


Deciphering signatures of natural selection via deep learning

Xinghu Qin , Charleston W. K. Chiang and Oscar E. Gaggiotti

Corresponding authors: Xinghu Qin, Centre for Biological Diversity, Sir Harold Mitchell Building, University of St Andrews, Fife KY16 9TF, UK, & CAS Key Laboratory of Genomics and Precision Medicine, Beijing Institute of Genomics, Chinese Academy of Sciences & China National Center for Bioinformation, Beijing 10010, China. E-mail: qin.xinghu@163.com; Oscar E. Gaggiotti, Centre for Biological Diversity, Sir Harold Mitchell Building, University of St Andrews, Fife KY16 9TF, UK. E-mail: oeg@st-andrews.ac.uk

Abstract

Identifying genomic regions influenced by natural selection provides fundamental insights into the genetic basis of local adaptation. However, it remains challenging to detect loci under complex spatially varying selection. We propose a deep learning-based framework, DeepGenomeScan, which can detect signatures of spatially varying selection. We demonstrate that DeepGenomeScan outperformed principal component analysis- and redundancy analysis-based genome scans in identifying loci underlying quantitative traits subject to complex spatial patterns of selection. Noticeably, DeepGenomeScan increases statistical power by up to 47.25% under nonlinear environmental selection patterns. We applied DeepGenomeScan to a European human genetic dataset and identified some well-known genes under selection and a substantial number of clinically important genes that were not identified by SPA, iHS, Fst and Bayenv when applied to the same dataset.

Keywords: deep learning, genome scan, genome-wide association studies, signatures of natural selection

Introduction

One of the main challenges of modern biology is to dissect and understand the molecular basis for naturally occurring phenotypic variation. Addressing this challenge is of fundamental importance not only for the field of evolutionary biology but also for a wide variety of applied fields involving human diseases, improvement of agricultural crops and breeds and biodiversity conservation.

The recent increases in genomic data generated by modern sequencing technologies have advanced our understanding of how natural selection, and its interactions with other evolutionary forces, shapes the genome and phenotype of species. Such technologies have now been applied to the study of a wide range of species, but the statistical methods used to analyze them tend to differ. The standard approach popularized over the last decade or so is to infer signature of selection at trait-associated loci identified through genome-wide association studies (GWAS; e.g. [1]), which linearly model the additive allelic effect of genotypes (the explicative variable) on phenotypes (the dependent variable). All of the recent methods for GWAS can take into account population stratification, using for example linear mixed models that incorporate the genetic relationship matrix (GRM) as a random effect (e.g. [2]). GWAS have been used to identify variants associated with

a wide range of phenotypic traits [3] and their results are used by several studies aimed at detecting polygenic adaptation [4–8], mainly through comparisons of allelic or haplotypic patterns between trait-associated loci and randomly drawn loci from the genome [9, 10].

An alternative approach deploys a wide range of methods to scan the genomes for the signature of selection, while generally without consideration of the phenotypic traits. These genome scans are often deployed to study wild species. Some of them are aimed at identifying genomic regions exhibiting outlier behaviour based on measures of among-population genetic differentiation [11–13] or some other descriptor of spatial origin of a sample (e.g. principal component axes; [14]). Other methods are focused on identifying variants associated with selective sweeps based on linkage disequilibrium patterns along the genome based on either haplotype structure [15–18] or distortions of the allele frequency spectrum [19, 20]. Another important family of methods aimed at establishing associations between loci and environmental variables, with the assumption that such variables could represent selective pressures acting upon genomic regions linked to the outlier loci [11–13, 21]. These methods (with the exception of De Villemereuil and Gaggiotti [22], which is based on F_{ST}) use linear mixed models that consider the genotype as the dependent

Xinghu Qin is a CAS Special Research Associate at CAS Key Laboratory of Genomics and Precision Medicine, Beijing Institute of Genomics, Chinese Academy of Sciences & China National Center for Bioinformation. His research mainly focuses on machine learning and deep learning for population genetic inference. **Charleston W. K. Chiang** is an Assistant Professor in the Department of Population and Public Health Sciences at Keck School of Medicine and Department of Quantitative and Computational Biology at Dornsife College of Letters, Arts, and Sciences, University of Southern California, USA. He is broadly interested in using genetic approaches to understand how natural selection and demographic history shaped the variations in complex traits within and between diverse human populations.

Oscar E. Gaggiotti is a MASTS Professor at the Center for Biological Diversity, School of Biology, University of St Andrews, UK. His research focuses on the study of spatial patterns of genetic diversity to better understand the evolutionary and ecological processes responsible for their origin and maintenance.

Received: May 19, 2022. **Revised:** July 11, 2022. **Accepted:** July 28, 2022

© The Author(s) 2022. Published by Oxford University Press.

This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (<https://creativecommons.org/licenses/by-nc/4.0/>), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com

variable, and the environmental factor as the explicative variable. Some of them are based on machine learning approaches such as latent factor mixed models (LFMMs) [12], and redundancy analysis (RDA) [23–25] and focus on selection gradients showing linear or monotonic patterns across environmental gradients.

Despite this apparent methodological dichotomy with respect to the genetic architecture of phenotypes, it is clear that a thorough understanding of how natural selection shapes the phenotype and genome of species requires some combination of these two different approaches. Indeed, the selective pressures exerted by the heterogeneous abiotic and biotic environment act upon individuals' phenotypes in a complex way and lead to changes in the spatial structuring of variation in the genomic regions underlying them. Thus, in order to fully characterize the action of natural selection, we need to answer three fundamental questions: (i) what are the environmental drivers of natural selection? (ii) what are the phenotypic traits upon which selective pressures act? and (iii) what are the genomic regions underlying those adaptive traits?

In the last few years, there has been an increasing interest in the application of machine learning approaches in population genomics [14, 23, 26] and GWAS [27, 28] but no single method applicable to both has been proposed. Here we present DeepGenomeScan, a unified deep learning approach for genome scan and GWAS, which can be used to answer questions (i) and (iii) above (question (ii) would require the use of a quantitative genetics method; see [29, 30]). The rationale underlying our method is that we can use the genotype of an individual to predict not only its phenotype but also some attribute of the habitat where they are sampled. This framework is akin to GWAS, but in our case, the response variable can be an individual phenotypic trait, an environmental variable associated with its habitat, the geographic location of the individual (latitude and longitude), or a variable describing spatial genetic structure (e.g. reduced features from a dimensionality reduction technique such as principal component analysis [PCA]).

Our approach leverages the power of deep neural networks to approximate arbitrarily complex functions linking dependent and explicative variables [31], as well as recent algorithms for model optimization [32], and for inferring the importance of each explicative variable in predicting the dependent variable [33]. It is important to note that as opposed to the prevalent use of neural networks, i.e. prediction and pattern recognition, here our main interest is in estimating the features (loci) that contribute the most to the predictive power of the neural network. To achieve this goal, we use the concept of 'feature importance' [33], which represents a proxy for the effect size of any given locus. In essence, our method estimates the effect size of genetic variants that explain a particular trait (phenotype or environmental variable), and identifies those with outlier values as pinpointing a QTL or a signature of natural selection.

An important advantage of our method when applied to spatial data is that it can consider any spatial selection pattern including the usually assumed linear environmental gradient as well as arbitrarily complex nonlinear spatial patterns and, importantly, coarse-grained heterogeneous selection with no clear spatial pattern. This represents an important advance as existing methods only consider linear or monotonic nonlinear patterns. In this particular application, our approach generalizes Yang et al.'s [34] idea of using geographic positioning of individuals to identify loci that exhibit particularly steep slopes of allele frequency change associated with recent positive selection. Our generalization is 2-fold: (i) instead of only considering monotonic spatial gradients,

we can detect loci associated with arbitrary and non-monotonic selection patterns, and (ii) the dependent variable can be the geographic location but also a phenotypic trait or an environmental factor.

In what follows, we introduce our method and evaluate its performance focusing on genome-scan applications under spatially complex selection scenarios, but we also present a preliminary evaluation of DeepGenomeScan performance as a tool to carry out GWAS (see Discussion and [Supplementary Material](#)).

Results

Underlying rationale for the deep learning approach

A neural network can be considered as a generalized regression approach used to learn complex functions expressing the association between the input data and a response variable [31, 35]. In our case, the input or predictor variables are the genotypes of the individuals and the response variables are observed phenotypes, environmental values, geographic locations or reduced features describing genetic structure. In what follows, we use the term traits to refer to these response variables.

Neural networks are closely related to standard regression; a linear regression fits a hyperplane to the data, while a neural network fits a space of hyperplanes in a transformed space, which allows it to be nonlinear. However, the increased model complexity of DNNs makes them more prone to overfitting than standard linear regression models. Additionally, their lack of interpretability has limited their use in population genetic applications. Here we present an interpretable neural network-based framework with adaptive hyperparameter optimization for detecting signatures of natural selection. We use a multilayer perceptron (MLP), which is composed of up to three layers of nodes (neurons) with connections between layers but not between nodes within a layer ([Figure 1](#) and [Supplementary Material](#)). Connection edges between nodes can have different weights and the main goal of the MLP is to learn the weights that best describe the complex function that associates inputs and outputs [36–38] by minimizing the difference between the predicted and observed traits. Once the optimal network is found, the weights connecting each input node with the output node can be used to devise a test to identify the loci that contribute the most to the trait under consideration. The underlying rationale for this test is that the absolute values of weights associated with the path linking an input node (i.e. a locus) with the output can be used to estimate the importance of an input variable [39, 40] (node or locus), which in turn is closely associated with the effect size of a genetic variant. Loci with extreme importance values are considered as outliers and good candidates for being under the influence of natural selection. Thus, the approach we implement tests the null hypothesis that the importance (effect size) of an SNP is zero. Full details of our approach are provided in [Methods](#) and [Supplementary Material](#).

Simulation study

We evaluated the performance of our method using simulated data generated by Capblancq et al. [23] (datasets are available from the Dryad Digital Repository at: <https://datadryad.org/review?doi=doi:10.5061/dryad.1s7v5>). The simulations assume a two-dimensional stepping-stone scenario and consider three quantitative traits (QTs), each coded by a distinct set of 10 loci (QTLs), resulting in a total of 30 causal SNPs, out of a total of 1000. Trait values are calculated simply as the sum of genotypic values of the causal loci. Each QT is influenced by environmental factors

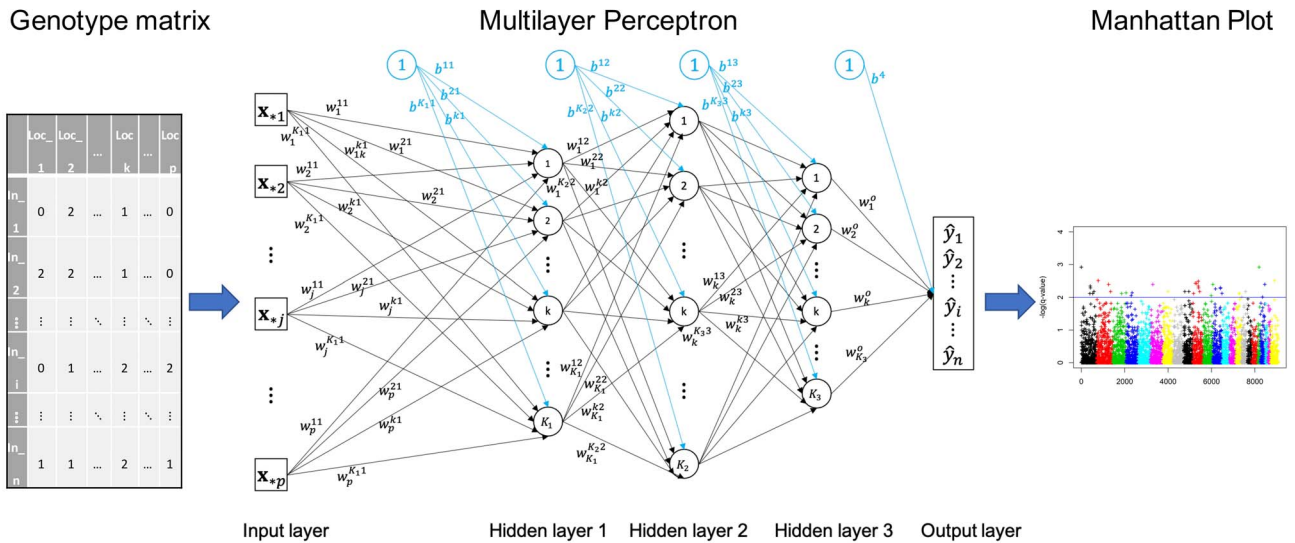


Figure 1. DeepGenomeScan framework. The first (input) layer receives the genotype matrix so that each of its input nodes contains the genotypes of all sampled individuals at a single locus. The last (output) layer contains the predicted trait values. Between these two layers, there is one or more hidden layers containing nodes that compute a nonlinear transformation of the previous layer outputs. Thus, the first hidden layer will transform the input data and feed a signal to the second hidden layer, which in turn will apply a transformation and feed the resulting signal to the third hidden layer, and so on and so forth until the last hidden layer, which will carry out a final transformation and feed the results (the predicted traits) to the output layer. After model training, optimization and hyperparameter tuning, the weights connecting each input node with the output are used to calculate p -values that are then used to produce a Manhattan plot. Note that all edges of the graph have associated weights but here we include only some of them to avoid cluttering the figure.

having a distinct spatial pattern. QT1 is influenced by factors with a quadratic spatial gradient, QT2 is influenced by factors with a linear gradient, and QT3 is influenced by factors with a coarse and patchy spatial pattern. There is a total of 10 environmental factors falling into one of these three categories (Supplementary Figure S2). Further details of the simulation approach are provided in Supplementary as well as in ref. [23].

We benchmarked the performance of our method with those of two recently proposed machine learning-based approaches: pcadapt [41] and RDA [23]. The pcadapt carries out a PCA of the genotype matrix and identifies outlier loci that have an unusually strong correlation with the top PCs, which explain most of the genetic differentiation [14]. RDA is an extension of multiple regression to the modelling of multivariate response data. It can be considered as a PCA constrained by environmental variables [25] as the PCA in this case is carried on the fitted values of a multivariate linear regression of the genotype matrix on the environmental variables. In other words, RDA detects the loci linearly associated with environmental variables by projecting the genetic variation between individuals that is explained by environmental data on a reduced space [23]. Figure 2 presents these results in terms of power (proportion of true positives), false discovery rate (FDR) and false-positive rate (FPR or type I error). Using a threshold $p = 10^{-8}$, as typically done in GWAS, our method has high power to detect all QTL types with a small FDR (0.117) and FPR (0.003). On the other hand, pcadapt and RDA have no power to detect loci associated with QTs 1 and 3, and very low power to detect those associated with QT2 (although they have very low error rates; Figure 2A–C). QQ plots (Supplementary Figure S3) and Manhattan plots (Figure 3) suggest using a threshold of 0.001 for pcadapt and RDA and a threshold of up to 10^{-10} for DeepGenomeScan when controlling for FPR lower than 0.005. Using a less stringent threshold for pcadapt and RDA ($P = 0.001$) but keeping $P = 10^{-8}$ for DeepGenomeScan still leads to a better performance of our method in terms of power when compared

to the other two methods (Figure 2D) and also in terms of FDR and FPR when compared to pcadapt (Figure 2E and F). Increasing the stringency of the test for DeepGenomeScan to the $P = 10^{-10}$ lowers the FDR of our method, making it similar to that of RDA while still having the highest power to detect QTL1 and QTL3 (Figure 2G–I). Overall, all three methods have high power to detect loci associated with QT2, which is influenced by a linear selective gradient. However, DeepGenomeScan with a threshold $P = 10^{-10}$ was the best for detecting QT1 and QT3 loci, which are influenced by nonlinear selection patterns (Figure 2G). In all cases, DeepGenomeScan outperforms both pcadapt and RDA in terms of statistical power while controlling the type I error. In particular, DeepGenomeScan increases statistical power, on average, by up to 47.25% compared with pcadapt, and up to 18.35% compared to RDA under nonlinear environmental gradient selection (QT1 and QT3). Similar results were obtained when we used q -value thresholds (Supplementary Figures S4 and S5). The higher power of DeepGenomeScan to detect loci under nonlinear selection patterns is expected as neural networks can model nonlinear functions while pcadapt and RDA only consider linear functions.

Application to a real data set

A common and longstanding framework to study the action of natural selection is to focus on clinal variation in phenotypic traits or allele frequencies along environmental gradients [42]. One approach to detect genomic regions under selection in these clinal variation scenarios is to identify loci exhibiting extreme frequency gradients across geographic space [34]. The underlying assumption of this approach is that the environmental gradient is continuous, leading to a monotonic but not necessarily linear change in allele frequency or phenotype. Although there may be several examples of such geographic variation, spatial patterns in selective pressures and the associated allele frequency can be non-monotonic. For example, it is possible that the maximum allele frequency is located in the middle of the geographic

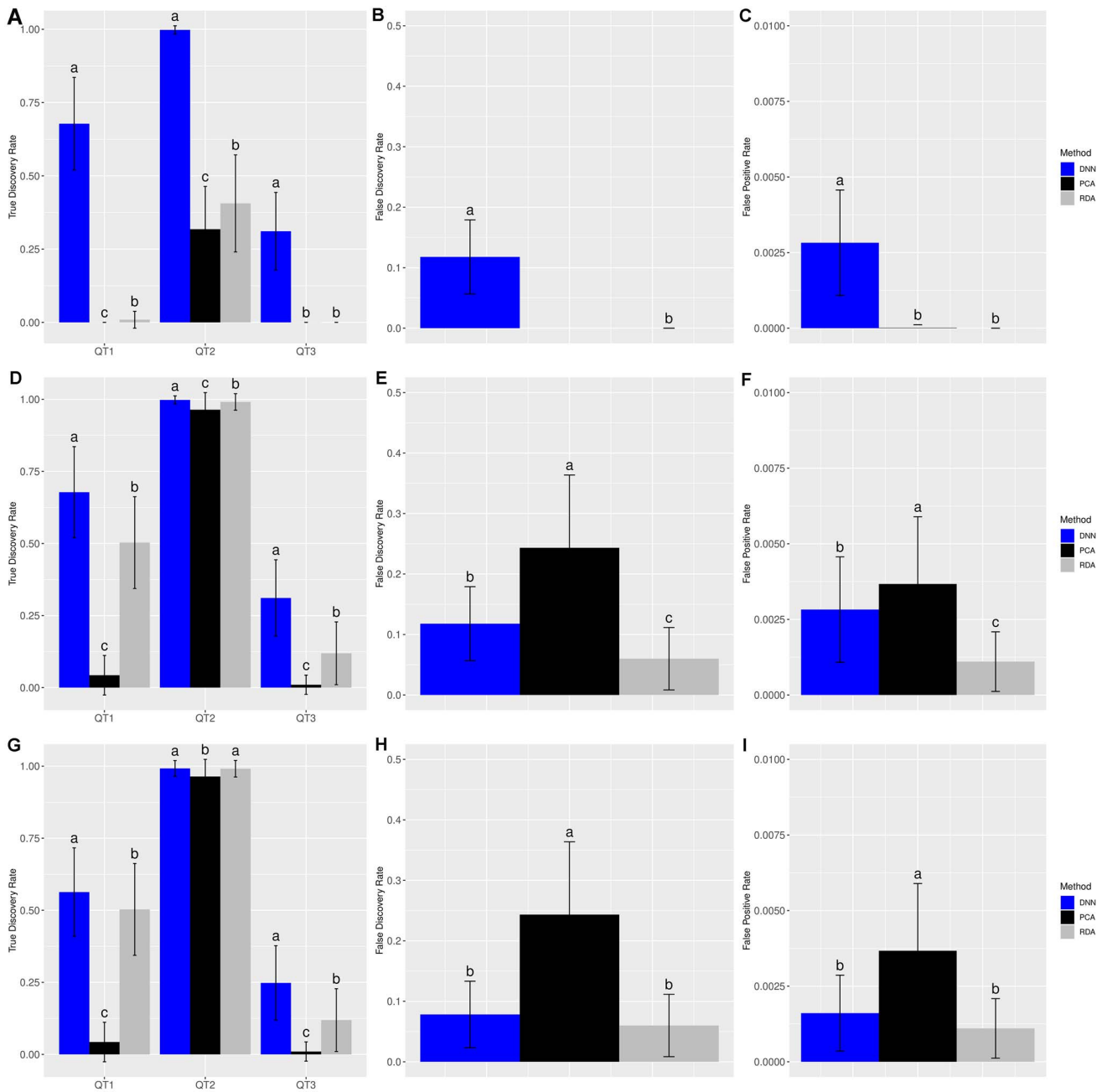


Figure 2. Performance of pcadapt, RDA and DeepGenomeScan (DNN) under different P -value thresholds. (A–C) $P=10^{-8}$ for all three methods; (D–F) $P=0.001$ for pcadapt and RDA and $P=10^{-8}$ for DeepGenomeScan; (G–I) $P=0.001$ for pcadapt and RDA and $P=10^{-10}$ for DeepGenomeScan. Error bars indicate the standard deviation and letters are used to indicate the statistical significance of the difference in performance based on a t -test with $P < 0.01$. Panels A, D and G present the power to identify loci underlying each quantitative trait separately (QT1–3). All other panels present the overall FDR and FPR (across all three types of QTLs). All estimates were based on 100 simulated datasets.

region under study [34]. As our simulations show, our approach is capable of identifying genomic regions subject to such nonlinear selection patterns.

In principle, approaches focused on clinal spatial patterns require the geographic coordinates of each individual sample, which is not always available. A potential solution to this problem is to carry out a PCA on the genotype matrix and use the first two PC axes as surrogates for geographic coordinates [43–45]. This approach, however, is based on a linear combination of genotypes and can lead to poor inference of spatial locations of admixed individuals [34]. A recently developed model-based approach can

be used to overcome this issue [34] but it requires the assumption of a smooth monotonic function to describe allele frequency behaviour as a function of geographic location, which may not be appropriate in all cases. Our implementation of DeepGenomeScan allows the use of a more general dimensionality reduction approach, kernel local discriminant analysis of principal components (KLFDA) [46; [Supplementary Material](#)] when spatial coordinates are not available.

Here we use a human dataset to determine if our approach can uncover new regions of the human genome that may be under the influence of nonlinear spatial selection patterns. We

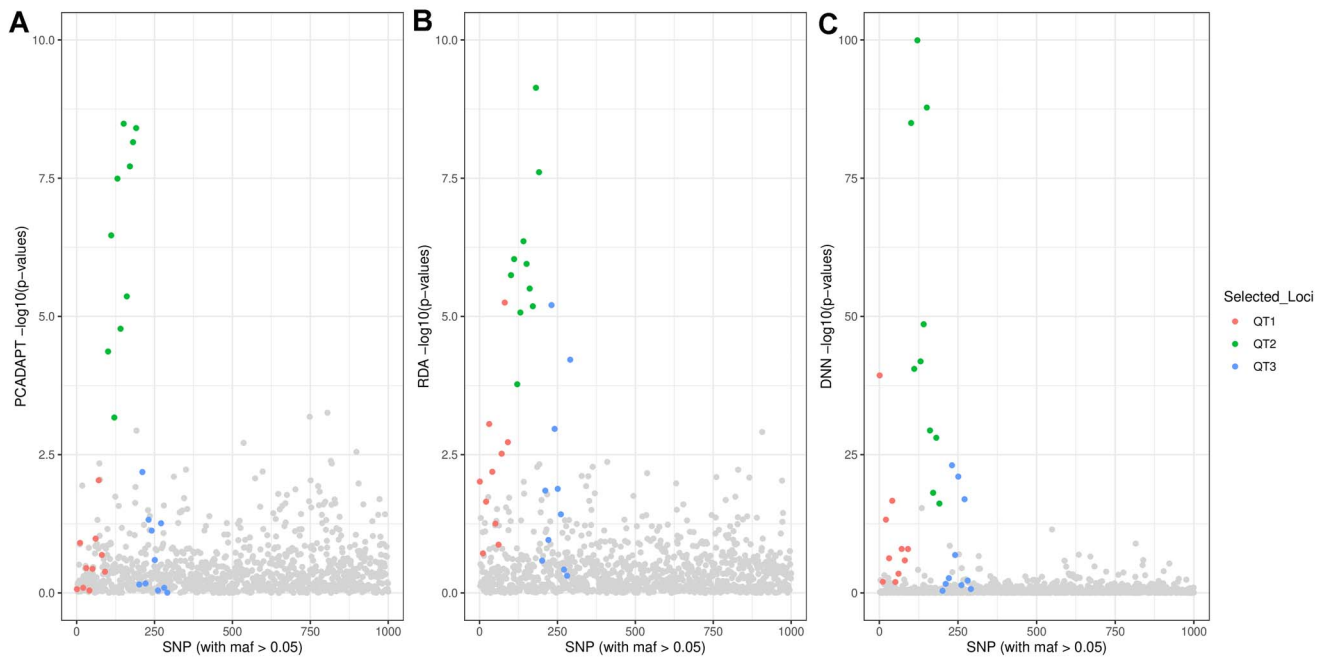


Figure 3. Manhattan plots of the results obtained with (A) pcadapt, (B) RDA and (C) DeepGenomeScan (DNN) for one simulated data set. Note that the scale of the y-axis in C differs from that used for A and B. The threshold for pcadapt and RDA is set to 0.001 and a threshold for DeepGenomeScan is set to 10^{-10} .

applied DeepGenomeScan to European individuals from the Population Reference Sample (POPRES) [44, 47] dataset using previously defined geographic coordinates and then using the two first reduced features from a KLFDA analysis. POPRES represents an excellent example of human genetic variation aligning with geography [44]. The dataset contains a total of 3192 European individuals genotyped at 500 568 loci using the Affymetrix 500 K SNP chip. Details about data filtering steps are provided in Online Methods.

We calculated p -values for each SNP (see Online Methods for details) and obtained a genomic inflation factor $GIF = 1$ for this data set. Using a threshold $P = 10^{-10}$ (Supplementary Figure S6), the analysis using geographic coordinates as the response variables detected 122 outlier SNPs located within the coding region of 33 known genes. The full list of genes we identified, and their chromosome positions are provided in Supplementary Table S1. Consistent with previous widely reported regions under selection, our method detects strong signals at the LCT region on chromosome 2, the ADH1C region on chromosome 4, the HLA region on chromosome 6, as well as the OCA2 and the HERC2 region on chromosome 15 (Figure 4 and Supplementary Table S1).

Besides the well-known genes under directional selection, our method also detects some disease-related genes that exhibited extreme variation across geographic space but were not identified by SPA [48] and other popular methods such iHS [49], Fst [50], Bayenv [21] when applied to the same database (cf., Supplementary Table S4 in ref. [34]); these include MGAT5, TMEM163, ACMSD, CCNT2, MAP3K19, R3HDM1, UBXN4, MCM6, DARS1, EHMT2, and CFB (Supplementary Table S1). For example, MCM6 is a regulatory element that controls the expression of the LCT gene. The MGAT5 gene (alpha-1,6-mannosylglycoprotein 6-beta-N-acetylglucosaminyltransferase) is one of the most important enzymes involved in the regulation of the biosynthesis of glycoprotein oligosaccharides and is associated with invasive malignancies and sclerosis [51–53] as well as visceral fat in women [54]. The TMEM163 gene (transmembrane protein 163), is

associated with Parkinson’s disease, ischemic stroke and coronary artery disease [55, 56]. The ACMSD gene (aminocarboxymuconate semialdehyde decarboxylase) is associated with Parkinson’s disease [55] and childhood obesity [57]. There are several other outlier SNPs detected by our method on chromosome 2, 4, 6, 10 and 11 (Supplementary Table S1) that are not within known genes. However, they may be linked to regulatory genes and, therefore, are of interest for human genetic studies.

In the analysis where we replaced the original geographic coordinates with the first two reduced features obtained from KLFDA, we first calculated the correlation between the first two reduced features and the original geographic coordinates. We compared these results to those of a similar analysis using the first two PC axes obtained from a PCA of the genotype matrix. The results confirm that KLFDA provides better estimates of geographic location than PCA (see Figure 4A and Supplementary Figure S7), as previously demonstrated [46].

The genome scan based on the first two KLFDA reduced features identifies a somewhat smaller number of outlier SNPs (116 loci, in 34 known genes) than when using geographic coordinates (Figure 4B and C and Supplementary Table S2). However, the $\log_{10}(p\text{-values})$ of the significant loci detected from these two different strategies showed a high correlation ($r = 0.995$, $P < 2.2 \times 10^{-16}$; Supplementary Figure S8). 88% of the outliers detected in this analysis were also identified by the analysis based on geographic coordinates (102 shared loci). This analysis also detected well-known selected genomic regions (LCT, HLA, ADH1, HERC2; Supplementary Table S2). However, there were six genes that were not identified when using latitude and longitude (Supplementary Table S3). This group includes genes associated with cancer (e.g. AK5), diabetes mellitus (e.g. HSPA1L, HCG26, BAG6, APOM), and a gene of unknown function, LOC101928978. On the other hand, there are also six genes that were not identified in this analysis but were highlighted by the analysis based on geographic coordinates. This group includes genes associated with pathogen recognition and activation of innate immunity (e.g. TLR10), speech-language

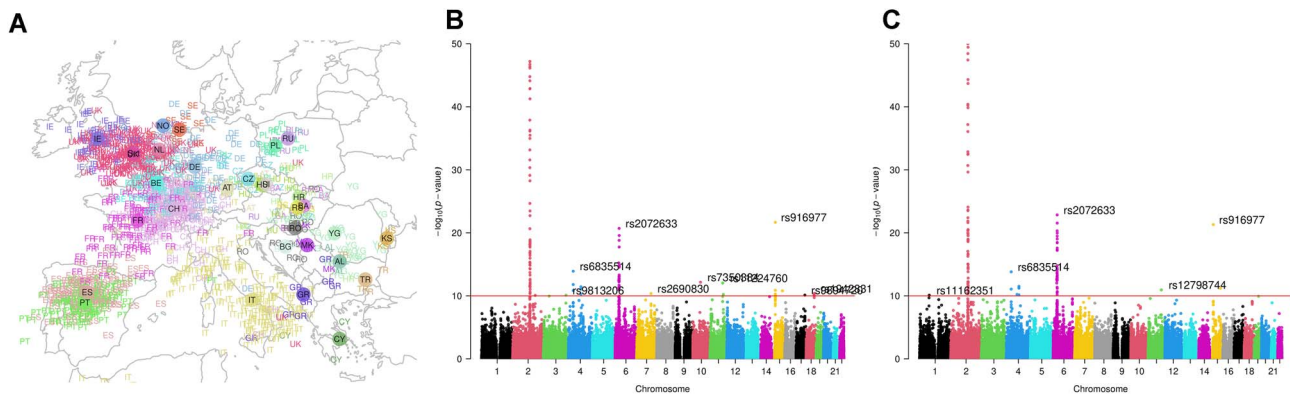


Figure 4. Spatial genetic structure of European populations and signals of selection detected by DeepGenomeScan. **(A)** Spatial genetic structure of European populations inferred via KLFDA with a $\sigma = 5$. **(B)** Manhattan plot indicating loci under spatial selection obtained from DeepGenomeScan using geographic coordinates as the response variables. **(C)** Manhattan plot indicating loci under spatial selection obtained from DeepGenomeScan using inferred spatial genetic structure (the first two reduced features of KLFDA) as the response variables. The outliers in B–C were identified with a threshold of $P = 10^{-10}$. The top hits are shown in panels B and C with their dbSNP Reference ID (rs ID). Country abbreviations: AL, Albania; AT, Austria; BA, Bosnia-Herzegovina; BE, Belgium; BG, Bulgaria; CH, Switzerland; CY, Cyprus; CZ, Czech Republic; DE, Germany; ES, Spain; FR, France; GB, United Kingdom; GR, Greece; HR, Croatia; HU, Hungary; IE, Ireland; IT, Italy; KS, Kosovo; MK, Macedonia; NO, Norway; NL, Netherlands; PL, Poland; PT, Portugal; RO, Romania; RS, Serbia and Montenegro; RU, Russia; Sct, Scotland; SE, Sweden; TR, Turkey; YG, Yugoslavia. A full list of outlier loci can be found in [Supplementary Tables S1 and S2](#).

disorder 1 (e.g. FOXP2), skin and eye colour (e.g. OCA2), expression of gamma-Glutamyltransferase (SOX9-AS1), bipolar disease or neuropsychiatric disorders (e.g. CDH7), as well as a gene of unknown function, LOC105373760 ([Supplementary Table S4](#)).

Discussion

In this study, we present DeepGenomeScan, a new deep learning method that can scan the genomes to identify loci under the influence of directional selection. Although model-based genome scans are still in use, they are increasingly being replaced by non-parametric methods such as pcadapt [14] and RDA [23]. However, both approaches have their shortcomings: pcadapt cannot explicitly associate those loci with environmental variables that may underlie the selective pressure [23], and RDA cannot detect associations influenced by nonlinear environmental gradients (cf., [Figure 2](#)). DeepGenomeScan overcomes these limitations by leveraging the ability of neural networks to model nonlinear functions. We found that DeepGenomeScan outperformed pcadapt and RDA-based genome scans in detecting signatures of selection under various spatial selection patterns. Markedly, DeepGenomeScan increases statistical power by up to 47.25% under nonlinear environmental selection patterns (quadratic environment gradients and grained heterogeneous spatial patterns). We applied DeepGenomeScan to POPRES [47] dataset to detect signals of natural selection. DeepGenomeScan identified a number of well-known genes under selection in the European population (e.g. LCT, HLA, ADH1, HERC2 and OCA2), as well as a number of previously unreported genes putatively under selection; these new candidate genes could not be identified by SPA [48], iHS [49], Fst [50] and Bayenv [21] when applied to the same dataset. Presumably, some of these genes were not previously detected because they are not subject to linear and monotonic selection gradient in space. Some of these genes have also been linked to complex diseases such as Parkinson's, obesity or various types of cancer, suggesting that natural selection plays a role in shaping the disease susceptibility of modern day human populations.

The insight upon which our method relies is the idea that we can use the genotypes of individuals to predict any associated trait, not limited to just their phenotype but also their

spatial location or the environmental attributes of the habitat they live in. Intuitively, the type of functions that can be used to describe the association between genotypes and any of these dependent variables is likely to be radically different depending on the 'trait' under consideration. Therefore, no single model-based approach can be used to take advantage of the above-mentioned insight. Additionally, no single model-free statistical method can be used to approximate very complex nonlinear functions linking genotypes and these disparate traits. In contrast, Deep Neural Networks can approximate arbitrarily complex nonlinear functions linking dependent and predictor variables [31]. For example, we applied our method to link genotypes to proxies of spatial locations based on the reduced features obtained from KLFDA, a nonlinear feature reduction method that has been shown to outperform PCA in recapitulating individual geography [46]. These characteristics allows DeepGenomeScan to identify loci associated with very complex spatial selection patterns and thus can be an invaluable tool for population and evolutionary genomics applications. While we focused on the genome-scan applications to detect signatures of natural selection, we note that our method can also be used to identify genomic regions associated with phenotypic traits (i.e. GWAS). In the [Supplementary Material](#), we present a preliminary evaluation of the performance of DeepGenomeScan when applied to detect genomic regions associated with QTs. This was done using the same set of simulations described for the genome-scan application but in this case, the response variable was a QT that influenced fitness (see details in [Supplementary Methods](#)). The results showed that DeepGenomeScan can achieve high power (80%) to identify QTs under a wide range of spatial scenarios (linear, nonlinear and coarse-grained heterogeneous environments) while maintaining very low error rates (FDR < 0.12; FPR < 0.001; see [Supplementary Figures S9 and S10](#)). A comprehensive evaluation on this framework in a more realistic GWAS setting needs to be tested independently, but is outside the scope of the current paper.

Our approach represents an integration of GWAS and genome scan approaches in the specific case of spatially structured populations. More precisely, it is focused on traits that vary spatially. Thus, it is not well adapted to study global selective sweeps unless

they are at an early stage where selected variants still exhibit clinal variation. Similarly, DeepGenomeScan is not appropriate to carry out GWAS of panmictic populations. On the other hand, it is perfectly suited to the study of local adaptation. There are three questions that need to be answered in this context: (i) what are the environmental drivers of natural selection? (ii) what are the phenotypic traits upon which selective pressures act? and (iii) what are the genomic regions underlying those adaptive traits? As a genome-scan method, DeepGenomeScan can identify the environmental drivers of natural selection while as a GWAS tool, it can identify the genomic regions underlying adaptive traits. Therefore, it can answer questions (i) and (iii). Question (ii), however, would require the application of a quantitative genetic method to identify phenotypic traits that are good candidates for being involved in local adaptation to heterogenous environmental conditions. In this regard, we note that a recent review [29] has highlighted the fact that the study of local adaptation requires combining population genomics and GWAS methods in the context of common garden experiments. Therefore, our unified approach provides an excellent tool to implement such frameworks.

It is noted that DeepGenomeScan significantly improved the power of detecting selection under nonlinear environmental gradients. Although less well studied empirically, nonlinear environmental selection gradients (e.g. coarse-grained heterogenous selection) with no clear spatial pattern has been an important focus of theoretical studies aimed at explaining observed patterns of genetic variation [58–63]. This scenario is particularly relevant to studies of protected polymorphisms [59, 61, 62] and hard versus soft selection [60]. Additionally, the importance of this type of selection pattern is particularly relevant for genome-wide association studies focused on diseases and phenotypic characters that do not exhibit clear spatial patterns. Deep learning genome-scan methods capable of uncovering genomic regions associated with this type of selection pattern can, therefore, provide a more general understanding of how prevalent these particular mechanisms are.

There are still several methodological challenges faced when implementing deep learning methods, which are associated with the computational cost of hyperparameter tuning and neural network training. However, recent adaptive resampling algorithms (cf., ref. [32]), which we implemented in our software, carry out an efficient exploration of hyperparameter space, allowing the use of deep learning to analyze large population genomic data sets.

Although the application of deep learning to population genetics problems is still in its infancy [32], there are already some examples of such applications [36, 64–69]. One of these applications is aimed at distinguishing between hard and soft sweeps and simultaneously incorporating the confounding effects of demographic history [36]. This application involves using simulated data generated under predefined evolutionary models. A drawback of this strategy is that it introduces model assumptions into a computational framework that could be completely free of them. Therefore, it introduces the same model mis-specification issues that affect statistical inference based on generative models [70]. This drawback is absent when the objective is to assign individuals to geographic locations as in ref. [64] or when scanning the genome in search of outlier loci, as in our study. Overall, DeepGenomeScan fully exploits the flexibility and power of deep neural networks, which makes it applicable to a wide range of problems including identification of genomic regions associated with diseases or economically important phenotypic traits, assignment of individuals to geographic locations, or identification of environmental factors associated with selective pressures.

Methods (online methods)

In what follows we first describe the architecture of the deep neural network and its implementation. We then describe the statistical approach to identify SNPs that are located in regions subject to positive (local adaptation or selective breeding) or negative selection (diseases). Finally, we explain the simulation approach used to test performance and the dataset used to provide a practical application.

MLP architecture

In this study, we constructed an MLP network with three hidden layers (Figure 1 and Supplementary Figure S1). To describe this MLP, we first assume that the genotypes of the n sampled individuals at p loci are described by a $n \times p$ matrix, $\mathbf{x} = (x_{ij})$, and are coded by the count of reference alleles, which in the case of a biallelic marker takes values $x_{ij} = 0, 1, 2$, where $i = 1, 2, \dots, n$, and $j = 1, 2, \dots, p$. Furthermore, the trait values of all individuals are arranged in a $n \times 1$ vector $\mathbf{y} = (y_i)$. Then the input layer of the MLP consists of p nodes, each one representing a distinct locus. The information contained in node j comprises the individual genotypes at locus j , $\mathbf{x}_{*j} = (x_{1j}, \dots, x_{nj})$. This information is transformed into a signal which is fed to the first hidden layer. The signal received by each node $k = 1, 2, \dots, K_1$ of the first hidden layer $\hat{\mathbf{s}}^{k1} = (\hat{s}_1^{k1}, \dots, \hat{s}_n^{k1}, \dots, \hat{s}_n^{k1})$ represents a weighted nonlinear regression of individuals' trait values on their multilocus genotypes. Here, the superscript $k1$ identifies the k th-node of the first hidden layer. Each element \hat{s}_i^{k1} of the signal vector received by a node k of the first hidden layer is given by

$$\hat{s}_i^{k1} = f^1(\mathbf{x}_{i*}) = G\left(\sum_{j=1}^p x_{ij} w_j^{k1} + b^{k1}\right) = G(\mathbf{x}_{i*} \mathbf{w}^{k1} + b^{k1}), \quad (1)$$

where w_j^{k1} is the weight of locus j , b^{k1} is a scalar representing a bias in the signal received by node k in the first hidden layer, and $G(\cdot)$ is the activation function used to transform the information sent by the input layer to the first hidden layer.

In a similar way, each node k of the second hidden layer will take the output signal of all nodes of the first hidden layer and apply a transformation before sending the signal to the third hidden layer:

$$\hat{s}_i^{k2} = f^2(\hat{\mathbf{s}}^{k1}) = G\left(\sum_{m=1}^{K_1} \hat{s}_i^{m1} w_m^{k2} + b^{k2}\right) = G(f^1(\mathbf{x}_{i*}) \mathbf{w}^{k2} + b^{k2}) \quad (2)$$

If there are only two hidden layers, then the neural network is represented by $f(\mathbf{x}) = f^2(f^1(\mathbf{x}))$. In general, there can be an arbitrary number L of hidden layers, each consisting of K_l nodes and the signal generated by the last hidden layer is the vector of predicted trait values, $\hat{\mathbf{y}} = (\hat{y}_i)$. Therefore, a MLP network can be represented by

$$f(\mathbf{x}) = f^L(f^{L-1}(\dots f^2(f^1(\mathbf{x}))), \quad (3)$$

where $\mathbf{f}^l(\cdot)$ is the signal received by the l^{th} layer.

Given the input data, i.e. the genotypes, and the observed trait values, the neural network learns the weights $\mathbf{w} = (\mathbf{w}^1, \dots, \mathbf{w}^L)$ with $\mathbf{w}^l = (\mathbf{w}^{l1}, \dots, \mathbf{w}^{lK_l})$ and biases $\mathbf{b} = (\mathbf{b}^1, \dots, \mathbf{b}^L)$ with $\mathbf{b}^l = (b^{l1}, \dots, b^{lK_l})$ that best describe the relationship between the inputs and the outputs by minimizing the difference between predicted and observed trait values [71, 72]. These weights represent an essential element of our genome

scan method because they are used to identify outlier loci (see below) using a procedure supported by previous studies [40, 73], which show that the importance of a variable (locus) can be estimated as the combination of the absolute values of weights associated with the graph edges connecting the predictor variable (locus) with the MLP output (predicted trait).

Implementation

DeepGenomeScan trains the deep neural network to estimate the nonlinear function mapping genotypes to individual traits. This process involves not only learning the vectors of weights and biases by iteratively adjusting these parameters but also the tuning of a large number of hyperparameters associated with structural components of the model, which impact model performance. These include those determined by the network architecture (activation functions, number of neurons per hidden layer) and those included in the optimization algorithm used for training (e.g. learning rate, batch size). For optimization, we used resilient backpropagation [74] with weight backtracking when analyzing both simulated and real datasets. However, the tuning algorithm used depended on the size of the input data. In the case of simulated datasets, which were of limited size, we used a full resampling strategy corresponding to Algorithm 1 in [32] based on a repeated k -fold cross validation with $k=5$ and five repetitions. In the case of the large POPRES dataset, we used an adaptive resampling approach corresponding to Algorithm 2 in [32], which incorporates fertility assessment into the model tuning process. The resampling approach in this case was the same as in the analysis of simulated data (five replications of a fivefold cross-validation). Detailed information about the settings used for the optimization and tuning algorithms is presented in [Supplementary Notes](#).

Identification of outlier loci

As mentioned before, the SNP importance can be estimated as combinations of the absolute values of connection weights [40, 73]. This is done once the optimal model is found. We used Olden and Jackson's [33] method, which is based on Garson's [75] algorithm to calculate the relative importance for each input node but adds a randomization step to identify non-significant connection weights.

DeepGenomeScan carries out separate runs for each trait and generates a vector of SNP importance for each of them. Given T traits, the position of each SNP in trait space is described by its associated vector of importance values and, therefore, it is possible to calculate the Mahalanobis distance between the focal locus and all other loci. Our approach leads to an intuitive definition of outlier loci as any locus with an 'extreme' Mahalanobis distance. Since the squared Mahalanobis distance follows a chi-squared distribution with T degrees of freedom, the P -values associated with each SNP can be obtained from a χ^2_T distribution [76] and used for identifying 'outlier' loci. This approach relies on the assumption that variables used to calculate Mahalanobis distance follow a normal distribution. Therefore, we used the arcsine transformation to normalize the weights before calculating the Mahalanobis distance. We set the p -value threshold by controlling FDR [77] under 0.1 with QQ plots as additional references [78]. SNPs at the point where p -values deviate from the expected distribution on the QQ plot are considered as highly significant SNPs [78]. Based on this principle, we determined the P -value threshold for the analysis of a real dataset described below.

Human dataset

We applied DeepGenomeScan to European populations from the POPRES project [47]. This project consists of 6000 individuals from worldwide populations. The subsample of European populations contains a total of 3192 individuals genotyped at 500 568 loci using the Affymetrix 500 K SNP chip. The sample collections and genotyping for POPRES are described in [47]. We removed individuals from outside of Europe and individuals whose grandparents had different geographic origins based on the criteria used by [44]. We also removed samples that only have one individual per country, such as Denmark, Finland, Latvia, Ukraine, Slovakia and Slovenia. Geographic coordinates for each individual corresponded to the central point of the geographic area of the individual's country of birth.

We removed the monomorphic SNPs and filtered out autosomal biallelic SNPs with a minor allele frequency (MAF) below 5%, and a missing rate of 5% or more. In the end, we kept 1,382 individuals from 32 countries carrying 283,499 autosomal SNPs. We annotated and updated the variant reference ID based on the comprehensive report of short human variations from the human variation database (dbSNP, GRCh37p13, b151, release 2018, https://ftp.ncbi.nih.gov/snp/organisms/human_9606/VCF/) using bcftools [79].

Data and code availability

The simulation data used in this study are available at <https://datadryad.org/review?doi=doi:10.5061/dryad.1s7v5>.

The datasets used for the analyses described in this manuscript were obtained from dbGap at https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000145.v4.p2 through dbGap accession number phs000145.v4.p2 (request approval #90291-1).

The DeepGenomeScan R code is hosted on GitHub at <https://github.com/xinghuq/DeepGenomeScan>. The scripts used in this study are available at https://github.com/xinghuq/DeepGenomeScan/tree/webpkg/DeepGenomeScan_simulation%2BPOPRES. The KLFDPAC package is available at <https://xinghuq.github.io/KLFDPAC/>.

Authors' contributions

XQ and OEG designed the study. XQ carried out the analyses and interpreted results with input from OEG and CWKC. OEG and XQ wrote the article with input from CWKC.

Acknowledgments

XQ was supported by a PhD scholarship from the China Scholarship Council and now is supported by International Postdoctoral Exchange Fellowship Program (Talent-Introduction Program) from China Postdoc Council. CWKC is supported in part by National Institute of General Medical Sciences (NIGMS) of the National Institute of Health (award number R35GM142783). Computation for this work is supported in part by USC's Center for Advanced Research Computing (<https://www.carc.usc.edu/>). This work benefited from discussions with Pierre de Villemeureuil. We thank Dr Thibaut Capablanca for providing script for [Supplementary Figure S2](#).

Ethics declarations

Ethics approval and consent to participate

Ethical approval was not needed for this study.

Conflict of interest

The authors declare no competing interests.

References

- Bush WS, Moore JH. Chapter 11: genome-wide association studies. *PLoS Comput Biol* 2012;**8**:e1002822.
- Kang HM, Sul JH, Service SK, et al. Variance component model to account for sample structure in genome-wide association studies. *Nat Genet* 2010;**42**:348–U110.
- Tam V, Patel N, Turcotte M, et al. Benefits and limitations of genome-wide association studies. *Nat Rev Genet* 2019;**20**:467–84.
- Edge MD, Coop G. Reconstructing the history of polygenic scores using coalescent trees. *Genetics* 2019;**211**:235–62.
- Field Y, Boyle EA, Telis N, et al. Detection of human adaptation during the past 2000 years. *Science* 2016;**354**:760–4.
- Racimo F, Berg JJ, Pickrell JK. Detecting polygenic adaptation in admixture graphs. *Genetics* 2018;**208**:1565–84.
- Speidel L, Forest M, Shi SN, et al. A method for genome-wide genealogy estimation for thousands of samples. *Nat Genet* 2019;**51**:1321.
- Turchin MC, Chiang CWK, Palmer CD, et al. Evidence of widespread selection on standing variation in Europe at height-associated SNPs. *Nat Genet* 2012;**44**:1015.
- Chen M, Chiang CW. Allele frequency differentiation at height-associated SNPs among continental human populations. *Eur J Hum Genet* 2021;**29**:1542–8.
- Chen M, Sidore C, Akiyama M, et al. Evidence of polygenic adaptation in Sardinia at height-associated loci ascertained from the biobank Japan. *Am J Hum Genet* 2020;**107**:60–71.
- de Villemereuil P, Gaggiotti OE. A new F-ST-based method to uncover local adaptation using environmental variables. *Methods Ecol Evol* 2015;**6**:1248–58.
- Frichot E, Schoville SD, Bouchard G, et al. Testing for associations between loci and environmental gradients using latent factor mixed models. *Mol Biol Evol* 2013;**30**:1687–99.
- Gaggiotti OE, Bekkevold D, Jorgensen HBH, et al. Disentangling the effects of evolutionary, demographic, and environmental factors influencing the genetic structure of natural populations: Atlantic herring as a case study. *Evolution* 2009;**63**:2939–51.
- Duforet-Frebourg N, Luu K, Laval G, et al. Detecting genomic signatures of natural selection with principal component analysis: application to the 1000 genomes data. *Mol Biol Evol* 2016;**33**:1082–93.
- Sabeti PC, Reich DE, Higgins JM, et al. Detecting recent positive selection in the human genome from haplotype structure. *Nature* 2002;**419**:832–7.
- Sabeti PC, Schaffner SF, Fry B, et al. Positive natural selection in the human lineage. *Science* 2006;**312**:1614–20.
- Voight BF, Kudaravalli S, Wen X, et al. A map of recent positive selection in the human genome. *PLoS Biol* 2006;**4**:e72.
- Stephan W, Li H. The recent demographic and adaptive history of *Drosophila melanogaster*. *Heredity* 2007;**98**:65–8.
- Chen H, Patterson N, Reich D. Population differentiation as a test for selective sweeps. *Genome Res* 2010;**20**:393–402.
- Fariello MI, Boitard S, Naya H, et al. Detecting signatures of selection through haplotype differentiation among hierarchically structured populations. *Genetics* 2013;**193**:929–41.
- Coop G, Witonsky D, Di Rienzo A, et al. Using environmental correlations to identify loci underlying local adaptation. *Genetics* 2010;**185**:1411–23.
- De Villemereuil P, Gaggiotti OE. A new FST-based method to uncover local adaptation using environmental variables. *Methods Ecol Evol* 2015;**6**:1248–58.
- Capblancq T, Luu K, Blum MG, et al. Evaluation of redundancy analysis to identify signatures of local adaptation. *Mol Ecol Resour* 2018;**18**:1223–33.
- Forester BR, Jones MR, Joost S, et al. Detecting spatial genetic signatures of local adaptation in heterogeneous landscapes. *Molecular ecology* 2016;**25**(1):104–20.
- Forester BR, Lasky JR, Wagner HH, et al. Comparing methods for detecting multilocus adaptation with multivariate genotype–environment associations. *Molecular Ecology* 2018;**27**(9):2215–33.
- Torada L, Lorenzon L, Beddis A, et al. ImaGene: a convolutional neural network to quantify natural selection from genomic data. *BMC bioinformatics* 2019;**20**(9):1–12.
- Yan Q, Jiang Y, Huang H, et al. Genome-wide association studies-based machine learning for prediction of age-related macular degeneration risk. *Transl Vis Sci Technol* 2021;**10**:29–9.
- Sun T, Wei Y, Chen W, et al. Genome-wide association study-based deep learning for survival prediction. *Stat Med* 2020;**39**:4605–20.
- de Villemereuil P, Gaggiotti OE, Mouterde M, et al. Common garden experiments in the genomic era: new perspectives and opportunities. *Heredity* 2016;**116**:249–54.
- de Villemereuil P, Mouterde M, Gaggiotti OE, et al. Patterns of phenotypic plasticity and local adaptation in the wide elevation range of the alpine plant *Arabis alpina*. *J Ecol* 2018;**106**:1952–71.
- Hornik K, Stinchcombe M, White H. Multilayer feedforward networks are universal approximators. *Neural Netw* 1989;**2**:359–66.
- Kuhn M. Futility analysis in the cross-validation of machine learning models. *arXiv:14056974* 2014.
- Olden JD, Jackson DA. Illuminating the “black box”: a randomization approach for understanding variable contributions in artificial neural networks. *Ecol Model* 2002;**154**:135–50.
- Yang WY, Novembre J, Eskin E, et al. A model-based approach for analysis of spatial structure in genetic data. *Nat Genet* 2012;**44**:725–U163.
- Specht DF. A general regression neural network. *IEEE transactions on neural networks* 1991;**2**(6):568–76.
- Sheehan S, Song YS. Deep learning for population genetic inference. *PLoS Comput Biol* 2016;**12**(3):e1004845.
- Attali J-G, Pagés G. Approximations of functions by a multilayer perceptron: a new approach. *Neural Netw* 1997;**10**:1069–81.
- Pal SK, Mitra S. *Multilayer perceptron, fuzzy sets, classification*. IEEE Transactions on Neural Networks, 1992;**3**(5).
- Gevrey M, Dimopoulos L, Lek S. Review and comparison of methods to study the contribution of variables in artificial neural network models. *Ecol Model* 2003;**160**:249–64.
- Olden JD, Joy MK, Death RG. An accurate comparison of methods for quantifying variable importance in artificial neural networks using simulated data. *Ecol Model* 2004;**178**:389–97.
- Luu K, Bazin E, Blum MG. Pcadapt: an R package to perform genome scans for selection based on principal component analysis. *Mol Ecol Resour* 2017;**17**:67–77.
- Endler JA. *Geographic variation, speciation and clines*. Princeton, NJ: Princeton University Press, 1977.
- Lao O, Lu TT, Nothnagel M, et al. Correlation between genetic and geographic structure in Europe. *Curr Biol* 2008;**18**:1241–8.
- Novembre J, Johnson T, Bryc K, et al. Genes mirror geography within Europe. *Nature* 2008;**456**:98–101.

45. Chiang CW, Mangul S, Robles C, et al. A comprehensive map of genetic variation in the world's largest ethnic group—Han Chinese. *Mol Biol Evol* 2018;**35**:2736–50.
46. Qin X, Chiang CWK, Gaggiotti OE. KLFDAFC: a supervised machine learning approach for spatial genetic structure analysis. *Brief Bioinform* 2022;**23**(4):bbac202. <https://doi.org/10.1093/bib/bbac202>.
47. Nelson MR, Bryc K, King KS, et al. The population reference sample, POPRES: a resource for population, disease, and pharmacological genetics research. *Am J Hum Genet* 2008;**83**:347–58.
48. Yang W-Y, Novembre J, Eskin E, et al. A model-based approach for analysis of spatial structure in genetic data. *Nat Genet* 2012;**44**:725.
49. Sabeti PC, Reich DE, Higgins JM, et al. Detecting recent positive selection in the human genome from haplotype structure. *Nature* 2002;**419**:832–7.
50. Lewontin RC, Krakauer J. Distribution of gene frequency as a test of the theory of the selective neutrality of polymorphisms. *Genetics* 1973;**74**:175–95.
51. Granovsky M, Fata J, Pawling J, et al. Suppression of tumor growth and metastasis in Mgat5-deficient mice. *Nat Med* 2000;**6**:306–12.
52. Brynedal B, Wojcik J, Esposito F, et al. MGAT5 alters the severity of multiple sclerosis. *J Neuroimmunol* 2010;**220**:120–4.
53. Wang R, Fan Q, Zhang J, et al. Hydrogen sulfide demonstrates promising antitumor efficacy in gastric carcinoma by targeting MGAT5. *Transl Oncol* 2018;**11**:900–10.
54. Fox CS, Liu Y, White CC, et al. Genome-wide association for abdominal subcutaneous and visceral adipose reveals a novel locus for visceral fat in women. *PLoS Genet* 2012;**8**:e1002695.
55. Nalls MA, Pankratz N, Lill CM, et al. Large-scale meta-analysis of genome-wide association data identifies six new risk loci for Parkinson's disease. *Nat Genet* 2014;**46**:989–93.
56. Dichgans M, Malik R, König IR, et al. Shared genetic susceptibility to ischemic stroke and coronary artery disease: a genome-wide analysis of common variants. *Stroke* 2014;**45**:24–36.
57. Comuzzie AG, Cole SA, Laston SL, et al. Novel genetic loci identified for the pathophysiology of childhood obesity in the Hispanic population. *PLoS One* 2012;**7**:e51954.
58. Bulmer MG. Multiple niche polymorphism. *Amer Natur* 1972;**106**:254–7.
59. Levene H. Genetic equilibrium when more than one ecological niche is available. *Amer Natur* 1953;**87**:331–3.
60. Levins R, MacArthur R. The maintenance of genetic polymorphism in a spatially heterogeneous environment: variations on a theme by Howard Levene. *Amer Natur* 1966;**100**:585–9.
61. Prout T. Sufficient conditions for multiple niche polymorphism. *Amer Natur* 1968;**102**:493.
62. Strobeck C. Haploid selection withn alleles in m niches. *Amer Natur* 1979;**113**:439–44.
63. Maynard S. Genetic polymorphism in a varied environment. *Amer Natur* 1970;**104**:487–90.
64. Battey CJ, Ralph PL, Kern AD. Predicting geographic location from genetic variation with deep neural networks. *Elife* 2020;**9**:e54507.
65. Fligel L, Brandvain Y, Schrider DR. The unreasonable effectiveness of convolutional neural networks in population genetic inference. *Mol Biol Evol* 2019;**36**:220–38.
66. Akesson M, Singh P, Wrede F, et al. Convolutional neural networks as summary statistics for approximate Bayesian computation. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 2021;(01):1–1.
67. Jiang B, Wu T-Y, Zheng C, et al. Learning summary statistic for approximate Bayesian computation via deep neural network. *Stat Sin* 2017;**27**(4):1595–1618.
68. Sanchez T, Cury J, Charpiat G, et al. Deep learning for population size history inference: design, comparison and combination with approximate Bayesian computation. *Mol Ecol Resour* 2021;**21**:2645–60.
69. Isildak U, Stella A, Fumagalli M. Distinguishing between recent balancing selection and incomplete sweep using deep neural networks. *Mol Ecol Resour* 2021;**21**:2706–18.
70. Schrider DR, Kern AD. Supervised machine learning for population genetics: a new paradigm. *Trends Genet* 2018;**34**:301–12.
71. Yang J, Ahmadi M, Jullien GA, et al. Model validation and determination for neural network activation function modeling. 1998 Midwest Symposium on Circuits and Systems (Cat. No. 98CB36268). IEEE, 1998, 548–51.
72. GOODFELLOW Ia. *Deep learning / Ian Goodfellow, Yoshua Bengio and Aaron Courville*. Cambridge, Massachusetts: The MIT Press, 2016, 0–5.
73. Gevrey M, Dimopoulos I, Lek S. Review and comparison of methods to study the contribution of variables in artificial neural network models. *Ecol Model* 2003;**160**:249–64.
74. Riedmiller M. Advanced supervised learning in multi-layer perceptrons—from backpropagation to adaptive learning algorithms. *Comput Standards Interf* 1994;**16**:265–78.
75. Garson DG. Interpreting neural network connection weights. *Artif Intell Exp* 1991;**6**:46–51.
76. Filzmoser P, Gregorich M. Multivariate outlier detection in applied data analysis: global, local, compositional and Cellwise outliers. *Math Geosci* 2020;**52**:1049–66.
77. Storey JD. A direct approach to false discovery rates. *J R Stat Soc Series B Stat Methodology* 2002;**64**:479–98.
78. Kaler AS, Purcell LC. Estimation of a significance threshold for genome-wide association studies. *BMC Genomics* 2019;**20**:1–8.
79. Li H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetic parameter estimation from sequencing data. *Bioinformatics* 2011;**27**:2987–93.