Received: 22 July 2020 | Revised: 1 March 2021 | Accepted: 5 March 2021

DOI: 10.1111/1755-0998.13379

SPECIAL ISSUE

MOLECULAR ECOLOGY RESOURCES WILEY

Distinguishing between recent balancing selection and incomplete sweep using deep neural networks

Ulas Isildak[1] | Alessandro Stella[2] | Matteo Fumagalli[3]

[1]Department of Biological Sciences, Middle East Technical University, Ankara, Turkey

[2]Laboratory of Medical Genetics, Department of Biomedical Sciences and Human Oncology, Università degli Studi di Bari Aldo Moro, Bari, Italy

[3]Department of Life Sciences, Silwood Park Campus, Imperial College London, London, UK

Correspondence
Matteo Fumagalli, Department of Life Sciences, Silwood Park Campus, Imperial College London, London SL5 7PY, UK.
Email: m.fumagalli@imperial.ac.uk

Funding information
Leverhulme Trust, Grant/Award Number: RPG-2018-208

Abstract

Balancing selection is an important adaptive mechanism underpinning a wide range of phenotypes. Despite its relevance, the detection of recent balancing selection from genomic data is challenging as its signatures are qualitatively similar to those left by ongoing positive selection. In this study, we developed and implemented two deep neural networks and tested their performance to predict loci under recent selection, either due to balancing selection or incomplete sweep, from population genomic data. Specifically, we generated forward-in-time simulations to train and test an artificial neural network (ANN) and a convolutional neural network (CNN). ANN received as input multiple summary statistics calculated on the locus of interest, while CNN was applied directly on the matrix of haplotypes. We found that both architectures have high accuracy to identify loci under recent selection. CNN generally outperformed ANN to distinguish between signals of balancing selection and incomplete sweep and was less affected by incorrect training data. We deployed both trained networks on neutral genomic regions in European populations and demonstrated a lower false-positive rate for CNN than ANN. We finally deployed CNN within the *MEFV* gene region and identified several common variants predicted to be under incomplete sweep in a European population. Notably, two of these variants are functional changes and could modulate susceptibility to familial Mediterranean fever, possibly as a consequence of past adaptation to pathogens. In conclusion, deep neural networks were able to characterize signals of selection on intermediate frequency variants, an analysis currently inaccessible by commonly used strategies.

KEYWORDS
adaptation, genomics/proteomics, molecular evolution, natural selection and contemporary evolution, population genetics—empirical, population genetics—theoretical

1 | INTRODUCTION

Balancing selection is a selective process that generates and maintains genetic diversity within populations, as firstly proposed by (Dobzhansky, (1951)). Many diverse mechanisms of balancing selection have been described (Charlesworth, 2006). Overdominance (or heterozygote advantage) occurs when heterozygote individuals at one locus have higher fitness than homozygotes. In sexually antagonistic selection, different alleles at the same locus have opposite effects in the two sexes creating a balanced polymorphism

at the population level. In negative frequency-dependent selection, rare alleles have a fitness advantage. Finally, spatially and temporally varying selection creates a scenario where different alleles are advantageous in different environments.

Until 2006, the general consensus was that only few loci in the human genome have been targets of balancing selection (Asthana et al., 2005; Bubb et al., 2006). Since then, the availability of large-scale population genomics data and the development of ad hoc statistical tests contributed to the current view that balancing selection is a widespread adaptive mechanism underlying a broad spectrum of features in the genetic architecture of phenotypes (Key et al., (2014); Llaurens et al., 2017).

In humans, balancing selection is responsible for shaping the diversity of genes involved in the adaptive and innate immune response (Andrés et al., 2009); DeGiorgio et al., 2014; Ferrer-Admetlla et al., 2008; Meyer et al., 2006), metabolism (Fumagalli et al., 2019) and other processes (Bitarello et al., 2018). Notably, variants targeted by pathogen-driven balancing selection have been found to be associated with susceptibility to several autoimmune diseases (Fumagalli et al., 2011). Therefore, by elucidating the genomic signals of balancing selection we have the ability to identify common alleles with critical functional consequences. For instance, balancing selection has been hypothesized to maintain a common variant in an angiotensin-converting enzyme (Cagliani, Fumagalli, Riva, Pozzoli, Comi, et al., 2010) which has been recently associated with increased susceptibility to SARS-CoV-2 (Delanghe et al., 2020).

Several methods to identify targets of balancing selection have been proposed (Fijarczyk & Babik, 2015). Genomic signatures of balancing selection have been detected by testing for an excess of heterozygous genotypes (Fumagalli et al., 2009a), a local increase in genetic diversity (Cagliani, Fumagalli, Riva, Pozzoli, Fracassetti, et al., 2010) and polymorphisms (Soni et al., 2021), a shift in the site frequency spectrum towards common frequencies (Andrés et al., 2009; Bitarello et al., 2018; Siewert & Voight, 2017), a population genetic differentiation lower or higher than expected under neutral evolution (Cagliani et al., 2008), presence of trans-species polymorphism (Leffler et al., 2013; Teixeira et al., 2015), by explicitly modelling the patterns of polymorphisms and substitutions (Cheng & DeGiorgio, 2020; DeGiorgio et al., 2014), and by correlating allele frequencies with environmental variables (Fumagalli et al., 2009b).

The application of such methods to large-scale human population genomic data has enabled the characterization of targets of long-term balancing selection (i.e. selection that predates the time to the most recent common ancestor in a species) in humans and their association to several diseases (Cagliani et al., 2008; Siewert & Voight, 2017). Nevertheless, all these studies contributed little to the understanding of the role of balancing selection in recent human evolution, despite short-term or transient balancing selection being predicted to be a common phenomenon in nature (Sellis et al., 2011). Recent balancing selection leaves traces that are almost indistinguishable from those left by recent positive selection (Fijarczyk & Babik, 2015), with beneficial alleles segregating at intermediate frequency in contemporary genomes in both cases (Charlesworth, 2006). Additionally, even when signatures of balancing selection are identified, the underlying evolutionary mechanism (e.g. overdominance or negative frequency-dependent selection) is often unknown (Llaurens et al., 2017). As such, current methods have only limited power to identify and characterize signatures of recent balancing selection in the human genome.

A promising solution to address this issue is provided by supervised machine learning (ML) which has been recently introduced in population genetics and successfully applied for evolutionary inferences (Schrider & Kern, 2018). For instance, several ML methods have been proposed and successfully applied to population genetic data to predict and classify neutral and selective events on genomic loci (Kern & Schrider, 2018; Lin et al., 2011; Mughal & DeGiorgio, 2019; Pavlidis et al., 2010; Ronen et al., 2013; Schrider & Kern, 2016; Sugden et al., 2018). Deep learning is a class of ML algorithms based on artificial neural networks (ANNs) which comprise nodes in multiple layers connecting features (input) and responses (output) (Lecun et al., 2015). ANNs have the potential to be used in population genetics to estimate parameters from genomic data using multiple summary statistics as input (Sheehan & Song, 2016).

Notably, deep learning algorithms can effectively learn which features (i.e. measurable properties of the data) are sufficient for the prediction (Krizhevsky et al., 2012; Lecun et al., 2015). Despite deep learning in population genetics being in its infancy, several studies have already introduced the use of convolutional neural networks (CNNs) to full population genomic data with convolutional layers automatically extracting informative features (Chan et al., 2018; Flagel et al., 2019; Sanchez et al., 2020; Torada et al., 2019; Xue et al., 2021). A convolution layer is comprised of several weight matrices that slide across the input image and perform a matrix convolutional to produce image matrices (Jiuxiang et al., 2018; Lecun et al., 1998). Recent reviews provide more detailed information on convolutional neural networks in population genetic inference (Flagel et al., 2019; Sanchez et al., 2020).

In this study, we aimed at developing and implementing deep neural networks to predict loci at intermediate allele frequency (i.e. between 40% and 60%) under natural selection (Test 1). By doing so, our goal is also to distinguish between signals of incomplete sweep (e.g. ongoing positive selection) and signals of balancing selection (Test 2), either due to overdominance or negative frequency-dependent selection. As mentioned above, these two types of selection are different biologically but leave similar signatures in genomes, making their discernment particularly challenging. Specifically, we compared the predictive power between ANNs (i.e. based on summary statistics) and CNNs (i.e. based on full population genomic data) to perform such classification.

Finally, we deployed the trained deep neural networks on population genomic data to identify and characterize signals of natural selection acting on the *MEFV* gene. Mutations in the *MEFV* gene cause familial Mediterranean fever (FMF), an autoinflammatory disease with recurrent episodes of fever, abdominal, joint and chest pain, with gradual development of nephropathic amyloidosis (kidney failure) in some cases (Touitou, 2001). FMF is highly prevalent in

populations of Mediterranean origin (Touitou, 2001), and the 3' terminal region of the *MEFV* gene has been hypothesized to be under balancing selection due to overdominance in some European populations (Fumagalli et al., 2009a). Recently, causative mutations in the *MEFV* gene have been reported as target of recent positive selection in the Turkish population as they confer resistance to *Yersinia pestis* (Park et al., 2020). By applying our deep neural networks on a large sample size of genomic data, we sought to establish which type of natural selection has been acting on *MEFV* with regard to susceptibility to FMF.

2 | MATERIALS AND METHODS

2.1 | Simulations of population genomic data

We performed extensive simulations both to assess the predictive power of summary statistics and to train deep neural networks. We generated synthetic population genomic data using *SLiM 3.2*, a forward-in-time genetic simulation software (Haller & Messer, 2019). We simulated four different scenarios: neutrality (NE), incomplete sweep (IS), overdominance (OD) and negative frequency-dependent selection (FD). A locus under balancing selection (BS) was considered to be under either OD or FD. All simulations were conditioned on a previously proposed demographic model for European populations (Jouganous et al., 2017) with a mutation rate of 1.44*e*−8, a generation time of 29 years and a recombination rate sampled from a normal distribution with mean 1*e*−8 and standard deviation 1*e*−9. Further details on the simulation model employed are available in Table S1 (Gravel et al., 2011).

For simulating scenarios of natural selection, we generated loci of 50 kbp (base pairs) with the selected variant at the centre of the simulated sequence. We assumed a model of selection on a de novo mutation. For illustrative purposes of this study, the selected mutation was introduced in the European population at 21 different times, ranging from 40 k to 20 kya (Figure S1). We classified these times into three categories: recent (20 k to 26 kya), medium (27 k to 33 kya) and old (34 k to 40 kya) selection.

To mimic the effect of a selected variant at intermediate frequency, we conditioned the final (i.e. contemporary) allele frequencies to be between 40% and 60% in the sample. If the final frequency of the selected allele was not within this range, the simulation restarted at the generation where the selected variant was introduced. For each selection scenario and time of onset of selection, we chose selection coefficients and parameters which maximized the probability of the final allele frequency being between 40% and 60% (Table S2). At the end of the simulations, we sampled 198 chromosomes (i.e. haploid individuals) to match the sample size of CEU (Central European) individuals in the 1000 Genomes Project (1000 Genomes Project Consortium, 2015).

In the neutral scenario, no selected variant was introduced. Instead, we generated data with a neutral variant at the centre of the sequence with a frequency between 40% and 60%. To achieve this, we (i) simulated a larger region of 500 kbp under neutral evolution, (ii) sampled 198 chromosomes, (iii) identified a variant with a frequency between 40% and 60%, and (iv) trimmed the large region to obtain a 50 kbp locus (Figure S2).

2.2 | Calculation of summary statistics and genomic images

We processed the simulated genomic data to be received as input to deep neural networks (i.e. both ANN and CNN). For ANN, we summarized each genomic sequence as a vector of all potentially informative summary statistics. Additionally, we divided each simulated 50 kbp sequence into two sub-regions: (i) proximal to the selection site (20–30 kbp) and (ii) distal from the selected site (0–20 kbp +30–50 kbp) (Figure S3), similarly to previous studies (Peter et al., 2012; Sheehan & Song, 2016). For each region, we calculated 32 summary statistics. The main statistics are as follows: nucleotide diversity π (Nei & Li, 1979), Watterson's estimator θ (Watterson, 1975), Tajima's *D* (Tajima, 1993), linkage disequilibrium (LD) r^2 (Hill & Robertson, 1968), Kelly's Z_{ns} (Kelly, 1997), Fu and Li's F^* and D^* (Fu & Li, 1993), H1, H12, H123, H2/H1 (Garud et al., 2015), iHS (Voight et al., 2006), EHH (Sabeti et al., 2002), Zeng et al.'s *E* (Zeng et al., 2006), Fay and Wu's *H* (Fay & Wu, 2000), nS_L (Ferrer-Admetlla et al., 2014), *NCD1* (Bitarello et al., 2018), raggedness index (Harpending, 1994), observed and expected heterozygosity, haplotype diversity, number of unique haplotypes and number of singletons. Finally, we included some derivatives of these main statistics, such as mean, median and maximum values of mean pairwise distances calculated for all chromosome pairs in a simulation (Figure S3, Table S3). All summary statistics were calculated using *scikit-allel* library (https://github.com/cggh/scikit-allel) and then scaled using the *StandardScaler* function from *sklearn* library (Pedregosa et al., 2011). All scaled summary statistics were considered as input features to the ANN.

For CNN, we created images from the alignment of sampled haplotypes, similar to previous studies (Chan et al., 2018; Flagel et al., 2019; Torada et al., 2019)). In this data representation, each row of the image is a sampled haplotype (i.e. individual chromosome) and each column corresponds to a specific segregating site. The colour coding indicates if a variant is derived or ancestral, or any other polarization of alleles (e.g. major/minor, reference/alternate). To disentangle the effect of random sorting of sampled haplotypes (Torada et al., 2019), we reordered rows of images as follows: (i) sampled haplotypes are divided into two groups based on the presence or absence of the targeted allele, (ii) haplotypes within each of the two groups are sorted separately based on haplotype frequency, (iii) the two sorted groups are combined to obtain the final reordered image. Lastly, to take into account the different dimensions of simulated loci, we resized images into 128 × 128 pixels (Torada et al., 2019) using the *Image* module from *Pillow* package (https://pypi.org/project/Pillow).

2.3 | Implementation and training of neural networks

Both ANN and CNN models were implemented in Python using *Keras* library with *Tensorflow* backend (Chollet, 2015). ANN model comprises one input, three hidden and one output fully connected (i.e. dense) layers. Similar to a previous study (Sheehan & Song, 2016), the hidden layers consist of 20, 20 and 10 neurons, respectively, all with a Rectified Linear Units (ReLU) activation function. The output layer, which performs the binary classification, consists of a single neuron with a sigmoid (i.e. logistic) activation function. To control for overfitting, in addition to batch normalization, we used a dropout rate of 0.5 and L2 weight decay of 0.005 across all but the output layers. Models were optimized using the Adam optimizer with a batch size of 64 and a learning rate of 0.005 (Kingma & Ba, 2014; Ruder, 2017).

The CNN model consisted of three sets of 2D convolution layers, each followed by a batch normalization layer and ReLU activation layer. A max-pooling layer was also applied after the first two convolution layers. All convolutional layers consisting of 32 filters had a kernel size of 3 x 3, applied at stride 1. The size of the pooling layers was 2 x 2, which were applied at stride 2. The convolutional layers were followed by a flatten layer, which transforms a two-dimensional feature matrix into a vector. Finally, we used a fully connected layer consisting of 128 units that uses the flattened feature vector as an input, followed by an output layer. Again, we used ReLU activation function on the output from the fully connected layer and the sigmoid function for the output layer. We performed extensive hyper-parameter tuning on training data over 25 epochs to optimize values of learning rate (Figure S4), number of units per layer (Figure S5), L2 regularization (Figure S6), dropout rates (Figure S7), batch normalization (Figure S8), image reshaping (Figure S9), to maximize accuracy for predicting loci under incomplete sweep or balancing selection (Test 2). A complete list of all hyper-parameter values used in the CNN model is available in Table S4. Further, we performed data augmentation during the training of CNN models by randomly flipping images horizontally (Figure S10) using the *ImageDataGenerator* function from *Keras* (Chollet et al., 2015). Similarly, we performed hyper-parameter tuning for ANN on 40 k training samples over 25 epochs to optimize values of learning rate (Figure S11), dropout and L2 weight decay (Figure S12), scaling (Figure S13), architecture (Figure S14), to maximize accuracy for predicting loci under incomplete sweep or balancing selection (Test 2).

We performed 480,000 simulations in total for training all deep neural networks. Each single model employed 80,000 simulated data samples, 64,000 of them for training and the remaining 16,000 for validation. All models were trained for 50 epochs each. Testing was performed on approximately 16,000 data samples. We trained both ANN and CNN to perform two classification tasks: predict loci under natural selection *vs*. neutral evolution (Test 1) and predict loci under balancing selection *vs*. incomplete sweep (Test 2). The predictive power of ANN and CNN for each test was quantified with a confusion matrix, where each row represents the instances of true class and each column the corresponding number of predicted instances.

2.4 | Prediction of natural selection from genomic data

We deployed the trained networks on phased population genomic data from the 1000 Genomes Project for the CEU population (1000 Genomes Project Consortium, 2015). We filtered all non-biallelic positions and selected all variants with a frequency between 40% and 60% in CEU populations within the *MEFV* gene region. We retrieved 41 such variants and, for each one, generated a haplotype matrix (Torada et al., 2019) of 50 kbp surrounding the putative target variant. We calculated summary statistics (for ANN) and generated images (for CNN) for each variant by applying the same pipeline used for training the networks. Test 2 was performed only on variants predicted to be under selection for Test 1. Genomic annotations were obtained using the *EnsDb. Hsapiens.v75 package* in R (Rainer, 2017), and *Gviz* package was used for visualization (Hahne et al., 2016). We also employed the same procedure on data from 99 randomly sampled individuals of Tuscans in Italy (TSI) from 1000 Genomes Project (1000 Genomes Project Consortium, 2015).

We further deployed the trained networks on genomic regions hypothesized to be neutrally evolving. We extracted two putative neutral regions (chr16: 62,852,764–62,944,210 and chr16: 63,651,950–63,684,341) predicted by the NRE Tool (Arbiza et al., 2012) which was run with default parameters for a large region proximal to *MEFV* gene on chromosome 16. We identified a total of 42 biallelic variants with intermediate allele frequency and applied the same procedure aforementioned to predict signals of selection using both trained networks.

2.5 | Software availability

A Python package called *BaSe* (Balancing Selection) that implements deep neural networks (both ANN and CNN) for the detection of selection and for discerning between incomplete sweep and balancing selection is available at https://github.com/ulasisik/balancing-selection. Data visualizations were performed in R, using *ggplot2* (Wickham, 2016), *ggpubr* (Kassambara, 2020) and *pheatmap* (Kolde, 2018) libraries. All remaining analyses were performed in Python.

3 | RESULTS

3.1 | Summary statistics are not sufficient to discriminate between balancing selection and incomplete sweep

Our first aim was to test whether commonly used summary statistics were sufficient to discriminate between loci under neutrality and natural selection, the latter comprising both incomplete sweep and balancing selection (Test 1). We calculated a total of 64 different summary statistics and compared their distributions calculated on simulated loci under either neutrality or selection, with the targeted allele at intermediate frequency (between 40% and 60%) in the centre of the region

(Figure S15). Figure 1 (upper panel a) shows a subset of these comparisons and indicates that the distribution of several summary statistics under neutral evolution or natural selection is statistically different. Therefore, these summary statistics can be used to predict loci under natural selection. This effect is particularly notable for haplotype-based summary statistics (Figure 1, upper left panel a), and it is consistent across all times of onset of selection (recent, medium and old), in line with the effect of recent selection on patterns of LD.

Next, we tested whether summary statistics were able to distinguish between loci under incomplete sweep and balancing selection (Test 2), and, again, we compared their distributions (Figure S16). Figure 1 (lower panel b) shows the same subset of comparisons. These results suggest that only few summary statistics can discern genomic patterns created by incomplete sweep from those created by balancing selection, and only marginally. This deficiency is particularly severe for allele frequency-based summary statistics and for medium to old times of selection onset.

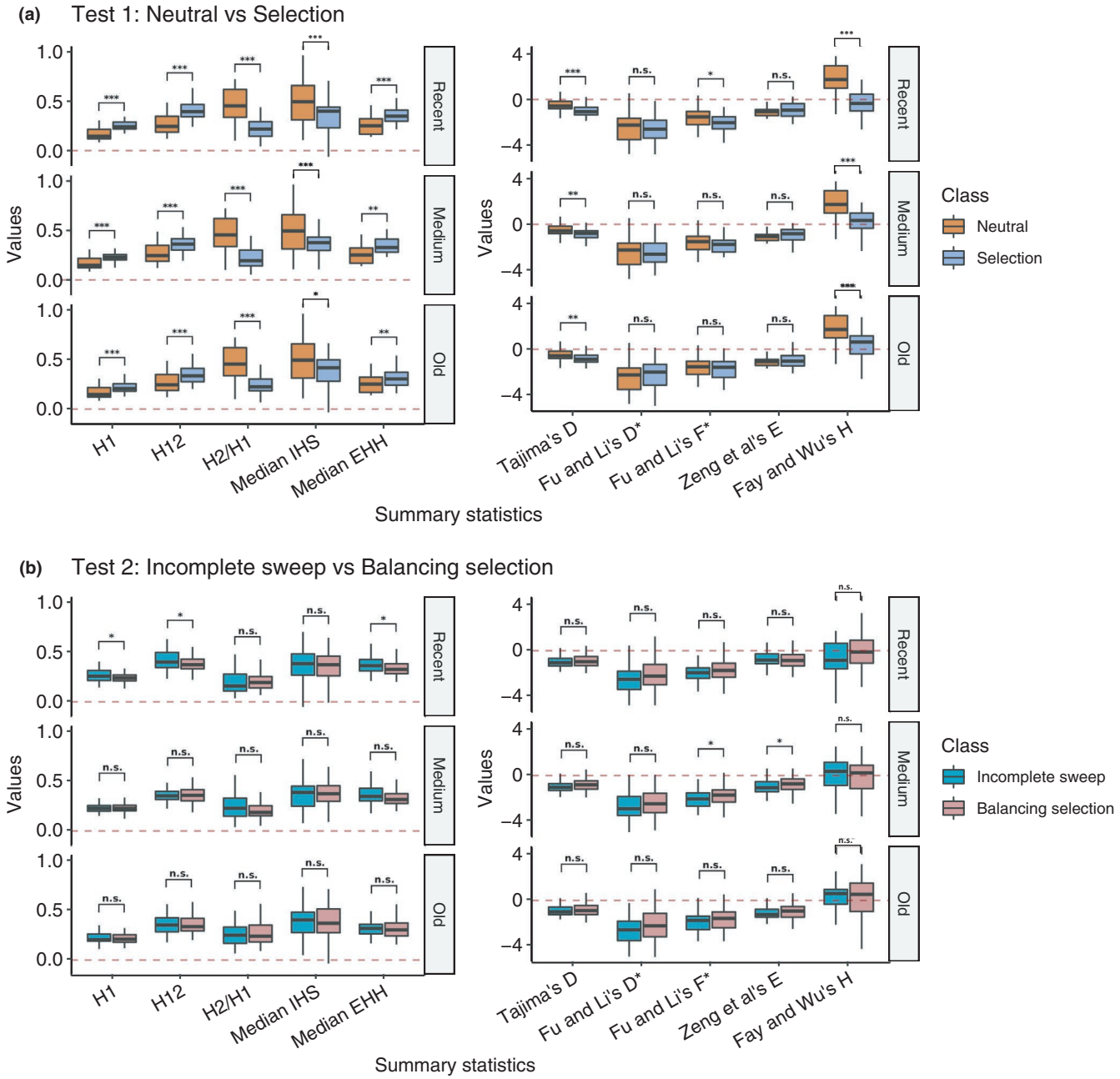

FIGURE 1 Distribution of a subset of summary statistics calculated on simulated loci under either neutral evolution or natural selection at different times of onset (recent, medium or old). Panel (a) shows the comparison between neutral evolution and natural selection (either ongoing positive selection or balancing selection). Panel (b) shows the comparison between incomplete sweep and balancing selection. Left panels group summary statistics based on haplotype diversity, while right panels group summary statistics based on allele frequency. Comparisons which are statistically significant (two-sided two-sample Mann–Whitney U-test) are depicted with *($p < 0.05$), **($p < 0.01$), ***($p < 0.001$), otherwise are depicted with n.s. (not significant)

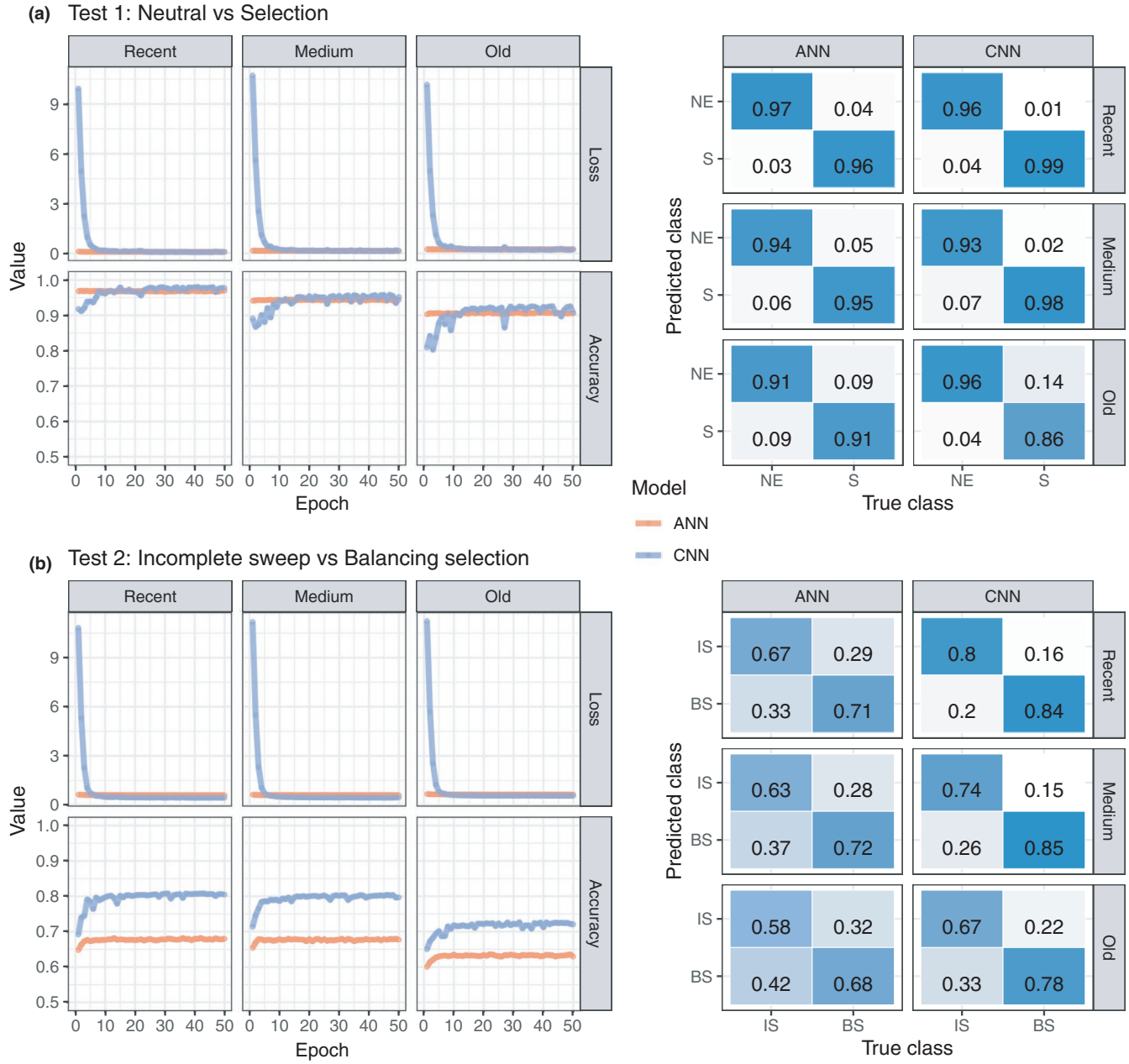

FIGURE 2 Performance of ANN (orange) and CNN (blue) to predict loci under selection (Test 1, upper panel a.) and to distinguish between incomplete sweep and balancing selection (Test 2, lower panel b.). For each category of time of onset of selection (recent, medium, old), training loss and accuracy (low-to-high gradient coloured in white-to-blue scale) over epochs are shown on the left side, while confusion matrices are shown on the right side. Different classes to predict are neutrality (NE), selection (S), incomplete sweep (IS), balancing selection (BS)

3.2 | Convolutional neural network has higher prediction accuracy than ANN to distinguish between incomplete sweep and balancing selection

As summary statistics do not have power to discriminate between incomplete sweep and balancing selection if considered individually, we then tested whether their predictive power increased when jointly integrated. Thus, we implemented a deep ANN which receives as input all calculated summary statistics (Sheehan & Song, 2016) and predicts whether a given locus is under either neutrality or natural selection, either due to an incomplete sweep or balancing selection (Test 1). We compared the predictive accuracy of ANN to an approach based on convolutional layers, in the form of a CNN applied to full population genomic data as an alignment of sampled haplotypes (Torada et al., 2019).

Figure 2 illustrates the performance of ANN and CNN to predict loci under different classes of evolution. The upper panel (a) on the left side shows the training loss and accuracy over epochs for classifying a locus under either neutral evolution (NE) or selection (S, Test 1). CNN showed a high loss and lower accuracy during the first few

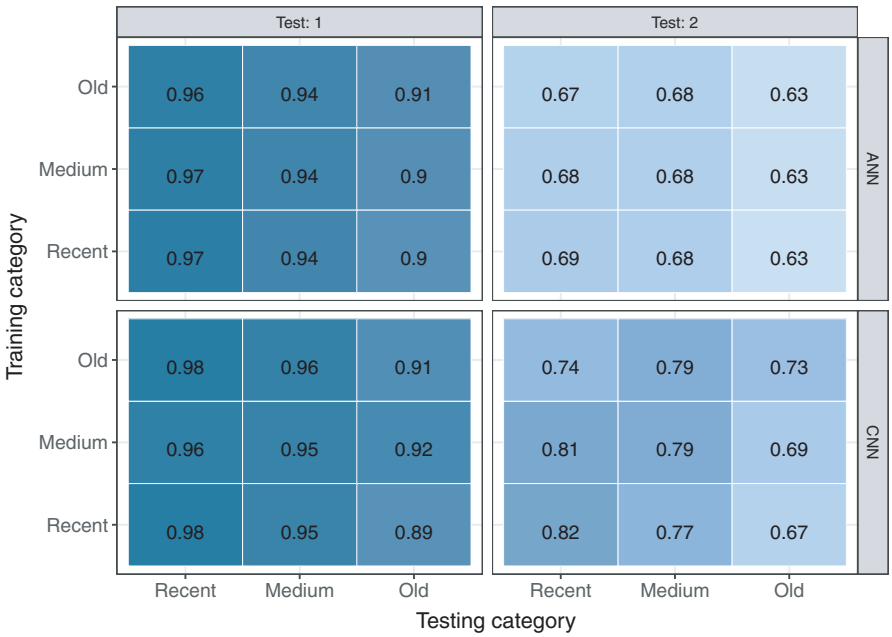

FIGURE 3 Prediction accuracy (low-to-high gradient coloured in white-to-blue scale) for classifying loci under different evolutionary events (Test 1 and Test 2, on columns) and methods (ANN and CNN, on rows) for all pairs of classes for time of onset of selection between training (y-axis) and testing data (x-axis). The antidiagonal shows accuracy values when the model used for both training and testing is the same, while accuracy values outside the antidiagonal are obtained when the models employed for training and testing differ

epochs, but both methods reached qualitatively similar levels of loss and accuracy after approximately ten epochs. Confusion matrices on testing data (top panel a on the right side of Figure 2) indicate similar predictive power for ANN and CNN. Recent selective events were more likely to be correctly classified than older events. For instance, we observed that the false-negative rate of identifying a gene under old selection is 9% for ANN and 14% for CNN, whereas it was 4% for ANN and 1% for CNN in case of recent selection (i.e. 20 kya).

The lower panel (b) of Figure 2 on the left side illustrates training loss and accuracy over epochs for classifying a locus under either incomplete sweep (IS) or balancing selection (BS, Test 2). The results recapitulated what was previously observed on the higher loss during the first few epochs for CNN. However, for this classification task, CNN exhibited a consistently higher prediction accuracy than ANN across all epochs. This observation was confirmed when investigating the confusion matrices calculated on testing data (Figure 2, right side of lower panel b). CNN consistently outperformed ANN for predicting loci under incomplete sweep or balancing selection, although the overall accuracy was lower than the one obtained for Test 1. For instance, we observed a false-negative rate of identifying a locus under old balancing selection of 32% for ANN and 22% for CNN, and 29% for ANN and 16% for CNN in case of recent selection. Again, recent selective events were more likely to be correctly classified than older events. However, we should stress that ANN will achieve better performance (and possibly similar prediction accuracy to the CNN) if a larger number of informative statistics are given as input. Overall, CNN had high power to identify loci under selection and substantial power to distinguish between incomplete sweep and balancing selection, two modes of evolution that leave extremely similar genomic patterns.

3.3 | Convolutional neural network is more robust than ANN to misspecified training data

The training of a neural network for population genetic inferences is conditional on a demographic and selection model to generate genomic data under different evolutionary scenarios. Therefore, we tested the robustness of both ANN and CNN to misspecified evolutionary parameters during training. Specifically, we used the already generated synthetic data and calculated the prediction accuracy for identifying loci under selection (Test 1) and for distinguishing between incomplete sweep and balancing selection (Test 2) when both ANN and CNN were trained on a specific time of onset of selection (recent, medium, old) but tested on a different value. By doing so, we were able to quantify any drop in accuracy when the training data did not reflect the underlying true evolutionary model.

Figure 3 shows the prediction accuracy for both tests (Test 1 and Test 2, on columns) and networks (ANN and CNN, on rows) for all possible pairs of time of onset of selection between training and testing data. Numbers on the antidiagonal represent accuracy values when the model used for both training and testing was the same. Numbers outside the antidiagonal indicate accuracy values when the models employed for training and testing differed. We observed a marginal decline in accuracy when using incorrect training data for Test 1 for both networks which performed similarly. These results were confirmed when investigating all corresponding confusion matrices (Figure S17). For Test 2, the drop in accuracy when employing a different model for training was more evident than for Test 1, although CNN outperformed ANN in most scenarios (Figure 3, Figure S18).

To further test the robustness of our inferences to a misspecified model, we tested both architectures trained on a European population to simulated data generated from a demographic model

of African and East Asian populations (Jouganous et al., 2017). Accuracy values for Test 1 are marginally affected by a misspecified demographic model (Figures S19 and S20), while we observed a slightly more pronounced decrease in performance for Test 2 (Figures S19 and S20).

3.4 | Convolutional neural network identifies signatures of recent natural selection in MEFV gene

We deployed the trained networks, both ANN and CNN, on genomic data for the *MEFV* gene from CEU population from the 1000 Genomes Project (1000 Genomes Project Consortium, 2015). MEFV gene has been previously associated with both balancing selection (Fumagalli et al., 2009a) and ongoing positive selection (Park et al., 2020). Here we tested whether MEFV gene has been targeted by natural selection and, if so, whether by balancing selection or incomplete sweep.

To assess the false-positive rate, we extracted flanking genomic regions to *MEFV* predicted to be under neutral evolution (Arbiza et al., 2012) and deployed both ANN and CNN algorithms on all intermediate frequency variants. We expected the networks not to predict signals of selection within these control neutral regions. ANN predicted 23 out of 42 sites to be under selection regardless of the time of onset of selection (Figure S21). Therefore, we decided not to use the ANN algorithm for inferences on the *MEFV* gene, as it showed a high false-positive rate when applied to putative neutral genomic regions. In contrast, CNN provided strong support for 39 out of 42 sites to be under neutral evolution, with only three sites possibly predicted to be under selection regardless of the time of onset (Figure S22).

Next, we aimed to identify signals of natural selection and deployed the trained CNN within the *MEFV* genomic region of European samples (CEU) from the 1000 Genomes Project database (1000 Genomes Project Consortium, 2015). We observed a large proportion of sites with intermediate allele frequency predicted to be under natural selection (Test 1) regardless of the time of onset of selection (Figure 4, upper panel). All sites under selection were predicted to be under incomplete sweep rather than balancing selection (Figure 4, second panel from top).

Sites predicted to be under selection (or in LD with the target of selection) encompass a haplotype block spanning from intron 2 to 3' UTR (untranslated region, Figure S23). Most of these variants are possibly functionally silent as they lay within introns or represent synonymous substitutions (Figure 4, third to fifth panels from top). However, two mutations within this region represent either missense (rs1231123, rs1231122) or stop-gained (rs1231122) substitutions, depending on the corresponding isoform. The predicted signals of selection in the *MEFV* gene were confirmed when deploying the trained network to genomic data from TSI samples (1000 Genomes Project Consortium, 2015), another European population (Figure S24). However, the results obtained using TSI population showed a higher false-positive rate when deployed to neutral genomic regions (Figure S25) than the ones obtained using CEU population, possibly because the network was trained on simulated data conditional on a demographic model inferred for the CEU population. In fact, 7, 14 and 10 out of 38 neutral sites were predicted to be under selection with recent, medium and old time of onset, respectively, using TSI population. In contrast, 3, 13 and 9 out of 42 neutral sites were labelled as targets of selection with recent, medium and old time of onset, respectively, using CEU population.

4 | DISCUSSION

In this study, we demonstrated the utility of deep learning to identify genomic signals of recent natural selection on intermediate frequency variants. We showed that algorithms based on either summary statistics (i.e. ANN) or full genomic data (i.e. CNN) had comparably high power to infer selective regimes (Figure 2) and exhibit lower false-positive and false-negative rates than commonly used neutrality tests (Figure S26). However, CNN had higher accuracy to distinguish between loci under balancing selection and incomplete sweep (Figure 2), it was generally more robust to incorrect training data (Figure 3), and it had a lower false-positive rate when deployed on neutral genomic regions than ANN (Figures S21 and S22). Finally, we illustrated the applicability of deep neural networks to detect and characterize signals of natural selection on common variants within the *MEFV* gene region (Figure 4).

Our results on the high predictive power offered by deep learning, and specifically by convolutional neural networks, to detect signals of natural selection expand previous findings (Chan et al., 2018; Flagel et al., 2019; Sanchez et al., 2020; Torada et al., 2019) to cases where the beneficial allele is at intermediate frequency. CNN outperformed ANN to distinguish between incomplete sweep and balancing selection, although, in our analyses, its training was slower by a factor of 300. In fact, CNN had more than 4 million parameters to estimate, in contrast to ANN which had approximately 2,000. Additionally, ANN received as input informative features (i.e. summary statistics) while convolutional filters in the CNN learned the optimal features from the raw data while training. In machine learning, the design of such features had been a major part of information engineering. As an illustration, in the field of computer vision, the 'features' used for many practical algorithms until the early 2000s consisted of hand-engineered gradient estimators (Shen & Bai, 2006), typically at multiple spatial scales (Gauch, 1999; Lowe, 1999), applied to images (arrays of pixels). The observation that features emerge within a deep network has been repeated in different domains. Therefore, we envisage that a novel area of research will focus on extracting informative features from trained networks for population genetic inference, possibly by analysing activation or saliency maps (Bahdanau et al., 2015). It is important to note that ANN will achieve higher performance with the inclusion of additional summary statistics not considered herein.

This study also contributes to ongoing efforts to design architecture and devise training techniques for deep learning algorithms in

FIGURE 4 Prediction of sites under natural selection (Test 1, upper panel) or balancing selection vs. incomplete sweep (Test 2, second panel from top) on intermediate frequency variants in the *MEFV* gene for a European population. For each tested variant, the predicted functional impact across all isoforms is reported (counts of functional consequences on third panel, genomic location on fourth panel and transcripts on fifth panel from the top)

population genetics (Sanchez et al., 2020). Resizing images to smaller dimensions appeared to reduce overfitting and learning time (Figure S9) and could be considered a complementary strategy to approaches based on cropping or padding (Flagel et al., 2019). The strategy to separately sort rows based on the presence or absence of the putative target variant is an alternative solution to adopt more general, but computational expensive, architectures based on exchangeable neural networks (Chan et al., 2018; Sanchez et al., 2020). We also

explored the applicability of forward-in-time simulations to train deep neural networks for population genetics and the usefulness of data augmentation (Figure S10) to reduce the computational time required to generate synthetic training data. The use of forward-in-time simulations should generate more realistic synthetic population genomic data and model more complex evolutionary scenarios than by using coalescent simulations. In any case, as suggested in this study (Figures S21 and S22), false-positive and false-negative rates should be assessed by deploying trained networks on loci previously identified as targets of selection or neutrally evolving. Further research on the design of neural networks for population genetics is in need, for instance by maximizing the prediction accuracy over a grid of hyperparameters (e.g. number of units and layers) using Neural Architecture Search (Wistuba et al., 2019).

We show that deep neural networks achieved higher prediction power to differentiate between the effects of neutral evolution, balancing selection and incomplete sweep for variants segregating at intermediate frequency (Figure 2) than commonly used summary statistics (Figure 1). However, the accuracy to distinguish between incomplete sweep and balancing selection using CNN ranges from 72% to 80% depending on the time of onset of the selection, with more recent events (around 20 kya) more accurately classified (Figure 3). While this accuracy is far higher than that achieved using summary statistics, higher accuracy could be achieved by employing a larger training data set, by using more extensive hyper-parameter tuning and architecture search, and by treating overdominance and negative frequency-dependent selection as separate prediction categories. In fact, future extensions of this study will include testing to distinguish between overdominance and negative frequency-dependent selection once a variant is predicted to be under balancing selection. It is likely that a different CNN architecture and training data is needed for this purpose as, for instance, information on heterozygosity (not considered herein given the simulation strategy) will likely emerge as an important feature. Additionally, a wider spectrum of times of the onset of selection should be considered to assess the power to predict balancing selection at different evolutionary scenarios. Finally, the CNN herein proposed requires fully resolved haplotypes with no missing data. Furthermore, we argue that such approach should be locus-specific as the network needs to be trained with the local characteristics of the region of interest (e.g. recombination rate). Therefore, this CNN is more suitable to be deployed to deep resequencing data on single loci of interest rather than to genome-wide low-coverage sequencing data. In the latter case, we argue that an ANN receiving as input statistical estimates of summary statistics from genotype likelihoods (Korneliussen et al., 2013) might be a valuable alternative. Nevertheless, the effect of data uncertainty should be further explored.

The analyses on the *MEFV* gene performed herein complement previous findings (Schaner et al., 2001) to suggest that this gene has been subjected to different evolutionary forces. The *MEFV* gene encodes for the Pyrin protein which plays an important role in inflammatory processes (Schnappauf et al., 2019). Five different functional domains have been identified within the Pyrin protein. The PYD domain (aa 1–92) is present in at least 20 human proteins involved in inflammatory pathways. However, in the analyses we performed the PYD domain seems to have neutrally evolved. The Pyrin central region hosts three domains: a bZIP domain (aa 266–280), a B-box domain (aa 370–412) and a coiled-coil domain (CC, aa 420–440). The role of these three domains has not been thoroughly elucidated and few FMF-causing variants localize to Pyrin's central region (Je Wook et al., 2007; Stella et al., 2019). Nevertheless, from our data this central region is apparently under recent selection (Figure 4) or is in LD with beneficial alleles (Figure S23). Similarly, the B30.2 domain (also known as PRY/SPRY domain), which is encoded by the *MEFV* exon 10 where most of the FMF-causing variants cluster (Accetturo et al., 2020), shows the same genetic patterns of ongoing selection.

A recent study demonstrated that the FMF-associated variants M694V, M680I and V726A, all localizing to the B30.2 region, decrease the binding of *Yersinia pestis* virulence factor YopM (Park et al., 2020). Further, the authors provided evidence that M694V and V726A variants were subject to recent positive selection in a cohort of Turkish individuals. Finally, FMF knock-in mice demonstrated survival advantage compared to wild-type mice. Thus, these experimental evidences suggest that mutations in the human Pyrin may have conferred resistance to *Yersinia pestis* (Park et al., 2020). However, the possibility that other pathogens could have concurred in conferring a selective advantage cannot be ruled out. Indeed, contrary to previous claims of overdominance acting on *MEFV* (Fumagalli et al., 2009a), our new results and Park et al's study suggest that the selection on human Pyrin is directional and either recent or possibly still ongoing. In fact, the frequency of M694V and V726A kept rising (Park et al., 2020) although no plague outbreaks rose to the scale of a pandemic after the 17th century.

The population sample we analysed in this study is different from the Turkish cohort investigated by Park et al which overlaps significantly with one of the plague outbreak sites. Nevertheless, even in the different population sample we analysed, the data presented herein suggest signals of recent selection on the human Pyrin. While our computational predictions are unable to identify the causal variant, it is possible to hypothesize that Pyrin, specifically its B30.2 region, could confer resistance to a broader range of pathogens including those causing more recent pandemics. Likewise, we cannot rule out more complex evolutionary scenarios, with *MEFV* being subjected to long-term balancing selection and recent positive selection on standing variation. A comprehensive picture of ongoing selection signatures in *MEFV* could be achieved by deploying deep neural networks trained on variants segregating at low or high frequency and to a wide range of Mediterranean populations. Finally, additional power to characterize recent selection in *MEFV* could be gained by integrating data from ancient genomes (Dehasque et al., 2020) as this would be particularly suitable to relate adaptation to past epidemics to current pathogenic threats (Patin, 2020).

In this study, we demonstrated how deep learning and, in particular, convolutional neural networks were able to perform predictions currently inaccessible by commonly used strategies based on

summary statistics. In particular, we showed that deep neural networks can differentiate between signals of incomplete sweep and balancing selection, despite the two evolutionary events leaving qualitatively similar patterns of genetic variation. Furthermore, our application to detect signals of selection on FMF-associated alleles highlighted the importance of a population genetic approach to understand the molecular basis of susceptibility and/or resistance to infectious diseases.

ACKNOWLEDGEMENTS

This work was supported by a Leverhulme Trust Research Grant (RPG-2018-208) and an Imperial College European Partners Fund to MF. We acknowledge the support offered by the Erasmus+ programme to UI. We are grateful to Aida Andrés, Anil A. Bharath and Mehmet Somel for discussions and Bárbara Bitarello and two anonymous reviewers for comments on the manuscript. We also thank Kivilcim Basak Vural and the METU Comparative and Evolutionary Biology Group for computational support.

AUTHOR CONTRIBUTIONS

MF and UI designed the research. UI performed the research with contributions from AS. MF, UI and AS analysed data and wrote the paper.

DATA AVAILABILITY STATEMENT

Detailed tutorials on pipelines for training and prediction, along with all the scripts used in this study, are available within *BaSe* package at https://github.com/ulasisik/balancing-selection. Sequencing data on human populations were retrieved from The International Genome Sample Resource (IGSR) at https://www.internationalgenome.org.

ORCID

Matteo Fumagalli https://orcid.org/0000-0002-4084-2953

REFERENCES

1000 Genomes Project Consortium (2015). A global reference for human genetic variation. *Nature*, *526*(7571), 68–74.

Accetturo, M., D'Uggento, A. M., Portincasa, P., & Stella, A. (2020). Improvement of MEFV gene variants classi_cation to aid treatment decision making in familial Mediterranean fever. *Rheumatology (Oxford)*, *59*(4), 754–761.

Andrés, A. M., Hubisz, M. J., Indap, A., Torgerson, D. G., Degenhardt, J. D., Boyko, A. R., Gutenkunst, R. N., White, T. J., Green, E. D., Bustamante, C. D., Clark, A. G., & Nielsen, R. (2009). Targets of balancing selection in the human genome. *Molecular Biology and Evolution*, *26*(12), 2755–2764.

Arbiza, L., Zhong, E., & Keinan, A. (2012). NRE: A tool for exploring neutral loci in the human genome. *BMC Bioinformatics*, *13*(1), 1.

Asthana, S., Schmidt, S., & Sunyaev, S. (2005). A limited role for balancing selection. *Trends in Genetics*, *21*(1), 30–32.

Bahdanau, D., Cho, K. H., & Bengio, Y. (2015). *Neural machine translation by jointly learning to align and translate*. (pp. 1–15).3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings

Bitarello, B. D., de Filippo, C., Teixeira, J. C., Schmidt, J. M., Kleinert, P., Meyer, D., & Andrés, A. M. (2018). Signatures of long-term balancing selection in human genomes. *Genome Biology and Evolution*, *10*(3), 939–955.

Bubb, K. L., Bovee, D., Buckley, D., Haugen, E., Kibukawa, M., Paddock, M., Palmieri, A., Subramanian, S., Zhou, Y., Kaul, R., Green, P., & Olson, M. V. (2006). Scan of human genome reveals no new loci under ancient balancing selection. *Genetics*, *173*(4), 2165–2177.

Cagliani, R., Fumagalli, M., Riva, S., Pozzoli, U., Comi, G. P., Bresolin, N., & Sironi, M. (2010). Genetic variability in the ACE gene region surrounding the Alu I/D polymorphism is maintained by balancing selection in human populations. *Pharmacogenetics and Genomics*, *20*(2), 131–134.

Cagliani, R., Fumagalli, M., Riva, S., Pozzoli, U., Comi, G. P., Menozzi, G., Bresolin, N., & Sironi, M. (2008). The signature of long-standing balancing selection at the human defensin β-1 promoter. *Genome Biology*, *9*(9):R143.

Cagliani, R., Fumagalli, M., Riva, S., Pozzoli, U., Fracassetti, M., Bresolin, N., Comi, G. P., & Sironi, M. (2010). Polymorphisms in the CPB2 gene are maintained by balancing selection and result in haplotype-preferential splicing of exon 7. *Molecular Biology and Evolution*, *27*(8), 1945–1954.

Chan, J., Perrone, V., Spence, J. P., Jenkins, P. A., Mathieson, S., & Song, Y. S. (2018). A likelihood-free inference framework for population genetic data using exchangeable neural networks. *Advances in Neural Information Processing Systems*, *2018*:8594–8605.

Charlesworth, D. (2006). Balancing selection and its effects on sequences in nearby genome regions. *PLoS Genetics*, *2*(4), 379–384.

Cheng, X., & DeGiorgio, M. (2020). Flexible mixture model approaches that accommodate footprint size variability for robust detection of balancing selection. *Molecular Biology and Evolution*, *37*(11), 3267–3291. https://doi.org/10.1093/molbev/msaa134

Chollet, F. (2015). *Keras*. https://keras.io

DeGiorgio, M., Lohmueller, K. E., & Nielsen, R. (2014). A model-based approach for identifying signatures of ancient balancing selection in genetic data. *PLoS Genetics*, *10*(8), e1004561.

Dehasque, M., Ávila-Arcos, M. C., Díez-del-Molino, D., Fumagalli, M., Guschanski, K., Lorenzen, E. D., Malaspinas, A.-S., Marques-Bonet, T., Martin, M. D., Murray, G. G. R., Papadopulos, A. S. T., Therkildsen, N. O., Wegmann, D., Dalén, L., & Foote, A. D. (2020). Inference of natural selection from ancient DNA. *Evolution Letters*, *4*(2), 94–108.

Delanghe, J. R., Speeckaert, M. M., & De Buyzere, M. L. (2020). COVID-19 infections are also affected by human ACE1 D/I polymorphism. *Clinical Chemistry and Laboratory Medicine*, *58*(7):1125–1126, pages 1–2.

Dobzhansky, T. (1951). *Genetics and the Origin of Species*. 3rd ed. Columbia Univ. Press.

Fay, J. C., & Wu, C. I. (2000). Hitchhiking under positive Darwinian selection. *Genetics*, *155*(3), 1405–1413.

Ferrer-Admetlla, A., Bosch, E., Martin Sikora, T., Marques-Bonet, A.-R.-S., Muntasell, A., Navarro, A., Lazarus, R., Calafell, F., Bertranpetit, J., & Casals, F. (2008). Balancing selection is the main force shaping the evolution of innate immunity genes. *The Journal of Immunology*, *181*(2), 1315–1322.

Ferrer-Admetlla, A., Liang, M., Korneliussen, T., & Nielsen, R. (2014). On detecting incomplete soft or hard selective sweeps using haplotype structure. *Molecular Biology and Evolution*, *31*(5), 1275–1291.

Fijarczyk, A., & Babik, W. (2015). Detecting balancing selection in genomes: Limits and prospects. *Molecular Ecology*, *24*(14), 3529–3545.

Flagel, L., Brandvain, Y., & Schrider, D. R. (2019). The unreasonable effectiveness of convolutional neural networks in population genetic inference. *Molecular Biology and Evolution*, *36*(2), 220–238.

Fu, Y. X., & Li, W. H. (1993). Statistical tests of neutrality of mutations. *Genetics*, *133*(3), 693–709.

Fumagalli, M., Cagliani, R., Pozzoli, U., Riva, S., Comi, G. P., Menozzi, G., Bresolin, N., & Sironi, M. (2009). A population genetics study of the familial mediterranean fever gene: Evidence of balancing

selection under an over dominance regime. *Genes and Immunity*, *10*(8), 678–686.

Fumagalli, M., Cagliani, R., Pozzoli, U., Riva, S., Comi, G. P., Menozzi, G., Bresolin, N., & Sironi, M. (2009). Widespread balancing selection and pathogen-driven selection at blood group antigen genes. *Genome Research*, *19*(2), 199–212.

Fumagalli, M., Camus, S. M., Diekmann, Y., Burke, A., Camus, M. D., Norman, P. J., Joseph, A., Abi-Rached, L., Benazzo, A., Rasteiro, R., Mathieson, I., Topf, M., Parham, P., Thomas, M. G., & Brodsky, F. M. (2019). Genetic diversity of CHC22 clathrin impacts its function in glucose metabolism. *eLife*, *8*, e41517. https://doi.org/10.7554/eLife.41517

Fumagalli, M., Sironi, M., Pozzoli, U., Ferrer-Admettla, A., Pattini, L., & Nielsen, R. (2011). Signatures of environmental genetic adaptation pinpoint pathogens as the main selective pressure through human evolution. *PLoS Genetics*, *7*(11), e1002355.

Garud, N. R., Messer, P. W., Buzbas, E. O., & Petrov, D. A. (2015). Recent selective sweeps in North American *Drosophila melanogaster* show signatures of soft sweeps. *PLoS Genetics*, *11*(2), 1–32.

Gauch, J. M. (1999). Image segmentation and analysis via multiscale gradient watershed hierarchies. *IEEE Transactions on Image Processing*, *8*(1), 69–79.

Gravel, S., Henn, B. M., Gutenkunst, R. N., Indap, A. R., Marth, G. T., Clark, A. G., Fuli, Y. U., Gibbs, R. A., & Bustamante, C. D. (2011). Demographic history and rare allele sharing among human populations. *Proceedings of the National Academy of Sciences*, *108*(29), 11983–11988.

Hahne, F., & Ivanek, R. (2016). Visualizing genomic data using Gviz and bioconductor. In E. Mathé, & S. Davis (eds.) *Statistical genomics: Methods and protocols* (pp. 335–351). Springer.

Haller, B. C., & Messer, P. W. (2019). SLiM 3: Forward genetic simulations beyond the wright-fisher model. *Molecular Biology and Evolution*, *36*(3), 632–637.

Harpending, H. C. (1994). Signature of ancient population growth in a low-resolution mitochondrial DNA mismatch distribution. *Human Biology*, *66*(4), 591–600.

Hill, W. G., & Robertson, A. (1968). Linkage disequilibrium in finite populations. *Theoretical and Applied Genetics*, *38*(6), 226–231.

Je Wook, Y. U., Fernandes-Alnemri, T., Datta, P., Jianghong, W. U., Juliana, C., Solorzano, L., McCormick, M., Zhang, Z. J., & Alnemri, E. S. (2007). Pyrin activates the ASC pyroptosome in response to engagement by autoinammatory PSTPIP1 mutants. *Molecular Cell*, *28*(2), 214–227.

Jiuxiang, G. U., Wang, Z., Kuen, J., Ma, L., Shahroudy, A., Shuai, B., Liu, T., Wang, X., Wang, G., Cai, J., & Chen, T. (2018). Recent advances in convolutional neural networks. *Pattern Recognition*, *77*, 354–377.

Jouganous, J., Long, W., Ragsdale, A. P., & Gravel, S. (2017). Inferring the joint demographic history of multiple populations: Beyond the diffusion approximation. *Genetics*, *206*(3), 1549–1567.

Kassambara, A. (2020). *Ggpubr: 'ggplot2' Based Publication Ready Plots*. R package version 0.3.0. https://rpkgs.datanovia.com/ggpubr/

Kelly, J. K. (1997). A test of neutrality based on interlocus associations. *Genetics*, *146*(3), 1197–1206.

Kern, A. D., & Schrider, D. R. (2018). DiploS/HIC: An updated approach to classifying selective sweeps. *G3: Genes, Genomes, Genetics*, *8*(6), 1959–1970.

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

Kingma, D. P., & Ba, J. (2014). Adam: A Method for Stochastic Optimization. *arXiv*, (1412.6980). https://arxiv.org/abs/1412.6980v9

Kolde, R. (2018). *pheatmap: Pretty Heatmaps*. R package version 1.0.12. https://github.com/raivokolde/pheatmap

Korneliussen, T. S., Moltke, I., Albrechtsen, A., & Nielsen, R. (2013). Calculation of Tajima's D and other neutrality test statistics from low depth next-generation sequencing data. *BMC Bioinformatics*, *14*(1), 289.

Krizhevsky, A., Sutskeverl, I., & Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems*, *25*(2), 1097–1105.

Lecun, Y., Bengio, Y., & Hinton, G. E. (2015). Deep learning. *Nature*, *521*(7553), 436–444.

Lecun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, *86*(11), 2278–2324.

Leffler, E. M., Gao, Z., Pfeifer, S., Ségurel, L., Auton, A., Venn, O., Bowden, R., Bontrop, R., Wall, J. D., Sella, G., Donnelly, P., McVean, G., & Przeworski, M. (2013). Multiple instances of ancient balancing selection shared between humans and chimpanzees. *Science*, *340*(6127), 1578–1582.

Lin, K., Li, H., Schlötterer, C., & Futschik, A. (2011). Distinguishing positive selection from neutral evolution: Boosting the performance of summary statistics. *Genetics*, *187*(1), 229–244.

Llaurens, V., Whibley, A., & Joron, M. (2017). Genetic architecture and balancing selection: The life and death of differentiated variants. *Molecular Ecology*, *26*(9), 2430–2448.

Lowe, D. G. (1999). Object recognition from local scale-invariant features. *Proceedings of the IEEE International Conference on Computer Vision*, *2*, 1150–1157.

Meyer, D., Single, R. M., Mack, S. J., Erlich, H. A., & Thomson, G. (2006). Signatures of demographic history and natural selection in the human major histocompatibility compex loci. *Genetics*, *173*(4), 2121–2142.

Mughal, M. R., & DeGiorgio, M. (2019). Localizing and classifying adaptive targets with trend filtered regression. *Molecular Biology and Evolution*, *36*(2), 252–270.

Nei, M., & Li, W. H. (1979). Mathematical model for studying genetic variation in terms of restriction endonucleases. *Proceedings of the National Academy of Sciences of the United States of America*, *76*(10), 5269–5273.

Park, Y. H., Remmers, E. F., Lee, W., Ombrello, A. K., Chung, L. K., Shilei, Z., Stone, D. L., Ivanov, M. I., Loeven, N. A., Barron, K. S., Hoffmann, P., Nehrebecky, M., Akkaya-Ulum, Y. Z., Sag, E., Balci-Peynircioglu, B., Aksentijevich, I., Gül, A., Rotimi, C. N., Chen, H., … Chae, J. J. (2020). Ancient familial Mediterranean fever mutations in human pyrin and resistance to Yersinia pestis. *Nature Immunology*, *21*(8), 857–867.

Patin, E. (2020). Plague as a cause for familial Mediterranean fever. *Nature Immunology*, *21*(8), 833–834.

Pavlidis, P., Jensen, J. D., & Stephan, W. (2010). Searching for footprints of positive selection in whole-genome SNP data from nonequilibrium populations. *Genetics*, *185*(3), 907–922.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikitlearn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Peter, B. M., Huerta-Sanchez, E., & Nielsen, R. (2012). Distinguishing between selective sweeps from standing variation and from a De Novo Mutation. *PLoS Genetics*, *8*(10), e1003011.

Rainer, J. (2017). *EnsDb.Hsapiens.v75: Ensembl based annotation package*. R package version 2.99.0. https://doi.org/10.18129/B9.bioc.EnsDb.Hsapiens.v75

Ronen, R., Udpa, N., Halperin, E., & Bafna, V. (2013). Learning natural selection from the site frequency spectrum. *Genetics*, *195*(1), 181–193.

Ruder, S. (2017). An overview of gradient descent optimization algorithms. *arXiv*, (1609.04747), https://arxiv.org/abs/1609.04747

Sabeti, P. C., Reich, D. E., Higgins, J. M., Levine, H. Z., Richter, D. J., Schaffner, S. F., Gabriel, S. B., Platko, J. V., Patterson, N. J., McDonald, G. J., Ackerman, H. C., Campbell, S. J., Altshuler, D.,

Cooper, R., Kwiatkowski, D., Ward, R., & Lander, E. S. (2002). Detecting recent positive selection in the human genome from haplotype structure. *Nature*, *419*(6909), 832–837.

Sanchez, T., Cury, J., Charpiat, G., & Jay, F. (2020). Deep learning for population size history inference: Design, comparison and combination with approximate Bayesian computation. *Molecular Ecology Resources*, 1–16. https://onlinelibrary.wiley.com/action/showCitFormats?doi=10.1111%2F1755-0998.13224

Schaner, P., Richards, N., Wadhwa, A., Aksentijevich, I., Kastner, D., Tucker, P., & Gumucio, D. (2001). Episodic evolution of pyrin in primates: Human mutations recapitulate ancestral amino acid states. *Nature Genetics*, *27*(3), 318–321.

Schnappauf, O., Chae, J. J., Kastner, D. L., & Aksentijevich, I. (2019). The Pyrin Inammasome in health and disease. *Frontiers in Immunology*, *10*, 1745.

Schrider, D. R., & Kern, A. D. (2016). S/HIC: Robust Identification of Soft and Hard Sweeps Using Machine Learning. *PLoS Genetics*, *12*(3), 1–31.

Schrider, D. R., & Kern, A. D. (2018). Supervised machine learning for population genetics: A new paradigm. *Trends in Genetics*, *34*(4), 301–312.

Sellis, D., Callahan, B. J., Petrov, D. A., & Messer, P. W. (2011). Heterozygote advantage as a natural consequence of adaptation in diploids. *Proceedings of the National Academy of Sciences of the United States of America*, *108*(51), 20666–20671.

Sheehan, S., & Song, Y. S. (2016). Deep learning for population genetic inference. *PLoS Computational Biology*, *12*(3), e1004845.

Shen, L., & Bai, L. (2006). A review on Gabor wavelets for face recognition. *Pattern Analysis and Applications*, *9*(2–3), 273–292.

Siewert, K. M., & Voight, B. F. (2017). Detecting long-term balancing selection using allele frequency correlation. *Molecular Biology and Evolution*, *34*(11), 2996–3005.

Soni, V., Vos, M., & Eyre-Walker, A. (2021). A new test suggests that balancing selection maintains hundreds of non-synonymous polymorphisms in the human genome. *bioRxiv*, *10.1101* (2021.02.08.430226), https://www.biorxiv.org/content/10.1101/2021.02.08.430226v2

Stella, A., Cortellessa, F., Scaccianoce, G., Pivetta, B., Settimo, E., & Portincasa, P. (2019). Familial Mediterranean fever: Breaking all the (genetic) rules. *Rheumatology (Oxford)*, *58*(3), 463–467.

Sugden, L. A., Atkinson, E. G., Fischer, A. P., Rong, S., Henn, B. M., & Ramachandran, S. (2018). Localization of adaptive variants in human genomes using averaged one-dependence estimation. *Nature Communications*, *9*(1), 703.

Tajima, F. (1993). Statistical analysis of DNA polymorphism. *Japanese Journal of Genetics*, *68*(6), 567–595.

Teixeira, J. C., de Filippo, C., Weihmann, A., Meneu, J. R., Racimo, F., Dannemann, M., Nickel, B., Fischer, A., Halbwax, M., Andre, C., Atencia, R., Meyer, M., Parra, G., Pääbo, S., & Andrés, A. M. (2015). Long-term balancing selection in LAD1 maintains a missense trans-species polymorphism in humans, chimpanzees, and bonobos. *Molecular Biology and Evolution*, *32*(5), 1186–1196.

Torada, L., Lorenzon, L., Beddis, A., Isildak, U., Pattini, L., Mathieson, S., & Fumagalli, M. (2019). ImaGene: a convolutional neural network to quantify natural selection from genomic data. *BMC Bioinformatics*, *20*(S9), 337.

Touitou I. (2001). The spectrum of Familial Mediterranean Fever (FMF) mutations. *European Journal of Human Genetics*, *9*(7), 473–483.

Voight, B. F., Kudaravalli, S., Wen, X., & Pritchard, J. K. (2006). A map of recent positive selection in the human genome. *PLoS Biology*, *4*(3), 446–458.

Watterson, G. A. (1975). On the number of segregating sites in genetical models without recombination. *Theoretical Population Biology*, *7*(2), 256–276.

Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis*. Springer-Verlag.

Wistuba, M., Rawat, A., Tejaswini Pedapati, A. (2019). A Survey on Neural Architecture Search. *arXiv*, (1905.01392). https://arxiv.org/abs/1905.01392v2

Xue, A. T., Schrider, D. R., & Kern, A. D. and Ag1000g Consortium. Discovery of ongoing selective sweeps within anopheles mosquito populations using deep learning. *Molecular Biology and Evolution*, 2021;*38*(3):1168–1183. https://doi.org/10.1093/molbev/msaa259

Zeng, K., Yun Xin, F. U., Shi, S., & Wu, C. I. (2006). Statistical tests for detecting positive selection by utilizing high-frequency variants. *Genetics*, *174*(3), 1431–1439.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section.

How to cite this article: Isildak U, Stella A, Fumagalli M. Distinguishing between recent balancing selection and incomplete sweep using deep neural networks. *Mol Ecol Resour*. 2021;00:1–13. https://doi.org/10.1111/1755-0998.13379