Deciphering signatures of natural selection via deep learning

Xinghu Qin[1*], Charleston W. K. Chiang[2], Oscar E. Gaggiotti[1*]

[1] Centre for Biological Diversity, Sir Harold Mitchell Building, University of St Andrews, Fife, KY16 9TF, UK

[2] Center for Genetic Epidemiology, Keck School of Medicine & Department of Quantitative and Computational Biology, University of Southern California, USA

* Correspondence should be addressed to XQ (xq5@st-andrews.ac.uk) and OE Gaggiotti (oeg@st-andrews.ac.uk)

Abstract

Identifying genomic regions influenced by natural selection provides fundamental insights into a wide range of problems including human health, animal and plant breeding, and the understanding of local adaptation. We propose a new method, DeepGenomeScan, that can be used to address all these problems. It is based on the principle that the genotypes of individuals can be used to predict any associated trait; not only their phenotype but also their spatial location or the environmental attributes of the habitat they live in. We, therefore, implemented a deep learning method to detect candidate regions under selection by identifying loci that contribute the most to the predictive power of the deep neural network. Using simulations, we show that our method can successfully identify loci underlying quantitative traits subject to complex spatial patterns of selection. We apply DeepGenomeScan to a European human genetic variation dataset and posit that the loci that contribute the most to the prediction of latitude and longitude are located in genomic regions under selection. Using this approach, we identified many SNPs located within well-known genes, some of which were not identified using existing population genetics approaches, e.g. MCM6, MGAT5, TMEM163.

34 **Main**

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

40 The recent increases in genomic data generated by modern sequencing technologies have
41 advanced our understanding of how natural selection, and its interactions with other
42 evolutionary forces, shapes the genome and phenotype of species. Such technologies have
43 now been applied to the study of a wide range of species, but the statistical methods used to
44 analyse them tend to differ depending on whether the focus is on humans and domesticated
45 species or on wild species. In the case of humans and domesticated species, there is a
46 particular interest in uncovering associations between well-defined phenotypes (traits
47 associated with diseases or, in the particular case of domesticated species, yield) and the
48 genomic variants that underlie them. The standard approach, in this case, is to use Genome
49 Wide Association Studies (GWAS; e.g. ref [1]), which linearly model the additive allelic effect
50 of genotypes (the explicative variable) on phenotypes (the dependent variable). All of the
51 recent methods can take into account population stratification, using for example linear
52 mixed models that incorporate the genetic relationship matrix (GRM) as a random effect (e.g.
53 ref. [2]). GWAS have been used to identify variants associated with a wide range of phenotypic
54 traits[3] and their results are used by several studies aimed at detecting polygenic adaptation[4-8].

55 In the case of evolutionary studies and in particular those focused on wild species, the most
56 commonly used approaches are a suite of methods known as the genome-scans [9], which
57 ignore phenotypic traits. Many of them are population-based and use summary statistic to

58 uncover genomic regions exhibiting outlier behaviour. A very popular approach is to identify
59 loci that exhibit unusually strong among-population genetic differentiation. This approach
60 assumes that, in the absence of selection, differences in allele frequencies among populations
61 should follow a neutral distribution driven by migration and genetic drift. Any locus that
62 departs from this expectation is considered a good candidate for being under the effect of
63 natural selection. This approach initially proposed by Lewontin & Krakauer [10] has been
64 implemented in popular Bayesian methods (e.g. [11,12]). More recently, there has been an
65 increased interest in establishing an association between a locus outlier behaviour and
66 environmental variables, with the assumption that such variables could represent selective
67 pressures acting upon genomic regions linked to the outlier loci[13-16]. Some of these methods
68 use raw genotypic data instead of summary statistics (e.g. [12]) but as opposed to GWAS, they
69 use linear mixed models that consider the genotype as the dependent variable, and the
70 environmental factor as the explicative variable.

71 Despite this apparent methodological dichotomy between studies focused on humans and
72 domesticated species and those focused on wild species, it is clear that a thorough
73 understanding of how natural selection shapes the phenotype and genome of species requires
74 some combination of these two different approaches. Indeed, the selective pressures exerted
75 by the heterogenous abiotic and biotic environment act upon individuals' phenotypes in a
76 complex way and lead to changes in the spatial structuring of variation in the genomic
77 regions underlying them. Thus, in order to fully characterise the action of natural selection,
78 we need to answer three fundamental questions: (i) what are the environmental drivers of
79 natural selection? (ii) what are the phenotypic traits upon which selective pressures act? and
80 (iii) what are the genomic regions underlying those adaptive traits?

81 Here we present DeepGenomeScan, a deep learning approach that can be used to answer all
82 three questions. The rationale underlying our method is that we can use the genotype of an
83 individual to predict not only its phenotype but also some attribute of the habitat where they
84 are sampled. This framework is akin to GWAS but in our case, the response variable can be
85 an individual phenotypic trait, an environmental variable associated with its habitat, or the
86 geographic location of the individual (latitude, longitude). Our approach leverages the power
87 of deep neural networks to approximate arbitrarily complex functions linking dependent and
88 explicative variables [17], as well as recent algorithms for model optimisation [18], and for
89 inferring the importance of each explicative variable in predicting the dependent variable [19].
90 It is important to note that as opposed to the prevalent use of neural networks, i.e. prediction
91 and pattern recognition, here our main interest is in estimating the features (loci) that
92 contribute the most to the predictive power of the neural network. To achieve this goal, we
93 use the concept of "feature importance"[19], which represents a proxy for the effect size of any
94 given locus.

95 An important advantage of our method when applied to spatial data is that it can consider any
96 spatial selection pattern including the usually assumed linear environmental gradient as well
97 as arbitrarily complex non-linear spatial patterns and, importantly, coarse-grained
98 heterogeneous selection with no clear spatial pattern. This represents an important advance as
99 existing methods only consider linear or monotonic non-linear patterns. In this particular
100 application, our approach generalises Yang et al.'s[20] idea of using geographic positioning of
101 individuals to identify loci that exhibit particularly steep slopes of allele frequency change
102 associated with recent positive selection. Our generalisation is two-fold: (i) instead of only
103 considering monotonic spatial gradients, we can detect loci associated with arbitrary and non-
104 monotonic selection patterns, and (ii) the dependent variable can be the geographic location
105 but also a phenotypic trait or an environmental factor.

Results

Underlying Rationale for the Deep Learning Approach

A neural network can be considered as a generalised regression approach used to learn complex functions expressing the association between the input data and a response variable[20]. In our case, the input or predictor variables are the genotypes of the individuals and the response variables are observed phenotypes, environmental values, or geographic locations. In what follows, we use the term traits to refer to these response variables.

As opposed to standard regression, which fits a fixed function to the data, a neural network learns the (non-linear) function most appropriate for the data at hand. Here 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 Notes). 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[21] by minimising 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[22,23] (node or locus). Loci with extreme importance values are considered as outliers and good candidates for being under the influence of natural selection. Full details of our approach are provided in Online Methods.

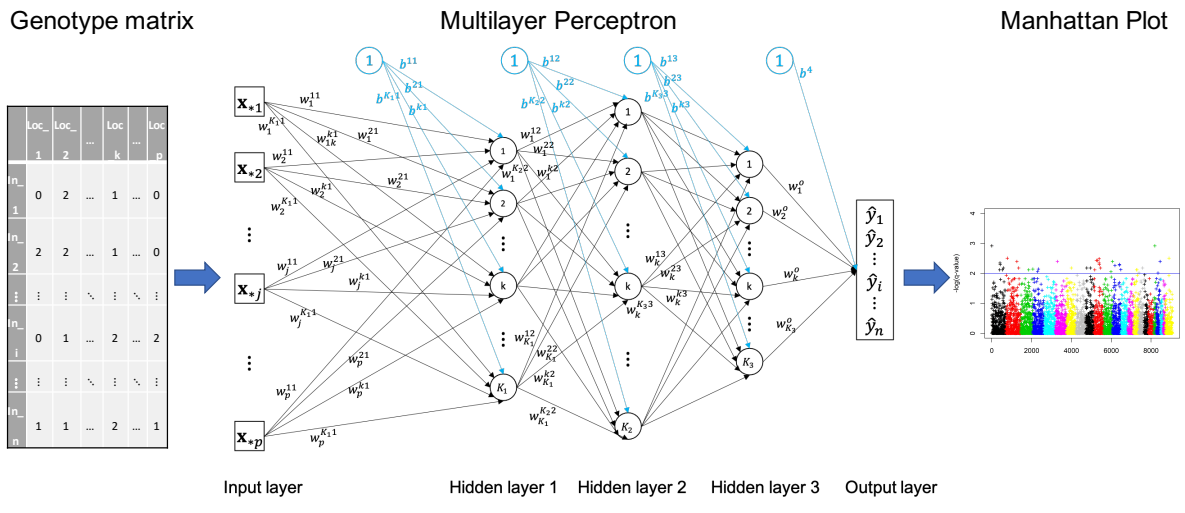

Figure 1: DeepGenomeScan framework. The first (input) layer receives the genotype matrix so that each of its input nodes contains the genotypes of all sampled individuals at a single locus. The last (output) layer contains the predicted trait values. Between these two layers there is one or more hidden layers containing nodes that compute a non-linear 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 q-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.

Simulation Study

We evaluated the performance of our method using simulated data generated by Capblancq et al.[24] (downloaded 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, 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 quantitative trait (QT) is influenced by environmental factors 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. Further details of the simulation approach are provided in Supplementary. Full details are given in ref. [24].

150 We compared the performance of our method with those of two recently proposed machine
151 learning-based approaches: pcadapt [25] and RDA[24]. Figure 2 presents these results in terms of
152 power (proportion of true positives), False Discovery Rate (FDR) and False Positive Rate
153 (FPR). Using a threshold p-value = 10^{-8}, as typically done in GWAS, our method has high
154 power to detect all QTL types, but the FDR is too high. On the other hand, pcadapt and RDA
155 have no power to detect loci associated with QTs 1 and 3, and very low power to detect those
156 associated with QT2 (Fig. 2a). We then progressively increased the stringency of the
157 threshold for DeepGenomScan (from p-value 10^{-16} up to 10^{-64}) and used a less stringent
158 threshold (p-value = 0.001) for pcadapt and RDA (Fig. 2d-l). It is possible to decrease the
159 FDR of our method below that of pcadapt using a threshold of 10^{-32}, and below that of RDA
160 using a threshold of 10^{-64} (Fig. 2h, i, k, l). Overall, all three methods have high power to
161 detect loci associated with QT2, which is influenced by a linear selective gradient.
162 DeepGenomeScan, however, has the lowest FDR under the most stringent threshold, and still
163 has much higher power than pcadapt and RDA to detect QT1 and QT3 loci, which are
164 influenced by non-linear selection patterns (Fig. 2j). It is important to note that this trade-off
165 between power and FDR only applies to cases with non-linear selection patterns. Indeed, the
166 power to detect loci underlying QT2 remains very high regardless of how stringent the
167 threshold we use is (Fig. 2). Qualitatively similar results were obtained when we used q-value
168 thresholds (Supplementary Notes).

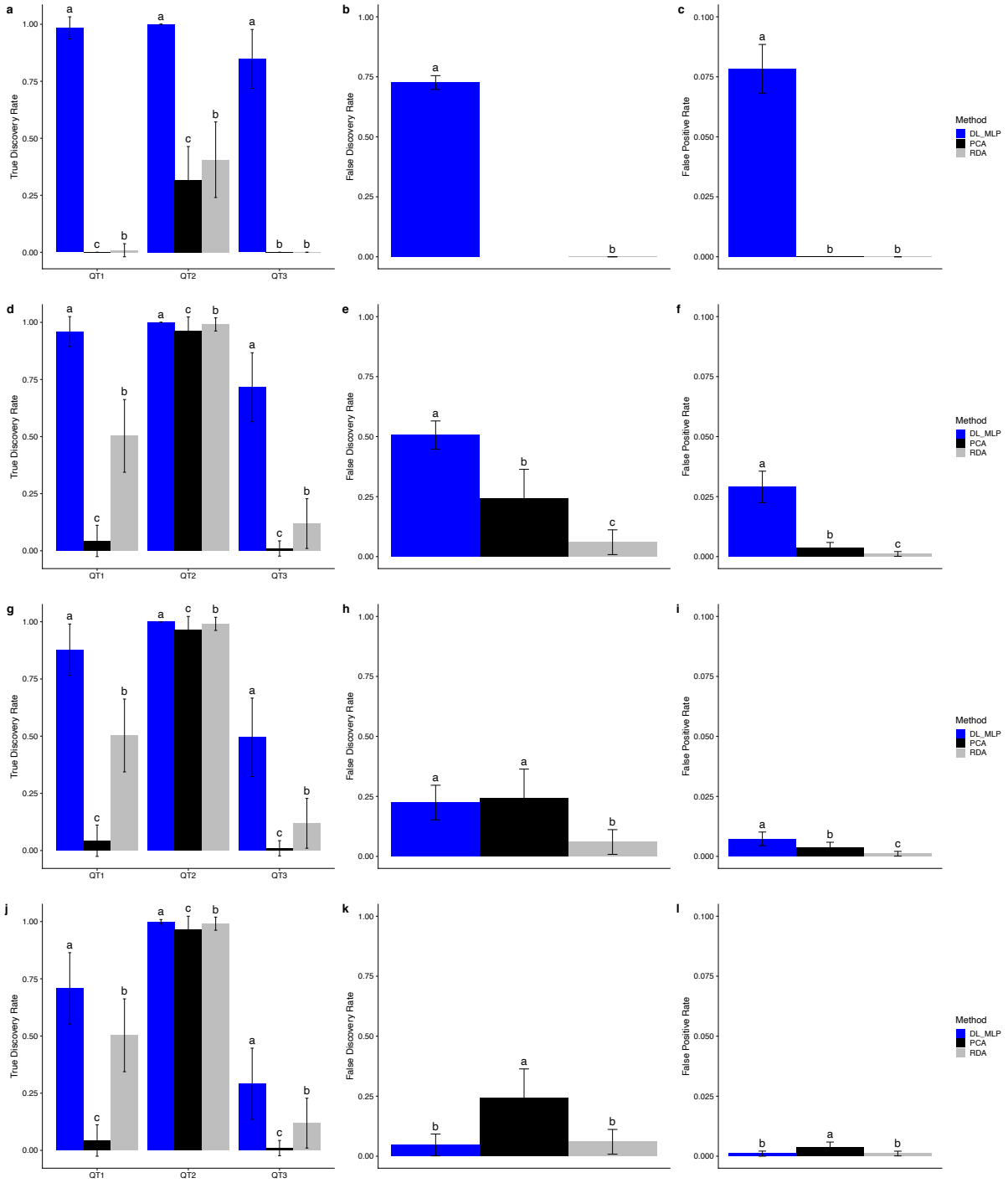

Figure 2: Performance of pcadapt, RDA and DeepGenomeScan (DL_MLP). (a-c): Initially, the p-value threshold used for all three methods was 10^{-8}. Subsequently, we used a threshold of 0.001 for pcadapt and RDA but we progressively increased the stringency of the threshold for DeepGenomeScan: (d-f): 10^{-16}; (g-i) 10^{-32}; (j-l) 10^{-64}. 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-value < 0.01. Panels a, d, g, j 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.

Figure 3 presents Manhattan plots for one simulated data set with a somewhat permissive threshold for pcadapt and RDA and a much more stringent threshold for DeepGenomeScan.

This figure illustrates the much higher power of DeepGenomeScan to detect outlier loci. Indeed, in the case of the linear selection gradient, all the outlier loci fall off the y-axis scale and only two are missed under quadratic selection gradients. Clearly, very discontinuous selection patterns (i.e., QTL3) represent a big challenge for all methods but DeepGenomeScan could still detect several of the outliers under this difficult scenario.

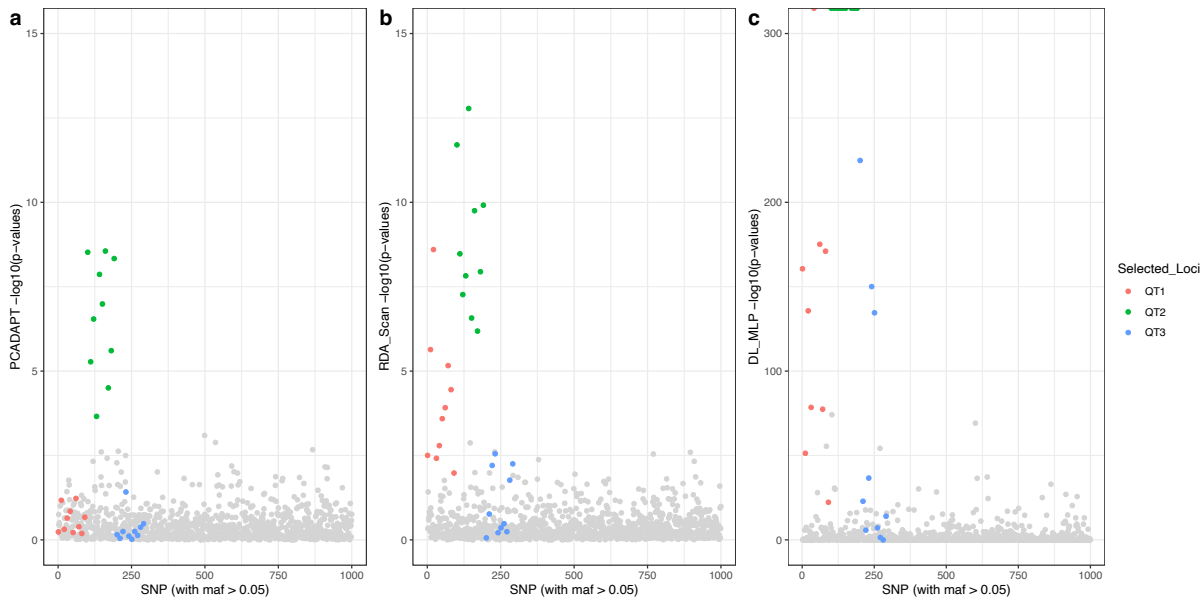

Figure 3: Manhattan plots of the results obtained with (a) pcadapt, (b) RDA and (c) DeepGenomeScan (DL_MLP) for one simulated data set. Note that in c, all QT2 loci and some QT1 loci fall out of the range of the y axis. Note also that the scale of the y axis in c differs from that used for a and b.

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 [26]. One approach to detect genomic regions under selection in these clinal variation scenarios is to identify loci exhibiting extreme frequency gradients across geographic space [20]. 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.

196 Although there may be several examples of such geographic variation, spatial patterns in
197 selective pressures and the associated allele frequency can be non-monotonic. For example, it
198 is possible that the maximum allele frequency is located in the middle of the geographic
199 region under study [20]. As our simulations show, our approach is capable of identifying
200 genomic regions subject to such non-linear selection patterns.

201 In principle, approaches focused on clinal spatial patterns require the geographic coordinates
202 of each individual sample. However, sometimes this information is not available. A potential
203 solution to this problem is to carry out a Principal Component Analysis on the genotype
204 matrix and use the first two PC axes as surrogates for geographic coordinates [27,28]. This
205 approach, however, is based on a linear combination of genotypes and can lead to poor
206 inference of spatial location of admixed individuals [20]. A recently developed model-based
207 approach can be used to overcome this issue [20] but it requires the assumption of a smooth
208 monotonic function to describe allele frequency behaviour as a function of geographic
209 location, which may not be appropriate in all cases. Our implementation of
210 DeepGenomeScan allows the use of a more general dimensionality reduction approach,
211 Kernel Local Discriminant Analysis of Principal Components (KLFDAPC[29]; Supplementary
212 Material) when spatial coordinates are not available.

213 Here we use a human dataset to determine if our approach can uncover new regions of the
214 human genome that may be under the influence of non-linear spatial selection patterns. We
215 applied DeepGenomeScan to the Population Reference Sample [30] (POPRES) dataset using
216 the known geographic coordinates and then using the two first reduced features from a
217 KLFDAPC analysis (Online Methods). POPRES represents an excellent example of human
218 genetic variation aligning with geography [28]. The dataset contains a total of 3,192 European

219 individuals genotyped at 500,568 loci using the Affymetrix 500K SNP chip. Details about
220 data filtering steps are provided in Online Materials.

221 We calculated p-values for each SNP (see Online Methods for details) and obtained a
222 genomic inflation factor GIF = 1 for this data set. Using a threshold p-value = 1e-100 for
223 identifying the outlier loci, the first analysis using geographic coordinates as the response
224 variables detected 245 outlier SNPs located within 84 known genes. The full list of genes we
225 identified, and their chromosome positions is provided in Supplementary Table S1.
226 Consistent with previous widely reported regions under selection, our method detects strong
227 signals at the LCT region on chromosome 2, the ADH1C region on chromosome 4, the HLA
228 region on chromosome 6, as well as the OCA2 and the HERC2 region on chromosome 15
229 (Fig. 4; Supplementary Table S1).

230 Besides the already known selected genes, our method also detects some disease related
231 genes that exhibited extreme variation across geographic space but were not identified by
232 existing methods (c.f. Supplementary Table 4 in ref. [20]); these include MGAT5, TMEM163,
233 ACMSD, CCNT2, MAP3K19, R3HDM1, UBXN4, MCM6, DARS1, EHMT2, and CFB
234 (Supplementary Table 1). For example, MCM6 is a regulatory element that controls the
235 expression of the LCT gene. The MGAT5 gene (alpha-1,6-mannosylglycoprotein 6-beta-N-
236 acetylglucosaminyltransferase), is one of the most important enzymes involved in the
237 regulation of the biosynthesis of glycoprotein oligosaccharides and is associated with
238 invasive malignancies and sclerosis [31-33] as well as visceral fat in women [34]. The TMEM163
239 gene (transmembrane protein 163), is associated with Parkinson's disease, ischemic stroke
240 and coronary artery disease [35,36]. The ACMSD gene (aminocarboxymuconate semialdehyde
241 decarboxylase) is associated with Parkinson's disease [35] and childhood obesity [37]. There are
242 many other genes with important functions that our method highlighted as contributing very

strongly to the estimation of geographic origin (Supplementary Table S1). There are also several other outlier SNPs detected by our method on chromosome 2, 4, 6, 8, 10, 11, 12, 15, 16, 18 (Supplementary Table 1) that are not within known functional genes. However, they may be linked to regulatory genes and, therefore, of interest for human genetic studies.

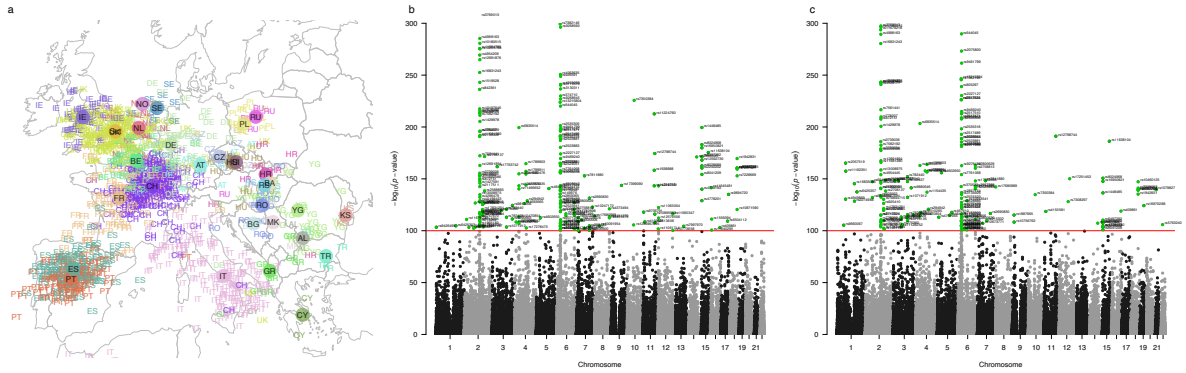

Figure 4: Spatial genetic structure of European populations and signals of selection detected by DeepGenomeScan. (A). Spatial genetic structure of European populations inferred via KLFDAPC with a σ = 5. (B). Manhattan plot indicating loci under spatial selection obtained from DeepGenomeScan using geographic 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 KLFDAPC) as the response variables. The outliers in B-C were identified with a threshold of p-value =1e-100. p-value lower than the threshold are highlighted in light green. 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. Note that in B and C, some of the outlier loci on Chromosome 2 and 6 fall out of the range of the y axis. A full list of outlier loci can be found in Supplementary Data Table 1-2.

In the case of the analysis for which we replaced the original geographic coordinates with the first two reduced features obtained from KLFDAPC, 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 KLFDAPC provides better estimates of geographic location than PCA (see Fig. 4a and Supplementary Notes Fig. S6).

The genome scan based on the first two reduced features identifies a somewhat smaller number of outlier SNPs (229, in 77 known genes) than when using geographic coordinates (Fig. 4c and Supplementary Table S2). However, 80% of the outliers detected in this analysis

269 were also identified by the analysis based on geographic coordinates (185 shared loci). This
270 analysis also detected well-known selected genomic regions (LCT, HLA, ADH1, CHERC2;
271 Supplementary Table S2). However, there were 28 genes that were not identified when using
272 latitude and longitude (Supplementary Table S3). Genes in this group include some
273 associated with cancer (SLC44A5, AK5, TNXB) and other diseases such as Behcet's disease
274 (IL12AS1-AS1) as well as obesity and diabetes mellitus (APOM). On the other hand, there
275 are also 29 genes that were not identified in this analysis but were highlighted by the analysis
276 based on geographic coordinates. This group includes genes associated with cognitive
277 disorders (e.g. TMEM132D), psychiatric disorders (e.g. TRIM26) but also five others that
278 have unknown function or associated phenotype (Supplementary Table S4).

279

280 **Discussion**

281 In this study, we present DeepGenomeScan, a new deep learning method to study how
282 natural selection shapes the phenotypes and genomes of species. Our approach represents an
283 integration of genome wide association studies and genome scan approaches and can be used
284 to address three fundamental evolutionary questions: (i) what are the environmental drivers
285 of natural selection? (ii) what are the phenotypic traits upon which selective pressures act?
286 and (iii) what are the genomic regions underlying those adaptive traits.

287 The insight upon which our method relies is the idea that we can use the genotypes of
288 individuals to predict any associated trait, not limited to just their phenotype but also their
289 spatial location or the environmental attributes of the habitat they live in. Intuitively, the type
290 of functions that can be used to describe the association between genotypes and any of these
291 dependent variables is likely to be radically different depending on the "trait" under

292 consideration. Therefore, no single model-based approach can be used to take advantage of
293 the above-mentioned insight. Additionally, no single model-free statistical method can be
294 used to approximate very complex non-linear functions linking genotypes and these disparate
295 traits. In contrast, Deep Neural Networks can approximate arbitrarily complex non-linear
296 functions linking dependent and predictor variables[17] and our study showed that they can be
297 an invaluable tool for population and evolutionary genomics applications.

298 The results of our analyses of an existing synthetic dataset[24] generated by simulations of a
299 polygenic trait under various spatial patterns of selection showed that DeepGenomeScan can
300 successfully identify outlier loci exhibiting complex non-linear associations with the target
301 trait. This represents an important advance over existing methods that can only consider
302 monotonic associations. For example, SPA uses a logistic function to describe allele
303 frequency behaviour as a function of geographic location and can only detect outlier loci
304 exhibiting monotonic geographic gradients [20]. Moreover, it cannot be applied to study
305 associations with phenotypic or environmental data.

306 Evolutionary biologists interested in understanding the mechanisms driving local adaptation
307 and speciation follow a longstanding tradition focused on the analysis of variation in
308 phenotype or allele frequencies along linear environmental gradients [26,38]. Here we showed
309 that recent machine learning methods such as pcadapt and RDA can successfully identify
310 genomic regions associated with those patterns (Figs. 2 & 3). However, their performance is
311 limited in the case of non-monotonic gradients and, moreover, they are completely incapable
312 of considering coarse-grained heterogeneous spatial patterns in allele frequencies (Fig 2 & 3).
313 Our method clearly outperforms pcadapt and RDA in uncovering the outlier loci that
314 contribute the most to spatial population structure, regardless of the underlying spatial pattern
315 in selective pressures. Nevertheless, it is clear that very coarse-grained heterogenous

316 selection with no clear spatial pattern represents a very difficult scenario, which reduces the
317 power of our method to detect genomic regions associated with the response of a species to
318 selection.

319 Although less well studied empirically, coarse-grained heterogenous selection with no clear
320 spatial pattern has been an important focus of theoretical studies aimed at explaining
321 observed patterns of genetic variation [39-44]. This scenario is particularly relevant to studies of
322 protected polymorphisms [40,42,44] and hard versus soft selection [41]. Additionally, the
323 importance of this type of selection pattern is particularly relevant for genome-wide
324 association studies focused on diseases and phenotypic characters that do not exhibit clear
325 spatial patterns. Deep learning genome-scan methods capable of uncovering genomic regions
326 associated with this type of selection pattern can, therefore, provide a more general
327 understanding of how prevalent these particular mechanisms are.

328 We also applied DeepGenomeScan to the Population Reference Sample [30] (POPRES) dataset
329 to determine if our method could uncover genomic regions undetected by SPA and other
330 popular methods such iHS [45], Fst [10], Bayenv [13]. Comparing Yang et al.'s [20] Supplementary
331 Table 4 with our Supplementary Table 1, reveals a large number of important genes that were
332 not detected by those other existing methods. Some of these genes have been linked to
333 complex diseases such as Parkinson, obesity or various types of cancer and may not be
334 associated with monotonic selection gradients.

335 DeepGenomeScan can be applied to spatial genetic data even when individuals' geographic
336 coordinates are not available. In this case, our method uses the first two reduced features
337 obtained from a very recent machine learning method (KLFDAPC[29]), as proxies for latitude
338 and longitude. A comparison of the results from analyses of the POPRES dataset based on
339 geographic coordinates with those based on the first two reduced features reveals very

340 extensive overlap in terms of the genes they uncover. The small discrepancy in genes
341 uncovered by the two options implemented in our software is due to differences in the type of
342 spatial information used. It is clear that reduced features could lead to a loss of detailed
343 spatial information provided by geographic location. On the other hand, the use of reduced
344 features could avoid biases introduced by low quality geographic information or could take
345 into account small scale spatial effects that are not detectable at the spatial scale used for the
346 geographic coordinates.

347 There are still several methodological challenges faced when implementing deep learning
348 methods, which are associated with the computational cost of hyperparameter tuning and
349 neural network training. However, very recent adaptive resampling algorithms (c.f. ref. [18]),
350 which we implemented in our software, carry out an efficient exploration of hyperparameter
351 space, allowing the use of deep learning to analyse large population genomic data sets.

352 Although the application of deep learning to population genetics problems is still in its
353 infancy [18], there are already some examples of such applications [21,46,47]. One of these
354 applications is aimed at distinguishing between hard and soft sweeps and simultaneously
355 incorporating the confounding effects of demographic history [21]. This application involves
356 choosing among two or more alternative models, an objective that is usually achieved using
357 Approximate Bayesian Computation (ABC) [48,49]. Although model choice in ABC is akin to a
358 classification application in machine learning, neural networks require a training set that is
359 not readily available when the target is a demographic or selection model. To overcome this
360 limitation, the authors train the neural network using synthetic data generated under
361 predefined evolutionary models. A drawback of this strategy is that it introduces model
362 assumptions into a computational framework that could be completely free of them and,
363 therefore, of very wide applicability. Additionally, instead of using the raw genotype matrix,

364 this framework requires the pre-processing of genetic data to generate summary statistics, as
365 input data, leading to a loss of information. These drawbacks are absent when the objective is
366 to assign individuals to geographic locations as in ref. [46] or when scanning the genome in
367 search of outlier loci, as in our study. Overall, DeepGenomeScan fully exploits the flexibility
368 and power of deep neural networks, which makes it applicable to a wide range of problems
369 including identification of genomic regions associated with diseases or economically
370 important phenotypic traits, assignment of individuals to geographic locations, or
371 identification of environmental factors associated with selective pressures.

372

373 **Methods**

374 MLP Architecture

375 In this study, we constructed a Multilayer Perceptron Network with three hidden layers
376 (Figure 1 and Supplementary Notes Figure S1). To describe this MLP, we first assume that
377 the genotypes of the n sampled individuals at p loci are described by a n×p matrix, $\mathbf{x} = (x_{ij})$,
378 and are coded by the number of derived alleles, which in the case of a biallelic marker takes
379 values $x_{ij} = 0, 1, 2$, where i = 1, 2, ..., n, and j=1, 2, ..., p. Furthermore, the trait values of all
380 individuals are arranged in a n×1 vector $\mathbf{y} = (y_i)$. Then the input layer of the MLP consists
381 of p nodes, each one representing a distinct locus. The information contained in node j
382 comprises the individual genotypes at locus j, $\mathbf{x}_{*j} = (x_{1j}, \cdots, x_{nj})$. This information is
383 transformed into a signal which is fed to the first hidden layer. The signal received by each
384 node k = 1, 2, …, K_1 of the first hidden layer $\hat{\mathbf{s}}^{k1} = (\hat{s}_1^{k1}, \cdots, \hat{s}_i^{k1}, \cdots, \hat{s}_n^{k1})$ represents a
385 weighted non-linear regression of individuals' trait values on their multilocus genotypes.

Here, the superscript k1 identifies the kth-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

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

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:

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

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,

$$f(\mathbf{x}) = f^L\left(f^{L-1}\left(\cdots f^2\left(f^1(\mathbf{x})\right)\right)\right), \qquad (3)$$

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

Given the input data, i.e. the genotypes, and the observed trait values, the neural network learns the weights $\mathbf{w} = (\mathbf{w}^1, \ldots, \mathbf{w}^L)$ with $\mathbf{w}^l = \left(w^{1l}, \cdots, w^{K_l l}\right)$ and biases $\mathbf{b} = (\mathbf{b}^1, \ldots, \mathbf{b}^L)$ with $\mathbf{b}^l = \left(b^{1l}, \cdots, b^{K_l l}\right)$ that best describe the relationship between the inputs and the outputs by minimising the difference between predicted and observed trait values [50,51].

406 These weights represent an essential element of our genome scan method because they are
407 used to identify outlier loci (see below) using a procedure supported by previous studies [23,52],
408 which show that the importance of a variable (locus) can be estimated as the combination of
409 the absolute values of weights associated with the graph edges connecting the predictor
410 variable (locus) with the MLP output (predicted trait).

411

412 Implementation

413 DeepGenomeScan trains the deep neural network to estimate the nonlinear function mapping
414 genotypes to individual traits. This process involves not only learning the vectors of weights
415 and biases by iteratively adjusting these parameters but also the tuning of a large number of
416 hyperparameters associated with structural components of the model, which impact model
417 performance. These include those determined by the network architecture (activation
418 functions, number of neurons per hidden layer) and those included in the optimisation
419 algorithm used for training (e.g. learning rate, batch size). For optimisation, we used resilient
420 backpropagation [53] with weight backtracking when analysing both simulated and real
421 datasets. However, the tuning algorithm used depended on the size of the input data. In the
422 case of simulated datasets, which were of limited size, we used a full resampling strategy
423 corresponding to Algorithm 1 in ref. [18] based on a repeated k-fold cross validation with k = 5
424 and five repetitions. In the case of the large POPRES dataset, we used an adaptive resampling
425 approach corresponding to Algorithm 2 in ref. [18], which incorporates futility assessment into
426 the model tuning process. The resampling approach in this case was the same as in the
427 analysis of simulated data (five replications of a five-fold cross-validation). Detailed
428 information about the settings used for the optimisation and tuning algorithms is presented in
429 Supplementary Notes.

430

431 Identification of outlier loci

432 As mentioned before, the SNP importance can be estimated as combinations of the absolute
433 values of connection weights (**w**) [23,52]. This is done once the optimal model is found. We used
434 Olden and Jackson's [19] method, which is based on Garson's[54] algorithm to calculate the
435 relative importance for each input node but adds a randomisation step to identify non-
436 significant connection weights.

437 DeepGenomeScan carries out separate runs for each trait and generates a vector of SNP
438 importance for each of them. Given T traits, the position of each SNP in trait space is
439 described by its associated vector of importance values and, therefore, it is possible to
440 calculate the Mahalanobis distance between the focal locus and all other loci. This leads to an
441 intuitive definition of outlier loci as any locus with an "extreme" Mahalanobis distance. Since
442 the squared Mahalanobis distance follows a chi-squared distribution with T degrees of
443 freedom, the p-values associated with each SNP can be obtained from a χ^2_T distribution [55] and
444 used for identifying 'outlier' loci.

445

446 Human dataset

447 We applied DeepGenomeScan to European populations from the POPRES project [30]. The
448 dataset contains a total of 3,192 European individuals genotyped at 500,568 loci using the
449 Affymetrix 500K SNP chip. The sample collections and genotyping for POPRES are
450 described in [30]. We removed individuals from outside of Europe and individuals whose
451 grandparents had different geographic origins based on the criteria used by [28]. We also

452 removed samples that only have one individual per country, such as Denmark, Finland,
453 Latvia, Ukraine, Slovakia and Slovenia. Geographic coordinates for each individual
454 corresponded to the central point of the geographic area of the individual's country of birth.

455 We removed the monomorphic SNPs and filtered out autosomal biallelic SNPs with a minor
456 allele frequency (MAF) below 5%, and a missing rate of 5% or more. In the end, we kept 1,
457 382 individuals from 32 countries carrying 283, 499 autosomal SNPs. We annotated and
458 updated the variant reference ID based on the comprehensive report of short human
459 variations from the human variation database (dbSNP, GRCh37p13, b151, release 2018,
460 https://ftp.ncbi.nih.gov/snp/organisms/human_9606/VCF/) using bcftools [56].

461

462 Data and code availability

463 The simulation data used in this study is available at
464 https://datadryad.org/review?doi=doi:10.5061/dryad.1s7v5.

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

468 The DeepGenomeScan R code is hosted on GitHub at
469 https://github.com/xinghuq/DeepGenomeScan. The scripts used in this study are available at
470 https://github.com/xinghuq/DeepGenomeScan/tree/webpkg/DeepGenomeScan_simulation%2
471 BPOPRES. The KLFDAPC package is available at https://xinghuq.github.io/KLFDAPC/.

472

Acknowledgments

XQ was supported by a PhD scholarship from the China Scholarship Council.

Author 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.

Competing Interests

The authors declare no competing financial interests.

References

1. Bush, W.S. & Moore, J.H. Chapter 11: Genome-Wide Association Studies. *Plos Computational Biology* **8**(2012).

2. Kang, H.M. *et al.* Variance component model to account for sample structure in genome-wide association studies. *Nature Genetics* **42**, 348-U110 (2010).

3. Tam, V. *et al.* Benefits and limitations of genome-wide association studies. *Nature Reviews Genetics* **20**, 467-484 (2019).

4. Edge, M.D. & Coop, G. Reconstructing the History of Polygenic Scores Using Coalescent Trees. *Genetics* **211**, 235-262 (2019).

5. Field, Y. *et al.* Detection of human adaptation during the past 2000 years. *Science* **354**, 760-764 (2016).

6. Racimo, F., Berg, J.J. & Pickrell, J.K. Detecting Polygenic Adaptation in Admixture Graphs. *Genetics* **208**, 1565-1584 (2018).

7. Speidel, L., Forest, M., Shi, S.N. & Myers, S.R. A method for genome-wide genealogy estimation for thousands of samples. *Nature Genetics* **51**, 1321-+ (2019).

8. Turchin, M.C. *et al.* Evidence of widespread selection on standing variation in Europe at height-associated SNPs. *Nature Genetics* **44**, 1015-+ (2012).

9. Oleksyk, T.K., Smith, M.W. & O'Brien, S.J. Genome-wide scans for footprints of natural selection. *Philosophical Transactions of the Royal Society B-Biological Sciences* **365**, 185-205 (2010).

10. Lewontin, R.C. & Krakauer, J. Distribution of gene frequency as a test of the theory of the selective neutrality of polymorphisms. *Genetics* **74**, 175-95 (1973).

11. Beaumont, M.A. & Balding, D.J. Identifying adaptive genetic divergence among populations from genome scans. *Molecular Ecology* **13**, 969-980 (2004).

12. Foll, M. & Gaggiotti, O. A Genome-Scan Method to Identify Selected Loci Appropriate for Both Dominant and Codominant Markers: A Bayesian Perspective. *Genetics* **180**, 977-993 (2008).

13. Coop, G., Witonsky, D., Di Rienzo, A. & Pritchard, J.K. Using Environmental Correlations to Identify Loci Underlying Local Adaptation. *Genetics* **185**, 1411-1423 (2010).

14. de Villemereuil, P. & Gaggiotti, O.E. A new F-ST-based method to uncover local adaptation using environmental variables. *Methods in Ecology and Evolution* **6**, 1248-1258 (2015).

15. Frichot, E., Schoville, S.D., Bouchard, G. & François, O. Testing for associations between loci and environmental gradients using latent factor mixed models. *Molecular biology and evolution* **30**, 1687-1699 (2013).

16. Gaggiotti, O.E. *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* **63**, 2939-2951 (2009).

17. Hornik, K., Stinchcombe, M. & White, H. Multilayer feedforward networks are universal approximators. *Neural networks* **2**, 359-366 (1989).

18. Kuhn, M. Futility analysis in the cross-validation of machine learning models. *arXiv:1405.6974* (2014).

19. Olden, J.D. & Jackson, D.A. Illuminating the “black box”: a randomization approach for understanding variable contributions in artificial neural networks. *Ecological modelling* **154**, 135-150 (2002).

20. Yang, W.Y., Novembre, J., Eskin, E. & Halperin, E. A model-based approach for analysis of spatial structure in genetic data. *Nature Genetics* **44**, 725-U163 (2012).

21. Sheehan, S. & Song, Y.S. Deep Learning for Population Genetic Inference. *Plos Computational Biology* **12**(2016).

22. Gevrey, M., Dimopoulos, L. & Lek, S. Review and comparison of methods to study the contribution of variables in artificial neural network models. *Ecological Modelling* **160**, 249-264 (2003).

23. Olden, J.D., Joy, M.K. & Death, R.G. An accurate comparison of methods for quantifying variable importance in artificial neural networks using simulated data. *Ecological modelling* **178**, 389-397 (2004).

24. Capblancq, T., Luu, K., Blum, M.G. & Bazin, E. Evaluation of redundancy analysis to identify signatures of local adaptation. *Molecular Ecology Resources* **18**, 1223-1233 (2018).

25. Luu, K., Bazin, E. & Blum, M.G. pcadapt: an R package to perform genome scans for selection based on principal component analysis. *Molecular ecology resources* **17**, 67-77 (2017).

26. Endler, J.A. *Geographic variation, speciation and clines*, (Princeton University Press, Princeton, NJ, 1977).

27. Lao, O. *et al.* Correlation between genetic and geographic structure in Europe. *Current Biology* **18**, 1241-1248 (2008).

28. Novembre, J. *et al.* Genes mirror geography within Europe. *Nature* **456**, 98-101 (2008).

29. Qin, X., Chiang, C.W.K. & Gaggiotti, O.E. Kernel Local Fisher Discriminant Analysis of Principal Components (KLFDAPC) significantly improves the accuracy of predicting geographic origin of individuals. *bioRxiv, 2021.2005.2015.444294* (2021).

30. Nelson, M.R. *et al.* The Population Reference Sample, POPRES: a resource for population, disease, and pharmacological genetics research. *The American Journal of Human Genetics* **83**, 347-358 (2008).

31. Granovsky, M. *et al.* Suppression of tumor growth and metastasis in Mgat5-deficient mice. *Nature medicine* **6**, 306-312 (2000).

32. Brynedal, B. *et al.* MGAT5 alters the severity of multiple sclerosis. *Journal of neuroimmunology* **220**, 120-124 (2010).

33. Wang, R. *et al.* Hydrogen sulfide demonstrates promising antitumor efficacy in gastric carcinoma by targeting MGAT5. *Translational oncology* **11**, 900-910 (2018).

34. Fox, C.S. *et al.* Genome-wide association for abdominal subcutaneous and visceral adipose reveals a novel locus for visceral fat in women. *PLoS Genet* **8**, e1002695 (2012).

35. Nalls, M.A. *et al.* Large-scale meta-analysis of genome-wide association data identifies six new risk loci for Parkinson's disease. *Nature genetics* **46**, 989-993 (2014).

36. Dichgans, M. *et al.* Shared genetic susceptibility to ischemic stroke and coronary artery disease: a genome-wide analysis of common variants. *Stroke* **45**, 24-36 (2014).

37. Comuzzie, A.G. *et al.* Novel genetic loci identified for the pathophysiology of childhood obesity in the Hispanic population. *PloS one* **7**, e51954 (2012).

38. Haldane, J.B.S. THE THEORY OF A CLINE. *Journal of Genetics* **48**, 277-284 (1948).

39. Bulmer, M.G. Multiple niche polymorphism. *Amer. Natur.* **106**, 254-257 (1972).

40. Levene, H. Genetic equilibrium when more than one ecological niche is available. *Amer. Natur.* **87**, 331-333 (1953).

41. Levins, R. & MacArthur, R. The maintenance of genetic polymorphism in a spatially heterogeneous environment: variations on a theme by Howard Levene. *Americal Naturalist* **100**, 585-589 (1966).

42. Prout, T. SUFFICIENT CONDITIONS FOR MULTIPLE NICHE POLYMORPHISM. *American Naturalist* **102**, 493-& (1968).

43. Smith, J.M. GENETIC POLYMORPHISM IN A VARIED ENVIRONMENT. *American Naturalist* **104**, 487-& (1970).

44. Strobeck, C. Haploid selection withn alleles in m niches. *American Naturalist* **113**, 439-444 (1979).

45. Sabeti, P.C. *et al.* Detecting recent positive selection in the human genome from haplotype structure. *Nature* **419**, 832-837 (2002).

46. Battey, C.J., Ralph, P.L. & Kern, A.D. Predicting geographic location from genetic variation with deep neural networks. *Elife* **9**(2020).

47. Flagel, L., Brandvain, Y. & Schrider, D.R. The Unreasonable Effectiveness of Convolutional Neural Networks in Population Genetic Inference. *Molecular Biology and Evolution* **36**, 220-238 (2019).

48. Beaumont, M.A., Zhang, W.Y. & Balding, D.J. Approximate Bayesian computation in population genetics. *Genetics* **162**, 2025-2035 (2002).

49. Csillery, K., Blum, M.G.B., Gaggiotti, O.E. & Francois, O. Approximate Bayesian Computation (ABC) in practice. *Trends in Ecology & Evolution* **25**, 410-418 (2010).

50. Jinming, Y., Ahmadi, M., Jullien, G.A. & Miller, W.C. Model validation and determination for neural network activation function modeling. 548-551 (1998).

51. Goodfellow, I.a. Deep learning / Ian Goodfellow, Yoshua Bengio and Aaron Courville. 0-5 (The MIT Press, Cambridge, Massachusetts :, 2016).

52. Gevrey, M., Dimopoulos, I. & Lek, S. Review and comparison of methods to study the contribution of variables in artificial neural network models. *Ecological modelling* **160**, 249-264 (2003).

53. Riedmiller, M. Advanced supervised learning in multi-layer perceptrons—from backpropagation to adaptive learning algorithms. *Computer Standards & Interfaces* **16**, 265-278 (1994).

54. Garson, D.G. Interpreting neural network connection weights. *Artificial Intelligence Expert* **6**, 46-51 (1991).

55. Filzmoser, P. & Gregorich, M. Multivariate Outlier Detection in Applied Data Analysis: Global, Local, Compositional and Cellwise Outliers. *Mathematical Geosciences* **52**, 1049-1066 (2020).

56. Li, H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* **27**, 2987-2993 (2011).