Uncovering the extensive trade-off between adaptive evolution and disease susceptibility

Graphical abstract

Highlights
- This study identifies and analyzes favored mutations in three human populations
- Favored and hitchhiking mutations are enriched in GWAS sites' population specifically
- Favored mutations make the carriers adapt to environments and susceptible to diseases

Authors
Ji Tang, Maosheng Huang, Sha He, Junxiang Zeng, Hao Zhu

Correspondence
zhuhao@smu.edu.cn

In brief
Tang et al. show the extensive trade-offs occurring between adaptive evolution and disease susceptibility. This relationship is critical for understanding and investigating humans’ genetic and phenotypic differences and the basis of human diseases.
Uncovering the extensive trade-off between adaptive evolution and disease susceptibility

Ji Tang,1,4 Maosheng Huang,1,3,4 Sha He,1,4 Junxiang Zeng,1 and Hao Zhu1,2,5,*

1Bioinformatics Section, School of Basic Medical Sciences, Southern Medical University, Guangzhou 510515, China
2Guangdong-Hong Kong-Macao Greater Bay Area Center for Brain Science and Brain-Inspired Intelligence, Southern Medical University, Guangzhou 510515, China
3School of Medical Information and Engineering, Guangdong Pharmaceutical University, Guangzhou 510006, China
4These authors contributed equally
5Lead contact
*Correspondence: zhuhao@smu.edu.cn
https://doi.org/10.1016/j.celrep.2022.111351

SUMMARY

Favored mutations in the human genome may make the carriers adapt to changing environments and lifestyles but also susceptible to specific diseases. The scale and details of the trade-off between adaptive evolution and disease susceptibility are unclear because most favored mutations in different populations remain unidentified. As no statistical test can discriminate favored mutations from nearby hitchhiking neutral ones, we report a deep-learning network (DeepFavored) to integrate multiple statistical tests and divide identifying favored mutations into two subtasks. We identify favored mutations in three human populations and analyzed the correlation between favored/hitchhiking mutations and genome-wide association study (GWAS) sites. Both favored and hitchhiking neutral mutations are enriched in GWAS sites with population-specific features, and the enrichment and population specificity are prominent in genes in specific Gene Ontology (GO) terms. These provide evidence for extensive and population-specific trade-offs between adaptive evolution and disease susceptibility. The unveiled scale helps understand and investigate differences and diseases of humans.

INTRODUCTION

The largely 1% genetic divergence between humans and chimpanzees suggests that many small changes in the human genome critically determine human-specific traits. During and after the out-of-Africa migration, single-nucleotide polymorphisms (SNPs) occurred further to make humans adapt to different environments and lifestyles but also susceptible or resistant to different pathogens and diseases (Lindesmith et al., 2003; Prohaska et al., 2019; Stobdan et al., 2017; Xue et al., 2006). The NCBI dbSNP database has 150 million SNPs documented for the human genome, but how many are favored mutations in different populations is unknown. Genome-wide association studies (GWASs) that detect trait/disease-SNP associations have identified abundant disease- or trait-associated SNPs (called GWAS sites). Some well-known favored mutations are both population and disease associated. For example, the mutations in the HBB gene make some Africans resistant to P. falciparum malaria but also susceptible to sickle cell anemia (Ackerman et al., 2005), and the mutations in genes encoding hypoxia-inducible factors (HIFs) make the carriers benefit from increased oxygen delivery but also suffer from increased blood viscosity (a contributing factor to the high incidence of stroke in Tibetans [Collaborators, 2021]). These examples highlight a trade-off between adaptive evolution and disease susceptibility (Benton et al., 2021; Prohaska et al., 2019). However, the scale, genomic/disease distribution, and population specificity remain unknown.

To address these issues, systematically identifying favored mutations in multiple populations is required. When a mutation is selected in a population, it generates multiple selection signals, especially high allele frequency and genetic hitchhiking (i.e., the frequencies of nearby neutral mutations increase). Many statistical tests have been developed to detect selection signals (Vitti et al., 2013); however, identifying the cause (favored) mutations is more difficult. Especially when a favored mutation arises and goes to a high frequency, linkage disequilibrium (LD) makes nearby neutral mutations become hitchhikers and go to high frequencies as well. These hitchhikers obtain the same signatures as, and thus are difficult to be discriminated from, the favored mutation (Figure 1A). Previous studies have tried to improve identifying favored mutations in two ways: developing statistical tests that directly identify them (e.g., iSAFE) (Akbari et al., 2018) and integrating statistical tests to locate them in selection signals (e.g., CMS) (Grossman et al., 2010). Since selection signals generated by a favored mutation...
are not independent of each other, SWIF(r) uses an averaged one-dependence estimator (AODE) to integrate four statistical tests into a probabilistic model (Sugden et al., 2018). However, even a combination of tests cannot equally powerfully discriminate favored mutations from hitchhiking neutral mutations (called hitchhiking mutations) and ordinary neutral mutations (called ordinary mutations) (Figure 1B). As reported, for each favored mutation identified by CMS, there are on average 20 nearly perfect proxies (LD score r^2 > 0.8) that are indistinguishable from the favored mutation (Grossman et al., 2010). Thus, integrating multiple statistical tests is necessary but not sufficient.

Deep-learning networks can capture undefined features in data and are robust to noise in data (LeCun et al., 2015; Rolnick et al., 2018). Deep learning was used to directly identify selection signals (especially the sweeps of genetic hitchhiking, called selective sweeps) from genomic data (Adrion et al., 2020; Flagel et al., 2019; Sheehan and Song, 2016). A convolutional neural network (CNN), upon a matrix of selection scan statistics, can be used to detect selective sweeps and also to distinguish between the targets of selection, regions/polymorphisms affected by linked positive selection, and regions/polymorphisms unaffected by sweeps (Kern and Schrider, 2018). This study tried to identify favored mutations using a deep-learning network (called DeepFavored) that integrates seven statistical tests and divides identifying favored mutations into two subtasks (Figures 1C–1G). After identifying favored mutations in the three human populations CEU, CHB, and YRI, we jointly analyzed favored mutations, hitchhiking mutations, and GWAS sites. Some Gene Ontology (GO) terms (especially metabolism- and immune-related ones) are significantly co-enriched in favored mutations and GWAS sites. These results provide systematic evidence supporting the extensive trade off between adaptive evolution and disease susceptibility.

RESULTS

Integrating multiple statistical tests using a deep-learning network

Hitchhiking mutations are very similar to favored mutations, and it is hard for any method to equally powerfully discriminate favored mutations from hitchhiking and ordinary mutations.
DeepFavored outperforms existing methods

We then systematically compared the performance of DeepFavored with iSAFE and SWIF(r), the two most recent methods of identifying favored mutations (Akbari et al., 2018; Sugden et al., 2018). First, we used these methods to identify favored mutations in diverse scenarios using simulated data. DeepFavored outperforms iSAFE and SWIF(r) when favored mutations are in both small and large genomic regions, are in soft sweeps, reach a high frequency in the target population, are in CEU and CHB, have a large selection coefficient, and are ancient (Figure S1). iSAFE performs very well if a favored mutation arose recently and its frequency is not high. Since favored and neutral mutations may have varied ratios in large genomic regions, we examined these methods using regions with different ratios of favored and neutral mutations. DeepFavored outperforms the other two when the ratio varies substantially (Figure S1).

Second, we ran DeepFavored, iSAFE, and SWIF(r) consecutively 10 times to identify the 10 strongly suspected favored mutations (which were widely used as gold-standard examples to test algorithms) in CEU, CHB, and YRI (Figure S2) (Akbari et al., 2018; Szpak et al., 2018). The mean and median of the ranks of the 10 favored mutations in 5 Mb regions computed by DeepFavored were smaller (Figure S1H), indicating DeepFavored’s better performance and stability. Third, we examined the ranks of all mutations in the 5 Mb regions containing the 10 gold-standard examples. In many cases, the favored mutations’ ranks computed by DeepFavored were lower than the ranks computed by the other two methods and were more distinct from the ranks of nearby neutral mutations (Figure S3).

Fourth, we examined the genomic regions containing APOL1, SLC22A4, FADS1, BNC2, and TRPM8 and, according to reports, having positive selection signals in these populations (Benton et al., 2021). DeepFavored and SWIF(r) identified the same favored mutations around FADS1, BNC2, and TRPM8, and the scores of iSAFE did not strongly differentiate favored mutations from nearby mutations. Finally, we developed three specific networks (called Network1/2/3) and compared them with DeepFavored and SWIF(r) to identify the factors that critically determine DeepFavored’s performance. These networks and comparisons indicate that dividing the identification of favored mutations into two subtasks is essential and that deep-learning networks can better extract information from selection signals than probabilistic models (Figure S4A–S4C).

Identifying favored mutations in CEU, CHB, and YRI

Using the phase III data of the 1000 Genome Project, Murga-Moreno et al. used eight statistical tests to scan selection signals in multiple human populations (Murga-Moreno et al., 2019). Their genomic scan identified 1,091 regions in CEU, CHB, and YRI (called PopHumanScan regions, which are either non-overlapping 10 Kb sliding windows or protein-coding genes) that contain a selective sweep. These regions have 9,965,378, 5,894,011, and 8,646,171 SNP sites in CEU, CHB, and YRI, respectively, and are good candidates for identifying favored mutations therein (as they had been detected by one or multiple statistical tests). We used the three methods, with the two criteria DFscore ≥ 0.5/iSAFE score ≥ 0.2/SWIF(r) score ≥ 0.5 and rank ≤ 5, to identify similar numbers of favored mutations in the 1,091 regions (each region was extended to 1 Mb). DeepFavored, iSAFE, and SWIF(r) identified 1,013 (454, 287, 272), 1,219 (560, 487, 172), and 789 (350, 334, 104) favored mutations in CEU, CHB, and YRI, respectively (Figure S4D; Table S1). The three sets of favored mutations have limited overlaps, partly because the criteria made favored mutations under-predicted.

Then, we checked whether favored mutations are enriched in supporting evidence. Convincing evidence remains rare. Since many adaptive mutations are expression quantitative trait loci (eQTLs) (De Maeyer et al., 2016; Kudaravalli et al., 2009), we assumed eQTLs and DNA methylation QTLs (mQTLs) as supporting evidence. We also assumed SNPs’ phenotype information (PHENO) and Pub Med records (PubMed) in the VEP database on the Ensembl website as supporting evidence (although published papers may provide just related, rather than supporting, information). For each kind of evidence, we calculated its percentages in all favored mutations and in all SNP sites and calculated the ratio of the two percentages. Favored mutations identified by DeepFavored have more PubMed records than those identified by the other methods, and the manual examination of some PubMed records confirms that they provide supporting evidence (Figure S4E). Favored mutations at eQTLs have different tissue distribution tendencies in different populations.

Fifty-five favored mutations were identified by all three methods, which should be more reliable than those identified by a single method (Figure S4D). We used the 55 favored mutations as test data to further evaluate DeepFavored’s performance. The mean rank of the 55 favored mutations in the target population (in which they were identified) is 2.42 (DeepFavored), 2.78 (iSAFE), and 2.64 (SWIF(r)). In parallel, the mean rank of the 10 gold-standard examples in the target population is 6.48 (DeepFavored), 9.62 (iSAFE), and 25.2 (SWIF(r)) (Figure S1H). The results of testing using these real data suggest that favored mutations identified by DeepFavored should be reasonable.

Population-specific favored mutations and GWAS sites are co-enriched in specific gene sets

A distinct difference between populations is diet structures. Since the high intake of sugars and salt is a key feature of
populations that entered agricultural civilization earlier, we conjecture that many population-specific favored mutations may be enriched in metabolism-related genes. To check the conjecture and uncover whether there is a potential trade-off, we identified favored mutations in multiple GO terms (Table S1) and examined the quadrilateral relationships among populations, GO terms, favored mutations, and GWAS sites (Table S1). For genes in each GO term, we computed a DFscore for each SNP and a mean DFscore for all SNPs. Whether the GO term is enriched in favored mutations in a population (compared with the other two) is judged jointly upon two criteria. (1) On the GO term level, the population has the largest mean DFscore, which is significantly larger than the smallest mean DFscore in the background population. The difference between the largest and smallest mean DFscores indicates the accumulated effect of all SNPs in the GO term. (2) On the SNP level, there is at least one SNP whose DFscore >0.4, which indicates very strong selection (as we used DFscore >0.5 to do the genome-wide prediction). Of note, sugar metabolism- and insulin-related GO terms are enriched in favored mutations in CEU, and salt metabolism- and taste-related GO terms are enriched in favored mutations in CHB and YRI (Table 1), reflecting the influence of diet structures on adaptive evolution.

We next addressed the trade-off by examining whether GO terms enriched in population-specific favored mutations are also enriched in population-specific GWAS sites using data from the GWAS Catalog. Confirmative results were obtained (Tables 1 and S1), suggesting that the trade-off between adaptive evolution and disease susceptibility may also occur on the GO term level.

Both favored mutations and hitchhiking mutations are enriched in GWAS sites

A mutation may be beneficial when some epidemic occurs but be otherwise deleterious. Thus, it could be eliminated if the epidemic intervals are long, but otherwise not (its fate also depends on factors including selection coefficient and the population size that influence the selection-drift balance). Such mutations (e.g., the rs334 in the HBB gene) are a kind of trade-off between adaptive evolution and disease susceptibility. Deleterious mutations near a favored mutation and becoming hitchhikers may indicate another kind of trade-off. To examine this scenario, we computed DFscores for SNPs in the 3,583 genes in the large GO: 0008152 and for SNPs in the 600 Kb regions centered on each favored mutation (Table S1). Upon DFscore ≥ 0.5 (rank = 1) and LD ≥ 0.6, we classified mutations into favored, hitchhiking, and ordinary ones. We then calculated the proportions of GWAS sites at favored, hitchhiking, and ordinary mutations and the ratios of the first and second proportion to the third. The proportions and ratios of GWAS sites are significantly higher at favored and hitchhiking mutations than at ordinary mutations. Examples include the hitchhitching mutations near the favored mutations rs5743618 in TLR1 and rs9614102 in MTMR3 (Figure 2) and near some other genes in GO: 0008152 (Figures S5 and S6). TLR1 and MTMR3 have important immune functions, and many GWAS sites overlapping hitchhiking mutations near the two favored mutations are also immune disease associated, indicating immune-related trade-offs. Similar results were obtained under varied DFscore and LD cutoffs (Table S1), supporting the trade-off characterized by GWAS sites at hitchhiking mutations.

To examine the potentially correlated population- and disease specificity of the trade-off, we examined whether the 454, 287, and 272 favored mutations identified by DeepFavored in the PopHumanScan regions in CEU, CHB, and YRI, respectively, and the nearby hitchhiking mutations are enriched in neurological, metabolic, and immune-related diseases/traits sites. The results indicate that the favored and hitchhiking mutations are enriched with GWAS sites, and the enrichment shows population- and disease-specific features (Figure 3A).

**Disease-related GWAS sites may hitchhike or not hitchhike**

Using non-synonymous SNPs, Chun and Fay found a higher ratio of deleterious SNPs to neutral SNPs in hitchhiking regions (Chun and Fay, 2011), suggesting that the scenario deleterious mutations can hitchhike with adaptive mutations. Di et al. identified another scenario in which recessive deleterious mutations (in Mendelian disease genes) may interfere with adaptive ones to slow down adaptation (Di et al., 2021). The second scenario, supported by a theoretical study (Assaf et al., 2015), indicates a second correlation between favored mutations and disease-related mutations. To further address which scenario (hitchhiking or not hitchhiking) is more prevalent, we computed the correlation between disease risk alleles (called risk alleles, at sites of hitchhiking and ordinary mutations) and favored mutations in the same haplotypes in genes in GO: 0008152 (Figure 3B) and examined the risk alleles’ frequencies (at sites of hitchhiking and ordinary mutations) (Figure 3C). The distributions of correlation coefficients and of allele frequencies indeed indicate two kinds of risk alleles at hitchhiking mutations. Of note, risk alleles that are positively/negatively correlated with favored mutations have higher/lower frequencies, respectively (Figure 3D), supporting the co-existence of the two scenarios. The large Pearson coefficient (0.87) between these risk alleles’ frequencies and correlation coefficients suggests that the features of risk alleles at hitchhiking mutations are unlikely identified by chance (Figure 3D). In contrast, risk alleles at ordinary mutations lack a clear correlation with favored mutations in the same haplotypes (Figure 3B) and can have very low frequencies (Figure 3C).

**DISCUSSION**

Drastic environmental changes, fast-evolving pathogens, and changed lifestyles are the motors of genome adaptation. However, many of those swept genomic regions could generate deleterious hitchhikers, and an adaptive mutation may become mal-adaptive when new external changes occur. Alleles that were beneficial just thousands of years ago may become disease associated now, causing “mismatch diseases” (e.g., diseases including type 2 diabetes in Lieberman, 2014). For example, some African-specific favored mutations have enabled Africans to respond to nutritional challenges by altering carbohydrate and lipid metabolism, but these mutations cause disease susceptibility when their lifestyles change (Jones and Rayner, 2021; Kilmentidis et al., 2011; Langenberg and Lotta, 2018). Increasing
Table 1. Some GO terms that are enriched in population-specific favored and disease/trait-associated mutations

| GO term                                                  | GO ID  | Target/background | The ratio of mean DFscores (target/background)* | The ratio of mean DFscores (GO/control)b | The ratio of GWAS site proportions (GO/control)c | Related diseases/trait in target/background populationsd |
|----------------------------------------------------------|--------|-------------------|-----------------------------------------------|------------------------------------------|-------------------------------------------------|----------------------------------------------------------|
| Sucrose transmembrane transporter activity               | 0008515| CEU/YRI           | 15.2*                                         | 3.4**                                    | 2.9***                                          | skin pigmentation traits; hair color; breast cancer; hair morphology traits; melanoma; blond versus brown/black hair color |
| Sucrose:cation symporter activity                         | 0009669| CEU/YRI           | 15.2*                                         | 4.2***                                   | 2.9***                                          | skin pigmentation traits; hair color; melanoma; blond versus brown/black hair color; basal cell carcinoma |
| Glucose import in response to insulin stimulus            | 0044381| CEU/YRI           | 75.7*                                         | 6.5***                                   | 1.0                                             | apolipoprotein B levels; gut microbiota |
| Positive regulation of insulin secretion                  | 0032024| CHB/CEU           | 6.5****                                       | 2.0****                                  | 1.4*                                            | type 2 diabetes; systemic lupus erythematosus; rheumatoid arthritis; cervical cancer; uterine fibroids; coronary artery disease; gout |
| Response to salt stress                                   | 0009651| CHB/YRI           | 54.2****                                      | 4.9****                                  | 1.1                                             | glomerular filtration rate; creatinine levels |
| Insulin-like growth factor I binding                      | 0031994| CHB/YRI           | 22.3**                                        | 1.4*                                     | 3.7****                                          | serum uric acid levels; creatinine levels; glomerular filtration rate |
| Negative regulation of insulin secretion                  | 0046676| CHB/YRI           | 7.4***                                        | 2.6****                                  | 1.6*                                             | low-density lipoprotein cholesterol levels; type 2 diabetes |
| Sour taste receptor activity                              | 0033040| YRI/CEU           | 2671.0*                                       | 6.2***                                   | 3.7*                                             | heel bone mineral density; blood metabolite levels; palmitoleic acid (16:1n-7) levels; blood protein levels; serum metabolite levels |
| Sensory perception of sweet taste                         | 0050916| YRI/CEU           | 31.6**                                        | 16.5****                                 | 3.6*                                             | insomnia; height; asthma and attention deficit hyperactivity disorder |

(Continued on next page)
findings highlight trade-offs between adaptive evolution and disease susceptibility; however, the trade-off notion remains controversial because its scale is unclear and because different studies also report conflicting results (Polimanti and Gelernter, 2017; Srinivasan et al., 2016; Yao et al., 2020). Abundant GWAS sites have been reported, but only a few trade-off cases have been annotated due to the lack of reliably and systematically identified favored mutations. To obtain systematic and convincing evidence for trade-offs, identifying favored mutations and analyzing them with GWAS sites in multiple human populations are required.

Identifying favored mutations from various selection signals they cause has been a long-standing challenge because of the confounding signals (especially hitchhiking mutations). Even a combination of statistical tests does not have adequate power to discriminate favored mutations from selection signals (Grossman et al., 2010; Szpak et al., 2018). Since deep learning can abstract fuzzy features from noisy data, it is advisable to use a deep-learning network to integrate statistical tests and learn favored mutations’ features from the statistical results. By specifically building and training the network DeepFavored, we divide identifying favored mutations into two parallel subtasks. This strategy contributes to DeepFavored’s performance greatly. We used it to systematically identify favored mutations in the human population CEU, CHB, and YRI and subsequently jointly analyzed favored mutations and GWAS sites in these populations. On both the SNP and GO term levels, favored mutations and hitchhiking mutations are significantly enriched in GWAS sites, and this enrichment shows population- and disease-specific features. These results provide systematic evidence for extensive trade-offs between adaptive evolution and disease susceptibility.

There may be three explanations for the enrichment of hitchhiking mutations in disease-associated mutations. First, when hitchhiking mutations obtain effects on some diseases under specific conditions, these effects may be accumulated instead of being cleared due to the strong LD with the favored mutation. However, if hitchhiking mutations are strongly deleterious, they can negatively influence favored mutations by deterring their frequency increase (Assaf et al., 2015). It may be hard to differentiate the two situations, but we find they are probably equal in number (Figures 3B and 3C). Second, if a favored mutation is associated with the susceptibility of some disease, the strong LD also makes nearby sites associated with the susceptibility of the disease. Third, some adaptive mutations may become maladaptive when new external changes (including lifestyle) occur. Many genomic regions contain numerous SNPs with small and even very small effects (Gibson, 2012), which contribute to the cause of diseases such as type 2 diabetes (Langenberg and Lotta, 2018). New findings, especially new GWAS sites in populations such as YRI, may provide more support for these explanations.

Rapid changes in lifestyles may cause many trade-offs or mismatches to occur. Uncovering these trade-offs and mismatches helps better understand and better treat diseases. A case in point is the mutations in genes encoding the epithelial sodium channel (ENaC) that have been selected for in Black Africans, probably due to selective pressure for sodium retention in traditional low-salt diets. Patients with increased activity of this sodium channel are more likely to respond to the selective sodium...
channel antagonist amiloride (Jones and Rayner, 2021). The scale of the trade-off may profoundly influence the understanding and study of human differences and diseases. Combined analysis of population-specific favored mutations and disease-associated mutations can provide valuable data and clues for precision medicine.

**Limitations of the study**

Favored mutations in other populations remain to be identified. Favored mutations in the three populations are probably under-predicted due to the high cutoff (0.5) of the DFscore. The lack of reports of GWAS sites in YRI probably makes the trade-offs and the correlation between adaptive evolution and disease susceptibility in YRI under-estimated.

**STAR METHODS**

Detailed methods are provided in the online version of this paper and include the following:

- **KEY RESOURCES TABLE**
- **RESOURCE AVAILABILITY**
Figure 3. GWAS sites at different mutations
(A) The enrichment of favored and hitchhiking mutations with GWAS sites is genome-wide and shows population- and disease-specific features. As in Figures 2C and 2D, the proportions of GWAS sites at favored and hitchhiking mutations (and the ratios of the first two proportions to the third) are significantly larger than those at ordinary mutations (one-sided two-proportion Z-tests; the only insignificant proportion is indicated by an asterisk [*]).
(B) The correlation between disease risk alleles (at hitchhiking and ordinary mutations, respectively) and favored mutations in the same haplotypes. Risk alleles at hitchhiking mutations are significantly different from risk alleles at ordinary mutations (two-sample Kolmogorov-Smirnov test); the former shows a positive or negative correlation with favored mutations in the same haplotypes, but the latter does not.
work, we also actively worked to promote gender balance in our reference list.

We worked to ensure diversity in experimental samples through the selection

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.celrep.2022.111351.

ACKNOWLEDGMENTS

The authors thank Dr. Akbari for the valuable comments on using iSAFE. This work was supported by the Department of Science and Technology of Guangdong Province (2020A1515010803) and the National Natural Science Foundation of China (31771456 and 31571348).

AUTHOR CONTRIBUTIONS

J.T. and H.Z. developed the concept and designed the study. J.T. implemented the network. J.T., M.H., S.H., J.Z., and H.Z. performed data analysis. H.Z. wrote the manuscript.

DECLARATION OF INTERESTS

The authors declare no competing interests.

INCLUSION AND DIVERSITY

We worked to ensure diversity in experimental samples through the selection of the genomic datasets. While citing references scientifically relevant for this work, we also actively worked to promote gender balance in our reference list.

Received: March 31, 2022
Revised: June 13, 2022
Accepted: August 23, 2022
Published: September 13, 2022

SUPPORTING CITATIONS

The following references appear in the supplemental information: Bryk et al., 2008, Cook et al., 2009, Donnelly et al., 2012, Enattah et al., 2002, Fujimoto et al., 2008, Heflebower et al., 2014, Mancus et al., 2017, Miller et al., 1976, Ohashi et al., 2011, Olds and Sibley, 2003, Peter et al., 2012, Sabeti et al., 2006, Sturm and Duffy, 2012, Tishkoff et al., 2001, Visser et al., 2012, Wong et al., 2010.

REFERENCES

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G.S., Davis, A., Dean, J., Devin, M., et al. (2016). TensorFlow: large-scale machine learning on heterogeneous distributed systems. Preprint at arXiv. https://doi.org/10.48550/arXiv.1603.04467.

Ackerman, H., Usen, S., Jallow, M., Sisay-Joof, F., Pinder, M., and Kwiatkowski, D.P. (2005). A comparison of case-control and family-based association methods: the example of sickle-cell and malaria. Ann. Hum. Genet. 69, 559–565.

Adirion, J.R., Galloway, J.G., and Kern, A.D. (2020). Predicting the landscape of recombination using deep learning. Mol. Biol. Evol. 37, 1790–1808.

Akbari, A., Vitti, J.J., Iranmehr, A., Bakhtiari, M., Sabeti, P.C., Mirarab, S., and Bafna, V. (2018). Identifying the favored mutation in a positive selective sweep. Nat. Methods 15, 279–282.

Assaf, Z.L., Petrov, D.A., and Blundell, J.R. (2015). Obstruction of adaptation in diploids by recessive, strongly deleterious alleles. Proc. Natl. Acad. Sci. USA 112, E2658–E2666.

Benton, M.L., Abraham, A., LaBella, A.L., Abbott, P., Rokas, A., and Capra, J.A. (2021). The influence of evolutionary history on human health and disease. Nat. Rev. Genet. 22, 269–283.

Bryk, J., Hardouin, E., Pugach, I., Hughes, D., Strotmann, R., Stoneking, M., and Myles, S. (2008). Positive selection in East Asians for an EDAR allele that enhances NF-kappaB activation. PLoS One 3, e2209.

Chun, S., and Fay, J.C. (2011). Evidence for hitchhiking of deleterious mutations within the human genome. PLoS Genet. 7, e1002240.

GBD 2019 Stroke Collaborators (2021). Global, regional, and national burden of stroke and its risk factors, 1990-2019: a systematic analysis for the Global Burden of Disease Study 2019. Lancet Neurol. 20, 795–820.

Cook, A.L., Chen, W., Thurban, A.E., Smit, D.J., Smith, A.G., Bladen, T.G., Brown, D.L., Duffy, D.L., Pastorino, L., Bianchi-Scarra, G., et al. (2009). Analysis of cultured human melanocytes based on polymorphisms within the SLC45A2/MATP, SLC24A5/NCKX5, and OCA2/P loci. J. Invest. Dermatol. 129, 392–405.

De Maeyer, D., Weytens, B., De Raedt, L., and Marchal, K. (2016). Network-based analysis of eQTL data to prioritize driver mutations. Genome Biol. Evol. 8, 481–494.

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, e69026.

Donnelly, M.P., Paschou, P., Grigorenko, E., Gurwitz, D., Barta, C., Lu, R.B., Zhukova, O.V., Kim, J.J., Siniscalco, M., New, M., et al. (2012). A global view of the OCA2-HERC2 region and pigmentation. Hum. Genet. 131, 683–696.

Enattah, N.S., Sahi, T., Savilahti, E., Terwilliger, J.D., Peltonen, L., and Jarvela, I. (2002). Identification of a variant associated with adult-type hypolactasia. Nat. Genet. 30, 253–257.
Flagel, L., Brandvain, Y., and Schrider, D.R. (2019). The unreasonable effectiveness of convolutional neural networks in population genetic inference. Mol. Biol. Evol. 36, 220–238.

Fujimoto, A., Ohashi, J., Nishida, N., Miyagawa, T., Morishita, Y., Tsunoda, T., Kimura, R., and Tokunaga, K. (2008). A replication study confirmed the EDAR gene to be a major contributor to population differentiation regarding head hair thickness in Asia. Hum. Genet. 124, 179–185.

Garud, N.R., Messer, P.W., Buzbas, E.O., and Petrov, D.A. (2015). Recent sweeps in North American Drosophila melanogaster show signatures of soft sweeps. PLoS Genet. 11, e1005004.

Gibson, G. (2012). Rare and common variants: twenty arguments. Nat. Rev. Genet. 13, 135–145.

Grossman, S.R., Shylakhter, I., Shylakhter, I., Karlsson, E.K., Byrne, E.H., Morales, S., Frieden, G., Hostetter, E., Angelino, E., Garber, M., et al. (2010). A composite of multiple signals distinguishes causal variants in regions of positive selection. Science 327, 883–886.

Heffelfinger, C., Pakstis, A.J., Speed, W.C., Clark, A.P., Haigh, E., Fang, R., Furtado, M.R., Kidd, K.K., and Snyder, M.P. (2014). Haplotype structure and positive selection at TLR1. Eur. J. Hum. Genet. 22, 551–557.

Jones, E., and Rayner, B. (2021). The importance of the epithelial sodium channel in determining salt sensitivity in people of African origin. Pediatr. Nephrol. 36, 237–243.

Kern, A.D., and Schrider, D.R. (2018). diploS/HIC: an updated approach to classifying selective sweeps. G3 8, 1959–1970.

Klimentidis, Y.C., Abrams, M., Wang, J., Fernandez, J.R., and Allison, D.B. (2011). Natural selection at genomic regions associated with obesity and type-2 diabetes: east Asians and sub-Saharan Africans exhibit high levels of differentiation at type-2 diabetes regions. Hum. Genet. 129, 407–418.

Kudaravalli, S., Veyrieras, J.B., Stranger, B.E., Dermitzakis, E.T., and Pritchard, J.K. (2009). Gene expression levels are a target of recent natural selection in the human genome. Mol. Biol. Evol. 26, 649–658.

Langenberg, C., and Lotta, L.A. (2018). Genomic insights into the causes of type 2 diabetes. Lancet 391, 2463–2474.

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. Nature 521, 436–444.

Lieberman, D. (2014). The Story of the Human Body - Evolution, Health, and Disease.

Lindesmith, L., Moe, C., Marionneau, S., Ruvoen, N., Jiang, X., Lindblad, L., Stewart, P., LePendu, J., and Baric, R. (2003). Human susceptibility and resistance to Norwalk virus infection. Nat. Med. 9, 548–553.

Malaria Genomic Epidemiology Network (2014). Reappraisal of known malaria resistance loci in a large multicenter study. Nat. Genet. 46, 1117–1204.

McLaren, W., Gil, L., Hunt, S.E., Riat, H.S., Ritchie, G.R.S., Thomann, A., Flicek, P., and Cunningham, F. (2016). The Ensembl variant effect predictor. Genome Biol. 17, 122.

McManus, K.F., Taravella, A.M., Henn, B.M., Bustamante, C.D., Sikora, M., and Cornejo, O.E. (2017). Population genetic analysis of the DARC locus (Duffy) reveals adaptation from standing variation associated with malaria resistance in humans. PLoS Genet. 13, e1006860.

Miller, L.H., Mason, S.J., Clyde, D.F., and McGinniss, M.H. (1976). The resistance factor to Plasmodium vivax in blacks. The Duffy-blood-group genotype, G6PD: recent origin of alleles that confer malarial resistance. Science 198, 1614–1620.

Ohashi, J., Naka, I., and Tsuchiya, N. (2011). The impact of natural selection on an ABCC11 SNP determining earwax type. Mol. Biol. Evol. 28, 849–857.

Olds, L.C., and Sibley, E. (2003). Lactase persistence DNA variant enhances lactase promoter activity in vitro: functional role as a cis regulatory element. Hum. Mol. Genet. 12, 2333–2340.

Polimanti, R., and Gelernter, J. (2017). Widespread signatures of positive selection in common risk alleles associated to autism spectrum disorder. PLoS Genet. 13, e1006618.

Polimanti, R., and Gelernter, J. (2017). Widespread signatures of positive selection in common risk alleles associated to autism spectrum disorder. PLoS Genet. 13, e1006618.

Prohaska, A., Racimo, F., Schork, A.J., Sikora, M., Stern, A.J., Ilardo, M., Allentoft, M.E., Folkesen, L., Bull, A., Moreno-Mayar, J.V., et al. (2019). Human disease variation in the light of population genomics. Cell 177, 115–131.

Rolnick, D., Veit, A., Belongie, S., and Shavit, N. (2018). Deep learning is robust to massive label noise. Preprint at arXiv. https://doi.org/10.48550/arXiv.1705.10694.

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.

Schipper, C., Moe, C., Ruvoen, N., Jiang, X., Lindblad, L., and Baric, R. (2003). Human susceptibility and resistance to Norwalk virus infection. Nat. Med. 9, 548–553.

Sheehan, S., and Song, Y.S. (2016). Deep learning for population genetic inference. PLoS Comput. Biol. 12, e1004845.

Srinivasan, S., Bettella, F., Mattingsdal, M., Wang, Y., Witoelar, A., Schork, A.J., Thompson, W.K., Zuber, V., Schizophrenia Working Group of the Psychiatric Genomics Consortium, et al. (2016). Genetic markers of human evolution are enriched in schizophrenia. Biol Psychiatry 80, 284–292.

Stobdan, T., Akbari, A., Azad, P., Zhou, D., Poulsen, O., Appenzeiler, O., Gonzales, G.F., Telenti, A., Wong, E.H.M., Saini, S., et al. (2017). New insights into the genetic basis of Monge’s disease and adaptation to high-altitude. Mol. Biol. Evol. 34, 3154–3168.

Sturm, R.A., and Duffy, D.L. (2012). Human pigmentation genes under environmental selection. Genome Biol. 13, 248.

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

Szpak, M., Mezzavilla, M., Ayub, Q., Chen, Y., Xue, Y., and Tyler-Smith, C. (2018). FineMAV: prioritizing candidate genetic variants driving local adaptations in human populations. Genome Biol. 19, 5.

Szipiecz, Z.A., and Hernandez, R.D. (2014). selscan: an efficient multithreaded program to perform EHH-based scans for positive selection. Mol. Biol. Evol. 31, 2824–2827.

Tishkoff, S.A., Varkonyi, R., Cahinhinan, N., Abbas, S., Argyropoulos, G., Destro-Bisol, G., Drousiotou, A., Dangerfield, B., Lefranc, G., Loiselet, J., et al. (2001). Haplotype diversity and linkage disequilibrium at human G6PD: recent origin of alleles that confer malarial resistance. Science 293, 455–462.

Visscher, M., Kayser, M., and Palstra, R.J. (2012). HERC2 rs12913832 modulates human pigmentation by attenuating chromatin-loop formation between a long-range enhancer and the OCA2 promoter. Genome Res. 22, 446–455.

Weir, B.S., and Cockram, C.C. (1984). Estimating F-statistics for the analysis of population structure. Evolution 38, 1358–1370.
Wong, S.H., Gochhait, S., Malhotra, D., Pettersson, F.H., Teo, Y.Y., Khor, C.C., Rautanen, A., Chapman, S.J., Mills, T.C., Srivastava, A., et al. (2010). Leprosy and the adaptation of human toll-like receptor 1. PLoS Pathog. 6, e1000979.

Xue, Y., Daly, A., Yngvadottir, B., Liu, M., Coop, G., Kim, Y., Sabeti, P., Chen, Y., Stalker, J., Huckle, E., et al. (2006). Spread of an inactive form of caspase-12 in humans is due to recent positive selection. Am. J. Hum. Genet. 78, 659–670.

Yao, Y., Yang, J., Xie, Y., Liao, H., Yang, B., Xu, Q., and Rao, S. (2020). No evidence for widespread positive selection signatures in common risk alleles associated with schizophrenia. Schizophr. Bull. 46, 603–611.

Zheng, Z., Huang, D., Wang, J., Zhao, K., Zhou, Y., Guo, Z., Zhai, S., Xu, H., Cui, H., Yao, H., et al. (2020). QTLbase: an integrative resource for quantitative trait loci across multiple human molecular phenotypes. Nucleic Acids Res. 48, D983–D991.
STAR+METHODS

KEY RESOURCES TABLE

| REAGENT or RESOURCE | SOURCE | IDENTIFIER |
|---------------------|--------|------------|
| Deposited data      |        |            |
| NHGRI-EBI GWAS Catalog (As of 2021-06-08) | NHGRI and EMBL-EBI | https://www.ebi.ac.uk/gwas/ |
| 1000 Genomes Project Phase 3 | EMBL-EBI | https://www.internationalgenome.org/ |
| PopHumanScan        | Sònia Casillas-lab | https://pophumanscan.uab.cat/ |
| QTLbase             | Mulinlab@Tianjin Medical University | http://www.mulinlab.org/qtlbase |
| GO                  | The Gene Ontology Consortium | www.geneontology.org |
| VEP                 | EMBL-EBI | http://grch37.ensembl.org/Homo_sapiens/Tools/VEP |

Software and algorithms

| Python 3.6.10         | Python Software Foundation | https://python.org/ |
| Tensorflow 1.14.0     | Google Brain Team           | https://www.tensorflow.org/ |
| DeepFavored           | https://github.com/GT269/DeepFavored | https://doi.org/10.5281/zenodo.6979852 |
| iSAFE 1.0.4           | Ali Akbari                  | https://github.com/alek0991/iSAFE |
| SWIF(r)               | ramachandran-lab            | https://github.com/ramachandran-lab/SWIFr |
| CMS 2.0               | Broadinstitute              | https://github.com/broadinstitute/cms |
| Selscan               | Szpiech lab                 | https://github.com/lasugden/selscan |
| Cosi2.0               | Broadinstitute              | https://software.broadinstitute.org/mpg/cosi2/ |
| ReportLab 3.5.36      | ReportLab, Inc              | https://www.reportlab.com/ |
| Seaborn 0.11.2        | Michael Waskom              | http://seaborn.pydata.org/ |
| Matplotlib 3.3.4      | Matplotlib development team | https://matplotlib.org/stable/index.html |
| PLINK                 | Christopher Chang           | https://www.cog-genomics.org/plink/2.0/ |

RESOURCE AVAILABILITY

Lead contact
Further information about this manuscript and requests for resources will be fulfilled by the Lead Contact, Hao Zhu (zhuhao@smu.edu.cn).

Materials availability
This study did not generate new unique reagents.

Data and code availability
This study did not generate any sequencing datasets. Source data are described in the Key resources table. Data generated by the computational analyses are in the Table S1.

The code of DeepFavored is publicly available on the GitHub website (https://github.com/GT269/DeepFavored); the DOI is https://doi.org/10.5281/zenodo.6979852 (listed in the Key resources table). The codes of iSAFE and SWIF(r) are publicly available (the information is given in the Key resources table).

Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

METHOD DETAILS

Generating and collecting genomic data
Simulating population genomic data
Population genomic data of neutral regions were simulated using the demographic models of the African, Asian, and European populations (YRI, CHB, and CEU) as previously reported (Sugden et al., 2018). Using the cosi2 and recosim programs in the cosi2 package, each simulation first generated a recombination map, then a genomic region (360 haplotypes, 120 from each population).
Multiple parameters were used to simulate genomic regions containing a selective sweep caused by a favored mutation, including allele frequency (0.2, 0.4, 0.6, 0.8, 1.0, but the max frequency for iSAFE was 0.99), target population (YRI, CHB, and CEU), selection coefficient, and start time (5, 10, 15, 20, 25, and 30 kya). These parameters were combined into 90 sets, with each set representing a selective sweep scenario and being used to simulate 100 genomic regions (1 Mb).

**Generating training data**

The numbers of favored mutations (M), hitchhiking neutral mutations (N), and ordinary neutral mutations (K) in the training dataset and the ld-cutoff value were hyper-parameters determined by 10-fold cross-validation grid searches (Table S2). We first simulated 100 sweep regions (1 Mb) for each of the 90 parameter sets (see section Simulating population genomic data) and then simulated 100 neutral regions (1 Mb). Next, for each neutral mutation in each sweep region, we computed its linkage disequilibrium (LD) with the favored mutation. Upon the parameter ld-cutoff, neutral mutations in a sweep region were classified into hitchhiking neutral mutations (LD > ld-cutoff) and ordinary neutral mutations (LD < ld-cutoff). Finally, we used sweep regions and neutral regions to constitute the training dataset. In this dataset, M favored mutations were sampled from the sweep regions, N hitchhiking neutral mutations were sampled from the hitchhiking neutral mutations (in the sweep regions), and K/2 ordinary neutral mutations were sampled from the ordinary neutral mutations (in the sweep regions), and K/2 ordinary neutral mutations were sampled from the 100 neutral regions.

We simulated 20 sweep regions (1 Mb) for each of the 90 selective sweep scenarios and 5 neutral regions (1 Mb) to generate the hold-out test dataset. We constituted this dataset by sampling 20'90 favored mutations and 20'90'X hitchhiking neutral mutations from the 20'90 sweep regions and sampling 20'90'X ordinary neutral mutations from the 5 neutral regions (we let X = 500) (these hitchhitching neutral mutations may not be independent of each other due to their LD with the favored mutations).

**Generating testing data**

To test the performance of DeepFavored, iSAFE, and SWIF(r), we simulated 50 sweep regions (1 Mb) for each of the 90 parameter sets and 1000 neutral regions (1 Mb) using the parameters mentioned above and methods mentioned in section Simulating population genomic data. For different testing purposes (Figures 2A–2G), the 50'90 = 4500 sweep regions were split into multiple subsets, and data with different ratios of favored and neutral mutations were generated.

**Collecting real population genomic data**

The real genomic data were from 1000 Genomes Project Phase 3 (ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/). These data also included genomic regions containing selection signals detected by PopHumanScan (Murga-Moreno et al., 2019).

**Classifying sweep regions into soft and hard ones**

Selective sweeps generated by a favored (adaptive) mutation in a population can be hard or soft. In the former case, a single adaptive haplotype rises to high frequency and leads to high haplotype homozygosity; in the latter case, multiple adaptive haplotypes sweep through the population simultaneously and produce distinct patterns of genetic variations instead of low haplotype homozygosity. Large effective population size or mutation rate can make a favored mutation generate a rather soft sweep (Garud et al., 2015). Therefore, favored mutations can occur in multiple haplotypes simultaneously in a large population. We used the number of adaptive haplotypes (i.e., haplotypes that carry a favored mutation) to indicate the softness of selective sweeps (Garud et al., 2015). We computed the Hamming distance between the selected and neutral loci can lead to multiple haplotypes linked to the beneficial allele (Schrider et al., 2013). Therefore, favored mutations can occur in multiple haplotypes simultaneously in a large population. We used the number of adaptive haplotypes (i.e., haplotypes that carry a favored mutation) to indicate the softness of selective sweeps (Garud et al., 2015). We computed the Hamming distance between every two adaptive haplotypes and assumed that if the Hamming distance cutoff was >100 the two haplotypes were of different types (Figure S2) (Pinheiro et al., 2005). We examined different empirical Hamming distance cutoffs, and these cutoffs generated the same results (Figure S3). Upon the Hamming distances, we defined the heterogeneity of adaptive haplotypes and used it to measure the softness of each selective sweep:

Here Nc denotes the number of adaptive haplotypes, Nc denotes the number of adaptive haplotype types, and these two parameters define a way to classify selective sweeps with varied softness. The simulated selective sweeps had different hc (Figure S2), indicating that our deep learning method can identify favored mutations in soft sweeps (Figure 2B).

**Computing component statistics**

DeepFavored computes seven statistical tests for each SNP site (Figure 1F), four of which were used in previous studies (Grossman et al., 2010; Sugden et al., 2018). The fixation index Fst is computed as previously reported for each pairwise population (Weir and Cockerham, 1984). The integrated haplotype score iHS is computed using the selscan program (Szpiech and Hernandez, 2014). The cross-population extended haplotype homozygosity XP-EHH is computed using the modified version of selscan (Sugden et al., 2018) (https://github.com/lasugden/SELSCAN). nSL (which detects sweeps for which the selected variant is still at low frequency) and ΔiHH (which is a modification of the iHS test that is less sensitive to the length of the ancestral haplotype) are computed using the CMS package (Grossman et al., 2010). ΔDAF (which tests for derived alleles that are at a high frequency relative to other populations) is calculated as ΔDAF = DAF1-1/2(DAF2+DAF3), where DAF1 is the derived allele frequency in the target population, and DAF2 and DAF3 are the derived allele frequencies in the other two populations (Grossman et al., 2010). iSAFE scores are computed using the authors’ code (Akbari et al., 2018). Each component test is computed independently, and the results of all tests form a matrix, with each row representing an SNP site and each column representing a component statistic. Thus, the computation of component statistics is similar to that used in the diploSHIC to detect selective sweeps (Kern and Schrider, 2018), but the component statistics are quite different. This method makes the integration of new component statistics easy. The matrix is inputted into DeepFavored. Calculating the statistics (especially XP-EHH) is the main factor of time consumption, and it needs about 2 h (a single CPU core) for a 1 Mb sequence.
We implemented SWIF(r) using the authors’ code and training data composition (Sugden et al., 2018) (https://github.com/alek0991/iSAFE). To reasonably compare the performance of the three methods, two new statistical tests (nSL and iSAFE) were added into SWIF(r), the same training data were used to SWIF(r) and DeepFavored, and the same testing data were used to test SWIF(r), iSAFE, and DeepFavored. We note that not all of our SWIF(r) results agree with the authors’ original results, because the way recombination hotspots were generated may make training data not identical and because SWIF(r) was originally tested using 1000 Genomes Project Phase 1 data.

Implementation of iSAFE and SWIF(r)

As previously reported (Akbari et al., 2018; Sugden et al., 2018), only biallelic SNP sites are considered. Any undefined statistics at an SNP site are imputed with the genome-wide mean computed from the neutral regions. When component statistics are computed at sites in neutral regions, the target population is determined arbitrarily, and the other two populations are used as reference populations (iHS and iSAFE do not need reference populations). The results of iHS and XP-EHH are normalized with the population-specific mean and standard deviation (which are computed for the neutral regions) to correct for inherent biases based on linkage disequilibrium structure. As previously reported (Grossman et al., 2010), iHS normalization is carried out within each of the 20 frequency bins because iHS scores are approximately normal for SNPs with comparable derived allele frequencies. The results of Fst, ΔDAF, nSL, ΔiHH, and iSAFE are normalized within 1 Mb regions. When computing statistics for the 1000 Genomes Project data, we normalize the above component statistics genome-wide using the method previously reported (Grossman et al., 2010), and undefined statistics at SNP sites are imputed with the genome-wide mean.

Euclidean distances were computed between the three kinds of mutations. Different situations of hitchhiking were simulated. The scores of the seven statistical tests (shown in Figure 1F) at each SNP site were inputted into the Euclidean distance formula.

Implementation of DeepFavored

Building and training DeepFavored

DeepFavored was built using TensorFlow (Abadi et al., 2016). It consists of hidden and output layers specific for processing hitchhiking and ordinary neutral mutations (indicated by ‘H’ and ‘O’) and hidden layers shared by the two subtasks (indicated by ‘S’). Key details are as follows. (a) The two kinds of mutation-specific hidden and output layers were trained alternatively, with the max number of epochs being 100. (b) 10-fold cross-validation was used to determine hyper-parameters. (c) The initial learning rate was 0.01, and was multiplied by 0.5 if validation loss was not improved for 10 epochs. (d) Mini-batch training was used, with Batch size being 32. (e) To avoid overfitting, we used Early stop to stop training after no improvement in validation loss for 15 epochs; this scheme, together with the large sample size, makes it unnecessary to use Dropout, L1, and L2 further. (f) Batch Normalization was used to normalize the outputs of each hidden layer. (g) After all candidate models (each set of hyper-parameters determines 10 candidate models) were trained, each model was examined using the hold-out test data by calculating a power-FDR AUC (Area Under the Curve). (h) Finally, we chose the set of hyper-parameters that generated the maximal mean power-FDR AUC as DeepFavored’s hyper-parameters (Table S2), and from the 10 models determined by this set of hyper-parameters, we chose the one with the maximal power-FDR AUC as DeepFavored.

The trained network outputs a score_H and a score_O for each SNP site. score_H is the result of the first subtask (discriminating favored mutations from hitchhiking mutations), and score_O is the result of the second subtask (discriminating favored mutations from ordinary mutations). Their product, called DFscore, indicates whether an SNP site is a favored mutation. If either score_H or score_O is low, the product will be very low, enabling DFscore to discriminate favored mutations from both hitchhiking and ordinary mutations.

Building and training the three variants of DeepFavored

Three networks (i.e., Network1/2/3), with hyper-parameters determined by 10-fold cross-validation grid searches, were built. They consisted only of shared hidden layers (all other aspects were identical to DeepFavored) and were trained using favored mutations plus the two kinds of neutral mutations, favored mutations plus hitchhiking mutations only, and favored mutations plus ordinary mutations, respectively (Figure S6).

Hyper-parameters

Three classes of hyper-parameters were used (Table S2): the first defined the generation of training data, the second defined the network architecture, and the third defined the learning algorithm. Different parameter values were used to define the three kinds of hidden layers (indicated by ‘H’, ‘O’, and ‘S’). When a candidate model without shared hidden layers was generated, the ‘H’ and ‘O’ optimizers (which optimize the ‘H’ and ‘O’ layers) were used to optimize all layers; otherwise, the ‘S’ optimizer was used to optimize all layers. The numbers and values of parameters in each bracket (e.g., [(16), (16, 16, 16)]) that defines hidden layers) specify the layers and units in each layer (Table S2). Optimal hyper-parameters were determined using 10-fold cross-validation grid searches.

Validation of DeepFavored

First, we validated DeepFavored by testing if it can identify the 10 well-characterized favored mutations in CEU, CHB, and YRI, which were used as gold standard examples in other studies (Table S1) (Akbari et al., 2018; Szpak et al., 2018). Second, we validated
DeepFavored by testing if it can identify favored mutations in genomic regions containing selection signals detected by PopHumanScan (Murga-Moreno et al., 2019). Specifically, we examined if DeepFavored better ranks the favored mutations identified by all of the three methods. Third, we examined if identified favored mutations were enriched in annotated functional sites, including expression quantitative trait loci (eQTL), DNA methylation quantitative trait loci (mQTL), phenotype (PHENO), and PubMed records (PUBMED). The data of eQTL and mQTL were from the QTLbase database (Zheng et al., 2020), and the data of PHENO and PubMed IDs were from the VEP database on the Ensembl website (McLaren et al., 2016). Finally, we examined favored mutations identified by the three methods in multiple genes containing positive selection signals reported in recent reviews.

Computing population-specific and GO term-specific enrichment of favored mutations and GWAS site

Population-specific enrichment of favored mutations

We downloaded GO terms from the GO website (www.geneontology.org). For all SNPs in genes (including gene body and the annotated promoter) in each GO term, we computed their DFscores. Whether a GO term is enriched in favored mutations in one population, compared with the other two populations, is determined jointly upon two criteria. (a) The ratio of the mean DFscore in the target population to the mean DFscore in the background population is significantly large. (b) There is at least one DFscore >0.4 in genes in the GO term (When performing the genome-wide scan, 0.5 was the threshold to make the three methods generate similar numbers of favored mutations so as to make their results comparable. It is advisable to use a stringent threshold when identifying favored mutations without any prior knowledge. When analyzing genes in GO terms, 0.5 is slightly too large). A large ratio of the mean DFscores indicates the difference between the populations due to the accumulated effect of SNPs. A large DFscore indicates that at least one SNP is strongly selected for. Statistical significance was evaluated using the one-sided two-sample T test. In Table 1, the number of asterisk indicates statistical significance; "*" to "****" indicate p < 5 × 10^{-2}, 5e-10, 5 × 10^{-15}, and 5e-20, respectively.

GO term-specific enrichment of favored mutations

To evaluate whether a GO term was significantly enriched with favored mutations, the GO term should be compared with a control. It is reasonable to compare the GO term with a "relevant" control (instead of an irrelevant one) because favored mutations are distributed unequally in the genome. To build such a control randomly, we searched for the GO term’s sisters nodes on the same level in the Gene Ontology tree, from the nearest to the farther ones, until the total length of obtained genes in these sister-GO terms was 100 times the length of genes in the GO term. Then, we concatenated all genes in these sister-GO terms randomly into a "super sequence" and sampled m (m was the gene number in the GO term) regions, each spanning n bp (n was the average length of genes in the GO term), randomly from the super sequence. The concatenating and sampling process was repeated 100 times, generating 100 "control" GO terms that provided a background level for evaluating whether the GO term was enriched in favored mutations (the functions of genes in these sister-GO terms should be related to, but different from, the functions of genes in the GO term). Enrichment was determined upon the mean DFscore of SNPs in the GO term and in the control. Statistical significance was evaluated using the one-sided two-sample T test. In Table 1, the number of asterisk indicates statistical significance; "***" to "*****" indicate p < 5 × 10^{-2}, 5e-10, 5 × 10^{-15}, and 5e-20, respectively.

GO term-specific enrichment of GWAS sites

We downloaded the GWAS data from the NHGRI-EBI GWAS Catalog database (https://www.ebi.ac.uk/gwas). For each GO term, we examined the GWAS sites in related genes. To evaluate whether a GO term is enriched with GWAS sites, we built a control GO term as described above (see 4.6.2). Enrichment was determined based on the proportion of GWAS sites in the GO term and in the control. Statistical significance was evaluated using the one-sided two-proportion Z-test. In Table 1, the number of asterisk indicates statistical significance; "***" to "*****" indicate p < 5 × 10^{-2}, 5e-10, 5 × 10^{-15}, and 5e-20, respectively.

Determining the enrichment of favored mutations and hitchhiking mutations in GWAS sites in specific populations

We examined the 3583 genes in the large metabolism-related GO term GO:0,008,152. First, we used DeepFavored to compute the DFscore for each SNP in these genes (gene body and the 2 Kb upstream sequence). Second, we identified genes containing favored mutations upon different DFscore cutoffs. Then, we classified neutral mutations into ordinary and hitchhiking upon their LD with the favored mutation for each of these genes. Finally, we calculated the proportion of GWAS sites at favored mutations, hitchhiking neutral mutations, and ordinary mutations. LD values were calculated using the PLINK program (https://www.cog-genomics.org/plink/2.0/) (Purcell et al., 2007).

Computing the frequencies of disease risk alleles

GWAS sites and their risk alleles were obtained from the NHGRI-EBI GWAS Catalog database (https://www.ebi.ac.uk/gwas). A GWAS site was classified as a disease GWAS site if its EFO terms include one of the following words: “disease”, “disorder”, “cancer”, or “inflammatory”. If a risk allele at a disease GWAS site was a reference allele or alternative allele in the 1000 Genomes Project Phase 3 data, its frequency in the population in which the GWAS site was identified was calculated using the 1000 Genomes data.

QUANTIFICATION AND STATISTICAL ANALYSIS

Measuring the distances between favored, hitchhiking, and ordinary mutations

Pairwise Euclidean distances between favored, hitchhiking, and ordinary mutations were calculated using the following equations:
\[
    d(i,j) = \sqrt{\sum_{k=1}^{7} \left( s_{ik} - s_{jk} \right)^2}
\]

\[
    s_{ni} = \frac{1}{n} \sum_{l=1}^{n} s_{nl} \text{(for a mutation type)}
\]

In the equations, \(i\) indicates a mutation site of a specific type; \(n\) indicates the mutation number of that type; \(k\) (from 1 to 7) indicates a component statistic; \(s_{nk}\) indicates the value of the statistic \(k\) at site \(l\); \(s_{ni}\) indicates the mean value of the statistic \(k\) computed at all sites of a mutation type; \(i\) and \(j\) indicate two mutation types; and \(d(i,j)\) indicates the Euclidean distance between mutation type \(i\) and \(j\) (Figure 1A).

**Measuring the performance of different models**

True positive rate-false discovery rate (power-FDR) curves and true positive rate-false positive rate (ROC) curves were used to evaluate and compare different models’ performance. True positive rate (TPR), false discovery rate (FDR), and false positive rate (FPR) were calculated using the following equations:

\[
    TPR = \frac{ff}{ff + fn}, \quad FDR = \frac{nf}{nf + ff}, \quad FPR = \frac{nf}{nf + nn}
\]

In these equations, \(ff\), \(fn\), \(nf\), and \(nn\) denote the number of favored mutations correctly predicted, the number of favored mutations wrongly predicted, the number of neutral mutations wrongly predicted, and the number of neutral mutations correctly predicted, respectively (Figure S1).

**Evaluating enrichment of favored mutations and GWAS sites**

A one-sided two-sample T test was used to evaluate whether the mean DFscore of SNPs in genes in a GO term is significantly large (compared with a constructed control) in the target population. A one-sided two-proportion Z-test was used to evaluate whether the proportion of GWAS sites in genes in a GO term is significantly large (compared with a constructed control). A one-sided two-sample T test was used to evaluate whether the mean DFscore of SNPs in genes in a GO term in a population is significantly large (compared with the background population). One-sided two-proportion Z-test and one-sided two-sample T test were used to evaluate whether the proportions of GWAS sites in favored and hitchhiking mutations, respectively, are significantly large (compared with in ordinary mutations) (Table 1; Figures 2 and 3). The Python package `statsmodel` was used for the one-sided two-proportion Z-test, and the python package `scipy.stats` was used for the one-sided two-sample T test.

**Measuring the correlation between, and frequency of, risk alleles and favored alleles**

Correlation coefficients computed using the following equation measure how much favored alleles and risk alleles are on the same haplotypes

\[
    r = \frac{f_{AB} - f_A f_B}{\sqrt{f_A f_a f_B f_b}}
\]

Here \(r\), \(f_A\), \(f_B\), \(f_a\), and \(f_{AB}\) denote the correlation coefficient, the frequency of favored alleles at favored mutation sites, the frequency of risk alleles at disease GWAS sites, the frequency of non-favored alleles at favored mutation sites, the frequency of non-risk alleles at disease GWAS sites, and the frequency of the haplotypes carrying both favored alleles and risk alleles. Two-sample Kolmogorov-Smirnov test was used to evaluate whether the correlation coefficients of risk alleles at hitchhiking mutation sites and ordinary mutation sites are significantly different. Violin plots were drawn upon the correlation coefficients’ probability density distributions (estimated by Kernel Density Estimation) (Figure 3B). The Python package `scipy.stats` was used for the two-sample Kolmogorov-Smirnov test.

A two-sample Kolmogorov-Smirnov test was used to test if the frequencies of risk alleles at hitchhiking and ordinary mutation sites are significantly different. Violin plots were drawn upon the frequencies’ probability density distributions (estimated by Kernel Density Estimation) (Figure 3C).

A Pearson correlation coefficient was computed to measure whether risk alleles’ frequency and their correlation with favored alleles were associated (Figure 3D).