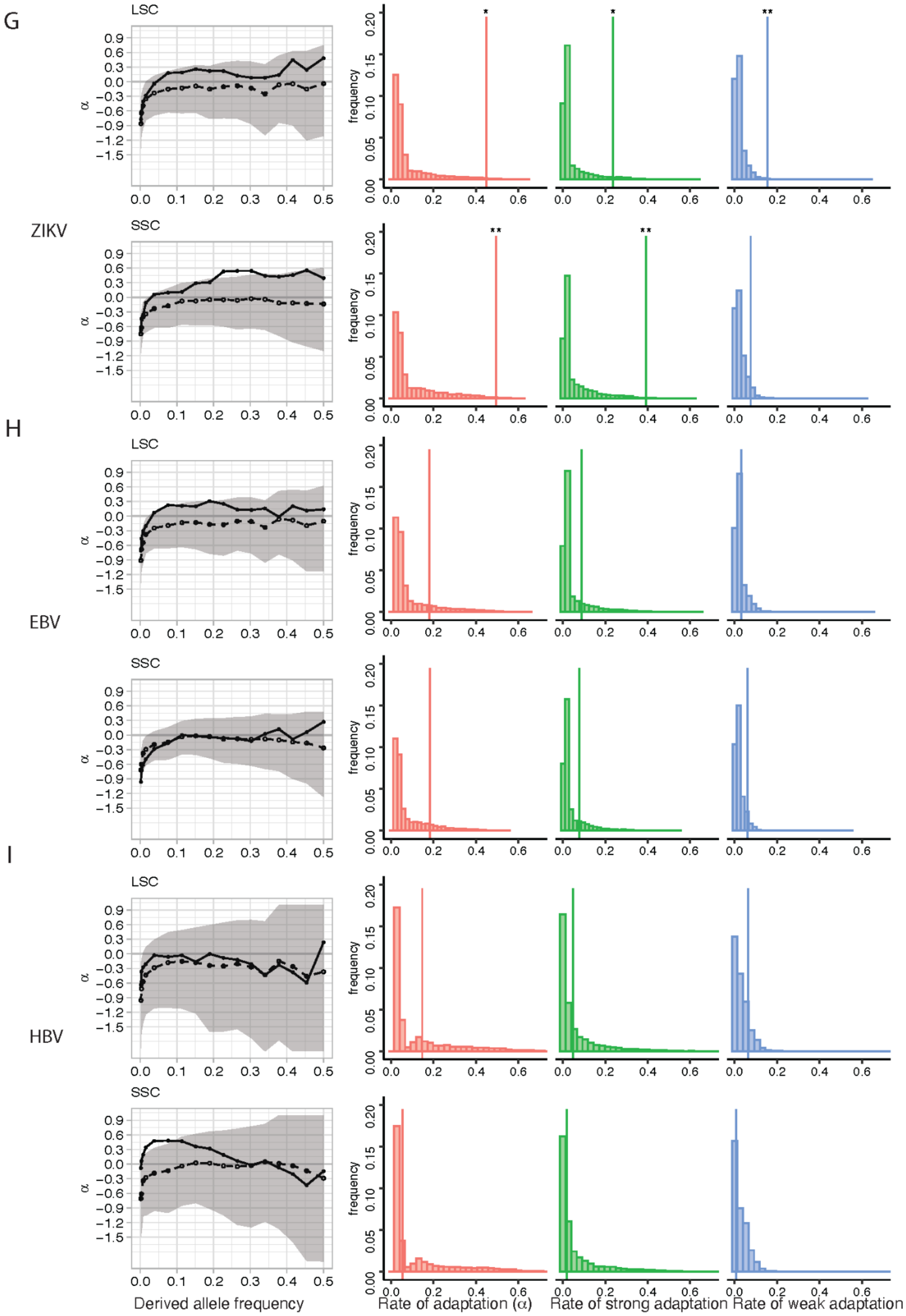


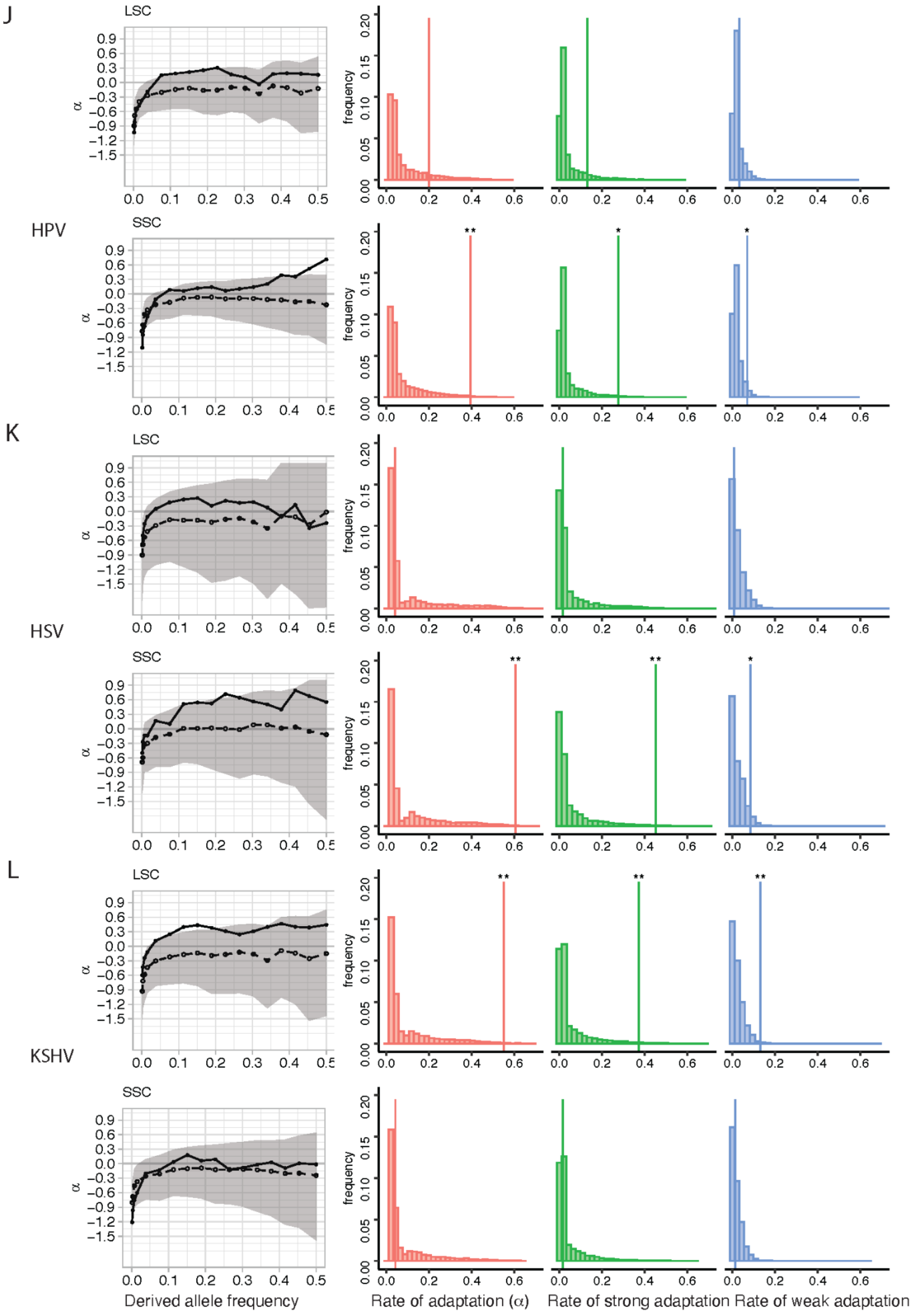
Figure S2. More adaptation in LSC in less exposed sites

Legend same as Figure 2. Instead of splitting mutations by the median RSA in LSC (RSA=0.23), here, we use the median of SSC (RSA=0.4) to split mutations into more exposed and less exposed parts. A) adaptation in LSC in more exposed parts of VIPs and control non-VIPs with RSA≥0.4, the median for SSC B) same as A) but for more buried parts with RSA<0.4.


A
DENV
B
HIV
C
IAV
LSC
SSC
α
0.9
0.6
0.3
0.0
−0.3
−0.6
−0.9
−1.2
−1.5
0.0 0.1 0.2 0.3 0.4 0.5
frequency
0.00 0.05 0.10 0.15 0.20
0.0 0.2 0.4 0.6
***
*
****
**
Derived allele frequency
Rate of adaptation (α)
Rate of strong adaptation
Rate of weak adaptation


D
WNV
LSC
SSC
E
EBOV
LSC
SSC
F
HCV
LSC
SSC
α
0.9
0.6
0.3
0.0
−0.3
−0.6
−0.9
−1.2
−1.5
0.0 0.1 0.2 0.3 0.4 0.5
frequency
0.00 0.05 0.10 0.15 0.20
0.0 0.2 0.4 0.6
*
**


G
ZIKV
H
EBV
I
HBV
LSC
SSC
α
0.9
0.6
0.3
0.0
−0.3
−0.6
−0.9
−1.2
−1.5
0.0 0.1 0.2 0.3 0.4 0.5
Derived allele frequency
frequency
0.00 0.05 0.10 0.15 0.20
0.0 0.2 0.4 0.6
*
**
Rate of adaptation (α)
Rate of strong adaptation
Rate of weak adaptation


J
HPV
LSC
SSC
K
HSV
LSC
SSC
L
KSHV
LSC
SSC
α
0.9
0.6
0.3
0.0
−0.3
−0.6
−0.9
−1.2
−1.5
0.0 0.1 0.2 0.3 0.4 0.5
frequency
0.00 0.05 0.10 0.15 0.20
0.0 0.2 0.4 0.6
**
*
Derived allele frequency
Rate of adaptation (α)
Rate of strong adaptation
Rate of weak adaptation

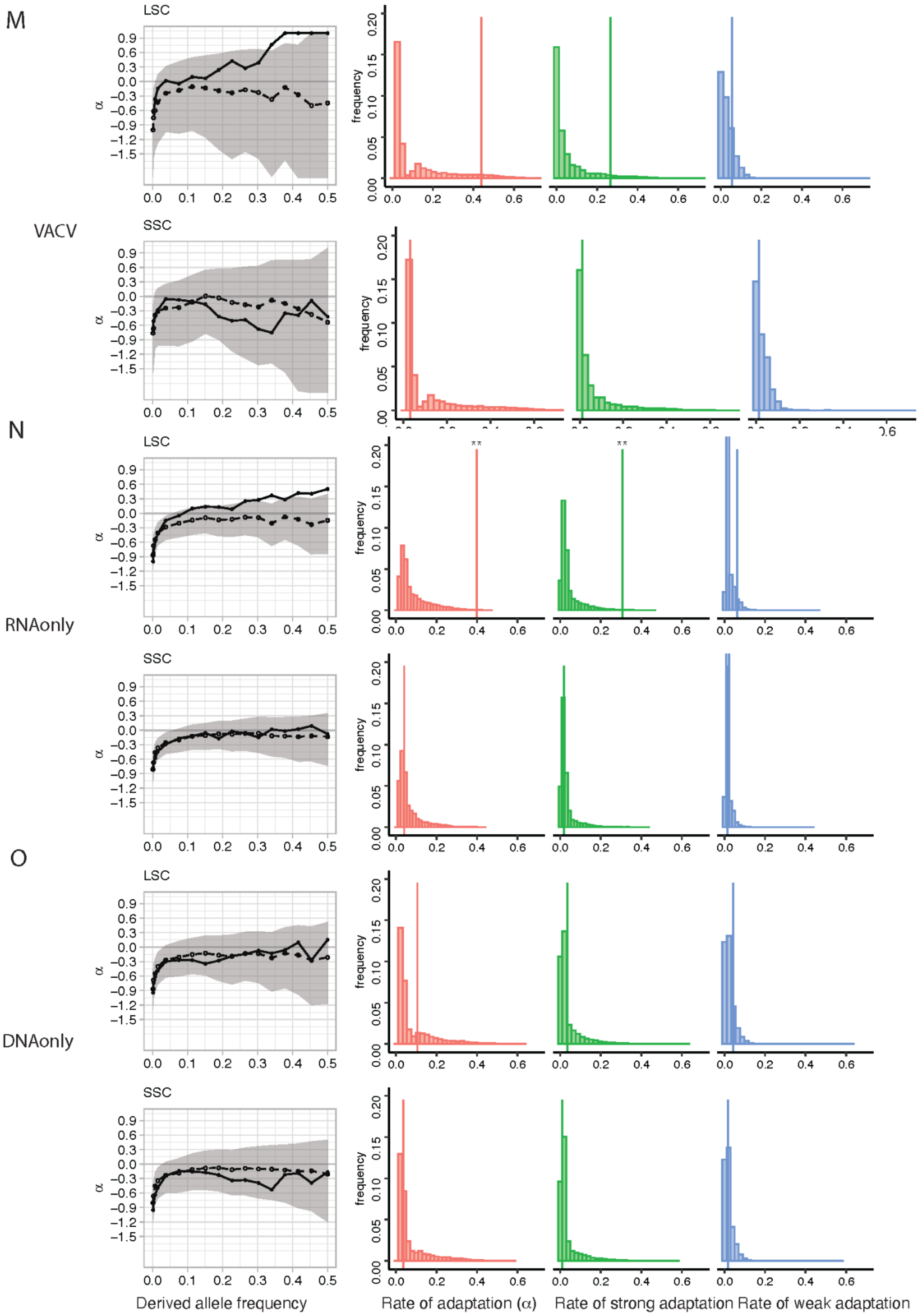

M
N
O
VACV
RNAonly
DNAonly
LSC
SSC
LSC
SSC
LSC
SSC
α
0.9
0.6
0.3
0.0
−0.3
−0.6
−0.9
−1.2
−1.5
0.0 0.1 0.2 0.3 0.4 0.5
frequency
0.00 0.05 0.10 0.15 0.20
0.0 0.2 0.4 0.6
**
**
Derived allele frequency
Rate of adaptation (α)
Rate of strong adaptation
Rate of weak adaptation

Figure S3. Adaptation in LSC and SSC for VIPs interacting with different viruses

Legend same as Figure 2. Adaptation in LSC and SSC for VIPs interacting with different viruses from A-O are DENV, HIV, IAV, WNV, EBOV, HCV, ZIKV, EBV, HBV, HPV, HSV, KSHV, VACV, RNAonly, DNAonly. The ranges of P-values are: <=0.05 (*), <=0.01 (**), <=0.001 (***), <=0.0001 (****).

References

Águeda-Pinto, Ana, Ana Lemos de Matos, Ana Pinheiro, Fabiana Neves, Patrícia de Sousa-Pereira, and Pedro J. Esteves. 2019. Not so Unique to Primates: The Independent Adaptive Evolution of TRIM5 in Lagomorpha Lineage. *PLOS ONE* 14 (12): e0226202. https://doi.org/10.1371/journal.pone.0226202.

Amberger, Joanna S, Carol A Bocchini, Alan F Scott, and Ada Hamosh. 2019. OMIM.Org: Leveraging Knowledge across Phenotype-Gene Relationships. *Nucleic Acids Research* 47 (D1): D1038–43. https://doi.org/10.1093/nar/gky1151.

Araya, Carlos L., Douglas M. Fowler, Wentao Chen, Ike Muniez, Jeffery W. Kelly, and Stanley Fields. 2012. A Fundamental Protein Property, Thermodynamic Stability, Revealed Solely from Large-Scale Measurements of Protein Function. *Proceedings of the National Academy of Sciences* 109 (42): 16858–63. https://doi.org/10.1073/pnas.1209751109.

Assaf, Zoe June, Dmitri A Petrov, and Jamie R Blundell. 2015. Obstruction of Adaptation in Diploids by Recessive, Strongly Deleterious Alleles. *Proceedings of the National Academy of Sciences* 112 (20): E2658 LP-E2666. https://doi.org/10.1073/pnas.1424949112.

Auton, Adam, Gonçalo R. Abecasis, David M. Altshuler, Richard M. Durbin, Gonçalo R. Abecasis, David R. Bentley, Aravinda Chakravarti, et al. 2015. A Global Reference for Human Genetic Variation. *Nature* 526 (7571): 68–74. https://doi.org/10.1038/nature15393.

Balick, Daniel J., Ron Do, Christopher A. Cassa, David Reich, and Shamil R. Sunyaev. 2015. Dominance of Deleterious Alleles Controls the Response to a Population Bottleneck. *PLoS Genetics* 11 (8): 1–23. https://doi.org/10.1371/journal.pgen.1005436.

Bayer, R., & Spitzer, R. L. (1982). Edited correspondence on the status of homosexuality in DSM-III. Journal of the history of the behavioral sciences, 18(1), 32–52. https://doi.org/10.1002/1520-6696(198201)18:1<32::aid-jhbs2300180105>3.0.co;2-0

Benton, M.L., Abraham, A., LaBella, A.L. et al. The influence of evolutionary history on human health and disease. Nat Rev Genet 22, 269–283 (2021). https://doi.org/10.1038/s41576-020-00305-9

Bersaglieri, Todd, Pardis C. Sabeti, Nick Patterson, Trisha Vanderploeg, Steve F. Schaffner, Jared A. Drake, Matthew Rhodes, David E. Reich, and Joel N. Hirschhorn. 2004. Genetic Signatures of Strong Recent Positive Selection at the Lactase Gene. *American Journal of Human Genetics* 74 (6): 1111–20. https://doi.org/10.1086/421051.

Beyer, Robert M., Mario Krapp, Anders Eriksson, and Andrea Manica. 2021. Climatic Windows for Human Migration out of Africa in the Past 300,000 Years. *Nature Communications* 12 (1): 4889. https://doi.org/10.1038/s41467-021-24779-1.

Blekhman, Ran, Orna Man, Leslie Herrmann, Adam R. Boyko, Amit Indap, Carolin Kosiol, Carlos D. Bustamante, Kosuke M. Teshima, and Molly Przeworski. 2008. Natural Selection on Genes That Underlie Human Disease Susceptibility. *Current Biology* 18 (12): 883–89. https://doi.org/10.1016/j.cub.2008.04.074.

Boorse, Christopher. 1975. On the Distinction between Disease and Illness. *Philosophy & Public Affairs* 5

(1): 49–68.

Boorse, Christopher. 1977. Health as a Theoretical Concept. *Philosophy of Science* 44 (4): 542–73. https://doi.org/10.1086/288768.

Buniello, Annalisa, Jacqueline A L Macarthur, Maria Cerezo, Laura W Harris, James Hayhurst, Cinzia Malangone, Aoife Mcmahon, et al. 2019. The NHGRI-EBI GWAS Catalog of Published Genome-Wide Association Studies, Targeted Arrays and Summary Statistics 2019. *Nucleic Acids Research* 47 (Database issue): 1005–12. https://doi.org/10.1016/0038-1098(79)91043-3.

Bustamante, Carlos D., Adi Fledel-Alon, Scott Williamson, Rasmus Nielsen, Melissa Todd Hubisz, Stephen Glanowski, David M. Tanenbaum, et al. 2005. Natural Selection on Protein-Coding Genes in the Human Genome. *Nature* 437 (7062): 1153–57. https://doi.org/10.1038/nature04240.

Cagiada, Matteo, Kristoffer E. Johansson, Audrone Valanciute, Sofie V. Nielsen, Rasmus Hartmann-Petersen, Jun J. Yang, Douglas M. Fowler, Amelie Stein, and Kresten Lindorff-Larsen. 2021. Understanding the Origins of Loss of Protein Function by Analyzing the Effects of Thousands of Variants on Activity and Abundance. *Molecular Biology and Evolution* 38 (8): 3235–46. https://doi.org/10.1093/molbev/msab095.

Castellano, David, Lawrence H. Uricchio, Kasper Munch, and David Enard. 2019. Viruses Rule over Adaptation in Conserved Human Proteins. *BioRxiv*, January, 555060. https://doi.org/10.1101/555060.

Charlesworth, B., M. T. Morgan, and D. Charlesworth. 1993. The Effect of Deleterious Mutations on Neutral Molecular Variation. *Genetics* 134 (4): 1289–1303.

Claussnitzer, Melina, Judy H. Cho, Rory Collins, Nancy J. Cox, Emmanouil T. Dermitzakis, Matthew E. Hurles, Sekar Kathiresan, et al. 2020. A Brief History of Human Disease Genetics. *Nature* 577 (7789): 179–89. https://doi.org/10.1038/s41586-019-1879-7.

Corbett, Stephen, Alexandre Courtiol, Virpi Lummaa, Jacob Moorad, and Stephen Stearns. 2018. The Transition to Modernity and Chronic Disease: Mismatch and Natural Selection. *Nature Reviews. Genetics* 19 (7): 419–30. https://doi.org/10.1038/s41576-018-0012-3.

Cuadros-Espinoza, Sebastian, Guillaume Laval, Lluis Quintana-Murci, and Etienne Patin. 2022. The Genomic Signatures of Natural Selection in Admixed Human Populations. *The American Journal of Human Genetics* 109 (4): 710–26. https://doi.org/10.1016/j.ajhg.2022.02.011.

Darwin, Charles, and Alfred Wallace. 1858. On the Tendency of Species to Form Varieties; and on the Perpetuation of Varieties and Species by Natural Means of Selection. *Journal of the Proceedings of the Linnean Society of London. Zoology* 3 (9): 45–62. https://doi.org/10.1111/j.1096-3642.1858.tb02500.x.

Daub, J. T., S. Moretti, I. I. Davydov, L. Excoffier, and M. Robinson-Rechavi. 2017. Detection of Pathways Affected by Positive Selection in Primate Lineages Ancestral to Humans. *Molecular Biology and Evolution* 34 (6): 1391–1402. https://doi.org/10.1093/molbev/msx083.

Daub, Josephine T., Isabelle Dupanloup, Marc Robinson-Rechavi, and Laurent Excoffier. 2015. Inference of Evolutionary Forces Acting on Human Biological Pathways. *Genome Biology and Evolution* 7 (6): 1546–58. https://doi.org/10.1093/gbe/evv083.

Di, Chenlu, Jesus Murga Moreno, Diego F Salazar-Tortosa, M Elise Lauterbur and David Enard. 2021. Decreased recent adaptation at human mendelian disease genes as a possible consequence of interference between advantageous and deleterious variants. *ELife*:10:e69026. https://doi.org/10.7554/eLife.69026

Di, Chenlu, Jesus Murga-Moreno and David Enard. 2022. Stability evolution as a major mechanism of human protein adaptation in response to viruses. *bioRxiv*: 12.01.518739. https://doi.org/10.1101/2022.12.01.518739

Elde, Nels C., Stephanie J. Child, Adam P. Geballe, and Harmit S. Malik. 2009. Protein Kinase R Reveals an Evolutionary Model for Defeating Viral Mimicry. *Nature* 457 (7228): 485–89. https://doi.org/10.1038/nature07529.

Enard, David, Le Cai, Carina Gwennap, and Dmitri A. Petrov. 2016. Viruses Are a Dominant Driver of Protein Adaptation in Mammals. *ELife* 5: 1–25. https://doi.org/10.7554/eLife.12469.

Enard, David, and Dmitri A. Petrov. 2018. Evidence That RNA Viruses Drove Adaptive Introgression between Neanderthals and Modern Humans. *Cell* 175 (2): 360-371.e13. https://doi.org/10.1016/j.cell.2018.08.034.

Enard, David, and Dmitri A Petrov. 2020. Ancient RNA Virus Epidemics through the Lens of Recent Adaptation in Human Genomes. *Philosophical Transactions of the Royal Society B: Biological Sciences* 375 (1812): 20190575. https://doi.org/10.1098/rstb.2019.0575.

Ferrer-Admetlla, Anna, Mason Liang, Thorfinn Korneliussen, and Rasmus Nielsen. 2014. On Detecting Incomplete Soft or Hard Selective Sweeps Using Haplotype Structure. *Molecular Biology and Evolution* 31 (5): 1275–91. https://doi.org/10.1093/molbev/msu077.

Galtier, Nicolas. 2016. Adaptive Protein Evolution in Animals and the Effective Population Size Hypothesis. *PLoS Genetics* 12 (1): 1–23. https://doi.org/10.1371/journal.pgen.1005774.

Gerasimavicius, Lukas, Xin Liu, and Joseph A Marsh. 2020. Identification of Pathogenic Missense Mutations Using Protein Stability Predictors. *Scientific Reports* 10 (1): 15387. https://doi.org/10.1038/s41598-020-72404-w.

Gibbs, Richard A., John W. Belmont, Paul Hardenbol, Thomas D. Willis, Fuli Yu, Huanming Yang, Lan-Yang Ch'ang, et al. 2003. The International HapMap Project. *Nature* 426 (6968): 789–96. https://doi.org/10.1038/nature02168.

Goldenzweig, Adi, and Sarel J. Fleishman. 2018. Principles of Protein Stability and Their Application in Computational Design. *Annual Review of Biochemistry* 87 (June): 105–29. https://doi.org/10.1146/annurev-biochem-062917-012102.

Goldman, N, and Ziheng Yang. 1994. A Codon-Based Model of Nucleotide Substitution for Protein-Coding DNA Sequences. *Molecular Biology and Evolution* 11 (5): 725–36. https://doi.org/10.1093/oxfordjournals.molbev.a040153.

Goldman, Nick, and Ziheng Yang. 1994. A Codon-Based Model of Nucleotide Substitution for Protein-Coding DNA Sequences. *Molecular Biology and Evolution*. https://doi.org/10.1093/oxfordjournals.molbev.a040153.

Gouy, Alexandre, Joséphine T. Daub, and Laurent Excoffier. 2017. Detecting Gene Subnetworks under Selection in Biological Pathways. *Nucleic Acids Research* 45 (16): e149. https://doi.org/10.1093/nar/gkx626.

Groucutt, Huw S., Michael D. Petraglia, Geoff Bailey, Eleanor M.L. Scerri, Ash Parton, Laine Clark-Balzan, Richard P. Jennings, et al. 2015. Rethinking the Dispersal of Homo Sapiens out of Africa. *Evolutionary Anthropology* 24 (4): 149–64. https://doi.org/10.1002/evan.21455.

Haldane, J.B.S. 1924. A Mathematical Theory of Natural and Artificial Selection. Part 1. 57 Transactions of the Cambridge Philosophical Society, 23: 19-41.

Høie, Magnus Haraldson, Matteo Cagiada, Anders Haagen Beck Frederiksen, Amelie Stein, and Kresten Lindorff-Larsen. 2022. Predicting and Interpreting Large-Scale Mutagenesis Data Using Analyses of Protein Stability and Conservation. *Cell Reports* 38 (2): 110207. https://doi.org/10.1016/j.celrep.2021.110207.

Jumper, John, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, et al. 2021. Highly Accurate Protein Structure Prediction with AlphaFold. *Nature* 596 (7873): 583–89. https://doi.org/10.1038/s41586-021-03819-2.

Katoh, Shigehiro, Yonas Beyene, Tetsumaru Itaya, Hironobu Hyodo, Masayuki Hyodo, Koshi Yagi, Chitaro Gouzu, et al. 2016. New Geological and Palaeontological Age Constraint for the Gorilla–Human Lineage Split. *Nature* 530 (7589): 215–18. https://doi.org/10.1038/nature16510.

Klunk, Jennifer, Tauras P. Vilgalys, Christian E. Demeure, Xiaoheng Cheng, Mari Shiratori, Julien Madej, Rémi Beau, et al. 2022. Evolution of Immune Genes Is Associated with the Black Death. *Nature* 611 (7935): 312–19. https://doi.org/10.1038/s41586-022-05349-x.

Konaté, Mariam M, Germán Plata, Jimin Park, Dinara R Usmanova, Harris Wang, and Dennis Vitkup. 2019. Molecular Function Limits Divergent Protein Evolution on Planetary Timescales. Edited by Nir Ben-Tal, Diethard Tautz, and Nir Ben-Tal. *ELife* 8 (September): e39705. https://doi.org/10.7554/eLife.39705.

Landrum, Melissa J., Jennifer M. Lee, George R. Riley, Wonhee Jang, Wendy S. Rubinstein, Deanna M. Church, and Donna R. Maglott. 2014. ClinVar: Public Archive of Relationships among Sequence Variation and Human Phenotype. *Nucleic Acids Research* 42 (Database issue): D980-985. https://doi.org/10.1093/nar/gkt1113.

Langergraber, Kevin E., Kay Prüfer, Carolyn Rowney, Christophe Boesch, Catherine Crockford, Katie Fawcett, Eiji Inoue, et al. 2012. Generation Times in Wild Chimpanzees and Gorillas Suggest Earlier Divergence Times in Great Ape and Human Evolution. *Proceedings of the National Academy of Sciences* 109 (39): 15716–21. https://doi.org/10.1073/pnas.1211740109.

Lauterbur, M. Elise, Kasper Munch, and David Enard. 2022. Versatile Detection of Diverse Selective Sweeps with Flex-Sweep. bioRxiv. https://doi.org/10.1101/2022.11.15.516494.

Lee, David, Oliver Redfern, and Christine Orengo. 2007. Predicting Protein Function from Sequence and Structure. *Nature Reviews Molecular Cell Biology* 8 (12): 995–1005. https://doi.org/10.1038/nrm2281.

Lewens, Tim, and John McMillan. 2004. Defining Disease. *The Lancet* 363 (9409): 664.

https://doi.org/10.1016/S0140-6736(04)15616-X.

Li, Bian, Yucheng T. Yang, John A. Capra, and Mark B. Gerstein. 2020. Predicting Changes in Protein Thermodynamic Stability upon Point Mutation with Deep 3D Convolutional Neural Networks. *PLoS Computational Biology* 16 (11): 1–24. https://doi.org/10.1371/journal.pcbi.1008291.

Macpherson, J. Michael, Josefa González, Daniela M. Witten, Jerel C. Davis, Noah A. Rosenberg, Aaron E. Hirsh, and Dmitri A. Petrov. 2008. Nonadaptive Explanations for Signatures of Partial Selective Sweeps in Drosophila. *Molecular Biology and Evolution* 25 (6): 1025–42. https://doi.org/10.1093/molbev/msn007.

Martelli, Pier Luigi, Piero Fariselli, Castrense Savojardo, Giulia Babbi, Francesco Aggazio, and Rita Casadio. 2016. Large Scale Analysis of Protein Stability in OMIM Disease Related Human Protein Variants. *BMC Genomics* 17 (Suppl 2). https://doi.org/10.1186/s12864-016-2726-y.

Mathieson, Sara, and Iain Mathieson. 2018. FADS1 and the Timing of Human Adaptation to Agriculture. *Molecular Biology and Evolution* 35 (12): 2957–70. https://doi.org/10.1093/molbev/msy180.

McDonald, John H., and Martin Kreitman. 1991. Adaptive Protein Evolution at the Adh Locus in Drosophila. *Nature* 354: 56–58.

Minster, Ryan L, Nicola L Hawley, Chi-Ting Su, Guangyun Sun, Erin E Kershaw, Hong Cheng, Olive D Buhule, et al. 2016. A Thrifty Variant in CREBRF Strongly Influences Body Mass Index in Samoans. *Nature Genetics* 48 (9): 1049–54. https://doi.org/10.1038/ng.3620.

Nielsen, Rasmus. 2001. Statistical Tests of Selective Neutrality in the Age of Genomics. *Heredity* 86 (6): 641–47. https://doi.org/10.1046/j.1365-2540.2001.00895.x.

Nielsen, Rasmus, Carlos Bustamante, Andrew G. Clark, Stephen Glanowski, Timothy B. Sackton, Melissa J. Hubisz, Adi Fledel-Alon, et al. 2005. A Scan for Positively Selected Genes in the Genomes of Humans and Chimpanzees. *PLoS Biology* 3 (6): 0976–85. https://doi.org/10.1371/journal.pbio.0030170.

Otto, Sarah P. 2004. Two Steps Forward, One Step Back: The Pleiotropic Effects of Favoured Alleles. *Proceedings. Biological Sciences* 271 (1540): 705–14. https://doi.org/10.1098/rspb.2003.2635.

Pavlidis, Pavlos, Jeffrey D. Jensen, Wolfgang Stephan, and Alexandros Stamatakis. 2012. A Critical Assessment of Storytelling: Gene Ontology Categories and the Importance of Validating Genomic Scans. *Molecular Biology and Evolution* 29 (10): 3237–48.

Pickrell, Joseph K., Graham Coop, John Novembre, Sridhar Kudaravalli, Jun Z. Li, Devin Absher, Balaji S. Srinivasan, et al. 2009. Signals of Recent Positive Selection in a Worldwide Sample of Human Populations. *Genome Research* 19 (5): 826–37. https://doi.org/10.1101/gr.087577.108.

Piñero, Janet, Núria Queralt-Rosinach, Àlex Bravo, Jordi Deu-Pons, Anna Bauer-Mehren, Martin Baron, Ferran Sanz, and Laura I. Furlong. 2015. DisGeNET: A Discovery Platform for the Dynamical Exploration of Human Diseases and Their Genes. *Database* 2015: 1–17. https://doi.org/10.1093/database/bav028.

Piñero, Janet, Juan Manuel Ramírez-Anguita, Josep Saüch-Pitarch, Francesco Ronzano, Emilio Centeno, Ferran Sanz, and Laura I. Furlong. 2020. The DisGeNET Knowledge Platform for Disease

Genomics: 2019 Update. *Nucleic Acids Research* 48 (D1): D845–55. https://doi.org/10.1093/nar/gkz1021.

Quintana-Murci, Lluis. 2016. Understanding Rare and Common Diseases in the Context of Human Evolution. *Genome Biology* 17 (1): 225. https://doi.org/10.1186/s13059-016-1093-y.

Radivojac, Predrag, Wyatt T. Clark, Tal Ronnen Oron, Alexandra M. Schnoes, Tobias Wittkop, Artem Sokolov, Kiley Graim, et al. 2013. A Large-Scale Evaluation of Computational Protein Function Prediction. *Nature Methods* 10 (3): 221–27. https://doi.org/10.1038/nmeth.2340.

Rappaport, Noa, Noam Nativ, Gil Stelzer, Michal Twik, Yaron Guan-Golan, Tsippi Iny Stein, Iris Bahir, et al. 2013. MalaCards: An Integrated Compendium for Diseases and Their Annotation. *Database* 2013: 1–14. https://doi.org/10.1093/database/bat018.

Rothenburg, Stefan, Eun Joo Seo, James S. Gibbs, Thomas E. Dever, and Katharina Dittmar. 2009. Rapid Evolution of Protein Kinase PKR Alters Sensitivity to Viral Inhibitors. *Nature Structural & Molecular Biology* 16 (1): 63–70. https://doi.org/10.1038/nsmb.1529.

Sabeti, Pardis C., John M. Higgins* , David E. Reich*, Stephen F. Schaffner* Haninah Z. P. Levine*, Daniel J. Richter*, Gavin J. McDonald* Stacey B. Gabriel*, Jill V. Platko*, Nick J. Patterson*, David Altshuler*§ Hans C. Ackerman‡, Sarah J. Campbell‡, Ryk Ward† & Eric S. Lander* Richard Cooperk, Dominic Kwiatkowski‡, and Thomas Eisner. 2002. Detecting Recent Positive Selection in the Human Genome from Haplotype Structure. *Nature* 419 (October). https://doi.org/10.1038/nature01027.

Sabeti, Pardis C., SF Schaffner, Fry B., Lohmueller J., P. Varilly, O. Shamovsky, A. Palma, T.S. Mikkelsen, D. Altshuler, and E.S. Lander. 2006. Positive Natural Selection in the Human Lineage 312: 1614–20. https://doi.org/10.1126/science.1124309.

Sarich, Vincent M., and Allan C. Wilson. 1967. Immunological Time Scale for Hominid Evolution. *Science* 158 (3805): 1200–1203. https://doi.org/10.1126/science.158.3805.1200.

Sawyer, Sara L., Lily I. Wu, Michael Emerman, and Harmit S. Malik. 2005. Positive Selection of Primate TRIM5α Identifies a Critical Species-Specific Retroviral Restriction Domain. *Proceedings of the National Academy of Sciences* 102 (8): 2832–37. https://doi.org/10.1073/pnas.0409853102.

Schrider, Daniel R., and Andrew D. Kern. 2017. Soft Sweeps Are the Dominant Mode of Adaptation in the Human Genome. *Molecular Biology and Evolution* 34 (8): 1863–77. https://doi.org/10.1093/molbev/msx154.

Scully, Jackie Leach. 2004. What Is a Disease? *EMBO Reports* 5 (7): 650–53. https://doi.org/10.1038/sj.embor.7400195.

Smith, John Maynard, and John Haigh. 1974. The Hitch-Hiking Effect of a Favourable Gene. *Genetical Research* 23 (1): 23–35. https://doi.org/10.1017/S0016672300014634.

Souilmi, Yassine, M. Elise Lauterbur, Ray Tobler, Christian D. Huber, Angad S. Johar, Shayli Varasteh Moradi, Wayne A. Johnston, Nevan J. Krogan, Kirill Alexandrov, and David Enard. 2021. An Ancient Viral Epidemic Involving Host Coronavirus Interacting Genes More than 20,000 Years Ago in East Asia. *Current Biology* 31 (16): 3504-3514.e9. https://doi.org/10.1016/j.cub.2021.05.067.

Stein, Amelie, Douglas M. Fowler, Rasmus Hartmann-Petersen, and Kresten Lindorff-Larsen. 2019. Biophysical and Mechanistic Models for Disease-Causing Protein Variants. *Trends in Biochemical Sciences* 44 (7): 575–88. https://doi.org/10.1016/j.tibs.2019.01.003.

Sugden, Lauren Alpert, Elizabeth G. Atkinson, Annie P. Fischer, Stephen Rong, Brenna M. Henn, and Sohini Ramachandran. 2018. Localization of Adaptive Variants in Human Genomes Using Averaged One-Dependence Estimation. *Nature Communications* 9 (1): 703. https://doi.org/10.1038/s41467-018-03100-7.

The UniProt Consortium. 2014. Activities at the Universal Protein Resource (UniProt). *Nucleic Acids Research* 42 (Database issue): 191–98. https://doi.org/10.1093/nar/gkt1140.

Tishkoff, Sarah A., Floyd A. Reed, Alessia Ranciaro, Benjamin F. Voight, Courtney C. Babbitt, Jesse S. Silverman, Kweli Powell, et al. 2007. Convergent Adaptation of Human Lactase Persistence in Africa and Europe. *Nature Genetics* 39 (1): 31–40. https://doi.org/10.1038/ng1946.

Tokuriki, Nobuhiko, Francois Stricher, Joost Schymkowitz, Luis Serrano, and Dan S. Tawfik. 2007. The Stability Effects of Protein Mutations Appear to Be Universally Distributed. *Journal of Molecular Biology* 369 (5): 1318–32. https://doi.org/10.1016/j.jmb.2007.03.069.

Torres, Raul, Zachary A. Szpiech, and Ryan D. Hernandez. 2018. Human Demographic History Has Amplified the Effects of Background Selection across the Genome. *PLOS Genetics* 14 (6): e1007387. https://doi.org/10.1371/journal.pgen.1007387.

Uricchio, Lawrence H., Dmitri A. Petrov, and David Enard. 2019. Exploiting Selection at Linked Sites to Infer the Rate and Strength of Adaptation. *Nature Ecology and Evolution* 3 (6): 977–84. https://doi.org/10.1038/s41559-019-0890-6.

Varadi, Mihaly, Stephen Anyango, Mandar Deshpande, Sreenath Nair, Cindy Natassia, Galabina Yordanova, David Yuan, et al. 2022. AlphaFold Protein Structure Database: Massively Expanding the Structural Coverage of Protein-Sequence Space with High-Accuracy Models. *Nucleic Acids Research* 50 (D1): D439–44. https://doi.org/10.1093/nar/gkab1061.

Voight, Benjamin F., Sridhar Kudaravalli, Xiaoquan Wen, and Jonathan K. Pritchard. 2006. A Map of Recent Positive Selection in the Human Genome. *PLoS Biology* 4 (3): 0446–0659. https://doi.org/10.1371/journal.pbio.0040072.

World Health Organization. 1994. Assessment of Fracture Risk and Its Application to Screening for Postmenopausal Osteoporosis : Report of a WHO Study Group [Meeting Held in Rome from 22 to 25 June 1992]. World Health Organization. https://apps.who.int/iris/handle/10665/39142.

Yang, Ziheng. 1998. Likelihood Ratio Tests for Detecting Positive Selection and Application to Primate Lysozyme Evolution. *Molecular Biology and Evolution* 15 (5): 568–73. https://doi.org/10.1093/oxfordjournals.molbev.a025957.

Yang, Ziheng, and Rasmus Nielsent. 2002. Codon-Substitution Models for Detecting Molecular Adaptation at Individual Sites along Specific Lineages. *Molecular Biology and Evolution* 19 (6): 908–17. https://doi.org/10.1093/oxfordjournals.molbev.a004148.

Yang, Ziheng, and Mario dos Reis. 2011. Statistical Properties of the Branch-Site Test of Positive

Selection. *Molecular Biology and Evolution* 28 (3): 1217–28. https://doi.org/10.1093/molbev/msq303.

ProQuest LLC
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106 - 1346 USA