Versatile detection of diverse selective sweeps with Flex-sweep

M. Elise Lauterbur[1], Kasper Munch[2], David Enard[1]

[1]Department of Ecology and Evolutionary Biology, University of Arizona, Tucson AZ 85719, USA
[2]Bioinformatics Research Centre, Aarhus University, DK-8000 Aarhus C, Denmark
*Corresponding author: M. Elise Lauterbur, lauterbur@gmail.com

ABSTRACT

Understanding the selection pressures influencing modern-day genomic diversity and their overall genomic impact is a major goal of evolutionary genomics. In particular, the contribution of selective sweeps to adaptation remains an open question, with persistent statistical limitations of sweep detection methods in terms of power and specificity. Sweeps with subtle genomic signals have been particularly challenging to detect. While many existing powerful methods are capable of detecting specific types of sweeps and/or those with obvious signals, their power comes at the expense of versatility. This means that these tools are likely failing to identify many sweeps. Thus it is valuable but difficult to be able to detect sweeps with diverse characteristics. We present Flex-sweep, a versatile machine-learning tool designed to detect sweeps with a variety of subtle signals, including those that are thousands of generations old. It is especially valuable for detecting sweeps in non-model organisms, for which we neither have expectations about the characteristics of sweeps present in the genome nor outgroups with population-level sequencing to otherwise facilitate detecting very old sweeps. We show that Flex-sweep is powerful at detecting selective sweeps with more subtle signals, even in the face of demographic model complexity and misspecification, recombination rate heterogeneity, and background selection. Flex-sweep is able to detect sweeps up to 5,000 generations old (approximately 125,000 years in humans), including those that are weak, soft, and/or incomplete; it is also capable of detecting strong, complete sweeps up to 10,000 generations old. Furthermore, we apply Flex-sweep to the 1000 genomes data set of Yoruba in Ibadan, Nigeria and, in addition to recovering previously identified selective sweeps, show that sweeps disproportionately occur within genic regions and close to regulatory regions. In addition, we show that virus-interacting proteins (VIPs) are strongly enriched for selective sweeps, recapitulating previous results that demonstrate the importance of viruses as a driver of adaptive evolution in humans.

INTRODUCTION

The genomic legacy of past natural selection sets the stage for understanding evolutionary trajectories and current diversity. However, establishing the specific genomic contributions of past natural selection to the modern melange of adaptations and their patterns within genomes remains a difficult task. Fully characterizing these patterns requires that we know the genomic contributions of past positive natural selection, either in the form of polygenic adaptation or in the form of selective sweeps. However, in the specific case of sweeps, this presents two primary challenges: First, patterns of positive selection through sweeps tend to fade over time, and it has been particularly challenging to detect older hitchhiking events. Second, since genomic patterns of positive natural selection include diverse types of selective sweeps with varying stages of completeness, varying starting frequency, and varying ages, existing methods often detect a specific type of sweep(s) at the expense of versatility.

Designing detection methods that are capable of detecting diverse sweeps remains a challenge (Kern and Schrider 2018). Thus, not only do the specific targets of old positive selection events remain poorly known even in well studied species such as humans, any effort to characterize sweeps genome-wide is susceptible to bias resulting from the sweep type specialization of many methods that can result in false negatives (Akey 2009). More specifically, the task of pushing back the earliest dates of detectable selective events has recently become an important goal of population geneticists, particularly for humans (Racimo, Kuhlwilm, and Slatkin 2014; Key et al. 2016; Racimo 2016; Marciniak and Perry 2017; Peyrégne et al. 2017; Cheng et al. 2022). In addition, while many disparate existing methods focus on detecting specific types of selective sweeps, the recent flourishing of machine learning and other composite statistical methods has brought the field to a methodological juncture that allows the development of a “Swiss Army Knife” approach to detect a greater diversity of sweeps with a single method.

A selective sweep describes the process of a beneficial allele increasing in frequency as a result of positive natural selection and the resulting patterns of genetic diversity around the swept locus (Maynard Smith and Haigh 1974). Statistical methods for identifying these patterns and distinguishing them from patterns caused by demographic changes, recombination, or background selection are both numerous and, for sweeps with strong signals, powerful (eg. (Kim and Nielsen 2004; Hermisson and Pennings 2005; Nielsen 2005; Voight et al. 2006).

These methods have been used to discover the genomic basis of a variety of adaptations in humans and other species, including adult lactase persistence (Enattah et al. 2002; Bersaglieri et al. 2004; Coelho et al. 2005), high altitude adaptation (Simonson et al. 2012; Hu et al. 2019; Y. Li et al. 2014), and Malaria adaptation through the Duffy null allele (Hamblin, Thompson, and Di Rienzo 2002; Hamblin and Di Rienzo 2000; McManus et al. 2017). In addition, these methods are used to characterize the nature of selection across the genome, including the relative frequencies of different types of sweeps (Hernandez et al. 2011; Schrider and Kern 2017) and overall strength and effects on diversity (Sattath et al. 2011).

However, as the time between a sweep and sampling grows, the genetic footprint dissipates (Przeworski, Coop, and Wall 2005). This erosion of signal with age makes old sweeps difficult to detect, which has so far limited the development of methods to detect the genomic signatures of selective events in the time frame between old, population-level selection, and speciation. (Here, we use “old” to refer specifically to selection that occurred more than 2,000 generations ago). In principle, these signals should be present thousands of generations after the end of a selective sweep. For example, one classic signature of a selective sweep is the reduction in nucleotide diversity at sites linked to the beneficial allele. As described in Librado and Orlando (2018), diversity will start to increase again immediately after the removal of the selective pressure that caused the sweep, at a rate of $\Delta\pi = 2\mu(1 - 1/N_e)^t$ (where $\Delta\pi$ is the change in nucleotide diversity, t is the time since the end of the sweep, in generations, N_e is the effective population size, and μ is the per-base mutation rate). Assuming an equilibrium demographic history and a sweep that causes a 10% reduction of diversity around the selected site with a conservative initial diversity of $\pi_0 = 0.03$ (W. H. Li and Sadler 1991), we would expect full recovery of diversity to take over 100,000 generations (at $N_e = 10{,}000$ and $\mu = 1e^{-8}$) (Librado and Orlando 2018). Despite the attenuation of signals of selective sweeps with time, old sweeps should still be detectable with sufficiently powerful and flexible methods (Bisschop, Lohse, and Setter 2021).

Existing methods designed to detect sweeps that occurred in this time frame are largely based on differences between populations or closely-related species (Prüfer et al. 2014; Racimo, Kuhlwilm, and Slatkin 2014; Racimo 2016; Peyrégne et al. 2017; Librado and Orlando 2018; Cheng et al. 2022). While powerful under some conditions, these methods typically rely on correlated allele frequency differences between populations, limiting selective sweep scans to

populations with genomic data for outgroups and/or admixture. With these additional populations, selection can be detected in the ancestral branch before the population split(s), and localized to the ancestral or daughter branches. Similar methods can also identify adaptive introgression (Gower et al. 2021). Despite this, these methods do not perform well with weak selection or soft sweeps (Peyrégne et al. 2017; Bisschop, Lohse, and Setter 2021), and the lack of an appropriate outgroup for many non-model species limits their utility. Ancient DNA is another useful avenue for detecting old sweeps (Key et al. 2016; Marciniak and Perry 2017), but it too is of limited availability, including in humans, especially beyond 10,000 years.

Largely-unknown old selective events set the stage for modern genomic diversity and phenotypes, and the importance of ancient selection for modern fitness and future adaptation is becoming more evident. The application of sweep detection methods that have poor power before 2,000 generations ago has the potential to cause false negatives and the inference of a lack of positive selection when it may indeed have existed. These limitations potentially impact our understanding of long-term patterns of positive selection across the genome. For example, we may wish to build a map of positive selection across the genome of a species of interest, integrated across time and populations. This would include knowledge such as variation in the rate of sweeps over time, contributions of adaptive introgression to the landscape of genomic diversity (eg. (Dannemann and Racimo 2018; Enard and Petrov 2018), relative contributions of selection on protein-coding and regulatory regions, functions of genes most commonly under selection, and prevalence of hitchhiking accompanying those sweeps (King and Wilson 1975; Carroll 2005; Halligan et al. 2013). Using this combined knowledge of the patterns and impacts of selective sweeps allows us to further explore build on observations the importance of genes involved in metabolic pathways (eg. Akey 2009; Hallmark et al. 2019; Choin et al. 2021) and pathogen adaptation (eg. Enard et al. 2016; Shultz and Sackton 2019; Choin et al. 2021) to the major historical ecological drivers of adaptation in humans and other vertebrates.

The most power and promise to detect sweeps with typically weaker signals, such as old, weak, and partial or soft sweeps, comes from machine learning, including methods that combine multiple statistics and, more recently, ancestral recombination graphs. While using ancestral recombination graphs (ARGs) in a deep-learning framework is a promising method to detect sweeps and estimate coefficients (Hejase et al. 2022), it is likely to suffer at older time scales as increased uncertainty in ARG reconstruction might begin to limit its power. It is currently

uncertain how great this effect is likely to be. Combining multiple statistics in machine learning (Pybus et al. 2015; Kern and Schrider 2018; Mughal and DeGiorgio 2018; Flagel, Brandvain, and Schrider 2019; Mughal et al. 2020), Approximate Bayesian Computation (Racimo, Kuhlwilm, and Slatkin 2014), or composite of multiple signals (Grossman et al. 2010; Sugden et al. 2018) approaches takes advantage of the strengths and ameliorate the weaknesses of individual statistics. Similar methods have been developed to distinguish between types of selection (Kern and Schrider 2018; Xue et al. 2021; Isildak, Stella, and Fumagalli 2021; Caldas, Clark, and Messer 2022).

In general, summary statistics developed to detect the signatures of selective sweeps fall into one of three categories: examining signatures in haplotype structure, the site frequency spectrum (SFS), and overall genetic diversity. These three types of signatures vary in how they are affected by standing variation, sweep completeness, background selection, recombination rate variation, and time since the beginning and/or end of the sweep (H. Chen, Patterson, and Reich 2010; Ferrer-Admetlla et al. 2014; Racimo 2016). For example, the power of haplotype structure-based statistics is susceptible to the breakdown of linkage disequilibrium (LD) over time, as well as sweeps from standing variation since the sweep does not start from a single haplotype (H. Chen, Patterson, and Reich 2010). In contrast, combining haplotype structure with overall diversity metrics increases the power to detect soft and incomplete sweeps (Ferrer-Admetlla et al. 2014). These approaches that combine multiple statistics can thus take advantage of these various complementary strengths and weaknesses.

Haplotype-based statistics (e.g. iHS (Voight et al. 2006), nS_L (Ferrer-Admetlla et al. 2014), iSAFE (Akbari et al. 2018), HAF (Ronen et al. 2015), and H12 (Garud et al. 2015)) are especially robust to background selection. Background selection, i.e. negative selection against deleterious alleles that reduces genetic diversity at linked sites as they are lost with the deleterious alleles (Charlesworth, Morgan, and Charlesworth 1993), can be particularly problematic because it has strong potential to confound selective sweep inference via other types of summary statistics (Stephan 2010). However, background selection does not increase the frequency of long haplotypes, as expected in genetic hitchhiking with a selective sweep (Enard, Messer, and Petrov 2014), so haplotype-based statistics are unlikely to be confounded. Thus, their inclusion in a multi-statistic based machine learning approach should increase its robustness to background selection (Schrider 2020).

Subtle signals are not limited to old sweeps but can also be caused by very incomplete sweeps, sweeps from standing genetic variation, and sweeps caused by weak selective pressure. Since genomic patterns of sweeps may include sweeps of all types, stages of completeness, and age; and because we don't typically know what type(s) of sweeps to expect during a targeted analysis (for example, in the case of an older selective pressure such as an ancient epidemic), detection methods need to be able to detect a diversity of different types of sweeps with the highest sensitivity and specificity possible. Once swept regions have been detected, other methods can be applied to determine their characteristics (eg. (Kern and Schrider 2018; Sugden et al. 2018; Torada et al. 2019; Caldas, Clark, and Messer 2022; Xue et al. 2021).

Here we develop a convolutional neural network (CNN) based machine learning method that is designed to detect diverse sweeps. This includes sweeps between 2,000 and 5,000 generations old, as well as those with subtle signals caused by other factors, including those that are very incomplete, those that started from a high level of standing variation, and those that are relatively weak. This method, Flex-sweep, makes this possible through the combination of six existing statistics and five new haplotype-based statistics, all with different strengths to detect selective sweeps with different characteristics.

These five new statistics (*hapDAF-o, hapDAF-s, Sratio, lowfreq, and highfreq*) take advantage of the hitchhiking to higher frequency of regions linked to an adaptive allele by comparing diversity between ancestral and derived haplotypes. This is the structure that was first used by the DIND statistic (Derived Intra-allelic Nucleotide Diversity, Barreiro et al. 2009) using π (pairwise genetic diversity) within a fixed window size around a focal variant. These new statistics instead calculate the number of shared variants between ancestral and derived backgrounds (*hapDAF-o* and *hapDAF-s*), the number of variants in a haplotype (*Sratio*), and the frequencies of derived alleles (*highfreq* and *lowfreq*). The pair of *hapDAF* statistics are similar to the Difference in Allele Frequency (DAF) statistic (Carneiro et al. 2014) in that they compare allele frequencies across two haplotypes to detect regions that are very different. But while the original DAF statistic compares haplotypes between populations, *hapDAF* compares the ancestral and derived haplotypes. *hapDAF-o* is particularly important because it has good power to detect old sweeps, including old, incomplete sweeps.

Flex-sweep doubles reliable sweep detection time from 2,000-3,000 generations ago (e.g. Sugden et al. 2018; Harris and DeGiorgio 2020) to at least 5,000 generations ago, and up to 10,000 generations ago for strong, hard sweeps. In addition, it accurately detects sweeps of a diverse range of time (including recent sweeps), completeness, initial allele frequency, and selection strength, making it a versatile tool. We show its robustness to demographic model misspecification, background selection, and recombination rate variation, all vital considerations for inference in any population genomic context (Johri et al. 2021). This expands our ability to study genome-wide patterns of selective sweeps, as well as the origins of phenotypically important loci, including adaptations to major historical selective pressures such as diet, climate, and ancient epidemics. Because Flex-sweep requires only one population being sequenced, we expect it to be particularly useful in the context of non-model species where outgroups with population-level sequencing are often not available.

NEW APPROACHES

Here, we present a novel selective sweep detection method that provides a versatile tool to identify genomic windows involved in a selective sweep up to 5,000 generations old, or a strong, hard sweep up to 10,000 generations old. This convolutional neural network-based method, Flex-sweep, uses a variety of types of statistics. In particular, the haplotype-based statistics provide robustness to background selection, a common confounder of sweep detection methods.

We introduce five new statistics designed to detect old, soft, and incomplete sweeps, which have sensitivities that complement those of existing statistics in detecting sweeps of different types. It is this combination of statistics that makes this method versatile in detecting sweeps of a variety of ages, completeness, starting allele frequency, and strengths. The combination of these statistics in a convolutional neural network, with an architecture optimized for versatility in sweep detection, makes this tool an all-purpose sweep detection method that is robust to background selection, recombination rate variation, and demographic model misspecification.

This versatility makes our new approach complementary to existing sweep classifiers and particularly useful for studies of non-model species for which we do not have strong empirical expectations of the characteristics of sweeps present in the genome. It can be used in conjunction with methods designed to classify or quantify sweep characteristics (for example, complete vs. incomplete, hard vs. soft, sweep age, or selection strength): Flex-sweep can first detect sweeps with high confidence, increasing the proportion of all sweeps detected in the genome; then the characteristics of these sweeps can be determined with greater certainty.

Flex-sweep is available at https://github.com/lauterbur/Flex-sweep. A singularity container and pre-trained equilibrium and Yoruba models has been deposited in Zenodo at DOI 10.5281/zenodo.7313705.

RESULTS

Flex-sweep combines 11 summary statistics, including five newly-defined statistics that we show have good power to detect selective sweeps with certain characteristics that complement the power of existing summary statistics. We show that the application of these statistics in a convolutional neural network framework can detect selective sweeps with subtle signals, and is robust to common confounders such as demographic model misspecification, recombination rate heterogeneity, and background selection.

New Statistics

We present five new statistics designed to detect the signatures of diverse sweeps and sweeps with attenuated signals designed to complement the published statistics (table 1).

hapDAF

The haplotype-derived allele frequency (*hapDAF*) statistic compares the frequency of shared variants on ancestral and derived backgrounds at a locus (see **Methods** for detailed description). The total frequency of the variant on both ancestral and derived backgrounds is set to a threshold tuned to focus the statistic on the goal to either identify older, incomplete sweeps (*hapDAF-o*), or more recent, incomplete sweeps from standing variation (*hapDAF-s*). Frequency thresholds were tested at 5% intervals. For *hapDAF-o*, the total frequency threshold with the most power to identify older, incomplete sweeps is a minimum of 25%, with the frequency of the variant on the ancestral background less than 25% (see **Methods**). For *hapDAF-s*, the total frequency threshold with the most power to identify more recent incomplete sweeps is a minimum of 10%, with the frequency of the variant on the ancestral background less than 10% (see **Methods**).

hapDAF-o has good power at selection times back to at least $\tau = 0.1 * 4N_e$ generations ago at $s = 0.01$ (Supplemental figure 1. This includes incomplete and soft sweeps. In comparison, *hapDAF-s* has more power to detect recent sweeps ($\tau = 0$) (Supplemental figure 2).

Sratio

The segregating sites ratio (*Sratio*) compares the number of variable sites around a focal derived allele with the number of variable sites around the corresponding ancestral allele (see

Methods for detailed description). This is designed to detect incomplete and soft sweeps. *Sratio* has comparable power to other statistics designed to detect sweeps under similar scenarios, including HAF, DIND, iHS, and nsl (Supplemental figure 3).

freq

The two *freq* statistics are designed to detect incomplete sweeps by looking for an excess of low or high-frequency alleles on the derived background (see **Methods** for detailed description). They complement each other by focusing on different ends of the site frequency spectrum (SFS). Frequency thresholds for derived backgrounds with the derived variant were tested at 5% intervals. For *lowfreq*, the frequency threshold with the highest power to detect recent, incomplete sweeps is a ceiling at 25%, while for *highfreq* it is a floor at 25%.

highfreq and *lowfreq* have the power to detect a variety of incomplete sweeps. This includes some power at weak (s = 0.005), older sweeps (τ = 0.05) that is comparable to that of existing statistics. While the power of *lowfreq* does not exceed that of existing statistics, it should add power in the synergistic context of a neural network or applications such as Approximate Bayesian Computation (Supplemental figure 4, Supplemental figure 5).

CNN Structure

The use of Convolutional Neural Networks (CNNs) for sweep detection and other population genetic inference has been comprehensively described by Kern and Schrider 2018 and reviewed in Schrider and Kern 2018 and Flagel, Brandvain, and Schrider 2019. Flex-sweep calculates multiple summary statistics across a locus to create a feature vector, which is then used as input data for the CNN. We explored multiple possible feature vector scaling strategies and CNN architectures to select one with both good performance and evidence that it did not overfit the training data (see **Methods**; Supplemental figure 7). The final model architecture is described in the Methods, Appendix and available on GitHub (https://github.com/lauterbur/Flex-sweep). In addition, we found that training data sets with 10,000 neutral and 10,000 sweep simulations provided the best power, with little improvement with larger data sets.

Performance

Flex-sweep has good power to find sweeps under multiple demographic scenarios when trained on simulations from an equilibrium demography with N_e = 10,000 (Figure 1), including sweeps with an age from 0 - 5,000 generations, incomplete sweeps, sweeps from standing variation, and sweeps resulting from weak selection. Sweep age, completeness, and starting allele frequency have the strongest impacts on power, with power declining as sweep age or starting allele frequency increase or sweep completeness decreases, especially at false positive rates below 1%. In particular, it struggles at the most restrictive false positive rate of 0.1% to detect very incomplete sweeps (20 - 40% complete), very old sweeps (4001 - 5000 generations old), sweeps from high starting allele frequency (7.5 - 10%), and very weak sweeps (0.001 <= *s* <= 0.0028), but is still able to detect most sweeps at a more relaxed false positive rate of 1% (Figure 2).

With training as described above, Flex-sweep is capable of detecting sweeps up to 10,000 generations old under the Yoruba demographic model, but its scope is limited (Supplemental figure 8). In particular, sweeps must be complete, or nearly so. At a false positive rate of 0.1%, its power to detect strong (s = 0.01), hard (SAF = 0), complete sweeps is 75%, which drops to <10% for incomplete sweeps (EAF = 0.7 and EAF = 0.9). At a more relaxed false positive rate of 1%, its power increases to 84% for complete sweeps and 52% for nearly complete sweeps (EAF = 0.9), but remains very low (11%) for less complete sweeps (EAF = 0.7), (Supplemental figure 8).

Figure 1

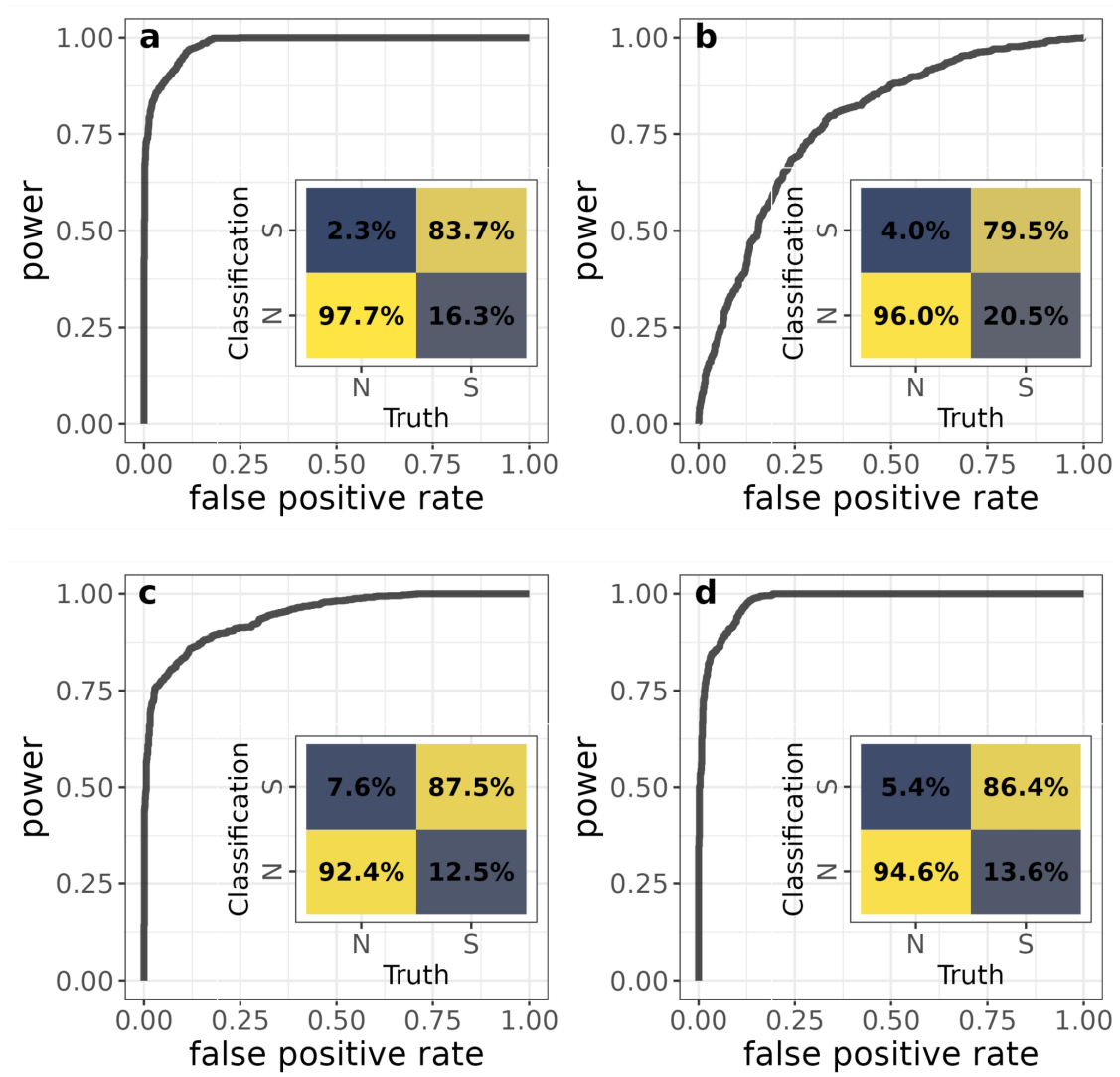

Figure 1 Power and false positive rates of the CNN trained on an equilibrium demographic scenario:

(a) Classifying simulations from the same equilibrium demographic scenario, (b) 5x decline, (c) 10x expansion, and (d) Yoruba demographic history after Speidel et al. 2019. The inset confusion matrices show, for each scenario, the classification rate of each region type (neutral or sweep, x axis) as neutral or sweep (y axis), with false positives (neutral regions classified as sweeps) in the upper left, and false negatives (sweep regions classified as neutral) in the bottom right. Darker blue indicates smaller values, brighter yellow indicates larger values.

Figure 2

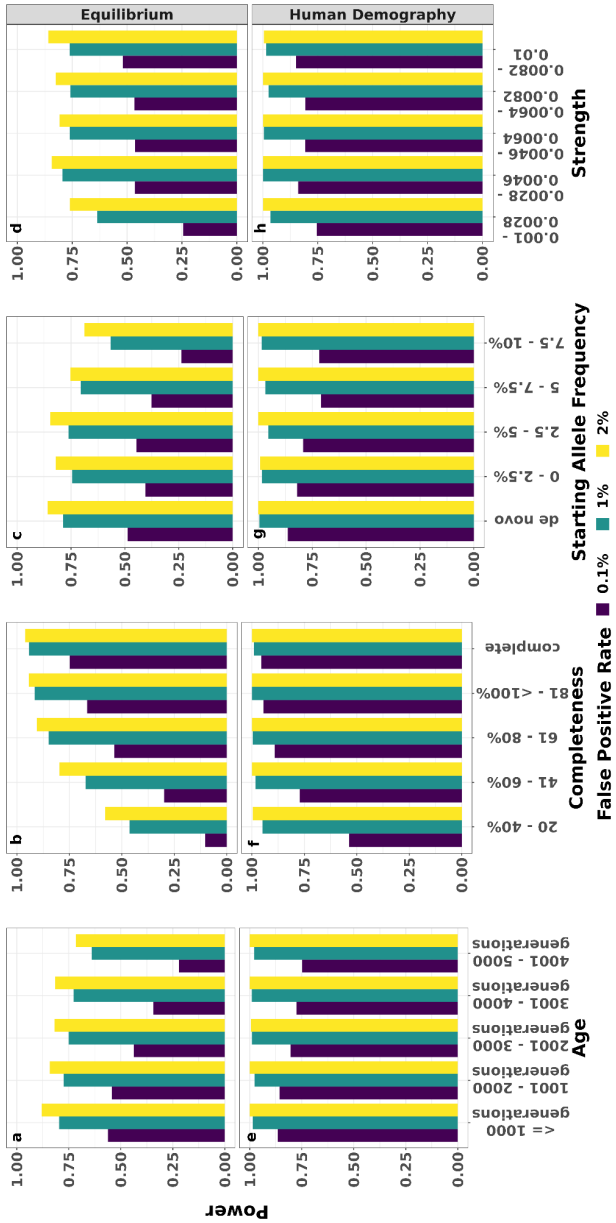

Figure 2 Power to detect sweeps of different types.

Power in a correctly specified equilibrium model (top) and the correctly specified human demographic model (bottom). Sweep types vary in age (a, e), completeness (b, f), starting allele frequency (c, g), and strength (d, h) at 0.1% (purple), 1% (teal), and 2% (yellow) false positive rates. In each panel, sweep characteristics otherwise come from the full distributions of each sweep parameter.

While the model has good power to detect sweeps even when misspecified with an equilibrium demographic history, power improved when the model was trained with correctly specified demographic models (Figure 3).

Figure 3

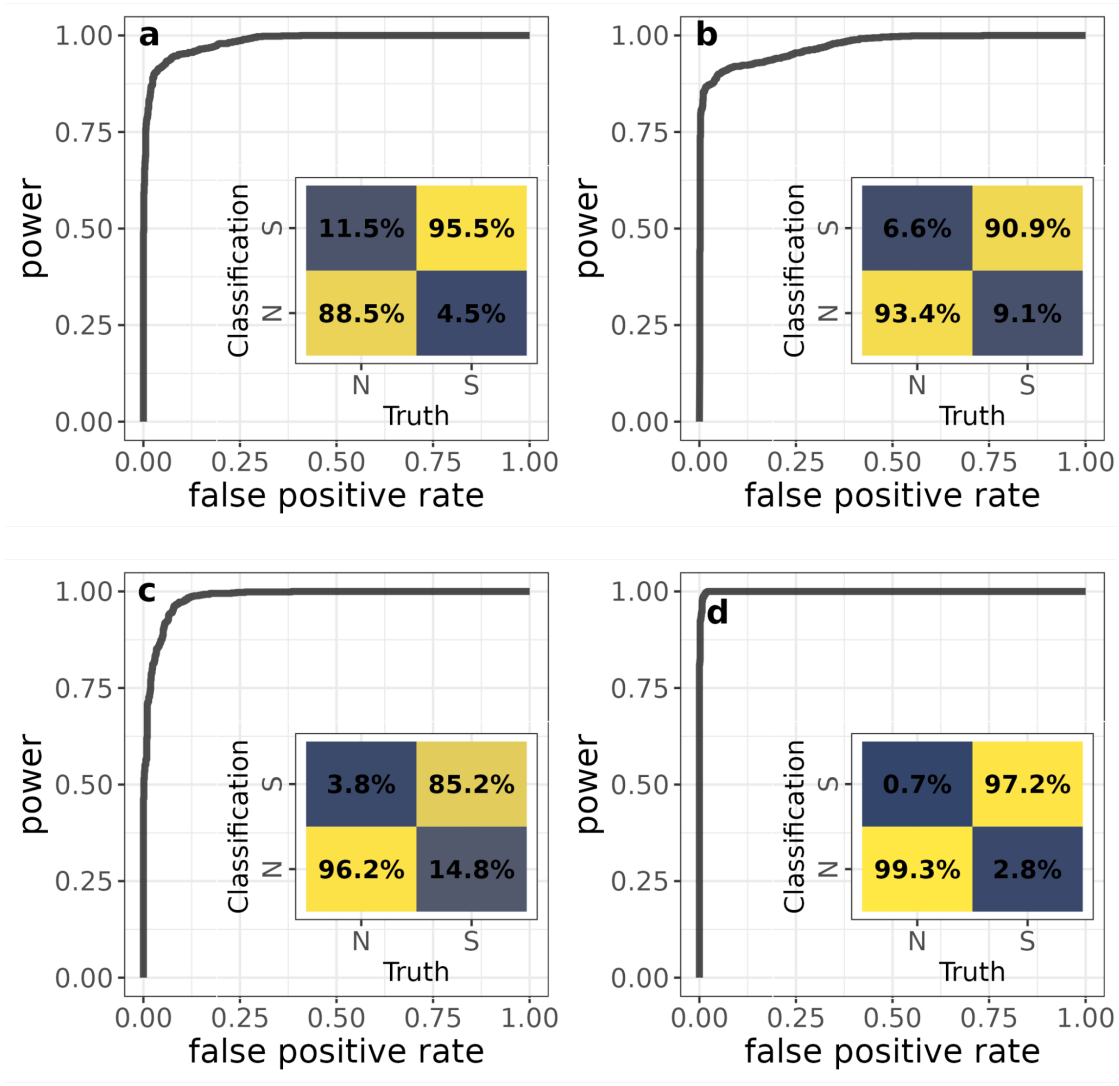

Figure 3 Power and false positive rates for Flex-sweep trained with non-equilibrium demographic scenarios.

(a) A combined data set made up of simulations with a history of population declines, expansions, and equilibrium demography, used to classify sweeps on the Yoruba population history (b) population decline (1x, 5x, 10x), used to classify sweeps on a history of 5x population decline (c) expansions (5x, 10x, 50x), used to classify sweeps on a history of 10x population expansion, and (d) the Yoruba demographic history, used to classify sweeps on the Yoruba demographic history.

To determine the robustness to recombination rate heterogeneity and background selection, we generated simulations under equilibrium demographic histories with either recombination maps subsetted from the deCODE map (Halldorsson et al. 2019; see **Methods**) with feature vectors calculated using the same recombination map; or background selection using a DFE modeled after human populations (Boyko et al. 2008) as in Schrider 2020 (DFE ~ γ(mean = -0.030, shape = 0.206; see **Methods**). Background selection increased the false negative rate (Figure 4a), but the effect on false positive rate was low (Figure 4a, b), and recombination rate heterogeneity had little effect on either false negative rate (Figure 4c) or false positive rate (Figure 4c, d). Because of the reduction in power caused by background selection, also previously observed (Schrider 2020), it may be valuable to explore the potential of creating a training data set that incorporates background selection. Such training data could improve power but is computationally intensive due to requiring forward-in-time simulations, so was not explored here.

Figure 4

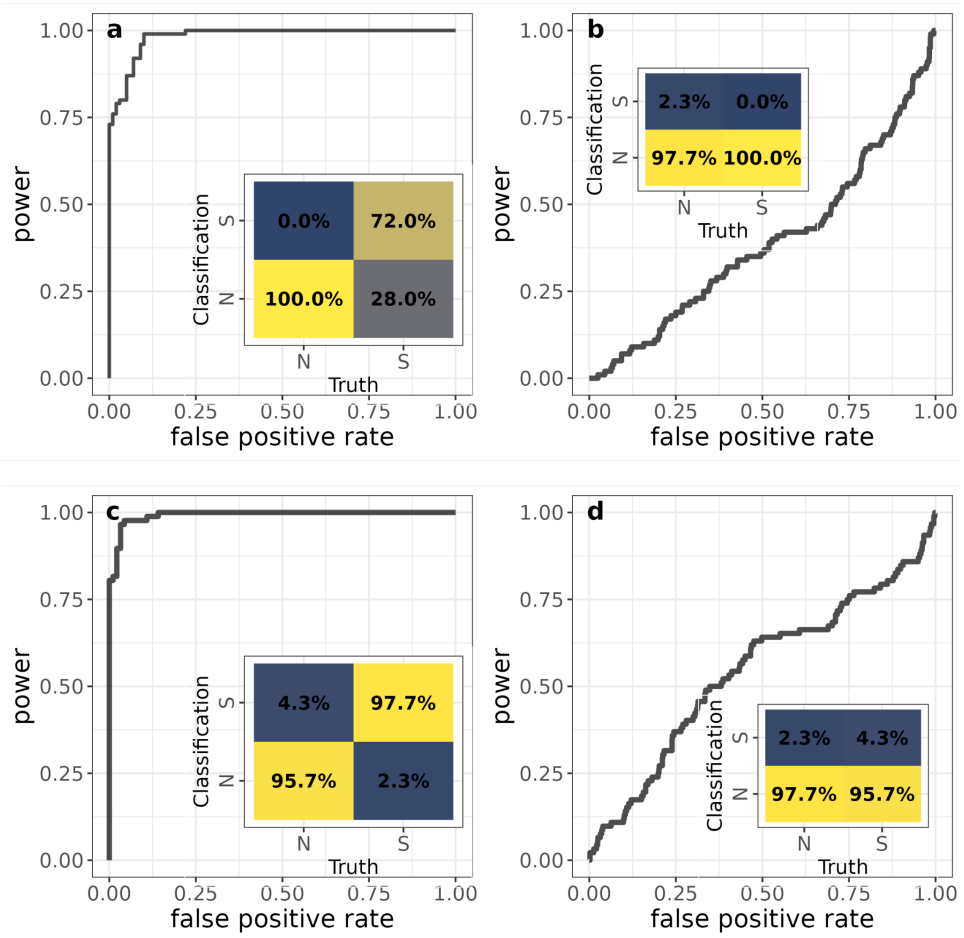

Figure 4 Power and false positive rates with background selection and recombination rate variation.

(a) The power to detect sweeps in the presence of background selection (DFE ~ γ(mean = -0.030, shape = 0.206)) and (b) the effect of background selection on false positive rate by using the CNN to classify neutral simulations with and without background selection; © The power to detect sweeps with recombination rate heterogeneity mimicking human chromosome 1 and (d) and the effect on false positive rate by using the CNN to classify neutral simulations with and without recombination rate heterogeneity. The diagonal ROC curves in (b) and (d) show that the CNN does not classify neutral simulations with background selection (b) or recombination rate heterogeneity (d) as sweeps more often that it does neutral simulations

without background selection or recombination rate heterogeneity.

We also compare Flex-sweep to diploS/HIC, another powerful CNN-based method (Kern and Schrider 2018) and one of the first to popularize CNNs for detecting selective sweeps. In this case, we include training and testing both diploS/HIC and our method using a data set with only complete sweeps because they are what diploS/HIC is designed to detect (see **Methods**). In all other respects the training and testing data sets are the same as the equilibrium data described above. In addition, because diploS/HIC is designed to detect and classify sweeps into four types (hard sweeps, soft sweeps, and regions linked to those respective types), we modified it to instead classify only the binary state - neutral or swept region. Flex-sweep has greater power and lower false positive rates than this modified, binary (sweep or neutral) diploS/HIC and, as expected because it was not diploS/HIC's purpose, this difference is more pronounced in data sets that include incomplete sweeps (Figure 5, Supplemental figure 10).

Figure 5

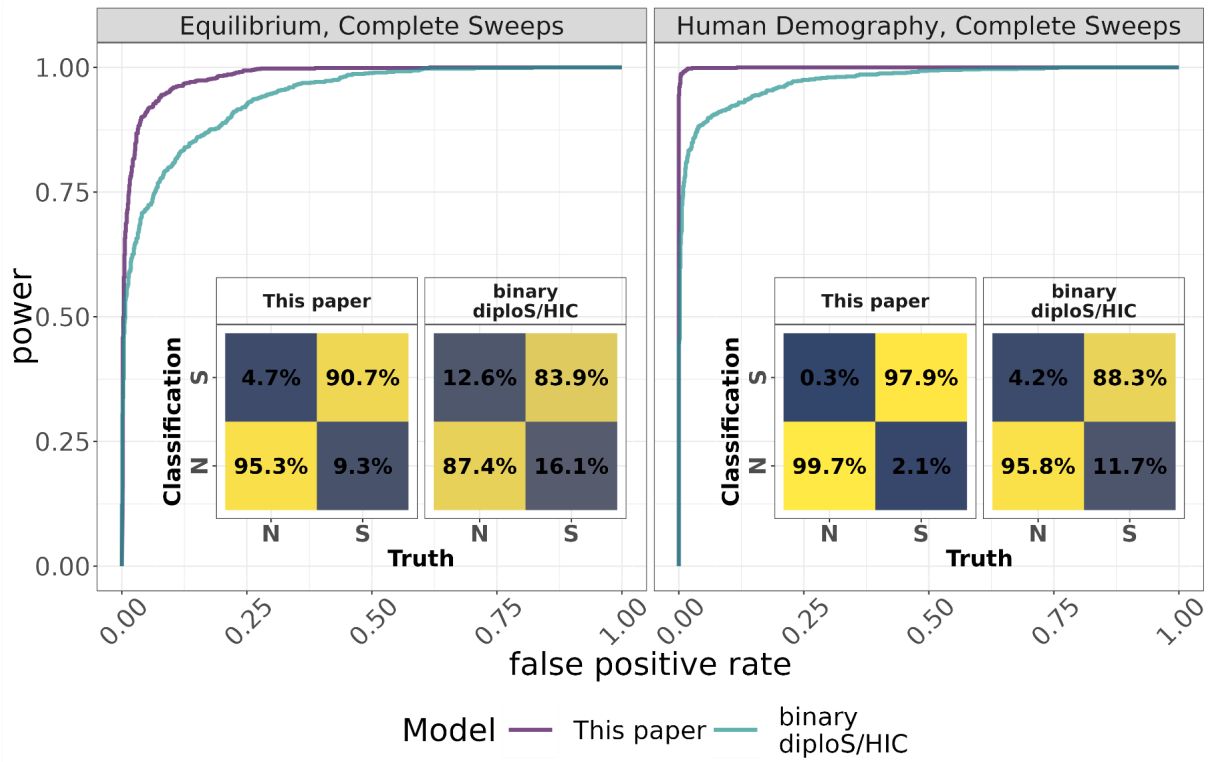

Figure 5 Power and false positive rates comparing Flex-sweep and diploS/HIC. Both were trained and tested on the same data sets using an equilibrium demography with only complete sweeps (left), and the human demography data set with only complete sweeps (right).

Application

We applied Flex-sweep to the 1000 Genomes data set of Yoruba in Ibadan, Nigeria (The 1000 Genomes Project Consortium et al. 2015). We chose this data set and population because it has been widely used in previous studies both to identify novel selective sweeps and to benchmark new selective sweep detection methods (eg. Ferrer-Admetlla et al. 2014; Racimo 2016). Because Flex-sweep is optimized to detect very old sweeps (up to 5,000 generations) of diverse types (both *de novo* and from standing variation, various levels of completeness, and a wide range of strengths, we use it to test the hypotheses that selective sweeps are expected to

occur 1. disproportionately in genic regions, and 2. disproportionately close to regulatory regions.

To understand the patterns of selective sweeps across the genome, we calculated the proportions of swept regions in genes and their distances from sites associated with regulatory activity. We find that selective sweeps are significantly more likely to be found within genes than would be expected by chance at all tested confidence thresholds (one-tailed randomization test (see **Methods**), all $p < 0.01$, Figure 6, Supplemental figure 11). Across the genome, 54% of sweeps are found within genes, while the expected mean by randomization is 27% at the default confidence threshold (Figure 6), and between 54% and 59% of sweeps are found within genes at higher confidence thresholds (Supplemental figure 11). In addition, the peaks of each sweep region (see **Methods**) are more likely to be closer to the transcription start site (one-tailed randomization test (see **Methods**), $p = 0.047$). While this pattern holds true for all confidence thresholds below a sweep probability of 1 (Supplemental figure 12), it is only significant at the 0.5, 0.6, and 0.95 confidence thresholds ($p_{60}= 0.045$, $p_{70} = 0.093$, $p_{80} = 0.081$, $p_{90} = 0.081$, $p_{95} = 0.048$, $p_{99} = 0.096$, $p_{100} = 0.227$). Given the reduced number of sweeps at higher thresholds and the sub-200bp differences between the observed means and the lower threshold of the expected means, it may require a higher resolution run of Flex-sweep (i.e. using a smaller step size between windows) to appropriately resolve the significance of these patterns. The peaks of each sweep region trend toward being closer to the 5' and 3' UTR start and end sites at confidence thresholds up to 0.99 as well, though these are not significant ($0.05 < p <= 0.204$), with the exception of the 3' UTR start and end sites at the 0.95 threshold ($p = 0.032$ and $p = 0.209$). Sweeps are not more likely to be close to transcription end sites than expected by chance (all $p > 0.05$ Figure 7, Supplemental figure 12).

Figure 6

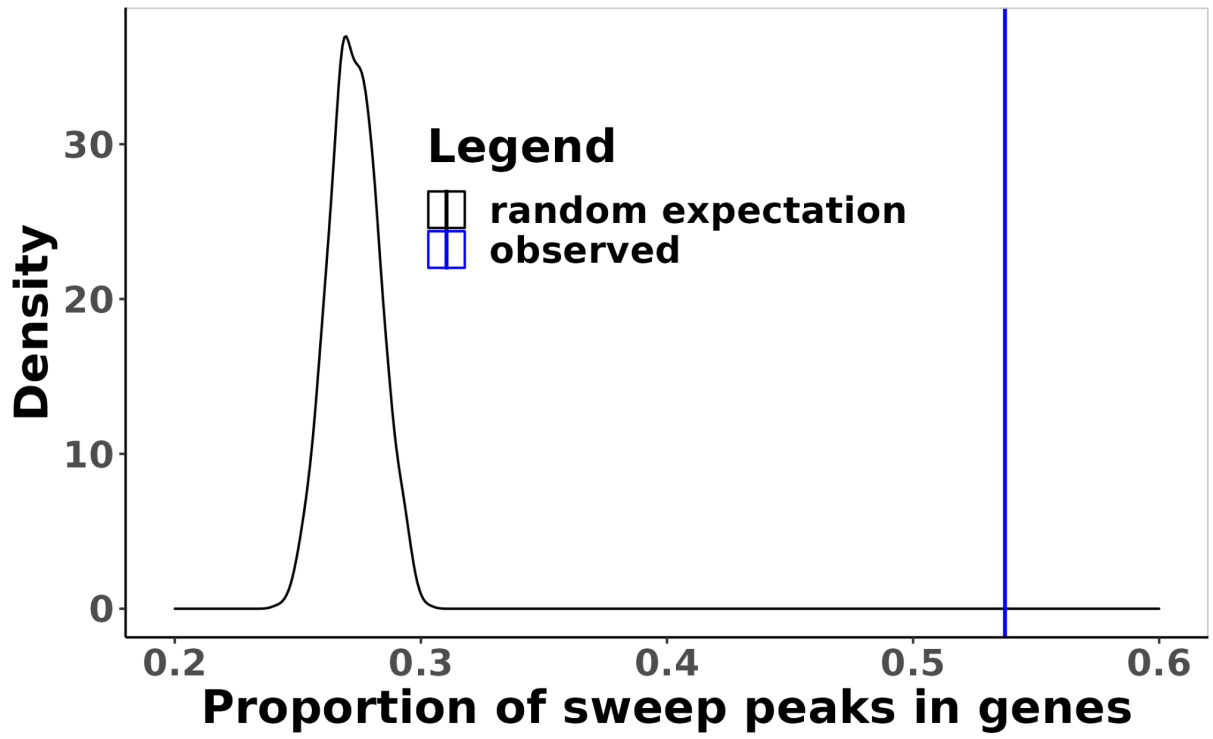

Figure 6 The proportion of sweeps inside genes.

The observed proportion of selective sweeps within genes (blue line) compared to the distribution expected by randomization (black line) at the default classification confidence threshold of 0.5.

Figure 7

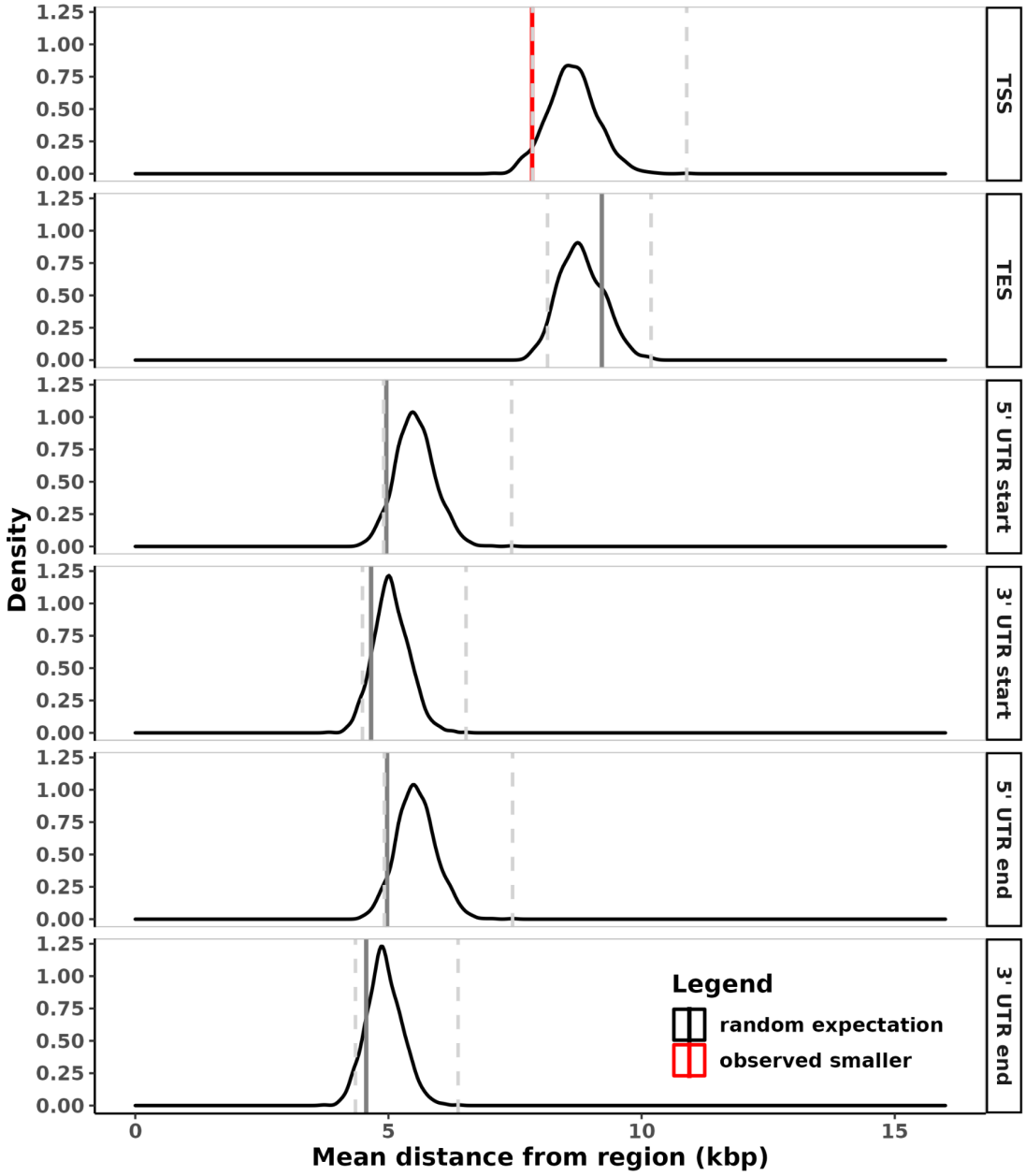

Figure 7 The distance of sweep peaks to regulatory regions.

The observed mean distance (red line) of the peak of each swept region to the nearest start and end site of each type of regulatory region: transcription, 5' UTR, and 3' UTR; compared to the distribution expected by randomization (black line) at the default classification confidence threshold of 0.5.

In addition, we estimated a false discovery rate (FDR) metric to demonstrate Flex-sweep's performance by comparing the proportion of windows classified as sweeps in the Yoruba data to the false positive rate from the validation tests using the Yoruba demographic model. The FDR is then calculated from these two rates as FDR = # false positive windows / total # positively classified windows (see **Methods**). Because Flex-sweep provides a classification confidence score for each window as a probability of containing a swept region or neutral, we can additionally tune the FDR to be more strict by increasing the confidence score at which a window is classified as a sweep. We can therefore use the confidence score to determine the threshold that would correspond to a tolerance for a specific false discovery rate.

72,749 of 258,665 windows (28%) were classified as sweeps at the default confidence threshold of 0.5, and 11,659 were unclassifiable. Windows are unclassifiable when there are too few SNPs to calculate one or more of the statistics in the feature vector over one or more window size. Most of those that were unclassifiable were near the centromeres of the chromosome. 1,654 swept regions were found, quantified as the number of unbroken streaks of windows classified. These regions, and the genes they contain, are provided in supplemental table 3, and make up 23% of the total genome (with the length of a region calculated from the peak of the first window classified as a sweep to the peak of the last). During training on the human demographic model, Flex-sweep had a 0.7% false positive rate (FPR = number of false positives / total number of positive classifications). The number of false positive windows in the Yoruba data is, therefore 1,214, giving a FDR = 1,214/72,749 = 0.017, or 1.7% (see **Methods**). Applied to the number of swept regions, we find that only 28 of the 1,654 regions classified as sweeps are likely to be false positives. Obtaining a false discovery rate of 1% would require increasing the confidence threshold to approximately 0.75 (Supplemental figure 13).

We must nevertheless consider the likely possibility that the false discovery rate may well be much higher when scanning actual sequencing data compared to simulation estimates. The proportion of the human genome classified in a sweep is indeed arguably high, but this may reflect the increased power to detect regions of the genome that overlap diverse types of sweeps, with more sensitivity and the ability to detect sweeps further back in time. Many windows also likely represent the impact of hitchhiking further away from the actually selected variants due to the increased detection power. When looking only at windows with the strongest

possible sweep classification probability of one, the percentage of genomic windows with such decisive sweep evidence falls from 28% to 1.9% (Supplemental figure 14).

To further test if Flex-sweep does identify true positive sweeps in the Yoruba genome, given the high proportion of the genome classified in sweeps, we compared the occurrence of sweeps at genes that interact with viruses (noted VIPs for Virus-Interacting-Proteins), where a strong excess of sweeps has previously been reported (Schrider and Kern 2017; Enard and Petrov 2020). If the sweep classifications in the Yoruba genome are over-run with false positives, we should not be able to detect an excess of sweeps at VIPs, because a high false positive rate would erase the true difference between VIPs and the rest of the genome. Using the same approach as reported in Enard and Petrov 2020, we find a very substantial sweep enrichment at VIPs (Figure 8). The observed enrichment is highly unlikely, with an estimated false discovery rate of 0.0006. The FDR estimation approach was previously described in Enard and Petrov 2020; Di et al. 2021. The enrichment is in particular greater than three-fold for sweeps with a very strong classification probability equal to one. This shows that, even if the false discovery rate may well be higher than estimated from simulation-based false positive rates, a substantial proportion of the sweep classifications have to be true positives. The very strong enrichment of VIPs in high-confidence sweep classifications is also in line with previous observations that viruses tend to drive strong selective sweeps reflecting strong selection (Enard and Petrov 2020; Souilmi et al. 2021).

Figure 8

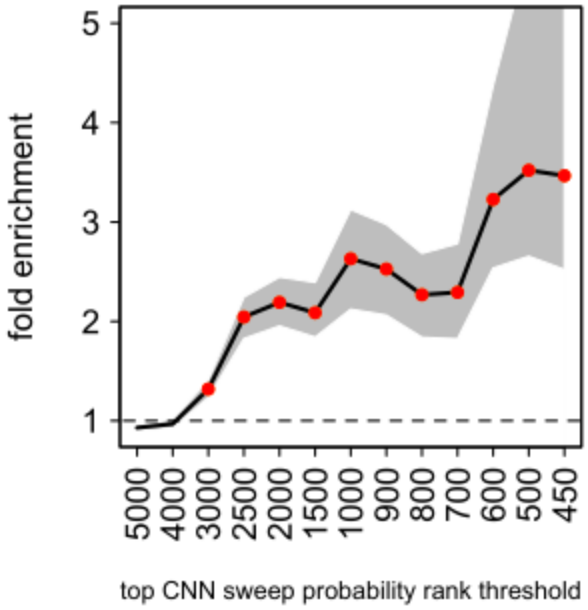

Figure 8 Flex-sweep sweep enrichment at VIPs relative to control non-VIPs.
The number of VIPs in sweeps is compared to the average number in random control sets of non-VIPs for increasing CNN sweep probability rank thresholds. The method to compare VIPs and non-VIPs, and the confounding factors taken into account, are already extensively described in (Enard and Petrov 2020). Grey area: 95% confidence interval of the fold enrichment. Red dots: fold enrichment test P<0.001.

Furthermore, we recover the known sweep at the high recombination locus of the Duffy Antigen Receptor for Chemokines gene (*DARC*), which is also the receptor for the malaria parasite *Plasmodium vivax*. This gene has been previously implicated as the target of a near-complete selective sweep, with the null (nonexpression) allele conferring malaria resistance (McManus et al. 2017). We find that a 430 kbp region surrounding the *DARC* locus and the causative promoter SNP rs2814778 is affected by a selective sweep at the default sweep confidence threshold of 0.5. At the more stringent threshold of 0.99, the region narrows to 370 kbp and still

includes the *DARC* locus and rs2814778 (Figure 9). The edges of the sweep correspond to recombination hotspots 5' and 3' from the *DARC* locus.

Figure 9

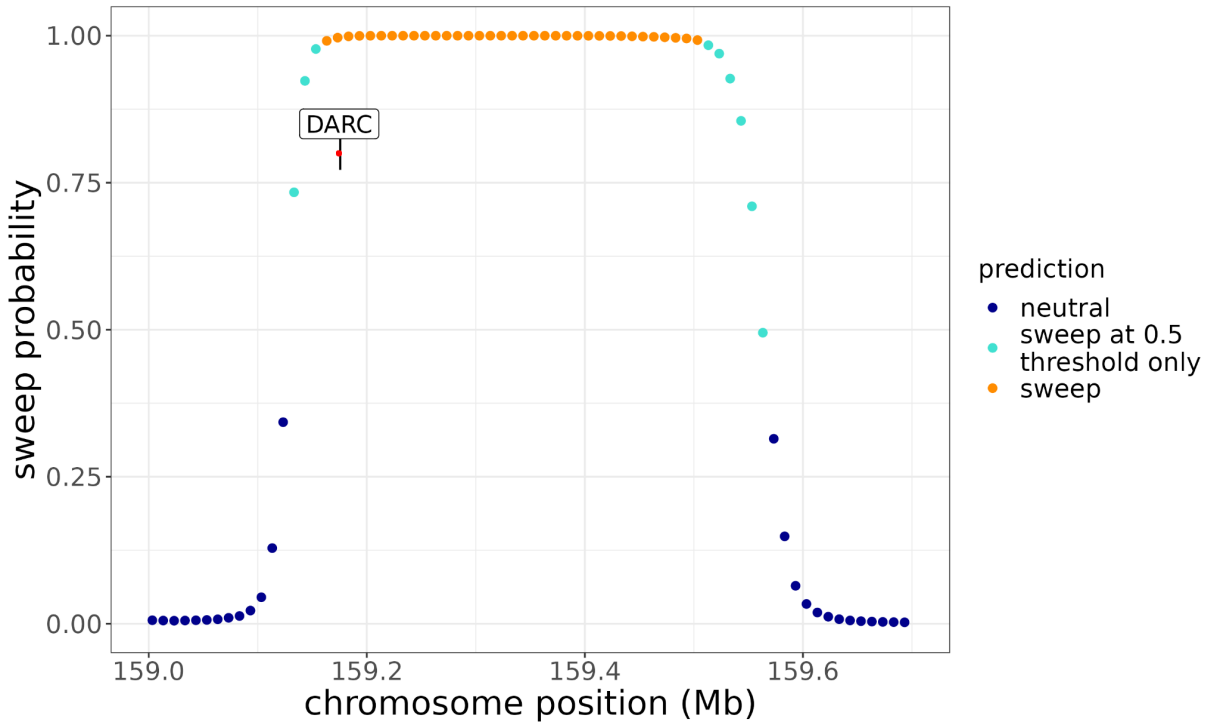

Figure 9 The region around the DARC gene on chromosome 1, using hg19 coordinates. The DARC gene is shown in black, with the causative polymorphism site rs2814778 shown as a red point. The entirety of the DARC gene, including rs2814778, lies within the swept region classified with high confidence by Flex-sweep.

DISCUSSION

Flex-sweep is a new and versatile deep-learning based method for detecting diverse selective sweeps of particular practical interest to studies of adaptation in non-model organisms. Because Flex-sweep is designed to detect sweeps of diverse types and characteristics, it is ideal for species for which we have few empirical expectations for the nature and types of selective sweeps to be detected. We show its power to detect sweeps under multiple demographic models, including demographic model misspecification, as well as its robustness to background selection and recombination rate heterogeneity. In addition to its broad versatility, it can detect sweeps at least 5,000 generations old, and up to 10,000 generations in limited cases, greatly expanding the potential to study the genomic legacy of ancient natural selection. We demonstrate Flex-sweep by identifying sweeps genome-wide using the 1000 genomes project data set of Yoruba in Ibadan, Nigeria (The 1000 Genomes Project Consortium et al. 2015) and characterize the patterns of sweeps relative to genic and gene regulatory regions, and at virus-interacting proteins (VIPs). We recapitulate previous results that sweeps are most likely to occur in genic regions and close to gene regulatory regions, as well as in VIPs.

Supervised machine learning methods have become a popular approach for detecting patterns in population genetic data, including selective sweeps (Kern and Schrider 2018; Xue et al. 2021; Torada et al. 2019), recombination (Gao et al. 2016; Adrion, Galloway, and Kern 2020), population demographic history and structure (Wang et al. 2021), and population assignment (K. Chen et al. 2018; Battey, Coffing, and Kern 2021). These methods typically benefit from the use of multiple powerful summary statistics to take advantage of their various complementary strengths and weaknesses. This gives such methods the particular potential to detect diverse patterns, including the broad diversity of selective sweeps present throughout the genome. However, most methods to detect sweeps focus on one or a few types, for example, hard vs. soft sweeps or partial sweeps (Kern and Schrider 2018; Xue et al. 2021). While distinguishing between sweep characteristics provides useful information, it may reduce the power to detect sweeps with attenuated signals.

We show that Flex-sweep is versatile in terms of the types and ages of sweeps it is able to detect, at the expense of determining the qualities of the sweep. To determine if a locus contains signatures of a sweep, it uses 11 statistics, including five new statistics, calculated over

nested windows of five sizes (see **Methods**). These five new statistics (*hapDAF-o, hapDAF-s, Sratio, lowfreq,* and *highfreq*) have good power to fill in some detection gaps of existing statistics, and otherwise complement these statistics in the context of composite, multi-statistic detection methods such as machine learning. Calculating the statistics over nested windows of multiple sizes allows the neural network to take advantage of the differences in power statistics can have depending on window size. Since we do not attempt to distinguish between sweep types or regions linked to selected loci in order to maintain the most power to detect a diversity of sweeps, other methods can be applied subsequently to determine its nature (Kern and Schrider 2018; Xue et al. 2021), location (Akbari et al. 2018; Sugden et al. 2018), strength (Torada et al. 2019), and/or timing (Speidel et al. 2019). Flex-sweep may be combined with one or more of these methods by first identifying a swept region using Flex-sweep and then applying a more targeted method on the resulting candidate regions to further investigate their qualities. This has the potential to increase our overall power to understand these qualities of old sweeps and others with subtle signals.

A potential avenue of improvement in the future would be to include ancestral recombination graphs (ARGs) as an element of Flex-sweep, or to combine the Flex-sweep convolutional neural network (CNN) with another deep-learning approach to sweep inference through ARGs. It is possible, for example, to combine CNNs with recurrent neural networks (RNNs) such as the ARG-based RNN developed by Hejase et al. 2022. Such a combination would have the advantage of taking into account both current patterns of diversity and the historical patterns they resulted from, and typically shows improved performance over CNNs or RNNs alone (as demonstrated by Gheisari et al. 2021).

An important feature of any sweep detection method is robustness to non-equilibrium demographic scenarios. Flex-sweep is trained using user-specified neutral and sweep simulations, allowing it to account for variations in demographic and selective sweep parameters. Even in populations with complex demographic histories, such as the sequence of expansions and bottlenecks in the Yoruba demographic history, it has good power to detect sweeps up to 5,000 generations old. We also demonstrate its robustness to demographic misspecification as well as recombination rate heterogeneity and background selection. It has good power at false positive rates of less than 1% in simulations, though this is reduced by background selection. Moreover, while this method does require knowledge of derived and

ancestral alleles, it is not dependent on having an outgroup sequenced population nor admixture within the population being studied.

As expected, there is less power to detect sweeps with extremely subtle signals, including very old, weak, or incomplete sweeps and sweeps from higher amounts of standing variation. However, we test Flex-sweep under a broader parameter space than commonly used in evaluating the performance of other methods, especially with respect to sweep age (up to 5,000 generations) and strength (as low as s = 0.001, a selection strength of γ = 20 at N_e = 10,000). While it has lower power to detect such very weak sweeps under a very stringent false positive rate of 0.1%, its performance improves at a more relaxed false positive rate of 1%. In addition, the presence of population declines reduces the power to detect sweeps because bottlenecks can weaken signatures of past sweeps and are well-known to create patterns of variation that mimic selective sweeps. This makes it challenging to distinguish between a real sweep and a rapid frequency increase due to sampling noise during the bottleneck. Training Flex-sweep under a demographic model that includes population declines partially ameliorates this effect. Conversely, it is easier to detect sweeps that happened before, during, or shortly after a population expansion because the expansion decreases drift, preserving the signatures of previous sweeps. In addition, ongoing sweeps are stronger because of the larger population size. These effects too are apparent in Flex-sweep's increased power to detect sweeps under conditions of population expansion.

Flex-sweep is capable of detecting some sweeps up to 10,000 generations old. This corresponds to the approximate time period of the emergence of modern humans (Schlebusch et al. 2017). As a result, there is the potential to detect adaptations that could correspond to shared traits among the ancestors of all modern humans. However, this is limited to strong, hard sweeps that are complete, or nearly so. Signals from sweeps in which the beneficial allele has gone to fixation will persist for a much longer time, as there are fewer ancestral variants remaining on the derived haplotype which will degrade signal through the action of recombination. The *hapDAF-o* statistic in particular will be particularly influential in detecting such sweeps. It is important that Flex-sweep is both able to detect these sweeps and robust to background selection, as while some statistics (such as π) may be able to detect old sweeps, they are also sensitive to background selection. We note that the CNN model used to classify the 10,000 generation old sweeps was the same Yoruba model described above, and was

trained only with sweeps up to 5,000 generations old. It is possible that training with even older sweeps would increase this power.

We find that Flex-sweep has better power under demographic scenarios that include population expansion, including the complex human demographic model tested (Speidel et al. 2019). Simulations that include expansions have less variation in genetic diversity overall, and stronger differences between neutral and sweep regions, than do simulations of equilibrium demography. This is likely related to the increase in the number of haplotypes carrying the swept allele that occurs during an expansion, which maintains the signal of an old sweep longer than in an equilibrium simulation. This is in contrast to population declines, which increase the likelihood of loss of haplotypes containing the swept allele, and thus reduce power. In addition, the larger population size will limit the impact of genetic drift, so there is less variation among neutral simulations. This amplifies the differences between neutral and swept regions.

We also compare Flex-sweep to diploS/HIC, another powerful CNN-based method that popularized deep machine-learning based sweep detection (Kern and Schrider 2018). Flex-sweep has greater power and lower false positive rates when trained and tested on data sets including only complete sweeps, which diploS/HIC is designed to classify. Flex-sweep should be a further improvement when attempting to detect sweeps genome-wide because of its ability to detect a greater diversity of sweeps with more power. However, Flex-sweep does not attempt to classify sweep type. In contrast, diploS/HIC (and its extension partialS/HIC (Xue et al. 2021)) classify sweeps by hard or soft, complete or incomplete, and whether the window includes the swept locus itself or is linked to the swept locus. It would be advantageous to use these tools in succession: Flex-sweep allows first detecting diverse sweeps with high confidence, and diploS/HIC, and/or other methods meant to determine the properties or specific locus of a selective sweep can then be applied to the high confidence regions where sweeps have already been detected to determine the properties of these selective sweeps with greater power. In these cases, if the successor tool classifies an entire high-confidence region as neutral, we could infer that it is more likely to be a region with a sweep that the tool has difficulty detecting. If the successor tool classifies partial sections on the ends of a high confidence region as neutral, we might instead infer that it is narrowing down the location of the specific swept locus.

Potential confounding influences such as recombination rate variation have little effect on Flex-sweep's ability to detect sweeps. While regions of low recombination rate and selective sweeps both reduce local diversity, the overall patterns of genetic variation they create are different (McVean 2007). Since the CNN calculates statistics in windows of five sizes to incorporate information from a range of distances from the site under selection, the CNN is able to take advantage of these differences in overall patterns to distinguish sweeps from recombination hotspots or recombination deserts. Similarly, while background selection reduces Flex-sweep's ability to identify swept regions, it does not cause the CNN to generate false positives. This is congruent with findings that selective sweeps and background selection cause different patterns of genetic diversity, especially when haplotype structure is taken into account (Enard, Messer, and Petrov 2014; Schrider 2020). In addition, we also find the surprising result that background selection causes false negatives for swept regions. This was also found by (Schrider 2020), who proposed that it may be the result of Hill-Robertson interference (Hill and Robertson 1966) and/or reduction in diversity locus-wide, including in the regions around selective sweeps that the CNN may use as distinguishing features. It may be valuable to explore the potential of using a training data set that incorporates background selection with an appropriate DFE to increase power.

Among other loci previously identified as being subject to a selective sweep, Flex-sweep classified a large swept region around the Duffy antigen/chemokine receptor (DARC) gene, which has a null allele that is partially responsible for malaria resistance in humans (McManus et al. 2017). This region contains three other genes that have previously been identified as candidates for positive selection in humans: high affinity immunoglobulin epsilon receptor subunit alpha (FCER1a) (Reiner et al. 2012), and two olfactory receptor genes (OR10J3 and OR10J1) (Reiner et al. 2012; Fernandes et al. 2019). Despite occurring in a region of relatively high recombination (up to 4 cM/Mbp), the length of this swept region suggests that Flex-sweep retains good power in high recombination regions, and that selection in this region was likely very strong (McManus et al. 2017).

Since it is designed to be versatile in terms of the types, strengths, and ages of sweeps it can detect, Flex-sweep will be particularly useful for scanning the genomes of non-model species for which we have few expectations for the nature of selection. Since it requires only a single species, requiring no outgroups save to polarize ancestral and derived alleles, it is also useful

for species that have limited data. At present, Flex-sweep requires good genome assemblies and phased haplotypes, which is currently a limitation for non-model species. However, we are in the midst of an accelerating explosion of both available genomic data (through efforts such as the Earth Biogenome Project (Lewin et al. 2018) and the Darwin Tree of Life Project (The Darwin Tree of Life Project Consortium et al. 2022) and of long-read sequencing technologies (Amarasinghe et al. 2020; Amarasinghe, Ritchie, and Gouil 2021). Together, these are lowering the financial and technical costs of sequencing and genome assembly, making genomic data of non-model species more attainable than ever. It will be possible to extend Flex-sweep to use unphased versions of the haplotype-based statistics, such as iHS (Voight et al. 2006) and nSL (Ferrer-Admetlla et al. 2014), as demonstrated and tested in Harris, Garud, and DeGiorgio 2018; Mughal and DeGiorgio 2018; Klassmann and Gautier 2022.

In the future, loci identified using Flex-sweep can be examined with existing or new methods to identify the timing of these sweeps, further illuminating the history of natural selection in humans and model species, but especially of non-model species that are more challenging to analyze with existing methods. This information can be combined with data about the timing of climate, diet, and biogeographic shifts, among other events of interest, to identify important selective pressures that impact modern diversity.

METHODS

We present Flex-sweep, a convolutional neural network-based sweep detection method, and describe its performance and application to the 1000 Genomes Project Yoruba population data (The 1000 Genomes Project Consortium et al. 2015). Flex-sweep uses six existing and five new (described below) summary statistics to quantify signals of selective sweeps (table 1).

Table 1. Summary statistics used in Flex-sweep.

summary statistic	description	summarizes	citation
iHS	integrated haplotype score	haplotype structure	(Voight et al. 2006)
nSL	number of segregating sites by length	haplotype structure	(Ferrer-Admetlla et al. 2014)
DIND	derived intra-allelic nucleotide diversity	diversity on derived background	(Barreiro et al. 2009)
iSAFE*	integrated selection of allele favored by evolution	haplotype structure	(Akbari et al. 2018)
HAF**	haplotype allele frequency	haplotype structure	(Ronen et al. 2015)
H12	frequencies of first and second most common haplotypes, modified t use 80% identity threshold	haplotype structure	(Garud et al. 2015)
hapDAF-o	haplotype derived allele frequency (old)	SFS	this paper

hapDAF-s	haplotype derived allele frequency (standing)	SFS	this paper
Sratio	segregating sites ratio	diversity on derived background	this paper
lowfreq	low frequency alleles on derived background	diversity on derived background	this paper
highfreq	high frequency alleles on derived background	diversity on derived background	this paper

* When a window included fewer than 300 SNPs (the minimum recommended by iSAFE (Akbari et al. 2018)), the iSAFE statistic was not calculated and the SAFE statistic was used instead.
** Only the top 10% of HAF values are used, as this provides a better signal for incomplete sweeps.

Definition of New Statistics

Here, we present five new statistics designed to detect the signatures of diverse sweeps and sweeps with attenuated signals designed to complement the published statistics. This includes incomplete sweeps from standing variation and incomplete sweeps older than can be detected with existing statistics. These statistics detect the signatures of positive selection using alterations in the frequency of variants on derived vs. ancestral backgrounds. In this respect, they are inspired by the DIND statistic (Derived Intra-allelic Nucleotide Diversity, Barreiro et al. 2009) that compares π (pairwise genetic diversity) within a fixed window size around a focal variant between the ancestral and derived backgrounds. The technique of comparing derived and ancestral backgrounds with a focal variant originates with the EHH statistics (Sabeti et al. 2002), from which other statistics such as iHS (Voight et al. 2006) and nS_L (Ferrer-Admetlla et al. 2014) are derived.

During a selective sweep, the haplotype containing the adaptive allele increases in frequency as it hitchhikes with the adaptive allele. This increases the frequency of variants linked to the

adaptive allele, thus reducing variation in the derived background relative to the ancestral background. These changes can be measured by comparing the frequencies of variants present on both the derived and the ancestral backgrounds (*hapDAF*), comparing the total number of variants on the derived and ancestral backgrounds (*Sratio*), and looking for an excess of low or high frequency alleles on the derived background (*lowfreq* and *highfreq*). Each statistic is standardized against the genome-wide distribution of focal variants with similar derived allele frequencies:

$$standardized\ statistic\ =\ \frac{statistic - E_p[statistic]}{SD_p[statistic]},$$

with E_p[statistic] and SD_p[statistic] estimated from an empirical distribution of variants with the same derived allele frequency (p) as the focal variant. This has the effect of making each statistic comparable across focal variants regardless of focal allele frequencies.

hapDAF

The haplotype derived allele frequency (*hapDAF*) statistic compares the frequency of shared variants on ancestral and derived backgrounds in a fixed window size around a focal variant. It is defined as follows: A focal variant with frequency between 0.25 and 0.95 is identified, and loci separated into those containing the derived allele and those containing the ancestral allele. Shared flanking variants that meet four criteria are identified in a window around the focal variant. These criteria are 1. the flanking variant is present on at least one derived background and at least one ancestral background; 2. the flanking variant is more common on the derived background; 3. the frequency of the flanking variant on the ancestral background is below a set threshold (see below); and 4, the total frequency of the variant on both ancestral and derived backgrounds is greater than a set threshold (see below). The frequency thresholds are determined by the goal of the scan, to identify older, incomplete sweeps (*hapDAF-o*), or more recent incomplete sweeps from standing variation (*hapDAF-s*).

The frequency of shared flanking variants is compared for shared flanking variants $i \in \{1,\ 2,\ 3,\ \dots\ ,\ k\}$, following the equations

$$hapDAF\ =\ \frac{\sum_{i=1}^{k}(f_{di}^2 - f_{ai}^2)}{k},$$

$$f_{di} = \frac{\#\ derived\ backgrounds\ containing\ variant\ i}{total\ \#\ derived\ chromosomes},$$

$$f_{ai} = \frac{\#\ ancestral\ backgrounds\ containing\ variant\ i}{total\ \#\ ancestral\ chromosomes},$$

given $f_{di} > f_{ai}$ (Supplemental figure 15).

hapDAF-o: This version of the *hapDAF* statistic is optimized to identify older, incomplete sweeps, by tuning the frequency thresholds of the shared flanking variants such that $f_{di} + f_{ai} > 0.25$ and $f_{ai} < 0.25$. Only variants that have hitchhiked to high frequency will stay informative as recombination and drift break up haplotypes and eliminate rare variants. By restricting the calculation to shared flanking variants with high combined frequency across both backgrounds, and allowing moderate frequency of shared flanking variants on the ancestral background, we increase the chance of finding focal alleles on backgrounds that have undergone considerable recombination and drift through time. While this increases the power to find old sweeps, it reduces the power to find sweeps from standing variation that did not consistently increase the frequency of shared variants on a single derived background.

hapDAF-s: This version of the *hapDAF* statistic is optimized to identify recent, incomplete sweeps from standing variation, by tuning the frequency thresholds such that $f_{di} + f_{ai} > 0.1$ and $f_{ai} < 0.1$. An excess of rare variants provides a good signal of a selective sweep in the recent past, i.e. before they have been eliminated by drift. By allowing the calculation to include shared flanking variants with low combined frequency across both backgrounds and restricting it to alleles that are rare on the ancestral background, we increase the number of informative variants for recent sweeps, thus increasing the power of the statistic. This increases the chance of finding focal alleles on backgrounds that have not consistently increased the frequency of shared variants on derived backgrounds (due to incomplete sweeps that do not strongly decrease variation on the derived background, or soft sweeps that have many derived backgrounds), but have increased the frequency of shared variants that are rare on ancestral backgrounds. These thresholds were chosen based on simulation-based power analyses.

Sratio

The segregating sites ratio (*Sratio*) compares the number of variable sites in a window of fixed size around a focal derived allele with the number of variable sites around the corresponding ancestral allele. A focal variant with a frequency between 0.25 and 0.95 is identified, and loci

are separated into those containing the derived allele and those containing the ancestral allele. Variable sites are identified in a window around the focal variant, and a ratio is calculated between the number of variable sites around the ancestral focal allele and the number around the derived focal allele: $Sratio = \frac{S_a}{S_d}$ (Supplemental figure 16).

Sratio thus detects incomplete and soft sweeps by identifying sites with fewer variable sites near derived vs. ancestral alleles. Both incomplete and soft sweeps are not expected to allow a single haplotype to hitchhike to high frequency but do reduce variation on the derived background relative to the ancestral background. Therefore *Sratio* does not look for similarities in haplotypes around a derived allele but instead for small differences in the amount of variation around the focal variant, as measured by the number of variable sites. Regions around selected alleles are expected to have lower variation than those around the ancestral allele, thus a low *Sratio*.

freq

The two *freq* statistics look for an excess of low or high-frequency alleles on the derived background by calculating the frequencies of derived alleles in a window of fixed size around a focal variant. After defining a focal variant with a frequency between 0.25 and 0.95, all loci containing the derived allele of the focal variant (derived backgrounds) are retained. We expect little power to detect sweeps at variants with a frequency less than 0.25. Variable sites are identified in a window around the focal variant, and each site is classified as having a derived allele or an ancestral allele. The frequency of the derived allele of each variant is calculated and these are averaged over the window around the focal variant.

These two statistics complement each other by focusing on different ends of the site frequency spectrum (SFS). That is, these statistics take advantage of the same expected excess of low and high-frequency alleles, as do Tajima's D (Tajima 1989) and Fay and Wu's H (Fay and Wu 2000), respectively. By taking the standardized difference between two SFS estimators, these classic statistics risk losing power when the variance is high, for example due to recombination (Zeng, Shi, and Wu 2007) or sequencing error (Korneliussen et al. 2013). Instead, the *freq* statistics take advantage of genome-wide data to standardize the frequency of the derived allele on the derived background against the genome-wide distribution of focal variants with similar

derived allele frequencies, avoiding the loss of power these classic statistics can experience under conditions with high variance. These excesses of low and high-frequency alleles are expected to be most pronounced after a complete, hard sweep. Restricting the calculation of the *freq* statistics to only flanking variant frequencies on the derived background approximates this expectation by treating the derived backgrounds as if they represent a complete sweep in which the derived focal allele has been fixed.

lowfreq: This statistic is calculated for flanking variants with the derived allele in the derived background $i \in \{1, 2, 3, \ldots, k\}$, following the equations

$$lowfreq = \frac{\sum_{i=1}^{k} (1 - f_{ddi})^2}{k},$$

$$f_{ddi} = \frac{\#\ derived\ backgrounds\ containing\ derived\ allele\ of\ variant\ i}{total\ \#\ derived\ chromosomes},$$

given $f_{ddi} < 0.25$. This statistic takes advantage of the same expected excess of rare variants as does Tajima's D (Tajima 1989). Regions linked to a selected, derived allele that has gone to fixation (complete sweep) are expected to have an excess of rare alleles, and we approximate that by examining only loci with the derived background in which the derived focal allele is "fixed." Because *lowfreq* is standardized by genome-wide variation data, it does not require the same theory-based standardization as Tajima's D, which was designed at a time when genome-wide variation data was not available.

highfreq: This statistic is calculated for flanking variants with the derived allele in the derived background $i \in \{1, 2, 3, \ldots, k\}$, following the equations

$$highfreq = \frac{\sum_{i=1}^{k} (f_{ddi})^2}{k},$$

$$f_{ddi} = \frac{\#\ derived\ backgrounds\ containing\ derived\ allele\ of\ variant\ i}{total\ \#\ derived\ chromosomes},$$

given $f_{ddi} > 0.25$. This statistic takes advantage of the same expected excess of high frequency variants as does Fay and Wu's H (Fay and Wu 2000) (Supplemental figure 17). Because *highfreq* is standardized by genome-wide variation data, it does not require the same theory-based standardization as Fay & Wu's H, which was designed at a time when genome-wide variation data was not available.

Power of New Statistics

The power of these new statistics was tested by simulating sweeps using discoal (Kern and Schrider 2016) for a 1.2 Mb locus at several values of sweep age (scaled by 4N generations, $\tau \in \{0, 0.05, 0.1, 0.15, 0.2\}$), selection strength (scaled by $2Ns$, $s \in \{0.0025, 0.005, 0.01, 0.05\}$), starting allele frequency ($f_{start} \in \{0, 0.01, 0.05, 0.1\}$), and ending allele frequency ($f_{end} \in \{0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1\}$). 100 simulations were performed for each combination of parameter values at a recombination rate $\varrho = 10^{-8}$, and each of the new statistics (*hapDAF-o, hapDAF-s, Sratio, highfreq, lowfreq*) was calculated for a relevant subset of those simulations across five window sizes ({50kb, 100kb, 200kb, 500kb, 1Mb}). In addition, the frequency thresholds used to determine variant frequency thresholds for the calculation of *hapDAF* and *freq* statistics were tested at 5% intervals.

Six additional statistics with known good power to detect sweeps of various ages, strengths, softness, and completeness were also calculated for these simulations: *DIND* (Barreiro et al. 2009), *iHS* (Voight et al. 2006), *iSAFE* (Akbari et al. 2018), nS_L (Ferrer-Admetlla et al. 2014), *HAF* (Ronen et al. 2015), and *H12* (Garud et al. 2015). These were calculated for the sweep parameter space in which each has been shown to perform with good power. The window sizes with the best power for each new and comparison statistic were chosen, and true and false positive rates were calculated for each statistic over the tested parameter value combinations.

Convolutional Neural Network for Sweep Detection

The use of Convolutional Neural Networks (CNNs) for sweep detection and other population genetic inference has been comprehensively described in (Kern and Schrider 2018) and reviewed in (Schrider and Kern 2018) and (Flagel, Brandvain, and Schrider 2019). Briefly, CNNs are a form of supervised deep learning that take input images labeled by class, process these images to “learn” common distinguishing traits of images in each, and use these learned traits to classify novel images. At their heart is the concept that only certain traits of an image are useful to distinguish its class from another class. While they were developed to classify images, under the hood they are processing numeric pixel values. The key innovation in the application of CNNs to population genomic data was that the numeric values processed by the CNN need not have their origin in an image but could be the values of various summary statistics calculated across a sample of loci (Kern and Schrider 2018) or genome alignment (Flagel, Brandvain, and

Schrider 2019). These summary statistics can be ordered in a feature vector (or alignment), which is then transformed into a two (or more) dimensional array before being passed to the CNN.

Simulations

Flex-sweep is trained using simulated data. While the simulations for the analyses presented were performed using discoal (Kern and Schrider 2016)and SLiM 3 (Haller and Messer 2019), training data could be generated by any simulation method capable of simulating the sweep scenarios the user wishes to include in the training data set, provided it can be converted to ms-style output. We simulated neutral and sweep data at a single 1.2 Mb locus, with the selected site at the center of the locus, under a broad parameter space of selective and demographic scenarios. One hundred chromosomes (50 diploid individuals) were sampled for each scenario.

We generated data sets for demographic equilibrium, population decline, population expansion, population bottleneck with recovery, and a complex scenario following the Yoruba population history inferred by (Speidel et al. 2019). Each data set had 10,000 neutral and 10,000 sweeps for training, and an additional 1,000 neutral and 1,000 sweep simulations for testing, with mutation and recombination rates as described in Supplemental table 1. Mutation and recombination rates for the human demographic simulations were based on (Speidel et al. 2019). The starting and ending frequencies for the swept allele and the selection strength were also drawn from the distributions described in Supplemental table 1. In the demographic equilibrium simulation set, the sweep timing was drawn from the distribution as described in Supplemental table 1, while in the simulations with population size changes, the sweep timing was drawn from the distributions described in Supplemental table 2. Simulations were also conducted to vary the time of sampling after the decline, expansion, or recovery, as well as the selection timing with respect to the demographic event(s) (before or during).

We also simulated data sets with 100 neutral and 100 sweep simulations in SLiM 3 (Haller and Messer 2019) under scenarios with recombination rate heterogeneity and background selection (DFE ~ γ(mean = -0.030, shape = 0.206)), with 75% deleterious and 25% neutral mutations, as described in (Schrider 2020). To generate simulations with recombination rate heterogeneity

similar to that observed in humans, we subsetted 100 random 1.2 Mb regions from chromosome 1 of the sex-averaged deCODE map (Halldorsson et al. 2019)) and provided these to SLiM 3.

Feature Vectors

Feature vectors represent the summarized variation at each simulated locus, with one vector per simulation. We divide each simulated 1.2 Mb locus into a series of 21 center points spaced at 10kb intervals, each with nested windows of five different sizes ({50kb, 100kb, 200kb, 500kb, 1000kb}) around the center point. We then calculate 11 summary statistics to summarize the characteristics of the sample with respect to site frequency spectrum (SFS), haplotype structure, and diversity on the derived background (Table 1) in each of the five nested windows.

Nested windows are used because sweep signals vary in strength at different distances from the focal variant. For example, recombination reduces signal more as the distance from the selected variant increases, so large windows are expected to have more noise and lower signal for older sweeps. However, for recent and strong sweeps, large windows include useful information and increase the power to identify these sweeps. By including windows of multiple sizes, the CNN is able to explore these aspects of sweep signals, retaining information that helps it to classify sweeps and discarding information that doesn't.

These statistics are normalized against the genome-wide distribution of focal variants with similar derived allele frequencies (as described in (Voight et al. 2006)) using the neutral simulations, with the exception of *HAF* and *H12*. These two statistics take a single value across the entire locus because they require a larger window to define haplotypes, so this value is repeated for each center point and window combination to maintain appropriate dimensions of the feature vector.

These normalized summary statistics are combined in a feature vector 1155 elements long, by statistic, center point (i), and window size (j), with each statistic, center point, and window represented by $stat_{i,j}$:

$$[iHS_{1,1}, iHS_{1,2} \dots iHS_{2,1}, iHS_{2,2} \dots iSAFE_{1,1}, iSAFE_{1,2} \dots H12_{21,3}, H12_{21,4}, H12_{21,5}]$$

Each feature vector is then translated into a two-dimensional image, with values of each statistic as a row, and each center point/window size combination as a column (Figure 10). Feature vectors can be generated with or without a recombination map.

Figure 10

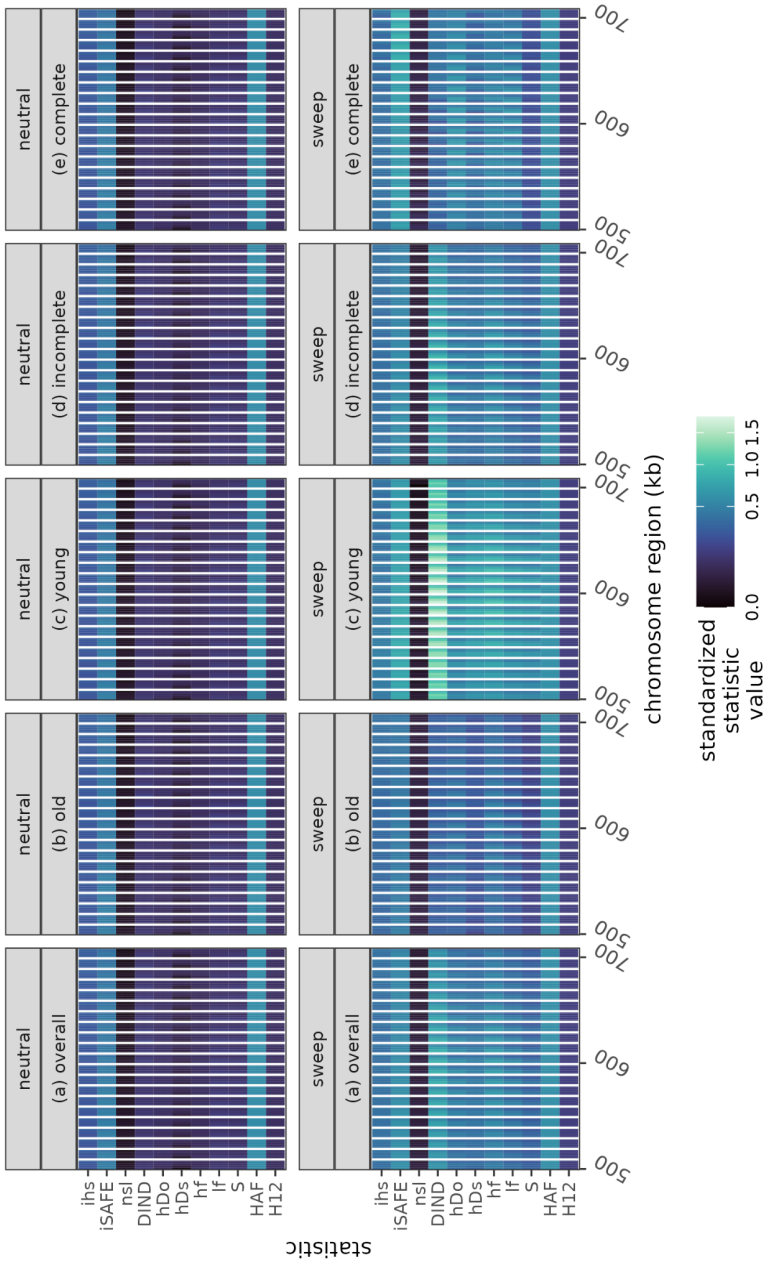

Figure 10. **Visualization of average feature vectors.**

Neutral (top) and sweep simulations (bottom), across (a) all combinations of parameter values, (b) sweeps 2000 to 5000 generations old, (c) sweeps up to 2000 generations old, (d) incomplete sweeps, and (e) complete sweeps. Each feature vector is represented as an image, with statistics as rows and center point/window size combinations as columns.

CNN Structure and Testing

To develop the final model, we tested a variety of model architectures (layer and filter numbers and sizes) in combination with a number of preprocessing steps, learning functions, activation functions, and regularization techniques. These strategies can interact in unpredictable ways, so combinations were tested comprehensively to find those that produced the best classification results, and thereafter specific parameters of each strategy were tuned to further improve classification and identify the best model.

We trained each model using sets of simulations as described above and tested them with balanced data sets of 1,000 neutral and 1,000 sweep simulations simulated under the same conditions (except for simulations for recombination rate heterogeneity and background selection, for which only 100 neutral and 100 sweep simulations were generated due to computational limitations). Models were tested under conditions of correct demographic specification (same demographic model for training and testing data), demographic misspecification (model tested with data generated under a different demographic scenario than it was trained on), and mixed demographic specification (model trained on data generated from a combination of demographic scenarios). We also examined the effect of misspecification of the timing of the sweep relative to the demographic event by testing models with sweeps occurring during or after a demographic event. Models were trained with 20% of the data held back for in-model validation. Once trained, the model was used to classify data from each testing data set.

We calculated true and false positive rates, accuracy, and precision for each combination of model architectures and strategies and constructed receiver-operating characteristic (ROC) curves for visualization and area under the curve (AUC) calculations. Initial evaluation was done by comparing ROC curves, and strategy combinations that consistently performed worse were discarded.

All CNN code was implemented in Python (3.6) using Tensorflow 2.4 and Keras 2.1.3. A singularity container and two pre-trained models (equilibrium and Yoruba) will be available at Zenodo, DOI 10.5281/zenodo.7313705..

Feature vector preprocessing

The characteristics of the feature vector, such as its variance among and between features, can have a significant effect on the ultimate performance of the CNN. Preprocessing includes strategies such as normalization and standardization (Bishop 1998). Multiple preprocessing strategies were tested to explore the impact of different types of feature vector rescaling and standardization on network success. All statistics except HAF and H12 were already of a similar order of magnitude because of the initial normalization against the genome-wide distribution of focal variants with similar derived allele frequencies. HAF could take on a much larger range of values, so feature vectors also were tested in three additional ways: with HAF/H12 normalization, with HAF scaled by a factor of 10 to bring its range of variation more in line with that of the other statistics, and without HAF and H12. This last is because HAF and H12 each take on a single value for the whole locus, repeated multiple times in the feature vector to maintain appropriate dimensions. Thus it is possible that these two statistics could overwhelm the signal from the other statistics.

The impact of rescaling feature vectors was tested in the following ways: Feature vectors without any rescaling, feature vectors rescaled proportionally such that each vector fell within the range 0-1 (both forcing all statistics to the same order of magnitude and to positive values which can be more compatible with some activation functions), and feature vectors with each statistic value rescaled by the mean and standard deviation of: 1. that statistic across all center points and window sizes; 2. that statistic across all center points, within window sizes; or 3. that statistic across all window sizes, within individual center points.

The most effective rescaling strategies across model architectures and data sets were: no rescaling, rescaling proportionately across the entire feature vector such that each vector fell within the range 0-1 (both forcing all statistics to the same order of magnitude and forcing them to positive values which can be more compatible with some activation functions), and rescaling subsetted by each statistic, proportionally within its center point, window size, or both. While the subsetted rescaling strategies gave the best performance under a few demographic scenarios, they were also more affected by demographic scenario and model misspecification

(Supplemental figure 6). Not rescaling provided the most consistently good performance across all tested scenarios.

Network architecture

To develop the final model, we tested model architectures using the simulations described above with a variety of: number of layers (3-4), number of convolution steps per layer (2-3), number of filters per convolution step (32-256), filter sizes (2x2, 3x3, 5x5, or with dilation 1x3, 1x4, 1x5), number of pooling steps (1-2), number of fully connected layers (2-3), and number of filters in each fully connected layer (32-512). In total, 96 different architectures were tested using these combinations. After limiting the possible architectures to the top 9 performing models on equilibrium, as well as correctly and mis-specified demographic scenarios, we tested different combinations of learning functions, activation functions, and regularization techniques. Learning optimization functions included Adam, Nadam, Cosine decay with restarts, AMSGrad, and Stochastic gradient descent (SGD) with and without Nesterov momentum. Activation functions included ReLU, Leaky ReLU, and PReLU, ELU. Addition of a squeeze and excitation layer was also tested, but this added no benefit. Regularization methods included early stopping, batch normalization, and L1 (lasso), L2 (ridge), and L1L2 regularization to improve generalization with multicollinearity.

These strategies can interact in unpredictable ways (Liao et al. 2022), so combinations were tested comprehensively to find those that produced the best classification results. Strategies that were found to improve the network, as determined by comparisons of AUC, accuracy, and false positive rate, were then tested with adjustments to one or more tuning parameters, including learning rate. Tensorflow's default learning rate of 0.01 provided inconsistent performance among runs, suggesting that it struggled to find optima. Learning rates below 0.001 performed less well, suggesting that they fell into local optima. A learning rate of 0.001 performed better than 0.01 or 0.0001, and all further tests using a constant learning rate (ie. not a learning rate scheduler) used this value. Models implementing early stopping with patience set to 5 avoided overfitting (Supplemental figure 7).

The best model architectures had a few things in common: three (rather than four) convolutional layers, an increasing number of filters per step, constant filter sizes for all steps, and one pooling step for each layer. Models with two or three fully connected layers both worked well.

Training set size

The size of the training data set can have a strong effect on CNN performance, but there is a trade-off with training time, so the marginal benefit of a larger training data set size may not be worth the increased cost of training time and computational resources. We compared the method's performance when trained on data sets of four different sizes, from 2,000 neutral and 2,000 sweep simulations to 20,000 of each. The false positive rate continues to improve to 20,000 simulations, but the biggest improvement is between 2,000 and 5,000 simulations. The power increases substantially until 10,000 simulations, after which it no longer improves (Supplemental figure 9).

Validation

Models were tested and compared, as described above, by training on a data set and then used to predict the held back test data sets, including those with recombination rate heterogeneity, background selection, and the human demographic model. ROC curves and AUC, false positive rates (FPR), accuracy, and precision were compared to determine the best networks.

In addition, we compared the stability of network predictions across training data set sizes. Increasing the size of the training data set increases the quality of a network by filling in the parameter space it uses to define class-distinguishing features but it also increases training time. While the expectation that more training data will improve network quality is true for CNNs in general, the marginal benefit of more training data is likely to vary depending on the network goals and structure. We thus compared CNN prediction accuracy across training data set sizes of 2,000, 10,000, 15,000, and 20,000 feature vectors for each of the combined data sets described above.

We compared the success of Flex-sweep to diploS/HIC, one of the state-of-the-art machine learning methods designed to detect selective sweeps (Kern and Schrider 2018). diploS/HIC uses a similar CNN-based architecture to detect and distinguish among hard sweeps, soft sweeps, and regions linked to each. In order to maintain the most power to detect ancient

sweeps, we do not attempt to distinguish between sweep types or linked regions, so for appropriate comparison to diploS/HIC we modified its original classification structure from five categories (neutral, hard sweep, soft sweep, linked hard, linked soft) to binary classification (neutral or sweep).

Application

We applied our method to data from the Yoruba population data in the 1000 Genomes Project (The 1000 Genomes Project Consortium et al. 2015). We divided each autosome into 1.2 Mb overlapping windows with a step size of 10kb and calculated a feature vector for each window using the deCODE recombination map (Halldorsson et al. 2019). We used Flex-sweep, trained on the Yoruba demographic model, to classify each window as neutral or containing a sweep. Furthermore, swept regions spanning multiple windows were identified as streaks of windows classified as sweeps.

We developed a false discovery rate (FDR) metric to demonstrate its actual performance in this real-world data set by comparing the proportion of windows classified as sweeps in the Yoruba data to the false positive rate from the validation tests using the Yoruba demographic model. We use the false positive rate (FPR = # false positives / (# false positives + # true negatives)) and false negative rate (FNR = # false negatives / (# false negatives + # true positives)) calculated during training to calculate the number of windows falsely classified as swept in the data set. The FDR is then calculated from the number of these false positive windows and the total number of windows classified as a sweep as: FDR = # false positive windows / total # windows classified as swept. Because Flex-sweep provides a classification confidence score for each window as a probability of containing a swept region or neutral, we can additionally tune the FDR to be more strict by increasing the confidence score at which a window is classified as a sweep. The FDR is then the ratio of these two rates. Because Flex-sweep provides a classification confidence score for each window as a probability of containing a swept region or neutral, we can additionally tune the FDR to correspond to a tolerance for a specific false positive rate by increasing the confidence score at which a window is classified as a sweep. The default confidence threshold is 0.5. This FDR is then applied to the number of swept regions to determine how many were likely to be spurious signals.

To understand the patterns of selective sweeps across the genome, we calculated the proportion of swept regions localized within genes, as well as the distances of swept regions from transcription start and sites, as well as 5' and 3' UTR start and end sites. These serve as proxies for regions affecting regulatory and expression patterns. A swept region was considered to be within a gene if the site of the peak of the sweep probability was within the interval bounds of any gene as defined by Ensembl (release 69) (Cunningham et al. 2022). Similarly, the site with the highest sweep probability within a swept region was taken as the point from which to measure the distance to transcription and UTR start and end sites. To test for statistical significance, we randomized the positions of the swept regions we identified while maintaining their internal features (length, location of peak relative to 5' and 3' ends), and recalculated the proportions and distances of the shuffled swept regions to these genomic features 1,000 times. For each, we conducted one-tailed tests with α = 0.05 - if the observed genic proportion was larger than 95% of the genic proportions calculated with the randomized regions, the difference was considered to be significant. Similarly, if the observed distance of a swept region to a transcription, 5', or 3' start or end site was smaller than 95% of the distances calculated with the randomized regions, the difference was considered to be significant. We repeated these analyses at sweep confidence thresholds of 0.5 (default), 0.6, 0.7, 0.8, 0.9, 0.95, and 0.99.

To understand the relationship of sweeps to virus-interacting proteins (VIPs) specifically, we compared the number of VIPs under selection to the number of control non-VIPs under selection, as described in (Enard and Petrov 2020), notably using the same controls for potential confounding factors. A gene (VIP or non-VIP) was considered to be subject to a sweep if the genomic center of the gene (as calculated by the midpoint of the most 5' start and the most 3' end) was covered by at least one window classified as a sweep.

Data and Code Availability

Flex-sweep is available at https://github.com/lauterbur/Flex-sweep, and a singularity container with all dependencies is available at Zenodo (DOI 10.5281/zenodo.7313705). To facilitate running Flex-sweep, two pre-trained Flex-sweep models for the equilibrium and Yoruba demographics described here are also provided at Zenodo (10.5281/zenodo.7313705). Code to replicate simulations and analyses presented here, and for the five new statistics presented, is also available at the above repository.

Acknowledgments

M. Elise Lauterbur is supported by an NSF Postdoctoral Research Fellowship #2010884. David Enard is supported by NIH NIGMS MIRA R35GM142677. This material is based upon High Performance Computing (HPC) resources supported by the University of Arizona TRIF, UITS, and Research, Innovation, and Impact (RII) and maintained by the UArizona Research Technologies department. We thank Dan Schrider and Parul Johri for valuable advice and discussions.

REFERENCES

Adrion, Jeffrey R, Jared G Galloway, and Andrew D Kern. 2020. “Predicting the Landscape of Recombination Using Deep Learning.” Edited by Claus Wilke. *Molecular Biology and Evolution* 37 (6): 1790–1808. https://doi.org/10.1093/molbev/msaa038.

Akbari, Ali, Joseph J Vitti, Arya Iranmehr, Mehrdad Bakhtiari, Pardis C Sabeti, Siavash Mirarab, and Vineet Bafna. 2018. “Identifying the Favored Mutation in a Positive Selective Sweep.” *Nature Methods* 15 (4): 279–82. https://doi.org/10.1038/nmeth.4606.

Akey, Joshua M. 2009. “Constructing Genomic Maps of Positive Selection in Humans: Where Do We Go from Here?” *Genome Research* 19 (5): 711–22. https://doi.org/10.1101/gr.086652.108.

Amarasinghe, Shanika L, Matthew E Ritchie, and Quentin Gouil. 2021. “Long-Read-Tools.Org: An Interactive Catalogue of Analysis Methods for Long-Read Sequencing Data.” *GigaScience* 10 (2): giab003. https://doi.org/10.1093/gigascience/giab003.

Amarasinghe, Shanika L., Shian Su, Xueyi Dong, Luke Zappia, Matthew E. Ritchie, and Quentin Gouil. 2020. “Opportunities and Challenges in Long-Read Sequencing Data Analysis.” *Genome Biology* 21 (1): 30. https://doi.org/10.1186/s13059-020-1935-5.

Barreiro, Luis B., Meriem Ben-Ali, Hélène Quach, Guillaume Laval, Etienne Patin, Joseph K. Pickrell, Christiane Bouchier, et al. 2009. “Evolutionary Dynamics of Human Toll-Like Receptors and Their Different Contributions to Host Defense.” Edited by Gil McVean. *PLoS Genetics* 5 (7): e1000562. https://doi.org/10.1371/journal.pgen.1000562.

Battey, C J, Gabrielle C Coffing, and Andrew D Kern. 2021. “Visualizing Population Structure with Variational Autoencoders.” Edited by A Sethuraman. *G3 Genes|Genomes|Genetics* 11 (1): jkaa036. https://doi.org/10.1093/g3journal/jkaa036.

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

Bishop, Christopher M. 1998. “Ch. 8 Pre-Processing and Feature Extraction.” In *Neural Networks for Pattern Recognition*, 295–329. Advanced Texts in Econometrics. United Kingdom: Elsevier Science.

Bisschop, Gertjan, Konrad Lohse, and Derek Setter. 2021. “Sweeps in Time: Leveraging the Joint Distribution of Branch Lengths.” *Genetics* 219 (2): iyab119. https://doi.org/10.1093/genetics/iyab119.

Boyko, Adam R., Scott H. Williamson, Amit R. Indap, Jeremiah D. Degenhardt, Ryan D. Hernandez, Kirk E. Lohmueller, Mark D. Adams, et al. 2008. “Assessing the Evolutionary Impact of Amino Acid Mutations in the Human Genome.” Edited by Mikkel H. Schierup. *PLoS Genetics* 4 (5): e1000083. https://doi.org/10.1371/journal.pgen.1000083.

Caldas, Ian V., Andrew G. Clark, and Philipp W. Messer. 2022. “Inference of Selective Sweep Parameters through Supervised Learning.” Preprint. Evolutionary Biology. https://doi.org/10.1101/2022.07.19.500702.

Carneiro, Miguel, Carl-Johan Rubin, Federica Di Palma, Frank W. Albert, Jessica Alföldi, Alvaro Martinez Barrio, Gerli Pielberg, et al. 2014. “Rabbit Genome Analysis Reveals a Polygenic Basis for Phenotypic Change during Domestication.” *Science* 345 (6200): 1074–79. https://doi.org/10.1126/science.1253714.

Carroll, Sean B. 2005. “Evolution at Two Levels: On Genes and Form.” *PLoS Biology* 3 (7): e245. https://doi.org/10.1371/journal.pbio.0030245.

Charlesworth, B, M T Morgan, and D Charlesworth. 1993. “The Effect of Deleterious Mutations on Neutral Molecular Variation.” *Genetics* 134 (4): 1289–1303. https://doi.org/10.1093/genetics/134.4.1289.

Chen, H., N. Patterson, and D. Reich. 2010. “Population Differentiation as a Test for Selective Sweeps.” *Genome Research* 20 (3): 393–402. https://doi.org/10.1101/gr.100545.109.

Chen, Kuan-Yu, Elizabeth A. Marschall, Michael G. Sovic, Anthony C. Fries, H. Lisle Gibbs, and Stuart A. Ludsin. 2018. “ *Assign POP :* An R Package for Population Assignment Using Genetic, Non-genetic, or Integrated Data in a Machine-learning Framework.” Edited by Timothée Poisot. *Methods in Ecology and Evolution* 9 (2): 439–46. https://doi.org/10.1111/2041-210X.12897.

Cheng, Jade Yu, Aaron J Stern, Fernando Racimo, and Rasmus Nielsen. 2022. “Detecting Selection in Multiple Populations by Modeling Ancestral Admixture Components.” Edited by Yuseob Kim. *Molecular Biology and Evolution* 39 (1): msab294. https://doi.org/10.1093/molbev/msab294.

Choin, Jeremy, Javier Mendoza-Revilla, Lara R. Arauna, Sebastian Cuadros-Espinoza, Olivier Cassar, Maximilian Larena, Albert Min-Shan Ko, et al. 2021. “Genomic Insights into Population History and Biological Adaptation in Oceania.” *Nature* 592 (7855): 583–89. https://doi.org/10.1038/s41586-021-03236-5.

Coelho, Margarida, Donata Luiselli, Giorgio Bertorelle, Ana Isabel Lopes, Susana Seixas, Giovanni Destro-Bisol, and Jorge Rocha. 2005. “Microsatellite Variation and Evolution of Human Lactase Persistence.” *Human Genetics* 117 (4): 329–39. https://doi.org/10.1007/s00439-005-1322-z.

Cunningham, Fiona, James E Allen, Jamie Allen, Jorge Alvarez-Jarreta, M Ridwan Amode, Irina M Armean, Olanrewaju Austine-Orimoloye, et al. 2022. “Ensembl 2022.” *Nucleic Acids Research* 50 (D1): D988–95. https://doi.org/10.1093/nar/gkab1049.

Dannemann, Michael, and Fernando Racimo. 2018. “Something Old, Something Borrowed: Admixture and Adaptation in Human Evolution.” *Current Opinion in Genetics & Development* 53 (December): 1–8. https://doi.org/10.1016/j.gde.2018.05.009.

Di, Chenlu, Jesus Murga Moreno, Diego F Salazar-Tortosa, M Elise Lauterbur, and David Enard. 2021. “Decreased Recent Adaptation at Human Mendelian Disease Genes as a Possible Consequence of Interference between Advantageous and Deleterious Variants.” *ELife* 10 (October): e69026. https://doi.org/10.7554/eLife.69026.

Enard, David, Philipp W Messer, and Dmitri a Petrov. 2014. “Genome-Wide Signals of Positive Selection in Human Evolution.” *Genome Research* 24 (6): 885–95. https://doi.org/10.1101/gr.164822.113.

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

———. 2020. “Ancient RNA Virus Epidemics through the Lens of Recent Adaptation in Human Genomes.” *Philosophical Transactions of the Royal Society B: Biological Sciences* 375 (November): 20190575. https://doi.org/10.1098/rstb.2019.0575.

Enattah, Nabil Sabri, Timo Sahi, Erkki Savilahti, Joseph D. Terwilliger, Leena Peltonen, and Irma Järvelä. 2002. “Identification of a Variant Associated with Adult-Type Hypolactasia.” *Nature Genetics* 30 (2): 233–37. https://doi.org/10.1038/ng826.

Fay, Justin C, and Chung-I Wu. 2000. “Hitchhiking Under Positive Darwinian Selection.” *Genetics* 155: 1405–13.

Fernandes, Veronica, Nicolas Brucato, Joana C Ferreira, Nicole Pedro, Bruno Cavadas, François-Xavier Ricaut, Farida Alshamali, and Luisa Pereira. 2019. “Genome-Wide Characterization of Arabian Peninsula Populations: Shedding Light on the History of a Fundamental Bridge between Continents.” Edited by Connie Mulligan. *Molecular Biology and Evolution* 36 (3): 575–86. https://doi.org/10.1093/molbev/msz005.

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

Flagel, Lex, Yaniv Brandvain, and Daniel R Schrider. 2019. “The Unreasonable Effectiveness of Convolutional Neural Networks in Population Genetic Inference.” Edited by Yuseob Kim. *Molecular Biology and Evolution* 36 (2): 220–38. https://doi.org/10.1093/molbev/msy224.

Gao, Feng, Chen Ming, Wangjie Hu, and Haipeng Li. 2016. “New Software for the Fast Estimation of Population Recombination Rates (FastEPRR) in the Genomic Era.” *G3: Genes|Genomes|Genetics* 6 (6): 1563–71. https://doi.org/10.1534/g3.116.028233.

Garud, Nandita R., Philipp W. Messer, Erkan O. Buzbas, and Dmitri A. Petrov. 2015. “Recent Selective Sweeps in North American Drosophila Melanogaster Show Signatures of Soft Sweeps.” Edited by Gregory P. Copenhaver. *PLOS Genetics* 11 (2): e1005004. https://doi.org/10.1371/journal.pgen.1005004.

Gheisari, Soheila, Sahar Shariflou, Jack Phu, Paul J. Kennedy, Ashish Agar, Michael Kalloniatis, and S. Mojtaba Golzan. 2021. “A Combined Convolutional and Recurrent Neural Network for Enhanced Glaucoma Detection.” *Scientific Reports* 11 (1): 1945. https://doi.org/10.1038/s41598-021-81554-4.

Gower, Graham, Pablo Iáñez Picazo, Matteo Fumagalli, and Fernando Racimo. 2021. “Detecting Adaptive Introgression in Human Evolution Using Convolutional Neural Networks.” *ELife* 10 (May): e64669. https://doi.org/10.7554/eLife.64669.

Grossman, S. R., I. Shylakhter, E. K. Karlsson, E. H. Byrne, S. Morales, G. Frieden, E. Hostetter, et al. 2010. “A Composite of Multiple Signals Distinguishes Causal Variants in Regions of Positive Selection.” *Science* 327 (5967): 883–86. https://doi.org/10.1126/science.1183863.

Halldorsson, Bjarni V., Gunnar Palsson, Olafur A. Stefansson, Hakon Jonsson, Marteinn T. Hardarson, Hannes P. Eggertsson, Bjarni Gunnarsson, et al. 2019. “Characterizing Mutagenic Effects of Recombination through a Sequence-Level Genetic Map.” *Science* 363 (6425): eaau1043. https://doi.org/10.1126/science.aau1043.

Haller, Benjamin C, and Philipp W Messer. 2019. “SLiM 3: Forward Genetic Simulations Beyond the Wright–Fisher Model.” Edited by Ryan Hernandez. *Molecular Biology and Evolution* 36 (3): 632–37. https://doi.org/10.1093/molbev/msy228.

Halligan, Daniel L., Athanasios Kousathanas, Rob W. Ness, Bettina Harr, Lél Eöry, Thomas M. Keane, David J. Adams, and Peter D. Keightley. 2013. “Contributions of Protein-Coding and Regulatory Change to Adaptive Molecular Evolution in Murid Rodents.” Edited by Bret A. Payseur. *PLoS Genetics* 9 (12): e1003995. https://doi.org/10.1371/journal.pgen.1003995.

Hallmark, Brian, Tatiana M Karafet, PingHsun Hsieh, Ludmila P Osipova, Joseph C Watkins, and Michael F Hammer. 2019. “Genomic Evidence of Local Adaptation to Climate and Diet in Indigenous Siberians.” Edited by Ryan Hernandez. *Molecular Biology and Evolution* 36 (2): 315–27. https://doi.org/10.1093/molbev/msy211.

Hamblin, Martha T., and Anna Di Rienzo. 2000. “Detection of the Signature of Natural Selection

in Humans: Evidence from the Duffy Blood Group Locus." *The American Journal of Human Genetics* 66 (5): 1669–79. https://doi.org/10.1086/302879.
Hamblin, Martha T., Emma E. Thompson, and Anna Di Rienzo. 2002. "Complex Signatures of Natural Selection at the Duffy Blood Group Locus." *The American Journal of Human Genetics* 70 (2): 369–83. https://doi.org/10.1086/338628.
Harris, Alexandre M, and Michael DeGiorgio. 2020. "A Likelihood Approach for Uncovering Selective Sweep Signatures from Haplotype Data." Edited by Yuseob Kim. *Molecular Biology and Evolution* 37 (10): 3023–46. https://doi.org/10.1093/molbev/msaa115.
Harris, Alexandre M, Nandita R Garud, and Michael DeGiorgio. 2018. "Detection and Classification of Hard and Soft Sweeps from Unphased Genotypes by Multilocus Genotype Identity." *Genetics* 210 (4): 1429–52.
Hejase, Hussein A, Ziyi Mo, Leonardo Campagna, and Adam Siepel. 2022. "A Deep-Learning Approach for Inference of Selective Sweeps from the Ancestral Recombination Graph." Edited by Tal Pupko. *Molecular Biology and Evolution* 39 (1): msab332. https://doi.org/10.1093/molbev/msab332.
Hermisson, Joachim, and Pleuni S Pennings. 2005. "Soft Sweeps : Molecular Population Genetics of Adaptation From Standing Genetic Variation" 2352 (April): 2335–52. https://doi.org/10.1534/genetics.104.036947.
Hernandez, Ryan D., Joanna L. Kelley, Eyal Elyashiv, S. Cord Melton, Adam Auton, Gilean McVean, 1000 Genomes Project, Guy Sella, and Molly Przeworski. 2011. "Classic Selective Sweeps Were Rare in Recent Human Evolution." *Science* 331 (6019): 920–24. https://doi.org/10.1126/science.1198878.
Hill, W. G., and Alan Robertson. 1966. "The Effect of Linkage on Limits to Artificial Selection." *Genetical Research* 8 (3): 269–94. https://doi.org/10.1017/S0016672300010156.
Hu, Xiao-Ju, Ji Yang, Xing-Long Xie, Feng-Hua Lv, Yin-Hong Cao, Wen-Rong Li, Ming-Jun Liu, et al. 2019. "The Genome Landscape of Tibetan Sheep Reveals Adaptive Introgression from Argali and the History of Early Human Settlements on the Qinghai–Tibetan Plateau." Edited by Rasmus Nielsen. *Molecular Biology and Evolution* 36 (2): 283–303. https://doi.org/10.1093/molbev/msy208.
Isildak, Ulas, Alessandro Stella, and Matteo Fumagalli. 2021. "Distinguishing between Recent Balancing Selection and Incomplete Sweep Using Deep Neural Networks." *Molecular Ecology Resources* 21 (8): 2706–18. https://doi.org/10.1111/1755-0998.13379.
Johri, Parul, Charles F. Aquadro, Mark Beaumont, Brian Charlesworth, Laurent Excoffier, Adam Eyre-Walker, Peter D. Keightley, et al. 2021. "Recommendations for Improving Statistical Inference in Population Genomics." Preprint. Evolutionary Biology. https://doi.org/10.1101/2021.10.27.466171.
Kern, Andrew D., and Daniel R. Schrider. 2016. "Discoal: Flexible Coalescent Simulations with Selection." *Bioinformatics* 32 (24): 3839–41. https://doi.org/10.1093/bioinformatics/btw556.
Kern, Andrew D, and Daniel R Schrider. 2018. "DiploS/HIC: An Updated Approach to Classifying Selective Sweeps," 12.
Key, Felix M., Qiaomei Fu, Frédéric Romagné, Michael Lachmann, and Aida M. Andrés. 2016. "Human Adaptation and Population Differentiation in the Light of Ancient Genomes." *Nature Communications* 7 (1): 10775. https://doi.org/10.1038/ncomms10775.
Kim, Yuseob, and Rasmus Nielsen. 2004. "Linkage Disequilibrium as a Signature of Selective Sweeps." *Genetics* 167 (3): 1513–24. https://doi.org/10.1534/genetics.103.025387.
King, Mary-Claire, and A. C. Wilson. 1975. "Evolution at Two Levels in Humans and Chimpanzees: Their Macromolecules Are so Alike That Regulatory Mutations May

Account for Their Biological Differences." *Science* 188 (4184): 107–16. https://doi.org/10.1126/science.1090005.
Klassmann, Alexander, and Mathieu Gautier. 2022. "Detecting Selection Using Extended Haplotype Homozygosity (EHH)-Based Statistics in Unphased or Unpolarized Data." Edited by Muhammad Abdul Rehman Rashid. *PLOS ONE* 17 (1): e0262024. https://doi.org/10.1371/journal.pone.0262024.
Korneliussen, Thorfinn Sand, Ida Moltke, Anders Albrechtsen, and Rasmus Nielsen. 2013. "Calculation of Tajima's D and Other Neutrality Test Statistics from Low Depth next-Generation Sequencing Data." *BMC Bioinformatics* 14 (1): 289. https://doi.org/10.1186/1471-2105-14-289.
Lewin, Harris A., Gene E. Robinson, W. John Kress, William J. Baker, Jonathan Coddington, Keith A. Crandall, Richard Durbin, et al. 2018. "Earth BioGenome Project: Sequencing Life for the Future of Life." *Proceedings of the National Academy of Sciences* 115 (17): 4325–33. https://doi.org/10.1073/pnas.1720115115.
Li, W H, and L A Sadler. 1991. "Low Nucleotide Diversity in Man." *Genetics* 129 (2): 513–23. https://doi.org/10.1093/genetics/129.2.513.
Li, Yan, Dong-Dong Wu, Adam R. Boyko, Guo-Dong Wang, Shi-Fang Wu, David M. Irwin, and Ya-Ping Zhang. 2014. "Population Variation Revealed High-Altitude Adaptation of Tibetan Mastiffs." *Molecular Biology and Evolution* 31 (5): 1200–1205. https://doi.org/10.1093/molbev/msu070.
Liao, Lizhi, Heng Li, Weiyi Shang, and Lei Ma. 2022. "An Empirical Study of the Impact of Hyperparameter Tuning and Model Optimization on the Performance Properties of Deep Neural Networks." *ACM Transactions on Software Engineering and Methodology* 31 (3): 1–40. https://doi.org/10.1145/3506695.
Librado, Pablo, and Ludovic Orlando. 2018. "Detecting Signatures of Positive Selection along Defined Branches of a Population Tree Using LSD." Edited by Ryan Hernandez. *Molecular Biology and Evolution* 35 (6): 1520–35. https://doi.org/10.1093/molbev/msy053.
Marciniak, Stephanie, and George H. Perry. 2017. "Harnessing Ancient Genomes to Study the History of Human Adaptation." *Nature Reviews Genetics* 18 (11): 659–74. https://doi.org/10.1038/nrg.2017.65.
Maynard Smith, John, and John Haigh. 1974. "The Hitch-Hiking Effect of a Favourable Gene." *Genetical Research* 23 (1): 23–35. https://doi.org/10.1017/S0016672300014634.
McManus, Kimberly F., Angela M. Taravella, Brenna M. Henn, Carlos D. Bustamante, Martin Sikora, and Omar E. Cornejo. 2017. "Population Genetic Analysis of the DARC Locus (Duffy) Reveals Adaptation from Standing Variation Associated with Malaria Resistance in Humans." Edited by Hopi E. Hoekstra. *PLOS Genetics* 13 (3): e1006560. https://doi.org/10.1371/journal.pgen.1006560.
McVean, Gil. 2007. "The Structure of Linkage Disequilibrium Around a Selective Sweep." *Genetics* 175 (3): 1395–1406. https://doi.org/10.1534/genetics.106.062828.
Mughal, Mehreen R, and Michael DeGiorgio. 2018. "Localizing and Classifying Adaptive Targets with Trend Filtered Regression." Edited by Joanna Kelley. *Molecular Biology and Evolution* 36 (2): 252–70. https://doi.org/10.1093/molbev/msy205.
Mughal, Mehreen R., Hillary Koch, Jinguo Huang, Francesca Chiaromonte, and Michael DeGiorgio. 2020. "Learning the Properties of Adaptive Regions with Functional Data Analysis." *PLOS Genetics* 16 (8): e1008896. https://doi.org/10.1371/journal.pgen.1008896.
Nielsen, Rasmus. 2005. "Molecular Signatures of Natural Selection."

https://doi.org/10.1146/annurev.genet.39.073003.112420.
Peyrégne, Stéphane, Michael James Boyle, Michael Dannemann, and Kay Prüfer. 2017. "Detecting Ancient Positive Selection in Humans Using Extended Lineage Sorting." *Genome Research* 27 (9): 1563–72. https://doi.org/10.1101/gr.219493.116.
Prüfer, Kay, Fernando Racimo, Nick Patterson, Flora Jay, Sriram Sankararaman, Susanna Sawyer, Anja Heinze, et al. 2014. "The Complete Genome Sequence of a Neanderthal from the Altai Mountains." *Nature* 505 (7481): 43–49. https://doi.org/10.1038/nature12886.
Przeworski, Molly, Graham Coop, and Jeffrey D Wall. 2005. "THE SIGNATURE OF POSITIVE SELECTION ON STANDING GENETIC VARIATION" 59 (11): 2312–23.
Pybus, Marc, Pierre Luisi, Giovanni Marco Dall'Olio, Manu Uzkudun, Hafid Laayouni, Jaume Bertranpetit, and Johannes Engelken. 2015. "Hierarchical Boosting: A Machine-Learning Framework to Detect and Classify Hard Selective Sweeps in Human Populations." *Bioinformatics*, August, btv493. https://doi.org/10.1093/bioinformatics/btv493.
Racimo, Fernando. 2016. "Testing for Ancient Selection Using Cross-Population Allele Frequency Differentiation." *Genetics* 202: 733–50.
Racimo, Fernando, Martin Kuhlwilm, and Montgomery Slatkin. 2014. "A Test for Ancient Selective Sweeps and an Application to Candidate Sites in Modern Humans." *Molecular Biology and Evolution* 31 (12): 3344–58. https://doi.org/10.1093/molbev/msu255.
Reiner, Alex P., Sandra Beleza, Nora Franceschini, Paul L. Auer, Jennifer G. Robinson, Charles Kooperberg, Ulrike Peters, and Hua Tang. 2012. "Genome-Wide Association and Population Genetic Analysis of C-Reactive Protein in African American and Hispanic American Women." *The American Journal of Human Genetics* 91 (3): 502–12. https://doi.org/10.1016/j.ajhg.2012.07.023.
Ronen, Roy, Glenn Tesler, Ali Akbari, Shay Zakov, Noah A. Rosenberg, and Vineet Bafna. 2015. "Predicting Carriers of Ongoing Selective Sweeps without Knowledge of the Favored Allele." Edited by Graham Coop. *PLOS Genetics* 11 (9): e1005527. https://doi.org/10.1371/journal.pgen.1005527.
Sabeti, Pardis C., David E. Reich, John M. Higgins, Haninah Z. P. Levine, Daniel J. Richter, Stephen F. Schaffner, Stacey B. Gabriel, et al. 2002. "Detecting Recent Positive Selection in the Human Genome from Haplotype Structure." *Nature* 419 (6909): 832–37. https://doi.org/10.1038/nature01140.
Sattath, Shmuel, Eyal Elyashiv, Oren Kolodny, Yosef Rinott, and Guy Sella. 2011. "Pervasive Adaptive Protein Evolution Apparent in Diversity Patterns around Amino Acid Substitutions in Drosophila Simulans." Edited by Gil McVean. *PLoS Genetics* 7 (2): e1001302. https://doi.org/10.1371/journal.pgen.1001302.
Schlebusch, Carina M., Helena Malmström, Torsten Günther, Per Sjödin, Alexandra Coutinho, Hanna Edlund, Arielle R. Munters, et al. 2017. "Southern African Ancient Genomes Estimate Modern Human Divergence to 350,000 to 260,000 Years Ago." *Science* 358 (6363): 652–55. https://doi.org/10.1126/science.aao6266.
Schrider, Daniel R. 2020. "Background Selection Does Not Mimic the Patterns of Genetic Diversity Produced by Selective Sweeps." *Genetics* 216: 499–519.
Schrider, Daniel R., and Andrew D. Kern. 2017. "Soft Sweeps Are the Dominant Mode of Adaptation in the Human Genome." *Molecular Biology and Evolution* 34 (8): 1863–77. https://doi.org/10.1093/molbev/msx154.
———. 2018. "Supervised Machine Learning for Population Genetics: A New Paradigm." *Trends in Genetics* 34 (4): 301–12. https://doi.org/10.1016/j.tig.2017.12.005.
Simonson, Tatum S, Donald a McClain, Lynn B Jorde, and Josef T Prchal. 2012. "Genetic

Determinants of Tibetan High-Altitude Adaptation." *Human Genetics* 131 (4): 527–33. https://doi.org/10.1007/s00439-011-1109-3.

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

Speidel, Leo, Marie Forest, Sinan Shi, and Simon R. Myers. 2019. "A Method for Genome-Wide Genealogy Estimation for Thousands of Samples." *Nature Genetics* 51 (9): 1321–29. https://doi.org/10.1038/s41588-019-0484-x.

Stephan, Wolfgang. 2010. "Genetic Hitchhiking versus Background Selection: The Controversy and Its Implications." *Philosophical Transactions of the Royal Society B: Biological Sciences* 365 (1544): 1245–53. https://doi.org/10.1098/rstb.2009.0278.

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

Tajima, Fumio. 1989. "Statistical Method for Testing the Neutral Mutation Hypothesis by DNA Polymorphism" 595 (3): 585–95.

The 1000 Genomes Project Consortium, Corresponding authors, Adam Auton, Gonçalo R. Abecasis, Steering committee, David M. Altshuler, Richard M. Durbin, et al. 2015. "A Global Reference for Human Genetic Variation." *Nature* 526 (7571): 68–74. https://doi.org/10.1038/nature15393.

The Darwin Tree of Life Project Consortium, Mark Blaxter, Nova Mieszkowska, Federica Di Palma, Peter Holland, Richard Durbin, Thomas Richards, et al. 2022. "Sequence Locally, Think Globally: The Darwin Tree of Life Project." *Proceedings of the National Academy of Sciences* 119 (4): e2115642118. https://doi.org/10.1073/pnas.2115642118.

Torada, Luis, Lucrezia Lorenzon, Alice Beddis, Ulas Isildak, Linda Pattini, Sara Mathieson, and Matteo Fumagalli. 2019. "ImaGene: A Convolutional Neural Network to Quantify Natural Selection from Genomic Data." *BMC Bioinformatics* 20 (S9): 337. https://doi.org/10.1186/s12859-019-2927-x.

Voight, Benjamin F, Sridhar Kudaravalli, Xiaoquan Wen, and Jonathan K Pritchard. 2006. "A Map of Recent Positive Selection in the Human Genome." Edited by Laurence Hurst. *PLoS Biology* 4 (3): e72. https://doi.org/10.1371/journal.pbio.0040072.

Wang, Zhanpeng, Jiaping Wang, Michael Kourakos, Nhung Hoang, Hyong Hark Lee, Iain Mathieson, and Sara Mathieson. 2021. "Automatic Inference of Demographic Parameters Using Generative Adversarial Networks." *Molecular Ecology Resources* 21 (8): 2689–2705. https://doi.org/10.1111/1755-0998.13386.

Xue, Alexander T, Daniel R Schrider, Andrew D Kern, and Ag1000g Consortium. 2021. "Discovery of Ongoing Selective Sweeps within *Anopheles* Mosquito Populations Using Deep Learning." Edited by Daniel Falush. *Molecular Biology and Evolution* 38 (3): 1168–83. https://doi.org/10.1093/molbev/msaa259.

Zeng, Kai, Suhua Shi, and Chung-I Wu. 2007. "Compound Tests for the Detection of Hitchhiking Under Positive Selection." *Molecular Biology and Evolution* 24 (8): 1898–1908. https://doi.org/10.1093/molbev/msm119.