GBE

Identifying Adaptive Footprints in the Presence of Demographic Uncertainty

Sandipan Paul Arnab [1,2], Mohammad Khan [1,2], Andre Luiz Campelo dos Santos [1], Matteo Fumagalli [3,4], Michael DeGiorgio [1,2,5,*]

[1]Department of Electrical Engineering and Computer Science, Florida Atlantic University, Boca Raton, FL, USA
[2]Center for Omics Technologies and Data Engineering, Florida Atlantic University, Boca Raton, FL, USA
[3]School of Biological and Behavioural Sciences, Queen Mary University of London, London, UK
[4]The Alan Turing Institute, London, UK
[5]Department of Biomedical Engineering, Florida Atlantic University, Boca Raton, FL, USA

*Corresponding author: E-mail: mdegiorg@fau.edu.

Accepted: March 09, 2026

Abstract

Identifying genomic regions shaped by natural selection is a central goal in evolutionary genomics. Existing machine learning methods for this task are typically trained on labeled data simulated according to specific evolutionary scenarios. While effective in controlled settings, these models are limited by their reliance on explicit class labels, detecting only the processes they were trained to recognize. This limitation makes it difficult to interpret predictions for regions shaped by other evolutionary forces, a problem especially acute when analyzing genomes influenced by mixtures of adaptive and demographic factors. One-vs-rest strategies offer a potential alternative but suffer from the complexity of modeling processes as a catch-all "rest" class. Here, we explore positive-unlabeled learning as a flexible framework for detecting adaptive events. This semi-supervised approach permits identification of a target class using only positive labels and an unlabeled background, without requiring explicit modeling of negatives. To assess its utility, we focus on a binary classification setting for detecting selective sweeps against a mixed background of unlabeled sweeps and neutrally evolving regions. We introduce *PULSe*, a method that trains only on labeled sweep observations while treating remaining data as unlabeled. By avoiding assumptions about background composition, *PULSe* enables robust sweep discovery in realistic genomic landscapes. We evaluate performance across demographic, adaptive, and confounding contexts, including domain shift from misspecified models, and find that *PULSe* delivers strong generalizability. Finally, analyzing European and Bengali genomes, we recapitulate known sweep candidates, demonstrating *PULSe* as a versatile tool for detecting adaptive regions across diverse genomic landscapes.

Key words: natural selection, positive-unlabeled learning, machine learning, model misspecification, Bengali in Bangladesh.

Significance

Identifying genomic regions shaped by natural selection is complicated by demographic uncertainty and the complexity of empirical genomes. We introduce a semi-supervised framework that avoids the need to define an explicit "neutral" reference set and show that it remains reliable even when trained and tested under different evolutionary scenarios. In European human genomes, the method recovered well-known targets of selection and showed strong consistency across contrasting analyses, increasing confidence in its predictions. When applied to Bengali individuals from Bangladesh, who have a demographic history that is complex and poorly characterized, the framework uncovered a large set of sweep candidates. These included voltage-gated potassium channel genes linked to neuronal excitability and psychiatric disorders, and *OAS1*, which is an antiviral gene associated with RNA virus defense and Neanderthal introgression.

Introduction

Detecting genomic regions influenced by adaptive evolution is critical for uncovering how organisms respond to their environments and for identifying loci that contribute to fitness-related traits (Pigllucci 1996; Ellegren and Sheldon 2008). Genomic regions influenced by adaptive processes often bear distinctive signals, including altered variation, or skewed allele frequency distributions that reflect their departure from neutrality (Lewontin and Krakauer 1973; Günther and Coop 2013). Detecting these regions offers a powerful means to infer the genetic basis of adaptation, illuminate species histories, and identify functionally important loci (Pardo-Diaz et al. 2015). However, the multiplicity of adaptive modes makes comprehensive identification difficult. Each mode can leave different, sometimes overlapping footprints in genetic data (Oleksyk et al. 2010), often entangled with confounding demographic effects (Saccheri and Hanski 2006; Byars et al. 2010). While empirical-based approaches (Voight et al. 2006) have the capacity to correct for demographic effects, they do not provide a framework for inferring the proportions of sites under adaptive evolution. The additional challenge lies not only in distinguishing adaptive from neutral regions, but in building models that can generalize beyond any single evolutionary process.

Modern machine learning methods have shown promise within this sphere by learning discriminative patterns across many features simultaneously, whether derived from summary statistics or raw genomic alignments (Sheehan and Song 2016; Schrider and Kern 2017; Sugden et al. 2018; Flagel et al. 2019; Torada et al. 2019; Mughal et al. 2020; Gower et al. 2021; Ahlquist et al. 2023; Cecil and Sugden 2023; Lauterbur et al. 2023; Whitehouse and Schrider 2023; Arnab et al. 2025). These tools can flexibly capture relationships between input data and evolutionary outcomes, providing an appealing alternative to single-test statistics. Some recent approaches have adopted deep learning architectures, such as convolutional neural networks (CNNs), to classify genomic windows by selection type or evolutionary history, often reporting superior performance over traditional methods (Kern and Schrider 2018; Hejase et al. 2022; Arnab et al. 2023). Despite these advances, most machine learning approaches in population genomics remain constrained by their dependence on simulated, explicitly labeled training data. Typically, models are trained to distinguish one or more types of selection from neutral evolution, relying on labels derived from simulations under predefined models. As a result, they tend to learn only the specific patterns they are trained to recognize. This limitation constrains their utility when applied to empirical data, where the true evolutionary history, which includes both demographic history and the full spectrum of adaptive processes, may be far more complex or entirely unmodeled.

The core limitation lies in how the "negative class" is defined. In many supervised machine learning approaches, the model is trained to classify between a selective process of interest and a negative (background) class, generally comprising neutrality. Yet, real genomes are shaped by a mixture of nonadaptive forces like mutation, migration, drift, population size changes, and demographic shifts, as well as adaptive forces acting through various mechanisms (Kimura 1979; Slatkin 1987; Barton and Charlesworth 1998; Lynch 2010). Attempting to simulate all nonsweep processes that could shape the background class would require substantial modeling effort and still fall short, as the full set of forces acting on empirical genomes is unknown. More importantly, the practical challenge is identifying which of these processes are most relevant for confounding sweep detection without introducing unnecessary processes that may not apply to the focal population. A classifier trained on an overly narrow or overly broad set of assumptions may fail when applied to data generated under different evolutionary histories. Even with expansive simulation sets, a classifier trained on specific assumptions may fail when applied

to data generated under different histories. This mismatch, which can be characterized as domain shift (Ganin et al. 2016; Luo et al. 2019; Stacke et al. 2020; Zhang et al. 2021), can degrade model accuracy and lead to overconfident and miscalibrated (Ovadia et al. 2019) predictions. One-vs-rest classifiers (Xu 2011; Bali and Mansotra 2021), which can treat all other forms of adaptive and nonadaptive processes except for the targeted mode of adaptation as a generic "rest" class, inherit this limitation and may confuse unknown selection modes with noise. Consequently, there is a need for detection frameworks that do not rely on explicitly modeling all possible evolutionary alternatives.

To address this issue, we introduce a strategy based on positive-unlabeled (PU) learning, a form of semi-supervised learning designed for classification problems where only a subset of positive samples are labeled, and no explicitly labeled negatives are available (Elkan and Noto 2008). PU learning has been successfully applied in domains such as fraud detection and medical diagnosis (Chen et al. 2020; Vinay et al. 2022), where true negatives are hard to define or curate. Its core assumption is that labeled positives represent one class of interest, while the unlabeled data contain a mixture of positives and negatives. The goal is to distinguish positives from this unlabeled background without ever specifying what a negative should look like. In many real-world PU learning applications, the unlabeled set is highly imbalanced, with positives representing only a small fraction of the total (Su et al. 2021). This assumption maps naturally to evolutionary genomics, where we may have only a few high-confidence examples of regions under selection, while the vast majority of the genome is expected to be neutrally evolving or shaped by other nonadaptive processes. PU learning, thus, offers a principled way to bypass the need for explicit negative labels, allowing models to learn what distinguishes known adaptive regions from the rest of the genome, regardless of the composition of the background.

Here, we present the method *PULSe* (*P*ositive-*U*nlabeled *L*earning for *Se*lection detection), which applies this framework to the task of identifying genomic regions shaped by adaptation. *PULSe* requires only a labeled set of adaptive samples and treats the remaining data as unlabeled, without making assumptions about the evolutionary processes shaping that background. While the method is intended to generalize to a wide range of adaptive scenarios, we use selective sweeps as a benchmark for validation. A selective sweep is a phenomenon that is introduced by positive natural selection. When a beneficial genetic mutation quickly spreads through a population, it reduces genetic variation in nearby regions of the genome (Przeworski 2002; Hermisson and Pennings 2005; Stephan 2016, 2019). Sweeps are a convenient test case because they produce well-characterized genomic signals of adaptation and can be readily simulated using widely available population genetic simulation software (Ewing and Hermisson 2010; Kern and Schrider 2016; Haller and Messer 2019). We assess the ability of *PULSe* to detect sweep regions when trained only on positive samples, in the presence of challenging confounders such as demographic model misspecification, and background selection.

We evaluate *PULSe* across an extensive array of simulation experiments, comparing its performance to that of fully supervised models. These evaluations include both in-domain and out-of-domain testing to assess the robustness of the method to domain shift. By design, *PULSe* is modular and compatible with a variety of feature generation protocols and classifier algorithms, making it adaptable to different organismal systems and data types. *PULSe* is also user-friendly and can be seamlessly integrated with common simulators, provided that the simulations can be exported in *ms* (Hudson 2002) or Variant Call Format (Danecek et al. 2011) files. As a demonstration of its practical utility, we apply *PULSe* to human polymorphism data from Europeans and Bengali in Bangladesh, treating the genomes as unlabeled backgrounds and using simulated sweeps as positive training samples. The model recovers several well-characterized sweep candidates previously identified in empirical scans and flags additional regions of interest for further investigation. These results suggest that *PULSe* can serve as a flexible tool for detecting regions shaped by selection in empirical genomes.

Results

Modeling Description

We utilized the coalescent simulator `discoal` (Kern and Schrider 2016) to generate neutral and sweep replicates under a nonequilibrium demographic history estimated for European (CEU) humans (Tennessen et al. 2012). This demographic model includes a recent severe population bottleneck followed by ongoing exponential growth, capturing key aspects of the complex evolutionary dynamics of this population. To simulate sweeps, we considered per-generation selection coefficients (s) ranging from 0.005 to 0.1, initial beneficial allele frequencies (f) spanning from $1/(2N_e)$ to 0.2, and fixation times (τ) varying from 0 to 2,000 generations before sampling. Additionally, both neutral and sweep replicates were generated with per-site per-generation mutation rates (μ) ranging from 2.21×10^{-9} to 2.21×10^{-8}, and per-site per-generation recombination rates (r) from 0 to 3×10^{-8}, with a mean rate of 10^{-8}.

These parameters were chosen to comprehensively capture the evolutionary forces acting on the CEU population while ensuring that a wide spectrum of selective and neutral genetic patterns was examined. For specific details about the simulations, refer to the *Simulation protocol* subsection of the *Materials and Methods*.

To further investigate the impact of demographic assumptions, we also generated neutral and sweep replicates based on a constant-size demographic model while maintaining the same genetic and adaptive parameter settings as in the CEU simulations. The only exception was the diploid effective population size, which we varied across a range from 3,000 to 30,000 in increments of 250. Given the breadth of selection strengths, beneficial allele frequencies, and fixation times explored, along with the inherent variation in mutation and recombination rates, distinguishing between neutral and selective scenarios remains challenging. This difficulty arises from the substantial overlap in the distributions of genetic variation between sweep and neutral scenarios, underscoring the need for robust and adaptable detection methods.

To construct the input for *PULSe*, we first generated images (Arnab et al. 2025) from the CEU and constant-size simulated replicates. These images were created by transforming minor allele count data into sorted representations across short, overlapping windows. The image generation process (Fig. 1a) from simulated data is detailed in the *Image generation* subsection of the *Materials and Methods*. To extract informative features from these images, we applied histogram of oriented gradients (HOGs) (Freeman and Roth 1995), which captures local gradient orientations and edge structures (Fig. 1b), effectively encoding spatial patterns indicative of changes in genomic diversity (see *Histogram of oriented gradients* subsection of the *Materials and Methods* for details). We then used these extracted HOG features in a positive-unlabeled (PU) learning framework (Fig. 1c), where logistic regression served as the base classifier (see *Positive-unlabeled learning algorithm* subsection of the *Materials and Methods* for details).

Unlike traditional supervised learning, PU learning does not require a fully labeled training dataset. Instead, it relies on a set of positively labeled samples (sweep regions in this case) while treating the remaining data as an unlabeled mixture of negative (nonsweep regions) and potentially undetected positive (sweep regions) samples. For simulated testing, we designed two scenarios to evaluate *PULSe*. In the first scenario, we used 10,000 CEU sweep replicates as the labeled set. We then used 20,000 CEU simulations as the unlabeled set, with the impact of varying ratios of labeled and unlabeled observations (detailed in the *Effect of the ratio of labeled to unlabeled samples* subsection below). Because most of the genome is expected to evolve neutrally (Kimura 1968; Jensen et al. 2019), 90% of the unlabeled set consisted of CEU neutral replicates, with the impact of specific proportions of sweep and neutral replicates detailed in the *Model performance as a function of unlabeled set class imbalance* subsection below. The second scenario followed the same structure but used 10,000 constant population size sweeps as the labeled set while keeping the same CEU unlabeled set as in the first scenario. We, respectively, denote these two scenarios as *CEU–CEU* (domain matched) and *Constant–CEU* (domain mismatched) for ease of reference. These experimental setups allow us to assess the robustness of PU learning in detecting selection in the CEU demographic context.

Heatmaps depicting mean images for the sweep and neutral classes in both the CEU and constant-size datasets reveal distinct patterns associated with positive selection. In contrast to the neutral image, the sweep image exhibits a dark vertical segment in the central columns, representing a loss of genetic variation due to high-frequency haplotypes (Fig. 2). This pattern arises because major alleles occur at high frequency near the center of the sweep replicates, reflecting the impact of selection on linked variation. By presenting mean images for both CEU and constant population size sweep and neutral replicates, we illustrate the consistency of these patterns across different demographic contexts.

Moreover, the PU learning framework does not follow the conventional separation between training data and a distinct test dataset used solely for evaluation. Instead, the unlabeled data serve a dual role. They are incorporated directly into the training objective, and after training, predictions are generated on the same unlabeled set (and during benchmarking, evaluated using their known ground-truth labels). We make use of simulated data in the unlabeled set to study HOG featurization of images supplied as input to *PULSe*, to assess *PULSe* performance under inherent genomic data limitations, and to conduct benchmarking comparisons. In empirical applications, the training procedure remains identical, but the unlabeled data are composed exclusively of images generated from real genomes, with no simulated background samples included. In this setting, simulations are used only to generate the labeled sweep replicates (positive samples), and predictions are produced on the same empirical unlabeled dataset used during training. It is important to note that, because the unlabeled set itself defines the learning problem, rather than training a single model and then applying it across multiple test scenarios, each *PULSe* experimental configuration, including empirical applications, corresponds to a separately trained model.

We thoroughly assessed various HOG parameters and pre- and postprocessing pipelines, with full details

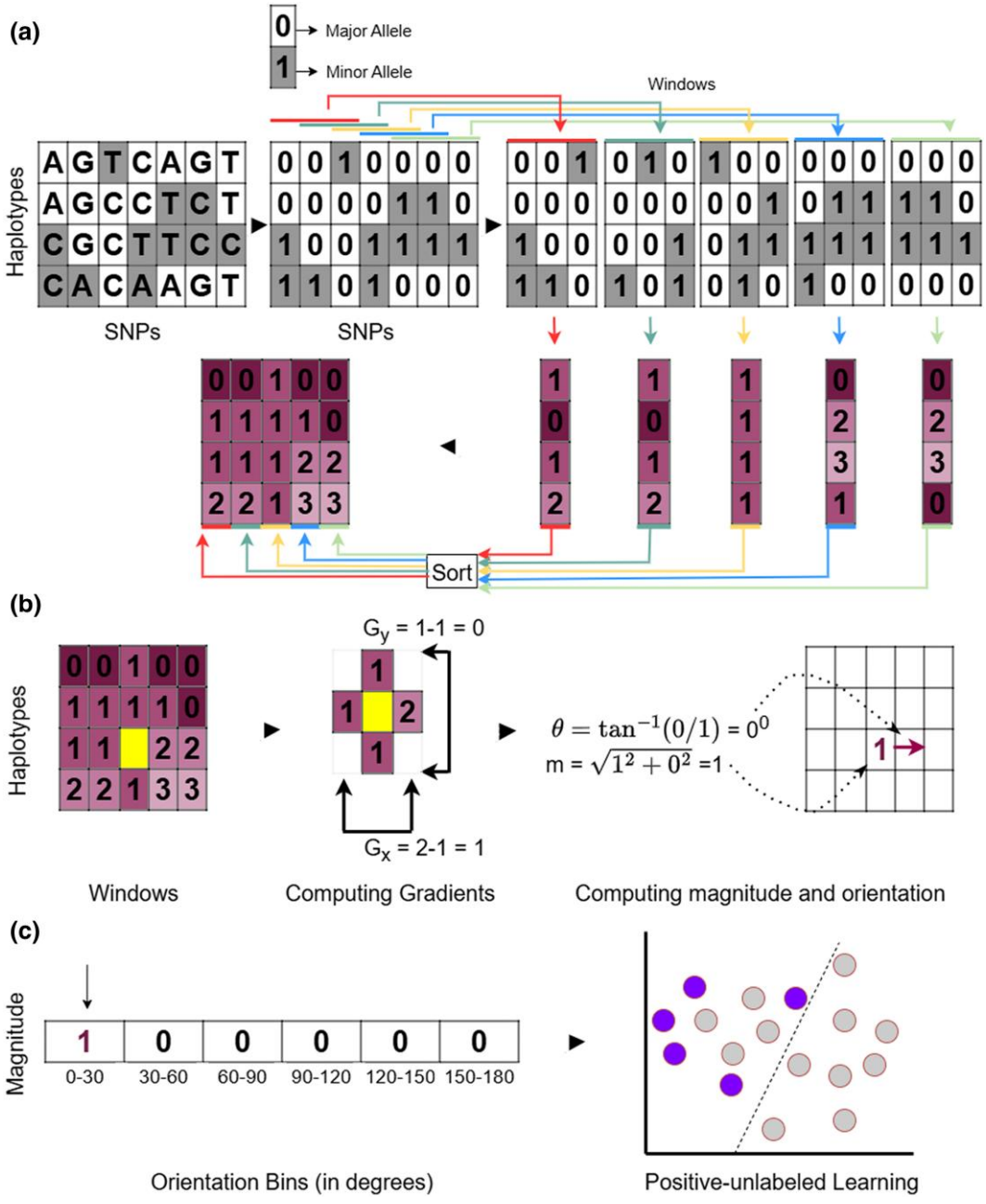

Fig. 1. Depiction of the *PULSe* model. a) As described in the *Image generation* subsection of the *Materials and Methods*, for a sample of haplotypes (rows) across a set of SNPs (columns), the major and minor alleles at each SNP are coded as zero and one, respectively. From this processed alignment, the number of minor alleles for each haplotype is counted within a window of fixed size (here of three SNPs) and window locations are shifted by a specific stride (here of one SNP). The minor allele counts for each window are then sorted, such that the top row has the smallest value and the bottom row has the largest value for a given window. A matrix is computed based on a specified number of consecutive windows (here five windows), which we consider as an input image. b) For a given pixel in this image, gradient magnitude m and orientation θ are calculated based on the differences in pixel intensities along the horizontal (G_x) and vertical (G_y) directions. c) These gradient magnitudes across the image are then accumulated into orientation bins, illustrated here with one such gradient, to build a histogram of oriented gradients vector that is used as input to a positive-unlabeled learning framework.

provided in the *Choosing the best feature extraction pipeline* subsection of the *Materials and Methods*. After extensive testing, we identified the *P1* and *P2* pipelines as performing best in the domain matched and mismatched scenarios, respectively, leading to the development of the *PULSe*[*P1*] and *PULSe*[*P2*] models

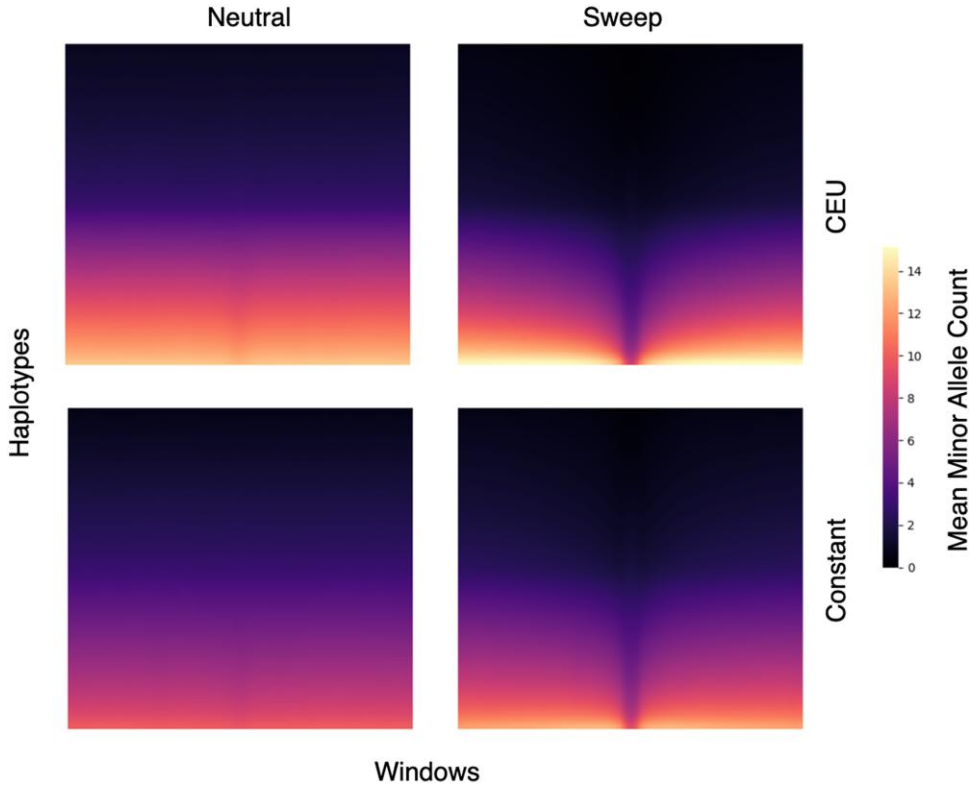

Fig. 2. Heatmaps depicting input images of size 198 × 238, averaged across 10,000 neutral and 10,000 sweeps replicates simulated under either the CEU (top row) or constant size demographic history (bottom row). Input images are processed as in the *Image generation* subsection of the *Materials and Methods*. Rows of the images represent locally sorted minor allele counts within haplotype windows, whereas columns represent genomic windows of 25 contiguous SNPs within a haplotype, with an equal number of windows flanking the center of a simulated genomic region. The colorbar indicates the number of minor alleles within the haplotype window, with darker colors representing higher numbers of major alleles and brighter colors representing higher numbers of minor alleles.

(see *Choosing the best feature extraction pipeline subsection* for details of these pipelines). To evaluate the effectiveness of different feature extraction pipelines, we used the area under the precision–recall curve (AUPRC) as our primary performance metric. This choice was motivated by the particular suitability of the AUPRC for imbalanced datasets, where it better reflects the ability of a model to identify the positive (minority) class by emphasizing precision (positive predictive value) and recall (true positive rate) over overall accuracy (Saito and Rehmsmeier 2015). For context, in datasets with only 10% positive samples, a baseline AUPRC expected from random guessing would fall around 0.1 (Saito and Rehmsmeier 2015). To complement AUPRC, we also evaluated the Matthews correlation coefficient (MCC) as an additional performance metric, as it is particularly robust for binary classification in imbalanced and partially labeled datasets (Boughorbel et al. 2017; Chicco and Jurman 2020). We report MCC as an alternative to overall accuracy because accuracy is misleading under severe class imbalance. AUPRC and MCC both address this issue but from complementary perspectives. AUPRC evaluates how effectively true sweeps are recovered from the pool of predicted sweeps across thresholds and does not explicitly account for true negatives, thereby emphasizing performance on the minority (sweep) class in our setting. In contrast, MCC is computed at a fixed decision threshold and incorporates all four outcomes: true positives, true negatives, false positives, and false negatives. Under extreme class imbalance, the large number of negative samples means that even low false positive rates can severely degrade MCC, making accurate classification of the majority class critical for good performance. Together, AUPRC and MCC provide a balanced assessment. AUPRC reflects ranking and sweep prioritization, whereas MCC captures the reliability of binary predictions under asymmetric class proportions.

Model Performance as a Function of Unlabeled Set Class Imbalance

To examine how the degree of class imbalance of the unlabeled set affects model performance, we

conducted experiments while fixing the number of labeled and unlabeled samples to 10,000 each. We varied the proportion of positive samples in the unlabeled set, testing levels of 2%, 4%, 6%, 8%, and 10%. Despite these changes, the neutral detection rate remained virtually unchanged across all tested conditions for both the *CEU–CEU* (ranging from 95.5% to 95.9%) and *Constant–CEU* (ranging from 88.3% to 88.4%) scenarios (Fig. S1).

Sweep detection rate, however, was heavily affected by the degree of class imbalance of the unlabeled set. As the proportion of positive samples increased, sweep detection rate dropped significantly, suggesting that at their present configurations, *PULSe* models tend to detect adaptive regions better with higher degrees of class imbalance. In the *CEU–CEU* scenario, the detection rate for *PULSe*[*P1*] decreased by 6.6%, whereas *PULSe*[*P2*] experienced a slightly larger drop of 7.3%. The effect was even more pronounced in the *Constant–CEU* scenario, where both models saw a decline of approximately 10.4%.

However, as the proportion of positive samples increased, the AUPRC noticeably improved. In the *CEU–CEU* scenario, AUPRC increased by 0.215 for *PULSe*[*P1*] and 0.220 for *PULSe*[*P2*], whereas for the *Constant–CEU* scenario, the increase was 0.107 for *PULSe*[*P1*] and 0.136 for *PULSe*[*P2*]. Additionally, though MCC shows an upward trend as the proportion of positive samples increases, its growth is not as sharp as AUPRC. In the *CEU–CEU* scenario, the MCC of *PULSe*[*P1*] increases by 0.103 and by 0.127 for *PULSe*[*P2*], whereas in the *Constant–CEU* scenario, the increases are 0.017 and 0.013, respectively.

To summarize the results, despite the unlabeled set size remaining constant, increasing the proportion of sweeps leads to a rise in both MCC and AUPRC, even though the sweep detection rate decreases sharply, while neutral detection remains unchanged. The increase in MCC and AUPRC is expected, because both of the metrics are affected by the balance between the total number of true and false predictions between the two classes. With a higher prevalence of sweeps, the absolute number of true positives increases substantially relative to false positives, resulting in higher MCC and AUPRC scores. However, MCC grows more conservatively because it remains bounded by symmetric error contributions from both classes. With an increasing proportion of sweep samples, the decline in sweep detection rate alongside an expected increase in AUPRC suggests that class separation in terms of ranking remains largely unchanged, while the predicted sweep probabilities become increasingly compressed, resulting in underconfident predictions.

Effect of the Ratio of Labeled to Unlabeled Samples

To evaluate how varying the ratio of labeled to unlabeled samples impacts *PULSe* model performance, we examined six test cases in which the number of labeled samples was fixed at 10,000, and the number of unlabeled samples was set to 10,000, 12,000, 14,000, 16,000, 18,000, and 20,000 (Fig. S2). In each scenario, 10% of the unlabeled samples consist of sweeps, whereas the remainder are neutral replicates.

PULSe[*P1*] under the *CEU–CEU* setting shows AUPRC values increasing from 0.7197 to 0.7669 as the number of unlabeled samples increases, whereas in the *Constant–CEU* scenario, it exhibits a smaller improvement in AUPRC, increasing from 0.3321 to 0.3411. On the other hand, though *PULSe*[*P2*] under the *CEU–CEU* setting showcases an increase in AUPRC values from 0.7106 to 0.7592, we notice a drop from 0.3794 to 0.3547 in the *Constant–CEU* scenario. Though the *P1* pipeline exhibits an upward trend in AUPRC and *P2*, a downward trend with increasing numbers of unlabeled samples in the *Constant–CEU* scenario, *P2* consistently achieves higher AUPRC than *P1* across all test cases.

We observe a similar trend in the MCC values. For instance, *PULSe*[*P1*] under the *CEU–CEU* scenario shows MCC values increasing from 0.6286 to 0.6655 as a function of the number of unlabeled samples, whereas in the *Constant–CEU* setting, it exhibits a rise from 0.2941 to 0.3112. However, though *PULSe*[*P2*] under the *CEU–CEU* scenario demonstrates an increase in MCC values from 0.6253 to 0.6733, in the *Constant–CEU* setting, MCC drops from 0.3368 to 0.3205. Just like AUPRC, MCC follows the same pattern, with *P2* outperforming *P1* in every *Constant–CEU* test case despite the opposite directional trends.

Across both scenarios, both *PULSe* model variants consistently exhibit increasing neutral detection accuracy as the number of unlabeled samples increases. In the *CEU–CEU* scenario, *PULSe*[*P1*] neutral detection accuracy rises from 94.42% with 10,000 unlabeled samples to 97.57% with 20,000 unlabeled samples, whereas *PULSe*[*P2*] demonstrates a similar improvement, increasing from 93.96% to 97.61%. The *Constant–CEU* setting presents a similar trend of increase in neutral detection accuracies, with *PULSe*[*P1*] rising from 86.86% to 93.88% and *PULSe*[*P2*] from 87.14% to 93.74% as a function of the number of unlabeled samples.

Sweep detection accuracy on the other hand decreases consistently as the number of unlabeled samples grows. In the *CEU–CEU* scenario, *PULSe*[*P1*] sweep detection accuracy declines from 74.90% to 64.85%, with *PULSe*[*P2*] harboring a slightly larger decrease from 76.30% to 64.70%, indicating that the ability of

the model to detect the minority sweep class diminishes with increased unlabeled to labeled ratio. The *Constant–CEU* scenario results further attest to this trend, with *PULSe*[*P1*] experiencing a drop in sweep accuracy from 49.80% to 35.95% and *PULSe*[*P2*] falling from 55.10% to 37.40%.

In both domain-matched and -mismatched scenarios, *PULSe* MCC and AUPRC exhibit only mild fluctuations as the unlabeled-to-labeled ratio is varied, indicating that this ratio has little influence on overall performance. Consequently, the resulting changes in MCC and AUPRC are minimal compared to the pronounced shifts observed when sweep prevalence in the unlabeled set is varied (see *Model performance as a function of unlabeled set class imbalance* subsection above). This difference arises because, though both MCC and AUPRC are sensitive to the prevalence of the positive class in the unlabeled set, when the proportion of sweeps in the unlabeled set is held constant, these metrics are largely invariant to the absolute numbers of positive and negative samples. Meanwhile, the concurrent decline in sweep detection rates and increase in neutral detection rates points to increasing compression of predicted sweep probabilities toward lower values, indicating growing underconfidence and further motivating explicit calibration analysis.

Improving Detection Performance with Model Calibration

To assess the reliability of the *PULSe* class probability estimates, we examined the calibration of our classifiers. Because calibration curves are sensitive to class imbalance (Zadrozny and Elkan 2002; Niculescu-Mizil and Caruana 2005), we constructed our calibration test set by selecting 50% of the sweep samples uniformly at random from the labeled set along with an equal number of random samples from the unlabeled set. We find that for both *PULSe*[*P1*] and *PULSe*[*P2*] models in both the *CEU–CEU* and *Constant–CEU* scenarios, the calibration curves deviate markedly from the ideal $y = x$ line, indicating significant miscalibration of the predicted probabilities (Fig. S3).

Performing calibration within the PU learning framework is not straightforward because the labeled set contains only positive samples, and in empirical applications, we cannot assume the true class labels for the unlabeled set. To address this challenge, we devised a novel approach for calibration of *PULSe* models (see *PULSe calibration process* in the *Materials and Methods* for details). Specifically, we chose to apply isotonic regression (Zadrozny and Elkan 2002) for calibration, though we acknowledge that Platt scaling can also be used, which as a parametric method, requires learning two parameters through training and validation, and can help reduce overfitting (Platt 1999).

To thoroughly evaluate the effectiveness of our calibration approach, we conducted experiments across multiple configurations. We fixed the number of positive samples in the labeled set to 10,000, while varying the number of unlabeled samples between 10,000 and 20,000, and choosing the proportion of sweeps in the unlabeled set to either 5% or 10%. Additionally, we explored an array of positive class probability thresholds for selecting negative samples from the unlabeled set, with these cutoff values drawn from 0.2, 0.3, 0.4, 0.5, 0.6, or 0.7. Though a positive class probability threshold of 0.5 would seem the most natural choice, we examined this broader range because, particularly in the *Constant–CEU* scenario, both *PULSe* models were notably weak in detecting sweeps.

Our results show that regardless of the configuration, the calibration approach improved the calibration of the models (Fig. S3). However, several general patterns emerged. With our calibration method, predictions that were initially underconfident become increasingly well-calibrated as the calibration threshold is lowered. This improvement continues until a specific threshold is reached, beyond which further lowering the threshold no longer improves alignment with the observed outcomes. Generally at lower thresholds, overconfident predictions are corrected, reducing bias in predicted probabilities, which improves overall calibration performance. Consequently, we could not identify a single pipeline that performs optimally across both scenarios. Rather, the findings suggest that in cases of domain match, the simpler *P1* pipeline performs better, whereas in cases of domain mismatch, the more complex *P2* pipeline is more effective. This difference may arise because, in a domain matched setting, the additional processing in *P2* introduces unnecessary complexity that can dilute discriminative features already well-aligned between training and test data. We hypothesize that with *P2*, histogram equalization (see *Choosing the best feature extraction pipeline* subsection of the *Materials and Methods* for details) helps preemptively mitigate the effect of domain shift, while the higher number of orientation bins is needed to capture more subtle changes in gradients under a domain mismatched setup. Additionally, when examining *PULSe*[*P1*] in the *CEU–CEU* scenario and *PULSe*[*P2*] in the *Constant–CEU* setting, we observed that the calibration performance improves as the number of unlabeled samples increases, regardless of the proportion of sweeps in the unlabeled set (Fig. S3). This behavior is expected, as isotonic regression is less prone to overfitting with a larger number of training samples (Niculescu-Mizil and Caruana 2005).

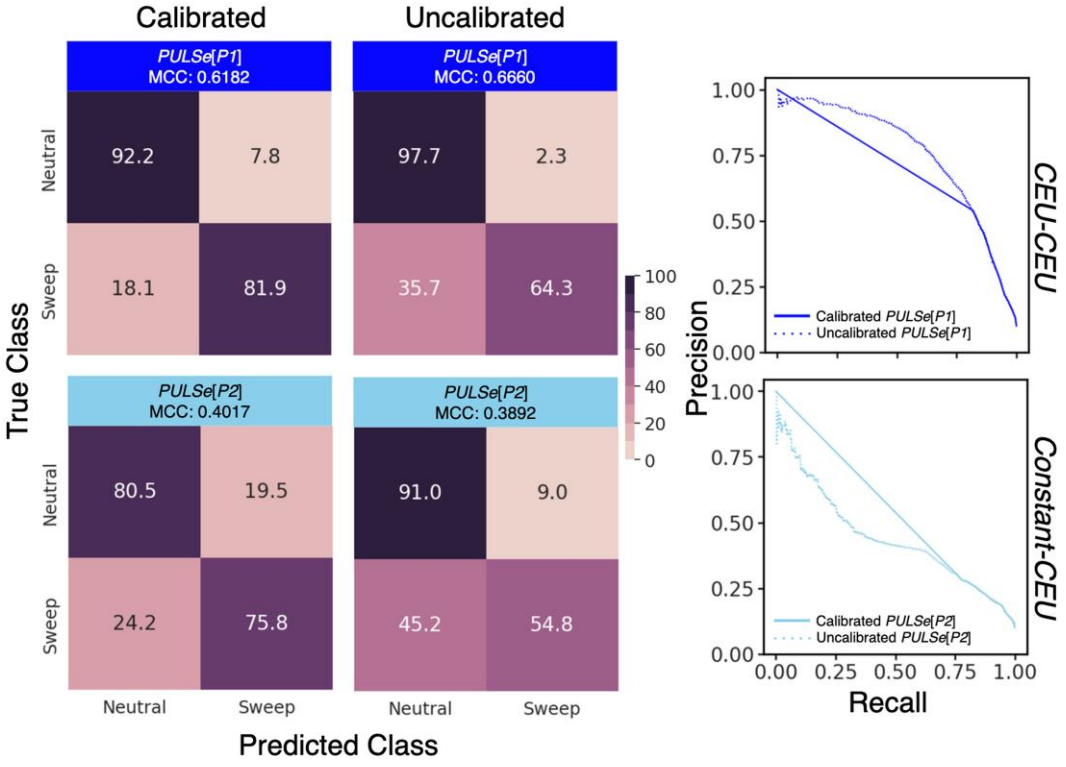

Fig. 3. Classification rates and accuracies as depicted by confusion matrices and reliability to detect sweeps as depicted by precision–recall curves to differentiate sweeps from neutrality on the unlabeled set for the *CEU–CEU* (top row) and *Constant–CEU* (bottom row) scenarios (see *Modeling Description* subsection of the *Results*) for calibrated versus uncalibrated *PULSe*[*P1*] and *PULSe*[*P2*] models. For both scenarios, the unlabeled set contains 1,000 sweep replicates and 9,000 neutral replicates simulated based on the recent strong bottleneck demographic history of the CEU human population from the 1000 Genomes Project dataset.

Comparing *PULSe* performance before and after calibration reveals an interesting contrast (Fig. 3). In the domain-matched scenario, AUPRC decreases from 0.7548 to 0.6895 following calibration, whereas in the domain-mismatched scenario, AUPRC increases from 0.4437 to 0.5464. More notably, in both cases, the PR curves exhibit near-linear behavior over the low-to-mid recall range. This behavior arises because calibration pushes many high-confidence predictions to values of exactly one and many low-confidence predictions to values of exactly zero, leaving a range of intermediate sweep probability values with few or no samples. As a result, adjusting the prediction threshold often does not change which samples are selected, leading to straight line segments in the PR curves. Concurrently, neutral detection rates decrease while sweep detection rates increase, yielding more balanced performance across classes. Overall, calibration mitigates the underconfidence of *PULSe* predictions (Fig. 4), but this comes at the cost of reduced discrimination between true and false positives among the highest-ranked samples.

Integrating Purifying Selection in Simulated Genomes

To evaluate the robustness of *PULSe* when the unlabeled set includes genomic regions affected by other evolutionary forces beyond only positive selection and neutral processes, we incorporated simulated replicates of background selection into the unlabeled pool. Background selection is a process by which purifying selection against deleterious mutations at a set of loci indirectly reduces genetic variation at nearby neutral sites due to linkage (Charlesworth et al. 1993; Hudson and Kaplan 1995; Charlesworth 2012). This setup presents a particularly informative test case, as background selection is a pervasive force (McVicker et al. 2009; Comeron 2014; Cvijović et al. 2018; Pouyet et al. 2018) that can produce patterns of genetic variation resembling those generated by selective sweeps (Keinan and Reich 2010; Seger et al. 2010; Nicolaisen and Desai 2013), thereby potentially complicating binary classification tasks.

We generated 1,000 background selection replicates using the forward-time simulator SLiM (Haller and

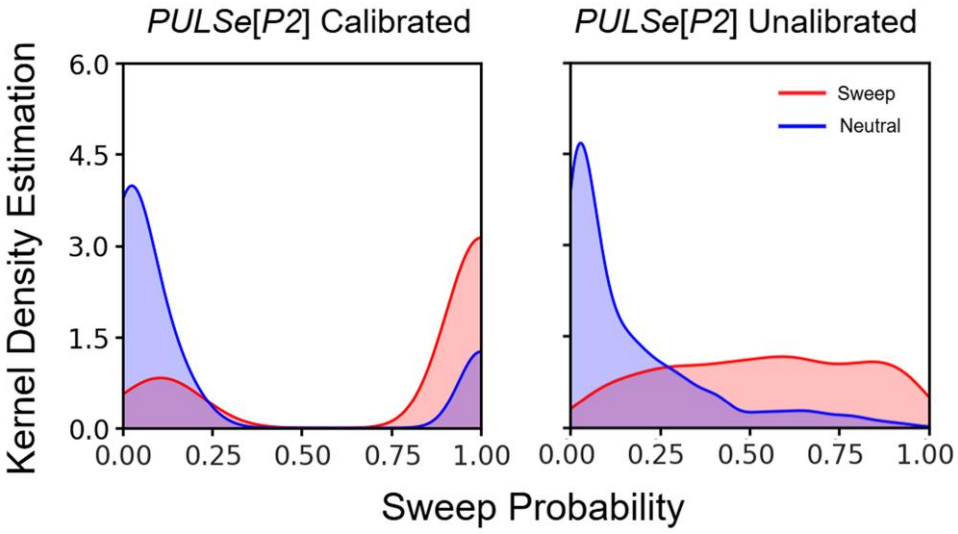

Fig. 4. Kernel density estimates (KDEs) of predicted *PULSe*[*P2*] sweep probabilities for 1,000 sweep (red) and 9,000 neutral (blue) replicates in the unlabeled set for the domain mismatched scenario. KDEs provide a nonparametric estimate of the underlying probability density function of prediction scores, with the area under each curve integrating to one. Panels compare calibrated and uncalibrated *PULSe*[*P2*] predictions, illustrating differences in prediction concentration and separation between sweep and neutral classes.

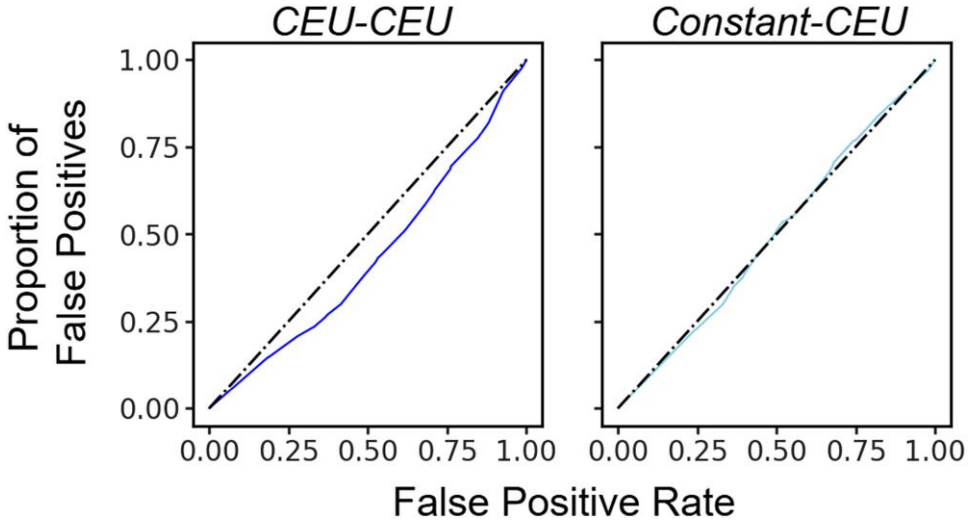

Fig. 5. The probability of falsely detecting background selection as a sweep as a function of the false positive rate for the *PULSe*[*P1*] and *PULSe*[*P2*] models applied to the *CEU–CEU* (left) and *Constant–CEU* (right) scenarios, respectively. Trained models are identical to those of Fig. 3, with the exception that 1,000 out of 9,000 neutral replicates were replaced by 1,000 background selection replicates. The proportion of false signals is calculated as the fraction of background selection replicates whose sweep probabilities exceed the threshold corresponding to a given false positive rate among neutral replicates.

Messer 2019), applying the same genetic and demographic parameters as those used in CEU simulations for training *PULSe*, but with the addition of background selection. Specifically, each 1.1 Mb simulated region contained a 55 kilobase protein-coding gene modeled after the simulated human gene architecture in Cheng et al. (2017), and composed of 50 exons (100 bases each), 49 introns (one kilobase each), 5′ and 3′ untranslated regions (UTRs; 200 and 800 bases, respectively), chosen to closely match the mean element lengths from the human genome (Mignone et al. 2002; Sakharkar et al. 2004). Within this gene, recessive ($h = 0.1$) deleterious mutations arose with selection coefficients drawn from a gamma distribution with mean of −0.0294 and shape parameter of 0.184 (Boyko et al. 2008; Schrider and Kern 2017). The percentage of mutations that were

deleterious in exons, introns, and UTRs were 75%, 10%, and 50%, respectively.

Our findings indicate that *PULSe* models under both domain matched and mismatched scenarios are robust toward misclassifying background selection replicates as selective sweeps (Fig. 5). This finding is unsurprising given that our image representations are based on haplotype structure, and unlike selective sweeps, background selection typically does not produce the same distinct haplotype patterns (Fagny et al. 2014; Schrider 2020). In the domain matched setting, the rate of classifying background selection replicates as nonsweeps (92.1%) nearly mirrors that of the neutral detection rate (92.2%). Interestingly, in the domain mismatched scenario, the rate of classifying background selection replicates as nonsweeps (81.6%) exceeds the neutral detection rate (80.3%). This result highlights that in the domain mismatched setting, *PULSe* is even less likely to mistakenly classify background selection as a sweep than it is to misclassify neutrally evolving regions, compared to domain matched setting.

Application to Unphased Data

Though *PULSe* models are originally trained on phased haplotype data, we tested their robustness when trained on unphased multilocus genotype (MLG) data. This is an important practical consideration, especially for nonmodel organisms where phasing is often unreliable or infeasible (Browning and Browning 2011). To generate *PULSe* input images from unphased data, we first isolate each pair of consecutive rows, *i.e.* haplotypes $2i-1$ and $2i$, where $i = 1, 2, \ldots, n/2$, from the original alignments of n haplotypes. We then transform each pair of isolated haplotypes into a single unphased MLG by counting the number of minor alleles at each diploid genotype, yielding genotypes coded as 0, 1, or 2. Next, we applied the same transformation and image construction procedures described in the *Image Generation* subsection of the *Materials and Methods* to produce the images, which results in an image of size 99×238 for our sample of $n = 198$ haplotypes. We resized these images to 198×238, and divided pixel values by a factor of two prior to computing HOG features to avoid altering the feature vector lengths and pixel value ranges used in *PULSe* models. Following image generation and resizing, we computed HOG representations using the *P1* and *P2* pipelines. Using these HOG representations, we trained two new models, denoted by *PULSe*[*P1*, *MLG*] under the domain matched scenario and *PULSe*[*P2*, *MLG*] under the domain mismatched scenario.

In the domain matched setting, *PULSe*[*P1*, *MLG*] achieved performance nearly identical to its haplotype-based counterpart *PULSe*[*P1*], with approximately 1.5% lower neutral detection rate and a slightly higher sweep detection rate (compare Fig. S4 with Fig. 3). Though, its MCC of 0.5912 and AUPRC of 0.6763 were also reduced, the overall results indicate that the model retains strong predictive power on MLG input in the domain match scenario. In contrast, *PULSe*[*P2*, *MLG*] showed a more notable drop under the domain mismatched setting, with a 1.7% reduction in neutral detection rate, an almost 10% loss in sweep detection rate, and a more pronounced decline in MCC (0.3364) and AUPRC (0.5049) compared to the haplotype setting (compare Fig. S4 with Fig. 3). Both *PULSe* MLG models were independently calibrated with a calibration threshold of 0.3. These results demonstrate that *PULSe* retains meaningful classification performance on unphased data in domain matched settings, though stronger performance when experiencing domain shift may require additional modeling modifications.

Effect of Lower Sample Size

Our assessment of *PULSe* models thus far has been performed with a fairly large sample size of 99 diploid individuals. To motivate the understanding of how a comparatively lower sample size affects *PULSe*, we evaluated its performance with roughly a quarter of the sample size of 50 haplotypes instead of 198. We then transformed these alignments into images using the same procedure detailed in the *Image generation* subsection of *Materials and Methods*. The resulting images are of size 50×238, which we resized to 198×238 prior to computing HOG features to facilitate direct comparison against *PULSe* models trained on 198 haplotypes without introducing confusion regarding different HOG feature vector sizes. Following image generation and resizing, we computed HOG representations using the *P1* and *P2* pipelines. Using these HOG representations, we trained two new models, denoted by *PULSe*[*P1*, *50h*] under the domain matched scenario and *PULSe*[*P2*, *50h*] under the domain mismatched scenario. Both *PULSe* 50h models were independently calibrated with a calibration threshold of 0.3. This setup allowed us to evaluate the effect of reduced population sample size on classification performance across both training scenarios.

In the domain matched setting, *PULSe*[*P1*, *50h*] exhibited a moderate decline in performance (Fig. S5) compared to *PULSe*[*P1*] (Fig. 3). The neutral detection rate dropped by approximately 7%, whereas the sweep detection rate improved slightly by approximately 3% compared to *PULSe*[*P1*]. However, both AUPRC (from 0.7548 to 0.6294) and MCC (from 0.6182 to 0.5073) declined, reflecting a loss in overall discriminative ability. Despite this loss, the results remain encouraging,

indicating that *PULSe* can still operate effectively with reduced sample sizes under domain matched conditions. In contrast, performance in the domain mismatched setting deteriorated substantially for *PULSe*[*P2*, *50h*]. Both neutral and sweep detection rates fell by over 10%, and AUPRC and MCC, respectively, dropped by around 8% (0.4616 compared to 0.5464) and 22% (0.2829 compared to 0.4017) relative to *PULSe*[*P2*] (compare Fig. S5 with Fig. 3). These findings closely mirror those observed for the *PULSe*[*P1*, *MLG*] and *PULSe*[*P2*, *MLG*] models in terms of the sharp decline in domain mismatch performance compared to *PULSe*[*P1*] and *PULSe*[*P2*] models.

Benchmarking Against Alternate Architectures

To contextualize the performance of *PULSe*, we benchmarked it against an alternate machine learning architecture under both domain match and mismatch scenarios. In an attempt to benchmark *PULSe* against a traditional domain-adaptive method, we first chose the small multi-branch CNN (*smbCNN*) architecture, which demonstrated strong performance in Arnab et al. (2025). For context, this architecture is the identical CNN used by Kern and Schrider (2018), except with two classes instead of five and operates on *PULSe* images (see *Image Generation* subsection of the *Materials and Methods*) we developed instead of the native summary statistic images.

Building on this foundation, we explored how gradient reversal layers (GRLs), as implemented in the dadaSIA framework (Mo and Siepel 2023), can render the *smbCNN* architecture domain adaptive. Mo and Siepel (2023) showed that equipping a previous model (termed SIA) (Hejase et al. 2022) with a GRL substantially improved performance under both domain matched and mismatched conditions. Prior work in domain adversarial learning has shown that GRLs allow neural networks to learn features that are both predictive of labels in the source domain and invariant to differences between source (training) and target (testing) domains (Ganin et al. 2016; Ding et al. 2019; Zügner et al. 2020). By inserting a GRL-connected domain discriminator into the *smbCNN* architecture, we adapted the network to jointly minimize classification loss on labeled source samples and maximize domain classification loss across both source and target samples, encouraging the feature extractor to obtain domain-invariant representations. This modification may allow the original architecture to function robustly across mismatched domain conditions without explicitly modeling population demographic history. Notably, we chose not to compare *PULSe* directly against dadaSIA, as it operates on ancestral recombination graphs (ARGs) as input that are fundamentally different from the *PULSe* images, thereby avoiding confounding differences arising from input types such as summary statistics or ARG features.

We denote the domain-adaptive alternate *smbCNN* architecture that takes *PULSe* images (*PI*) as input by *smbCNN_GRL*[*PI*], with two variants corresponding to a model with a hyperparameter tuned domain adaptation parameter (λ) variant (*smbCNN_GRL*[*PI, h*]) and a model with placing λ on a rate schedule (*smbCNN_GRL*[*PI, ls*]). A direct comparison with *PULSe* is not possible as *smbCNN_GRL*[*PI*] requires labeled samples from both classes for training, whereas *PULSe* operates with the labeled positive sweep class only. Nevertheless, this test provides a useful performance baseline even though *smbCNN_GRL*[*PI*] benefits from a significant advantage in training supervision. A detailed description of the *smbCNN_GRL*[*PI*] model architecture and training procedure is provided in the *Training the smbCNN_GRL[PI] architectures* subsection of the *Materials and Methods*.

In the domain-matched setting, both *smbCNN_GRL*[*PI*] models, *smbCNN_GRL*[*PI, h*] and *smbCNN_GRL*[*PI, ls*], comfortably outperformed *PULSe* across all evaluated performance metrics, including AUPRC, MCC, and class-wise detection accuracies. Among the two, *smbCNN_GRL*[*PI, ls*] exhibits slightly more balanced classification performance and achieves approximately a 2.5% improvement in class separation as measured by AUPRC (Fig. 6).

In contrast, in the domain mismatched setting, the scenario for which domain adaptive CNN approaches are explicitly designed for, the performance differences between *smbCNN_GRL*[*PI*] variants and *PULSe* are considerably mitigated. Compared to *PULSe*, *smbCNN_GRL*[*PI, h*] achieves only a modest 1% increase in AUPRC (0.5564 compared to 0.5464), while *smbCNN_GRL*[*PI, ls*] exhibits a slight reduction in AUPRC (0.5450 compared to 0.5464). Despite these marginal differences in AUPRC, both *smbCNN_GRL*[*PI*] models yield substantially lower MCC values (0.3420 and 0.1760, respectively, compared to 0.4017), driven primarily by poor neutral detection accuracies (67.1% and 34.5%, respectively; Fig. 6).

Examination of the kernel density estimates (KDEs) of sweep prediction values (Fig. 7) reveals distinct failure modes for the two *smbCNN_GRL*[*PI*] models. For *smbCNN_GRL*[*PI, h*], the score distribution lacks low-confidence predictions, with both neutral and sweep classes exhibiting density peaks around 0.4. However, the sweep density peak is lower than that of neutrality. Though the distributions remain partially overlapping there on, sweep predictions gradually dominate at intermediate predictions. Notably, at high prediction values, neutrality becomes more frequent than sweeps, indicating a substantial number of high-ranking false positives. This finding suggests that CNN overconfidence under

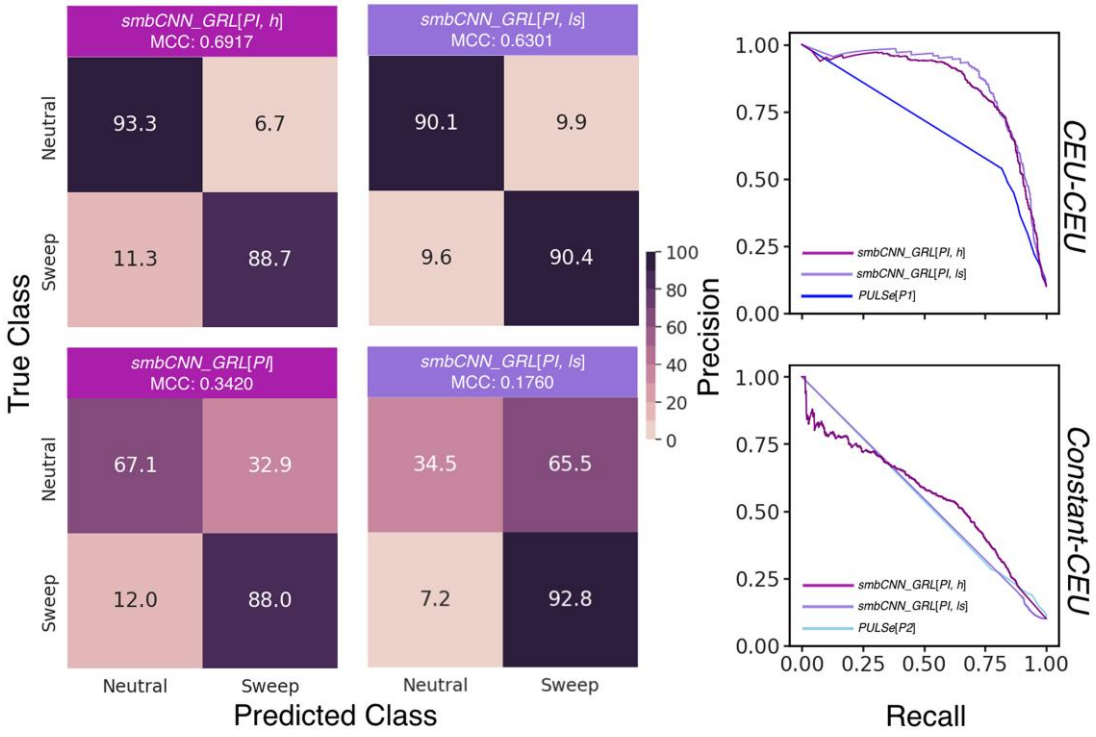

Fig. 6. Classification rates and accuracies as depicted by confusion matrices and reliability to detect sweeps as depicted by precision–recall curves to differentiate sweeps from neutrality on the unlabeled set for the *CEU–CEU* (top row) and *Constant–CEU* (bottom row) scenarios (see *Modeling Description* subsection of the *Results*) for the *PULSe*[*P1*] and *PULSe*[*P2*] models, compared against *smbCNN_GRL*[*PI*] (see *Benchmarking against alternate architectures* subsection of the *Results* for details). For both scenarios, the unlabeled set (test set in the case of *smbCNN_GRL*[*PI*] models) contains 9,000 neutral and 1,000 sweep replicates simulated based on the recent strong bottleneck demographic history of the CEU human population from the 1000 Genomes Project dataset. The precision–recall curves for the *PULSe* models are identical to those in Fig. 3.

pronounced train-test distribution shift (Hein et al. 2019; Wei et al. 2022) is not effectively mitigated by the inclusion of a domain adaptive discriminator branch (Fig. S6, bottom right panel). In contrast, *smbCNN_GRL*[*PI, ls*] does not exhibit the same overconfidence behavior. Instead, predictions are concentrated near extreme low or high score thresholds, suggesting that, in enforcing domain invariance through scheduled adaptation (Fig. S6, bottom left panel), discriminative structure is largely lost and overall classification performance is severely degraded. In contrast to these failure patterns, *PULSe* maintained coherent separation between sweep and neutral predictions in the domain-mismatched setting, albeit with reduced performance relative to the domain-matched scenario. However, the *smbCNN_GRL*[*PI*] variants exhibited distinct breakdowns in predictive behavior and did not retain stable discriminative performance under domain mismatch.

Application to Human Genomes of known European Demographic History

To explore empirical signals of adaptation, we applied *PULSe* to phased haplotype data from 99 individuals in the CEU population from the 1000 Genomes Project dataset (1000 Genomes Project Consortium 2015). As detailed in the *Empirical data filtration and image generation* subsection of the *Materials and Methods*, we first excluded low-mappability regions and partitioned each chromosome into overlapping windows of length 499 single nucleotide polymorphisms (SNPs) with a stride of 50 SNPs. Each window was rendered as an image of spatial haplotype variation, and HOG features were extracted using both domain matched (*P1*) and domain mismatched (*P2*) pipelines. These windows were then used as the unlabeled set to retrain the calibrated logistic regression models of *PULSe*[*P1*] and *PULSe*[*P2*] models, following the same calibration procedures used in the simulation experiments (see *PULSe calibration process* subsection of the *Materials and Methods* for details). After calibration, we identified an optimal threshold of 0.2 for the domain matched model and 0.3 for the domain mismatched models (Fig. S7).

Using the *P1* pipeline, trained under a matched demographic model, 3.359% of windows exceeded the sweep probability threshold of 0.5, with 3.357% of windows assigned a probability of exactly 1.0. In contrast, the *P2* pipeline, trained under a mismatched

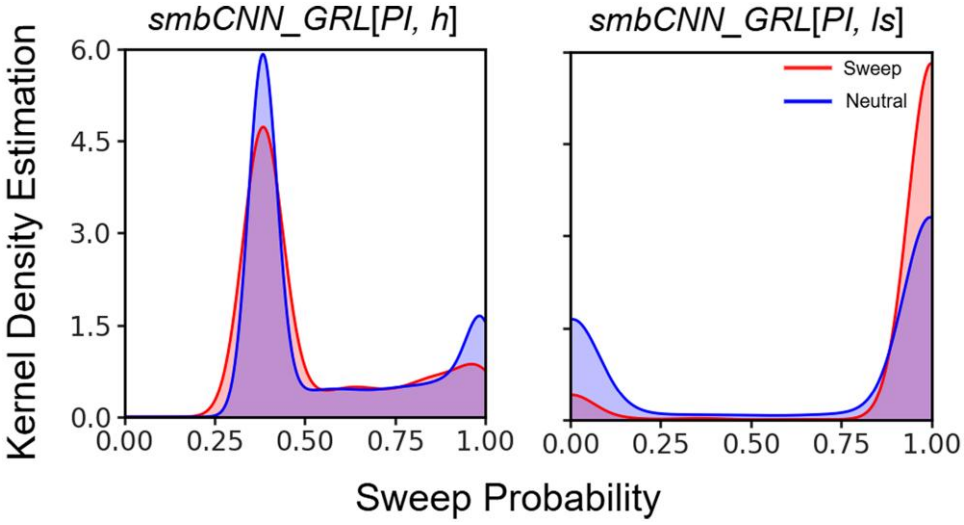

Fig. 7. Kernel density estimates (KDEs) of predicted *smbCNN_GRL*[*PI, h*] and *smbCNN_GRL*[*PI, ls*] sweep probabilities for 1,000 sweep (red) and 9,000 neutral (blue) replicates in the unlabeled set for the domain mismatched scenario. KDEs provide a nonparametric estimate of the underlying probability density function of prediction scores, with the area under each curve integrating to one. Panels compare two versions of domain-adaptive *smbCNN_GRL*[*PI*] predictions under different input representations, illustrating differences in prediction concentration and separation between sweep and neutral classes.

demographic model, classified 7.889% of windows above the 0.5 threshold, and 7.887% received a predicted probability of 1.0. In total, the high sweep probability windows with predicted sweep probability exceeding 0.5 found by *PULSe*[*P1*] fall within the bounds of 961 genes in our genome-wide scan, whereas the peaks identified by *PULSe*[*P2*] fall within the bounds of 2,122 genes.

Of the genes identified by *PULSe*[*P1*] and *PULSe*[*P2*], a total of 603 genes, which is around 63% of the *PULSe*[*P1*] discoveries, are shared between the two scans. The fact that *PULSe*[*P2*], despite being handicapped by domain mismatch, still recovered the majority of discoveries of *PULSe*[*P1*] reinforces confidence in the reliability of the *PULSe*[*P2*] results. The CEU scan, thus, provides a valuable opportunity to assess the utility of *PULSe*[*P2*] in a true domain mismatched application, as the CEU population has been extensively studied (Szpiech and Hernandez 2014; Chen et al. 2016; Duforet-Frebourg et al. 2016; Kern and Schrider 2018; Mughal et al. 2020; Wang et al. 2021; Arnab et al. 2025), and offers a strong comparative basis for evaluating our findings. Among the *PULSe*[*P2*] identified candidates are several well-established targets of positive selection in European populations, including *LCT*, which underlies lactase persistence and has long been recognized as a hallmark of recent adaptation to dairy farming (Bersaglieri et al. 2004; Sabeti et al. 2006). The finding of this signal lends credence to the ability of *PULSe* to recapitulate known adaptive loci. Another notable candidate is *ABCA12*, a gene essential to skin barrier function through its role in lipid transport in the epidermis (Akiyama 2014; Colonna et al. 2014). Adaptation at this locus has been linked to variation in environmental ultraviolet exposure and skin physiology, particularly among Eurasian populations. *PULSe*[*P2*] also detected selection at *HLA-DRB6*, a class II pseudogene located in the major histocompatibility complex (MHC) region, where long-term balancing and intermittent positive selection events are hypothesized to maintain diversity for pathogen defense (Cree et al. 2010). Though *HLA-DRB6* is nonfunctional, its genomic context suggests it may be linked to selective pressures acting on nearby regulatory or coding regions within the MHC.

It is important to note that all of the above candidates were also found by the *PULSe*[*P1*] scan. Interestingly, the common gene set also includes previously identified, but less well-characterized, candidates such as *RBMS3* and *NKAIN2*, both of which were also highlighted in earlier scans by the *TrIdent* classifier of Arnab et al. (2025). This convergence is particularly notable given that *PULSe* and *TrIdent* operate on almost identical image representations of haplotype variation, differing primarily in their downstream modeling architectures. *RBMS3* has been implicated as a tumor suppressor in breast cancer and may play roles in transcriptional regulation and cell cycle arrest (Yang et al. 2018), whereas *NKAIN2* has been linked to tumor inhibition and neurological function (Zhao et al. 2015). The recovery of these cancer-associated genes from independent scans suggests the possibility that some adaptive signals may be linked to pleiotropic effects involving disease resistance

or cell proliferation. Taken together, these overlapping results reinforce the capacity of *PULSe* to uncover both canonical and novel candidates of positive selection and offer a biologically meaningful set of targets for downstream functional and evolutionary investigation.

This consistency across pipelines further underscores the ability of *PULSe* to recover biologically meaningful candidates of positive selection and reinforces its ability to identify our targeted mode of selection across modeling conditions. Notably, over half of the genes identified by *PULSe*[*P1*] were also recovered by *PULSe*[*P2*], demonstrating a degree of robustness despite demographic model misalignment. The increased number of candidate sweeps detected in *PULSe*[*P2*] probably reflects a higher false positive rate, which is expected given that the labeled sweeps from the constant-sized demographic history lack the recent population size bottleneck present in the CEU population.

To investigate whether identified sweep candidates were enriched for particular biological functions or structures, we conducted Gene Ontology (GO) enrichment analysis using GOrilla (Eden et al. 2009) on two unranked lists of genes, where the target list contains candidate genes detected by the domain mismatched scenario of *PULSe*[*P2*], and the background list includes all genes that are not removed by our filtration process. Significant enrichment was determined using an false-discovery rate-adjusted q-value threshold of 0.05. Among the 29 significant biological process terms identified (Table S1), several top-ranked categories highlight regulatory processes. In particular, terms such as "regulation of multicellular organismal development," "regulation of neuron projection development," and "regulation of synaptic signaling" point to regulatory control over developmental and neuronal processes. We do not find any enriched molecular function terms. Meanwhile, enriched cellular component terms (Table S2) emphasized structures like "neuron part," "cell projection part," and "plasma membrane bounded cell projection part," reinforcing the notion that many identified sweep candidates may affect cell morphology and intercellular communication. The GO analyses highlight a broad but consistent enrichment of genes involved in cellular structure organization, signaling, and membrane-associated components, suggesting that positive selection in the CEU population may have acted on pathways essential for maintaining and regulating cell architecture and communication.

Application to Human Genomes of unknown South Asian Demography

To extend our evaluation of *PULSe* beyond the CEU population, we conducted a genome-wide scan of the Bengali in Bangladesh (BEB) population containing phased haplotype data from 86 individuals from the 1000 Genomes Project dataset (1000 Genomes Project Consortium 2015) using the same domain mismatched pipeline *PULSe*[*P2*] that was previously applied to the CEU. In this analysis, the model was retrained and calibrated using the BEB windows as the unlabeled set, and the identified optimal calibration threshold was 0.3 (Fig. S7).

The populations within Bangladesh exhibit high levels of genetic variation, characterized by contributions from a broad array of ancestral groups (Reich et al. 2009; Gazi et al. 2013; Narasimhan et al. 2019). These contributions include overlapping ancestries from South and East Asia, the Middle East, and Europe, which together contribute to a complex admixture landscape. This ancestry mixture has been shaped by centuries of trade, invasions, warfare, colonization, and religious migration, all of which introduced novel ancestral lineages into the region (Reich et al. 2009; Prakash 2014; Basu et al. 2016; Narasimhan et al. 2019). Despite this heterogeneity, the population shows minimal substructure that would reflect isolated population clusters (Chatterjee 2020). Instead, ancestral components are distributed broadly and continuously across the population, reflecting a long history of migration, admixture, and demographic change. A recent population genetic study on individuals from the Indian subcontinent revealed a complex pattern of ancestry, archaic introgression, and the maintenance of potentially deleterious variants (Kerdoncuff et al. 2025). Additionally, the population has experienced repeated natural disasters and recurrent exposure to infectious diseases, both of which may have exerted further selective and demographic pressures (Sen 1982; Haque 1997; Karlsson et al. 2013). Such demographic complexity makes a BEB genomic dataset an excellent testbed for assessing robustness of *PULSe* under domain mismatch, especially given admixture and history of environmental and sociopolitical disruptions experienced by the population.

PULSe identified 8.55% of windows exceeding the sweep probability threshold of 0.5, with a majority of those windows receiving a predicted probability of 1.0. The L_2-norm of the difference between the mean HOG feature vectors of the 1,000 windows with the highest sweep probabilities and the 1,000 windows with the lowest sweep probabilities, scaled by the length of the HOG feature vector, is 9.11×10^{-5} for the BEB population and 7.08×10^{-5} for CEU. This larger value for BEB indicates a greater separation between sweep and nonsweep candidates on average, which may explain the higher number of sweep signals detected in the BEB population. Overall, the identified

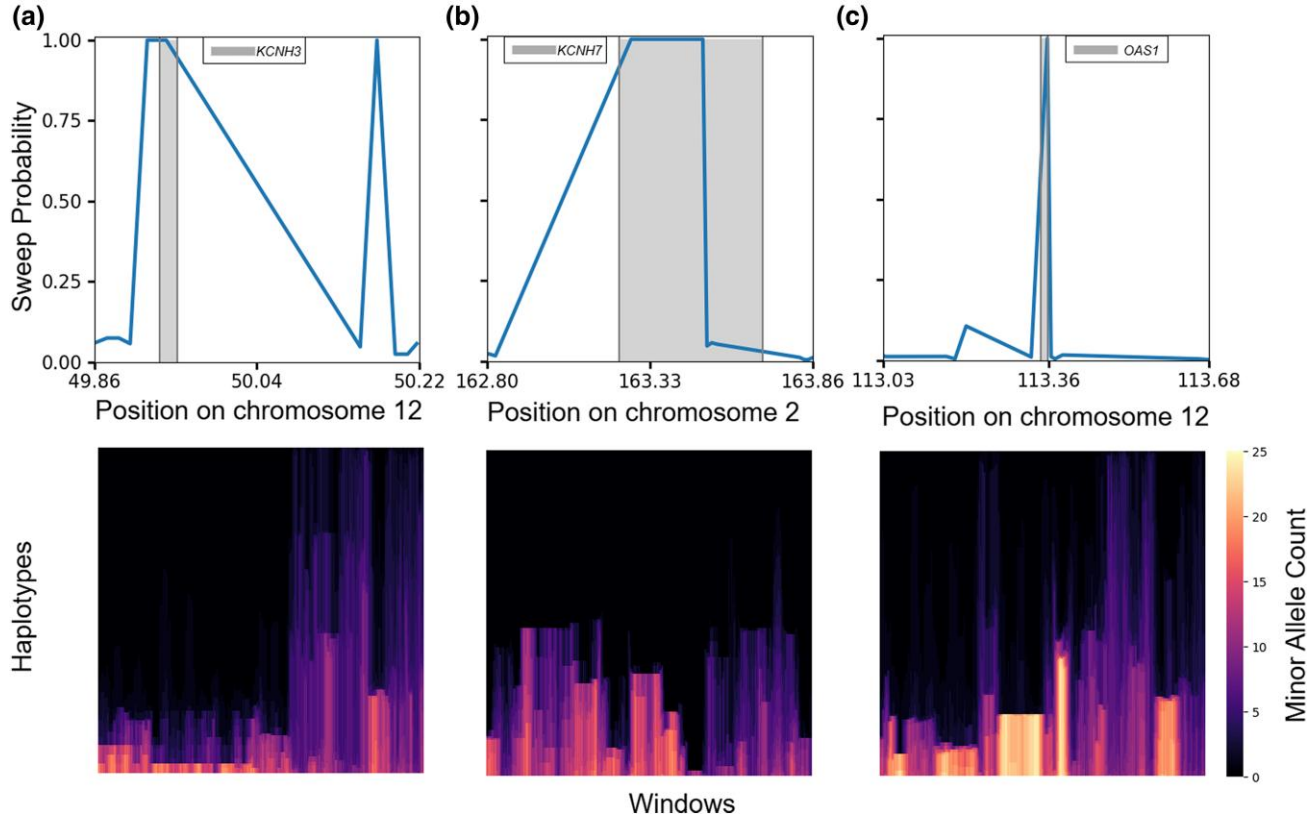

Fig. 8. Candidate selective sweep regions identified by *PULSe* in the Bengali in Bangladesh (BEB) population from the 1000 Genomes Project dataset. A total of 1,978 genes across 22 autosomes exhibit qualifying signs of selective sweeps, of which three (top row) of the most interesting candidates are reported here. In the plots, the candidate window within the gene of interest with the highest sweep probability is shown along with its five flanking windows on each side, giving a total of 11 windows. The bottom row presents corresponding *PULSe* images (see *Image generation* subsection of the *Methods* for details) for the same genomic regions reported in the top row, produced using all SNPs in the reported windows. Together, these panels link the statistical signal of elevated sweep probability (top row) with the image-based representation used by the model for classification (bottom row).

sweep probability peaks intersect the boundaries of 2,232 genes.

Among the most compelling sweep candidates detected are *KCNH3* (Fig. 8a) and *KCNH7* (Fig. 8b) on chromosomes 12 and 2, respectively, both of which encode voltage-gated potassium channels that regulate the flow of potassium ions across cell membranes (Jensen et al. 2012). These channels are essential for maintaining neuronal excitability and proper synaptic transmission, playing a critical role in electrical signaling within the brain (Singh and Auerbach 2024). Their involvement in neurodevelopmental processes is well established, and they have been implicated in susceptibility to various cognitive and psychiatric disorders, including schizophrenia, bipolar disorder, and dementia (Strauss et al. 2014; Wang et al. 2019; Rexach et al. 2023). *KCNH3*, in particular, has previously been identified as a shared sweep candidate between closely related South and East Asian populations (Liu et al. 2017). This finding is especially notable in light of recent studies reporting increased prevalence of certain mental health disorders among South Asian populations (Vidyasagaran et al. 2023; Islam et al. 2024). Just as our CEU scan identified cancer associated genes, we hypothesize that certain adaptations in South Asians, initially advantageous under past environmental stressors, may now predispose individuals to deleterious conditions in modern environments (Di Rienzo and Hudson 2005; Lea et al. 2023). In addition to its neurological role, *KCNH7* has also been linked to susceptibility to severe cholera in the BEB population (Karlsson et al. 2013), suggesting a possible dual role in adaptation that balances neurological function and resistance to infectious diseases in the pathogen-rich environment of Bangladesh.

Situated on chromosome 20, *DOK5* (Docking Protein 5) is another sweep candidate identified by *PULSe* that plays a role in insulin and neurotrophin signaling pathways (Cai et al. 2003). *DOK5* has been associated with

type II diabetes and metabolic disorders and has been identified as a candidate for positive selection in a scan of Indian populations (Metspalu et al. 2011), who are also from South Asia. Moreover, located on chromosome 15, *SLC24A5* emerged as another strong candidate signal of selection in our scan. This gene encodes a cation exchanger involved in melanosome function and is one of the most robustly documented targets of positive selection in human populations outside of Africa (Lamason et al. 2005). Furthermore, *PULSe* also detected *OAS1* (Fig. 8c) on chromosome 12 as a potential candidate, a gene that plays a central role in the innate immune response. The antiviral mechanism of *OAS1* (Melchjorsen et al. 2009) is critical for controlling infections such as dengue, hepatitis, and other RNA viruses, many of which are endemic to South Asia (Kristiansen et al. 2010). *OAS1* is also a hypothesized target of selection in primates (Mozzi et al. 2015), and in humans, it is thought to have been introgressed from Neanderthals (Sams et al. 2016). *PULSe* also detected multiple candidate genes within the MHC region, including *HLA-DRB1*, *HLA-DRB5*, and *HLA-DQB1*. The MHC region is a well-established hotspot of selection in human populations due to its central role in host–pathogen coevolution and pathogen defense (Trowsdale 2011). The identification of sweep signals in this region within BEB further supports the involvement of this region in adaptation to local pathogenic environments.

Our BEB scan also revealed sweep candidates that overlapped with those identified in the CEU population with the *P2* pipeline, including *ABCA12*, *RBMS3*, and *NKAIN2*. The discovery of these genes across populations with distinct evolutionary histories suggests that they may represent robust targets of positive selection, either due to convergent pressures or shared ancestral adaptations. In particular, the replication of candidates like *RBMS3* and *NKAIN2* indicates that certain adaptive signals may persist across different environments. These shared findings bolster the hypothesis that some adaptive variants may have broad functional relevance, potentially tied to disease resistance.

Our testing on the CEU demographic history (see *Improving detection performance with model calibration* subsection above) indicated that increasing the threshold reduces false positives but may reduce sensitivity to detect true sweeps. To assess the robustness of our BEB scan under alternative calibration parameters, we tested thresholds of 0.4 and 0.5. With a threshold of 0.4, the number of genes with predicted sweep signals decreased to 1,211, and with a cutoff of 0.5, the total declined further to 591. Notably, even under this strictest setting, only *ABCA12* and *SLC24A5* of the candidate loci highlighted above were excluded, suggesting that these genes likely contained weaker signals within the BEB scan.

To better understand the biological context of sweep candidates in the BEB, we carried out GO enrichment analysis using GOrilla (Eden et al. 2009). Candidate genes identified by *PULSe* were compared against a background set of all retained genes following our filtration procedure, with significance defined by an FDR-adjusted q-value threshold of 0.05. We did not identify any notable overlaps between the CEU and BEB GO terms. Among the 13 significantly enriched biological process terms (Table S3), several of the top categories reflected processes involved in structural organization, including "regulation of cell projection organization," "regulation of plasma membrane bounded cell projection organization," and "regulation of multicellular organismal process." These findings suggest that some candidate sweep signals may indicate potential selection on genes influencing cell morphology. Similar to the CEU scan, we do not find any enriched molecular function terms. In terms of cellular component enrichment (Table S4), top terms such as "neuron part", "cell projection part", and "plasma membrane" indicate a strong representation of genes associated with cellular extensions and membrane-associated functions. These results suggest that many of the BEB sweep candidates may play roles in shaping how cells interact with their environment, particularly in the context of neural and structural function.

Discussion

PU learning provides a natural framework for detecting selective sweeps from empirical genome variation, where only putative positive samples can be simulated and the vast majority of samples remain unlabeled. In developing *PULSe*, we used simulation-based analyses not as a prerequisite for applying the method to real data, but as proof of concept experiments to validate the training and calibration procedure under conditions where true class labels and class proportions are known. To enable direct comparison with standard binary classifiers and to ensure interpretability, we designed the benchmarking setup as a binary classification task. This controlled setting was not intended to mimic the full complexity of empirical genomic backgrounds. Extending the benchmarking to a three-class setting, where the unlabeled set included sweeps, neutral regions, and background selection, did not substantially alter detection of the adaptive class, indicating that the method is robust to additional sources of adaptation. Having established how *PULSe* behaves under controlled conditions, we next deployed it in its intended empirical setting. In this stage, simulated unlabeled

data were replaced with empirical genomic windows, and the calibrated scores were used to identify windows exhibiting patterns consistent with the simulated samples in the labeled set. Importantly, this design reflects a deliberate trade-off. By avoiding explicit simulation and labeling of nonsweep classes in empirical application, and instead treating the heterogeneous genomic landscape as unlabeled, *PULSe* reduces reliance on detailed assumptions about the underlying evolutionary processes. As expected, relaxing these assumptions comes at the cost of performance relative to fully specified, domain-matched models. However, our results show that *PULSe* remains comparatively stable under misspecification, highlighting its utility when prior knowledge of the system is limited.

An important consideration in applying *PULSe* to empirical data is the relationship between the labeled and unlabeled sets in the PU framework. In our simulated benchmarking experiments, the selected completely at random (SCAR) assumption theoretically holds by design, as labeled sweeps are randomly drawn from the same pool as the unlabeled sweeps, ensuring a representative sampling of the positive class. In contrast, when moving to empirical applications, SCAR cannot be assumed. The labeled positives originate exclusively from simulations, whereas the unlabeled set consists of real genomic windows, creating a fundamental mismatch in how positive examples enter the two sets. This mismatch reflects precisely the types of SCAR-violation effects documented in recent work, such as systematic miscalibration (Kumar and Lambert 2024) and validation bias (Molotkov and Artomov 2023). We generated a broad range of sweep scenarios spanning diverse selection coefficients, fixation times, sweep softness, and demographic histories. Nevertheless, it remains impossible to guarantee that the sweep configurations represented in the empirical sample are mirrored in the simulated labeled set, and some sweep types may be over- or under-represented relative to reality. Such imbalance can shift the distribution of predicted sweep probabilities upward or downward when applied to empirical data. However, this challenge is not unique to PU learning, as any supervised classifier trained on simulated sweeps, regardless of whether labels are fully observed, would encounter the same fundamental problem when the labeled simulations and the empirical genome differ in the underlying distribution of sweep types.

PULSe exhibited strong performance in the selective sweep detection task, serving as a proof of concept for evaluating the framework. In the domain-matched setting, analogous to the single-set training protocol described by Elkan and Noto (2008), where the demographic history of the training samples mirrors that of the test data, the *smbCNN_GRL*[*PI*] models clearly outperform *PULSe*, an expected outcome given their access to explicit labeled background (neutral) samples. When extended to the domain-mismatched scenario, analogous to the case-control design in Elkan and Noto (2008), the relative advantage of the domain-adaptive models is substantially reduced. In this regime, both *smbCNN_GRL*[*PI*] variants achieve only AUPRC values close to that of *PULSe*, suggesting similar power in terms of class separability under demographic shift. However, this apparent parity in AUPRC does not translate into comparable overall classification performance. Both *smbCNN_GRL*[*PI*] models suffer from highly skewed sweep probability distributions, leading to markedly reduced MCC and elevated false detection rates. Though *smbCNN_GRL*[*PI, h*] offers a more reasonable comparison, given the spread and shape of its sweep probability distributions on the unlabeled set, its outputs are best interpreted at present as ranks rather than as binary predictions and would require further refinement to function reliably in a binary classification setting.

It is also important to clarify that the ability of the *smbCNN_GRL*[*PI*] models to detect domain shift does not necessarily imply failure of the domain discriminator branch. In practice, domain-adversarial training optimizes a trade-off between maximizing predictive performance and minimizing domain-specific information in the learned representation. As a result, invariance is typically approximate rather than complete, and residual domain signal may persist, particularly when fully eliminating domain cues would also degrade task-relevant structure. Thus, discriminator success in identifying domain shift should be interpreted as evidence of incomplete invariance under strong distributional differences, rather than as a binary indicator that the adversarial branch has failed.

In contrast, *PULSe* initially exhibits the opposite tendency, under-predicting based on raw score distributions. Prior to calibration, though underconfident overall, the highest-ranked predictions were almost exclusively true positives, as evidenced by the sweep probability density plots (Fig. 4). However, true sweep instances were broadly and nearly uniformly distributed across the sweep probability spectrum, resulting in many sweeps remaining undetected at conventional thresholds. The calibration procedure restructures this distribution by introducing clearer separation between high- and low-confidence predictions, enabling the sweep probabilities to function effectively as binary predictions. As a consequence, calibration increases sweep detection sensitivity but also elevates the density of false positives at higher score thresholds. This behavior reflects an inherent tradeoff in which improved sweep detection and interpretable binary predictions are

achieved at the cost of an increased risk of false positive calls.

Several factors might underlie the pronounced performance drop of *smbCNN_GRL*[*PI*] models in the domain mismatched scenario. First, a key reason may lie in how domain adversarial networks function, as they aim to align the marginal distributions of the source and target feature representations globally, effectively forcing the feature extractor to produce domain-invariant representations. This strategy works well when the domain shift is expected to be relatively uniform across classes (Ganin et al. 2016). However, the interplay between natural selection and demographic history means that domain shifts may often manifest differently for non-neutral and neutral loci. Demographic bottlenecks or expansions can accentuate or obscure non-neutral regions in ways that do not uniformly affect all genomic regions, often altering patterns of variation and linkage disequilibrium that resemble the signatures of selection (Johri et al. 2022). As a result, regions that are truly neutral in the target domain may still display features that look like sweeps under the feature representations learned from the source domain training data. Because domain adversarial networks primarily align the overall feature distributions without explicitly disentangling class-specific shifts, this may cause the model to confuse these neutral regions for sweeps. This drawback likely contributed to *smbCNN_GRL*[*PI*] disproportionately misclassifying many neutral samples under the domain mismatched scenario.

Given the complexities of natural populations, the unlabeled dataset we used in this study intentionally featured a strong imbalance, with the majority of samples representing nonsweep regions. This imbalance is supported by several empirical scans, which suggest that loci evolving under positive selection leading to sweeps are relatively rare (Hernandez et al. 2011; Lohmueller et al. 2011; Granka et al. 2012; Jensen et al. 2019). Due to the significant class imbalance in our unlabeled sets, we prioritized the AUPRC and MCC as our primary and secondary evaluation metrics, respectively, when selecting the optimal *PULSe* architecture and comparing its performance against *smbCNN_GRL*[*PI*] models. Though both metrics are sensitive to class imbalance, they capture complementary aspects of model behavior that are critical in our extremely imbalanced setting. AUPRC is driven primarily by performance on the minority (sweep) class and, therefore, reflects the ability of a model to rank true sweep regions ahead of a large background of neutral regions. In contrast, MCC places greater emphasis on balanced performance across all outcome types and is strongly influenced by correct classification of the majority (neutral) class. By considering AUPRC and MCC jointly, we are able to disentangle ranking power from decision reliability, distinguishing models that effectively separate sweep signals and background variation versus those that also produce stable and usable binary predictions.

In contrast, receiver operating characteristic (ROC) curves and the associated area under the curve (AUROC) summarize classifier performance through the tradeoff between true positive rate (TPR; recall) and false positive rate (FPR) across all thresholds (Carter et al. 2016; Chicco and Jurman 2023). Under severe class imbalance, AUROC can remain artificially high even when a model produces a large number of false positives, rendering it comparatively uninformative for assessing practical detection performance in our setting. PR curves, by contrast, directly reflect the proportion of predicted positives that are correct and therefore explicitly penalize an increase in false positives. Consequently, models with similar TPR-FPR behavior can exhibit markedly different precision at the same recall level, particularly under extreme class imbalance, leading ROC- and PR-based summaries to diverge even when both nominally assess class separation. However, there is a formal connection between the two representations as well. When the ROC curve of one model consistently dominates that of another across all operating points, the corresponding PR curve will also dominate in precision–recall space (Davis and Goadrich 2006). In this context, ROC curves for *smbCNN_GRL*[*PI*] models on the unlabeled set show markedly higher power than *PULSe* in the domain matched setting (Fig. S8). Under domain mismatch, ROC performance declines for the *smbCNN_GRL*[*PI*] models, but remains higher than *PULSe* for *smbCNN_GRL*[*PI, h*], whereas *smbCNN_GRL*[*PI, ls*] exhibits poor ROC performance, consistent with its overall performance degradation. In the case of *smbCNN_GRL*[*PI, h*], the divergence reflects improved global separation across thresholds for *smbCNN_GRL*[*PI, h*], which elevates AUROC through stronger TPR-FPR tradeoffs at intermediate sweep probabilities. However, because high-score predictions remain substantially contaminated by neutral regions (Fig. 7), this improvement does not translate into higher precision among top-ranked sweep candidates, resulting in AUPRC values comparable to those of *PULSe*.

To extract useful features for input to *PULSe*, we employed HOG (Freeman and Roth 1995), a legacy computer vision technique that encodes spatial autocovariance by capturing local gradients and edge directions, effectively preserving spatial relationships of pixel values in *PULSe* images that simple image flattening would not be able to preserve (Elkan and Noto 2008). Moreover, HOG produces feature arrays of fixed size for a given image dimension, a property that many contemporary approaches do

not guarantee without additional pooling or resizing (Lowe 1992; Bay et al. 2006; Rublee et al. 2011). Modern alternatives such as SIFT (Scale Invariant Feature Transfrom; Lowe 1992), SURF (Speeded-Up Robust Features; Bay et al. 2006), or ORB (Oriented FAST and Rotated BRIEF; Rublee et al. 2011) offer robust scale and rotation invariance. However, these benefits come at a higher computational cost and are unnecessary for *PULSe* images, where neither scale nor orientation is expected to greatly vary. On the other hand, if genomic variation is summarized into one-dimensional vectors, such as summary statistics, then we do not anticipate any need for further transformation to fit the base classifier of *PULSe*. Conversely, if working with multidimensional arrays, then users are not restricted to using HOG. Any feature extraction method can be employed, provided it yields a consistent feature length across samples and ensures that each feature position represents the same characteristic for all samples.

To showcase its practical utility and flexibility, we applied *PULSe* to two distinct empirical contexts. We first focused on the well-studied CEU population, which served as a baseline to understand how *PULSe* behaves when deployed through its two pipelines. Using *PULSe*[*P1*], with labeled positives derived from sweep replicates simulated under an inferred CEU demographic model, we successfully recovered many well-established selective sweep candidates. Meanwhile, *PULSe*[*P2*], trained on sweep replicates generated from a constant population size demographic model, was still able to recapitulate a substantial portion of the sweep candidates identified by *PULSe*[*P1*]. Even with the introduction of the signature population bottleneck of the CEU population, *PULSe*[*P2*] identified the majority of literature-supported candidates found by *PULSe*[*P1*], albeit with an increased number of sweep discoveries. This high degree of overlap in their candidate genes provides support for the robustness of the assessed pipelines, lending confidence that the two *PULSe* pipelines do not fundamentally alter the empirical signals detected. Furthermore, it is worth emphasizing that these analyses were conducted using calibration thresholds optimized for the best calibrated state of each pipeline. However, calibration thresholds can be adjusted to identify only the strongest candidates or to cast a wider net with less strict cutoffs.

Our application of *PULSe* to the BEB population was potentially the more challenging, yet more significant of the two test cases, particularly because of its intricate demographic history. Positioned between South and Southeast Asia, Bengal has experienced complex layers of ancestral contributions, from early Austroasiatic and Dravidian settlers (Basu et al. 2016; Silva et al. 2017) to moderate Indo-Aryan migration (Metspalu et al. 2011; Narasimhan et al. 2019), centuries of rule by Persian, Turkic, and Central Asian dynasties (Eaton 1993), and more recent European colonial contact (Prakash 2014; Chatterjee 2020). These historical movements were further compounded by distinct gene flow events that left Bengali individuals with particularly high proportions of East Asian ancestry among South Asians (Reich et al. 2009; Gazi et al. 2013). The population has also endured repeated, though relatively more recent demographic upheavals through famines, cyclones, and seasonal floods (Sen 1982; Haque 1997; Mirza 2002), massive displacements during the partition of India, periods of forced starvation under the British Empire (Mallik 2024), and the 1971 Bangladesh Liberation War (Chatterji 2007; Ranjan 2016), all of which likely induced bottlenecks, founder effects, and admixture signals that complicate genetic landscapes. Despite this rich and challenging demographic backdrop, and the fact that they represent a large fraction of the global population, the Bengali people remain underexplored in population genomics. By applying *PULSe* in this context, we showcase its ability to operate in complex populations, providing insights that could inform more targeted studies on adaptation and population history in this and similar groups.

An important caveat of our approach relates to the level of biological resolution inherent to any window-based featurization strategy. Because *PULSe* represents genomic regions as fixed-size SNP windows encoded through minor allele counts along a haplotype, it cannot localize specific causal variants or determine whether signals of adaptation across populations originate from the same underlying mutation. This limitation is not unique to *PULSe*, as it is shared broadly by methods that use input summarized by genomic windows. Moreover, differences in SNP density across populations, together with our filtration process, naturally cause window boundaries to shift slightly, making perfect alignment across populations unattainable. Thus, this property reflects an inherent trade-off of using windowed genomic representations, one that prioritizes robustness to confounding factors and generalizability, rather than a shortcoming of *PULSe* itself. Future extensions that incorporate position-aware inputs could complement this framework, but the current behavior is expected and typical for methods of this category.

Fundamentally, it is of high importance to place *PULSe*, and more broadly, PU learning, within its practical application space, highlighting both its strengths and limitations. In our domain matched experiments, the supervised *smbCNN_GRL*[*PI*] model outperformed *PULSe*, as expected given its explicit use of labeled information about the neutral (negative) class. However,

PULSe required no such negative class data, relying solely on samples of the class of interest, which substantially simplifies study design and minimizes assumptions about the demographic or selective forces shaping the rest of the genome. This advantage makes *PULSe* particularly appealing when simulating or defining realistic negative classes is infeasible. Its straightforward design also means that predictions are easier to interpret, especially when the negative class fails to encompass all evolutionary and nonevolutionary dynamics expected to impact genetic variability in a population. However, most notably, our results indicate that *PULSe* is particularly helpful under domain mismatch scenarios—settings where demographic shifts or sampling differences might undermine more traditional supervised or domain-adaptive approaches. Here, the ability to adjust the calibration threshold becomes invaluable, lending control over the balance between stringency and flexibility, from only capturing the strongest candidates of the target class to broadly flagging all regions suggestive of positive selection. Taken together, these characteristics underscore the versatility of *PULSe* and its promise for evolutionary inference across an array of study systems. Looking forward, an intriguing avenue for future exploration is whether rather than relying exclusively on simulations for the labeled set, one could consider generating augmented copies of well-characterized adaptive regions from empirical genomes of the analyzed population. Such an approach would effectively place the analysis in a domain matched setting, as the empirical images provide direct examples from the expected positive class distribution, which is a scenario where *PULSe* achieves its highest power. However, this strategy carries an important limitation. Relying on augmented empirical sweeps could bias the labeled set toward only the strongest or most easily detectable signals, which would reduce the benefit of applying a machine-learning classifier. Generative approaches (Booker et al. 2023) may help alleviate this issue by combining the controllability of parameterized sweep models with the realism of empirical data, which remains an open direction for future research.

Materials and Methods

Simulation Protocol

We employed the coalescent simulator `discoal` (Kern and Schrider 2016) to generate both neutral and sweep replicates under two demographic models: a nonequilibrium demographic history estimated for European (CEU) humans (Tennessen et al. 2012) and a constant population size model (*Constant*). The CEU demographic model incorporates a recent severe population bottleneck, capturing the complex evolutionary dynamics of this population, whereas the *Constant* model permits investigation of selection under a stable population size assumption.

To simulate selective sweeps, we considered per-generation selection coefficients (s) drawn uniformly at random within the range [0.005, 0.1], initial beneficial allele frequencies (f) sampled uniformly at random within $[1/(2N_e), 0.2]$ (where $N_e = 18,750$ for the CEU population Schrider and Kern 2017), and fixation times (τ) uniformly distributed within [0, 2,000] generations before sampling (Schrider and Kern 2017). These choices ensure that our simulations capture a broad array of sweep dynamics, including varying levels of selection strength, hard versus soft sweeps, and different temporal stages of fixation, which all contribute to the detectability of selective events in genetic data.

Both neutral and sweep replicates were simulated with per-site per-generation mutation rates (μ) sampled uniformly at random within the interval $[2.21 \times 10^{-9}, 2.21 \times 10^{-8}]$, with a mean of 1.21×10^{-8} (Scally and Durbin 2012; Schrider and Kern 2017). The per-site per-generation recombination rate (r) was drawn from an exponential distribution with a mean of 10^{-8} and truncated at three times the mean (Payseur and Nachman 2000; Schrider and Kern 2017). These settings ensure that our simulations incorporate realistic levels of mutation and recombination heterogeneity.

For each replicate simulated under the inferred CEU demographic history, we sampled 198 haplotypes of length 1.1 megabases (Mb), reflecting the empirical dataset structure. Similarly, for the *Constant* demographic model, we maintained the same simulation parameters and number of haplotypes, except that the effective population size was drawn uniformly at random from $N_e \in \{3,000, 3,250, \ldots, 30,000\}$. For the domain-mismatched scan of the BEB population, *Constant* demographic model sweep simulations were generated in an analogous manner, using the same parameters but with 172 haplotypes to match the target population sample size.

Image Generation

Our image generation process closely follows the approach described in Arnab et al. (2025). Unlike the original framework in Arnab et al. (2025), we did not perform image resizing, as our approach did not require adapting images to fit predefined neural network input sizes. We omitted this step while retaining the core methodology.

For each simulated replicate, we retained only bi-allelic SNPs with a minor allele count of at least three

(*i.e.* singletons and doubletons were removed). We then generated a matrix **M** of dimension $n \times 499$, where n represents the number of haplotypes in a replicate. Each row of **M** corresponds to a sampled haplotype, whereas each column represents one of the 499 SNPs. These SNPs were selected as follows: the SNP closest to the central position, the 249 closest SNPs upstream of this central SNP, and the 249 closest SNPs downstream of this central SNP. The central position in a simulated genomic region is the site that beneficial mutations are introduced in sweep simulations. The element $\mathbf{M}_{ij}$ takes a value of zero if the haplotype in the ith row contains the major allele at the SNP in the jth column and a value of one if it contains the minor allele.

To create a structured representation of genomic variation for improved pattern recognition, we constructed a transformed matrix **X** of dimension $n \times 237$. This transformation was achieved by applying a sliding window approach over the columns of **M**, using windows of 25 SNPs with a stride of two between each window. Specifically, for $j \in \{1, 2, \ldots, 237\}$, we computed the minor allele counts for each haplotype within a genomic window of length 25 SNPs, starting at column $2j - 1$ of **M**. These minor allele count values were then sorted in increasing order and assigned to column j of **X**. As a result, rows toward the top of **X** represent haplotypes with a greater number of major alleles, whereas rows toward the bottom contain haplotypes with more minor alleles. This transformation provides a condensed summary of minor allele distributions across haplotypes in a structured manner, enhancing downstream image-based learning approaches.

Choosing the Best Feature Extraction Pipeline

To determine the optimal feature extraction pipeline for a downstream PU learning architecture, we evaluated different preprocessing techniques, HOG parameters, and postprocessing methods. This process of choosing the best pipeline ensures that extracted features effectively capture distinguishing patterns. We evaluated the feature extraction pipelines under both *CEU–CEU* and *Constant–CEU* scenarios. To evaluate and select among HOG feature generation pipelines, we used a logistic regression model as the base classifier. We also opted for 10,000 labeled sweeps and 10,000 unlabeled samples for both scenarios. In both cases, we maintained a consistent imbalanced ratio of sweep and neutral replicates in the unlabeled set, with 1,000 sweep replicates (10% of the unlabeled set). The effect of additional imbalance ratios is assessed in the *Model performance as function of unlabeled set class imbalance* and *Improving detection performance with model calibration* subsections of the *Results*, respectively.

We considered preprocessing input images before computing HOG features to aid in providing consistency across images. We tested four approaches: no preprocessing, histogram equalization (Gonzalez 2009), image normalization (scaling pixel values in an image to range from zero to one), and division by standard deviation (scaling pixel values in an image so that its pixel values have a standard deviation of one). While no preprocessing serves as a baseline, histogram equalization improves contrast by redistributing pixel intensities, whereas both image normalization and division by standard deviation reduces brightness differences among images while preserving relative intensity information within images.

After preprocessing, we extracted features using HOG with different parameter settings, chosen based on prior studies (Lowe 2004; Dalal and Triggs 2005). The working principle of HOG is detailed in the *Histogram of oriented gradients* subsection in *Supplemental methods*. In HOG, images are divided into cells, where gradient orientations are binned, and neighboring cells are grouped into blocks for normalization. The number of gradient orientations controls angular resolution, smaller cells capture finer spatial detail while larger cells capture coarser patterns, and increasing the number of cells per block trades off finer localization for greater stability in the feature representation. These parameters together determine the length of the resulting feature vector and balance the overall tradeoff between capturing fine-grained local patterns and maintaining computational efficiency. We varied the number of gradient orientations (6, 9, or 12), pixel sizes per cell (8×8 or 16×16), and cells per block (2×2 or 3×3) to determine which configuration best captures local patterns for better discrimination between sweep and nonsweep. These choices impact the level of detail retained in the feature representation, with smaller cells capturing finer details and larger cells providing coarser patterns.

We applied postprocessing to the extracted HOG features to further refine the input for the PU learning model. We tested no postprocessing, sample-wise mean subtraction (ensuring a mean of zero across the features of a single sample), feature-wise standardization (ensuring a mean of zero and standard deviation of one across samples), and sample-wise normalization (ensuring feature values from zero to one across a single sample). These transformations help mitigate biases in feature distributions and improve comparability across different images.

Based on AUPRC scores, we found that two postprocessing approaches, feature-wise standardization and sample-wise normalization, were the least effective in both *CEU–CEU* and *Constant–CEU* scenarios (Tables S5 and S6). Consequently, we excluded these methods

from further analysis. In the *CEU–CEU* scenario, the highest AUPRC of 0.7327 was achieved using a feature extraction pipeline with no preprocessing, six gradient orientations, a pixel size of 16×16 per cell, a block size of 3×3 cells, and sample-wise mean subtraction as postprocessing. The second-highest AUPRC of 0.7326 was obtained with the same configuration, but without any postprocessing. Given that these two pipelines performed nearly identically, we selected the second configuration due to its comparative simplicity. We refer to this pipeline as *P1* going forward.

In the *Constant–CEU* scenario, the overall performance was noticeably poorer compared to the *CEU–CEU* scenario, as expected. The best AUPRC achieved was 0.3544 using a pipeline with histogram equalization as preprocessing, nine gradient orientations, a pixel size of 16×16 per cell, a block size of 3×3 cells, and no postprocessing. The second-best AUPRC was 0.3522, obtained with the same configuration but incorporating sample-wise mean subtraction as postprocessing. Given the marginally higher performance and the reduced complexity of the best-performing configuration, we selected the first one as the optimal pipeline for this scenario, which we denote as the *P2* pipeline.

Comparing these results to the *CEU–CEU* scenario, we observe that the inclusion of histogram equalization and a higher number of gradient orientations (nine instead of six) led to improved performance in the *Constant–CEU* setting. Histogram equalization enhances contrast by redistributing pixel intensities, which likely helps mitigate variation in image brightness and improves edge detection in a dataset with more heterogeneous samples. Similarly, increasing the number of gradient orientations allows the HOG descriptor to capture more detailed directional patterns, which may be particularly beneficial when dealing with the greater domain mismatch in the *Constant–CEU* scenario.

To assess the generalizability of the selected pipelines, we compared the performance of *PULSe* across both scenarios using the *P1* and *P2* pipelines and, respectively, denote these method variants by *PULSe*[*P1*] and *PULSe*[*P2*]. When evaluated on the *CEU–CEU* scenario, *PULSe*[*P2*] exhibited a slight decline in performance, with an AUPRC drop of 0.014 and an MCC drop of 0.0124 compared to *PULSe*[*P1*] (Fig. S9). Conversely, in the *Constant–CEU* scenario, *PULSe*[*P1*] demonstrated a larger AUPRC decline of 0.0274 relative to *PULSe*[*P2*], and an MCC decline of 0.0142 (Fig. S9). These results suggest that, in the specific scenarios tested here, the *P2* pipeline showed slightly stronger performance under domain mismatch, whereas the *P1* pipeline performed better under domain match. Whether these patterns hold more broadly would require evaluation across a wider range of demographic and selection scenarios.

PULSe Calibration Process

To improve the reliability of *PULSe* predictions, we devised a novel calibration step tailored to our PU setting, in which the labeled set contains only sweeps while the unlabeled set contains a mixture of sweeps and nonsweeps. Standard calibration approaches are not directly applicable because negative labels are unavailable. Therefore, we construct an approximate calibration set from the available data and fit a nonparametric mapping from raw to calibrated probabilities.

Specifically, we use isotonic regression for calibration (Zadrozny and Elkan 2002). Isotonic regression makes minimal functional assumptions and fits a monotonic mapping between raw scores and observed event rates. We preferred isotonic regression for its flexibility, though parametric alternatives such as Platt scaling (Platt 1999) remain valid options. Our calibration set is constructed by first identifying all unlabeled samples whose uncalibrated sweep probabilities fall below a cutoff threshold T, where $T \in \{0.2, 0.3, \ldots, 0.7\}$. These samples form the negative calibration candidates. We then randomly choose an equal number of known simulated sweeps from the labeled set to serve as the positive calibration examples. This procedure yields an approximately balanced calibration set suitable for isotonic regression. We then fit isotonic regression on this constructed calibration set and apply the resulting mapping to calibrate *PULSe* probabilities across the full unlabeled dataset (Fig. S10).

Training the *smbCNN_GRL*[*PI*] Architectures

To train the *smbCNN_GRL*[*PI*] model under the *Constant–CEU* scenario, we assembled a training set consisting of 9,000 sweep and 9,000 neutral replicates each from both CEU and constant population size demographic histories. The model jointly learned to classify sweeps and neutral samples while also determining the source domain of each input, requiring two forms of supervision: class labels and domain labels. The constant population size training replicates were annotated with both class labels (sweep or neutral) and a domain label of zero. In contrast, the CEU replicates were only assigned a domain label of one, and their class labels were masked during training, meaning they were excluded from the classification loss calculation. Construction of the validation set followed a similar procedure, using 1,000 sweep and 1,000 neutral replicates each from the CEU and constant population size histories. Early stopping (Prechelt 2002) based on lowest validation loss was used to prevent overfitting.

An identical training strategy was applied for the *CEU–CEU* scenario, except that all replicates (both domain label zero and one) were drawn from the CEU population. No constant population size replicates were used in this test, allowing us to evaluate domain-invariant performance in a demographically matched scenario. The test set used for evaluation was identical to the unlabeled samples employed in both *PULSe* models, enabling a consistent comparison across models.

To ensure stable optimization, we follow the adaptation scheduling approach of Ganin et al. (2016), where the domain adaptation parameter $\lambda \in (0, 1)$ is updated at each training minibatch according to

$$\lambda_j = \frac{2}{1 + \exp(-10 \cdot j/B)} - 1,$$

where $j \in \{1, 2, \ldots, B\}$ is the index of the current minibatch out of the B minibatches that are processed during training. For a dataset with N samples, a minibatch size of 64 samples, and i training epochs that each process $\lceil N/64 \rceil$ minibatches, we have that the total number of minibatches during training is $B = i \cdot \lceil N/64 \rceil$. This schedule suppresses the influence of the domain classifier during early training, allowing the feature extractor to first learn discriminative features before gradually enforcing domain invariance. We denote this scheduled formulation as *smbCNN_GRL*[*PI, ls*].

In addition to this scheduled strategy, we explored an alternative formulation in which λ is treated as a fixed hyperparameter. Specifically, we evaluated $\lambda \in \{0.1, 0.2, \ldots, 1.0\}$ and selected the optimal value based on performance on the class-labeled portion of the validation set. This variant is denoted *smbCNN_GRL*[*PI, h*], with the highest AUPRC achieved at $\lambda = 0.8$.

Empirical Data Filtration and Image Generation

To apply *PULSe* on empirical data, we used phased haplotype data from the 99 CEU and 86 BEB individuals of the 1000 Genomes Project dataset (1000 Genomes Project Consortium 2015). Consistent with our simulation protocol, we retained only biallelic SNPs with a minor allele count of at least three, thereby removing singletons and doubletons. To mitigate the potential influence of technical artifacts, we followed the filtering protocol of Mughal et al. (2020) and excluded genomic segments of length 100 kilobases with mean CRG (Consensus Reference Genomes) mappability and alignability scores below 0.9 (Talkowski et al. 2011). These scores reflect the confidence with which sequencing reads can be accurately mapped to specific genomic regions, and their use ensures that downstream analyses are restricted to reliably interpretable regions of the genome.

We then applied the *PULSe* image generation approach described in the *Image Generation* subsection to all 22 autosomes. For each autosome, we began with the first 499 SNPs and generated the image representation of the spatial change in genomic variation using the same transformation pipeline used for simulated samples. Specifically, the haplotype matrix across these 499 SNPs was converted into a structured $n \times 237$ matrix that summarizes haplotypic variation across overlapping genomic windows, where $n = 198$ is the number of haplotypes from the CEU sample and $n = 172$ is the number of haplotypes from the BEB sample. We then advanced the SNP window by 50 SNPs and repeated the process until reaching the end of the chromosome. Each image was assigned a genomic coordinate corresponding to the 250th SNP within the 499-SNP window. This procedure generates a genome-wide series of *PULSe* images, which are constructed from large overlapping regions, capture changes in haplotypic variation, and serve as the unlabeled input set for *PULSe*.

Supplementary Material

Supplementary material is available at *Genome Biology and Evolution* online.

Acknowledgments

Computations for this research were performed using the services provided by Research Computing at Florida Atlantic University.

Funding

This work was supported by National Institutes of Health grant R35GM128590, by National Science Foundation grants DBI-2130666 and DEB-1949268, and by the UKRI Natural Environment Research Council grant NE/Y003519/1.

Data Availability

We release the source code for *PULSe* under the MIT open source license, and this repository can be accessed on GitHub (https://github.com/sandipanpaul06/PULSe/). To facilitate reuse and encourage further analysis, we have released the simulated replicates and empirical *PULSe* image datasets through Zenodo (https://doi.org/10.5281/zenodo.17156259). The CEU and BEB data from the 1,000 Genomes Project can be accessed from the project website (https://www.internationalgenome.org/category/phase-3/). The *PULSe* software is organized into modular components, separating image generation, feature extraction, and model training into independent

stages. Users may substitute their own feature-extraction routines in place of the default HOG implementation, provided that the resulting feature vectors have consistent dimensionality across images. Documentation in the repository outlines the expected input formats and naming conventions required for seamless integration with the downstream training scripts.

Literature cited

1000 Genomes Project Consortium. A global reference for human genetic variation. Nature. 2015:526:68–74. https://doi.org/10.1038/nature15393.

Ahlquist K, Sugden L, Ramachandran S. Enabling interpretable machine learning for biological data with reliability scores. PLoS Comput Biol. 2023:19:e1011175. https://doi.org/10.1371/journal.pcbi.1011175.

Akiyama M. The roles of ABCA12 in epidermal lipid barrier formation and keratinocyte differentiation. Biochim Biophys Acta Mol Cell Biol Lipids. 2014:1841:435–440. https://doi.org/10.1016/j.bbalip.2013.08.009.

Arnab S, Amin M, DeGiorgio M. Uncovering footprints of natural selection through spectral analysis of genomic summary statistics. Mol Biol Evol. 2023:40:msad157. https://doi.org/10.1093/molbev/msad157.

Arnab SP, Campelo dos Santos AL, Fumagalli M, DeGiorgio M. Efficient detection and characterization of targets of natural selection using transfer learning. Mol Biol Evol. 2025. 42: msaf094.https://doi.org/10.1093/molbev/msaf094.

Bali A, Mansotra V. Transfer learning-based one versus rest classifier for multiclass multi-label ophthalmological disease prediction. Int J Adv Comput Sci Appl. 2021:12. https://doi.org/10.14569/IJACSA.2021.0121269.

Barton N, Charlesworth B. Why sex and recombination? Science. 1998:281:1986–1990. https://doi.org/10.1126/science.281.5385.1986.

Basu A, Sarkar-Roy N, Majumder P. Genomic reconstruction of the history of extant populations of India reveals five distinct ancestral components and a complex structure. Proc Natl Acad Sci U S A. 2016:113:1594–1599. https://doi.org/10.1073/pnas.1513197113.

Bay H, Tuytelaars T, Van Gool L. Surf: Speeded up robust features. In: Computer Vision–ECCV 2006: 9th European Conference on Computer Vision, Graz, Austria, May 7–13, 2006. Proceedings, Part I 9. Springer; 2006. p. 404–417.

Bersaglieri T et al. Genetic signatures of strong recent positive selection at the lactase gene. Am J Hum Genet. 2004:74: 1111–1120. https://doi.org/10.1086/421051.

Booker W, Ray D, Schrider D. This population does not exist: learning the distribution of evolutionary histories with generative adversarial networks. Genetics. 2023:224:iyad063. https://doi.org/10.1093/genetics/iyad063.

Boughorbel S, Jarray F, El-Anbari M. Optimal classifier for imbalanced data using Matthews correlation coefficient metric. PLoS One. 2017:12:e0177678. https://doi.org/10.1371/journal.pone.0177678.

Boyko AR et al. Assessing the evolutionary impact of amino acid mutations in the human genome. PLoS Genet. 2008:4: e1000083. https://doi.org/10.1371/journal.pgen.1000083.

Browning S, Browning B. Haplotype phasing: existing methods and new developments. Nat Rev Genet. 2011:12:703–714. https://doi.org/10.1038/nrg3054.

Byars S, Ewbank D, Govindaraju D, Stearns S. Natural selection in a contemporary human population. Proc Natl Acad Sci U S A. 2010:107:1787–1792. https://doi.org/10.1073/pnas.0906199106.

Cai D, Dhe-Paganon S, Melendez P, Lee J, Shoelson S. Two new substrates in insulin signaling, IRS5/DOK4 and IRS6/DOK5. J Biol Chem. 2003:278:25323–25330. https://doi.org/10.1074/jbc.M212430200.

Carter J, Pan J, Rai S, Galandiuk S. ROC-ing along: evaluation and interpretation of receiver operating characteristic curves. Surgery. 2016:159:1638–1645. https://doi.org/10.1016/j.surg.2015.12.029.

Cecil RM, Sugden LA. On convolutional neural networks for selection inference: revealing the effect of preprocessing on model learning and the capacity to discover novel patterns. PLoS Comput Biology. 2023:19:e1010979. https://doi.org/10.1371/journal.pcbi.1010979.

Charlesworth B. The effects of deleterious mutations on evolution at linked sites. Genetics. 2012:190:5–22. https://doi.org/10.1534/genetics.111.134288.

Charlesworth B, Morgan M, Charlesworth D. The effect of deleterious mutations on neutral molecular variation. Genetics. 1993:134:1289–1303. https://doi.org/10.1093/genetics/134.4.1289.

Chatterjee P. The nation and its fragments: colonial and postcolonial histories. Princeton University Press; 2020.

Chatterji J. The spoils of partition. Cambridge University Press; 2007.

Chen G, Lee S, Zhu Z, Benyamin B, Robinson M. EigenGWAS: finding loci under selection through genome-wide association studies of eigenvectors in structured populations. Heredity. 2016:117:51–61. https://doi.org/10.1038/hdy.2016.25.

Chen X et al. Self-PU: self boosted and calibrated positive-unlabeled training. In: International Conference on Machine Learning. PMLR; 2020. p. 1510–1519.

Cheng X, Xu C, DeGiorgio M. Fast and robust detection of ancestral selective sweeps. Mol Ecol. 2017:26:6871–6891. https://doi.org/10.1111/mec.14416.

Chicco D, Jurman G. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. BMC Genomics. 2020:21:1–13. https://doi.org/10.1186/s12864-019-6413-7.

Chicco D, Jurman G. The Matthews correlation coefficient (MCC) should replace the ROC AUC as the standard metric for assessing binary classification. BioData Min. 2023:16:4. https://doi.org/10.1186/s13040-023-00322-4.

Colonna V et al. Human genomic regions with exceptionally high levels of population differentiation identified from 911 whole-genome sequences. Genome Biol. 2014:15:1–14. https://doi.org/10.1186/gb-2014-15-6-r88.

Comeron J. Background selection as baseline for nucleotide variation across the Drosophila genome. PLoS Genet. 2014:10:e1004434. https://doi.org/10.1371/journal.pgen.1004434.

Cree B et al. A major histocompatibility class I locus contributes to multiple sclerosis susceptibility independently from HLA-DRB1* 15: 01. PLoS One. 2010:5:e11296. https://doi.org/10.1371/journal.pone.0011296.

Cvijović I, Good B, Desai M. The effect of strong purifying selection on genetic diversity. Genetics. 2018:209:1235–1278. https://doi.org/10.1534/genetics.118.301058.

Dalal N, Triggs B. Histograms of oriented gradients for human detection. In: 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05). Vol. 1. 2005. p. 886–893. https://doi.org/10.1109/CVPR.2005.177.

Danecek P et al. The variant call format and VCFtools. Bioinformatics. 2011:27:2156–2158. https://doi.org/10.1093/bioinformatics/btr330.
Davis J, Goadrich M. The relationship between precision-recall and roc curves. In: Proceedings of the 23rd International Conference on Machine Learning. Association for Computing Machinery (ACM); 2006. p. 233–240.
Ding X et al. Learning multi-domain adversarial neural networks for text classification. IEEE Access. 2019:7:40323–40332. https://doi.org/10.1109/Access.6287639.
Di Rienzo A, Hudson R. An evolutionary framework for common diseases: the ancestral-susceptibility model. Trends Genet. 2005:21:596–601. https://doi.org/10.1016/j.tig.2005.08.007.
Duforet-Frebourg N, Luu K, Laval G, Bazin E, Blum M. Detecting genomic signatures of natural selection with principal component analysis: application to the 1000 genomes data. Mol Biol Evol. 2016:33:1082–1093. https://doi.org/10.1093/molbev/msv334.
Eaton R. The rise of islam and the Bengal frontier, 1204-1760. Vol. 17. University of California Press; 1993.
Eden E, Navon R, Steinfeld I, Lipson D, Yakhini Z. Gorilla: a tool for discovery and visualization of enriched go terms in ranked gene lists. BMC Bioinformatics. 2009:10:1–7. https://doi.org/10.1186/1471-2105-10-1.
Elkan C, Noto K. Learning classifiers from only positive and unlabeled data. In: Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining. Association for Computing Machinery (ACM); 2008. p. 213–220.
Ellegren H, Sheldon B. Genetic basis of fitness differences in natural populations. Nature. 2008:452:169–175. https://doi.org/10.1038/nature06737.
Ewing G, Hermisson J. MSMS: a coalescent simulation program including recombination, demographic structure and selection at a single locus. Bioinformatics. 2010:26:2064–2065. https://doi.org/10.1093/bioinformatics/btq322.
Fagny M et al. Exploring the occurrence of classic selective sweeps in humans using whole-genome sequencing data sets. Mol Biol Evol. 2014:31:1850–1868. https://doi.org/10.1093/molbev/msu118.
Flagel L, Brandvain Y, Schrider DR. The unreasonable effectiveness of convolutional neural networks in population genetic inference. Mol Biol Evol. 2019:36:220–238. https://doi.org/10.1093/molbev/msy224.
Freeman W, Roth M. Orientation histograms for hand gesture recognition. In: International workshop on automatic face and gesture recognition. Vol. 12. Institute of Electrical and Electronics Engineers (IEEE); 1995. p. 296–301.
Ganin Y et al. Domain-adversarial training of neural networks. J Mach Learn Res. 2016:17:1–35. https://doi.org/10.1007/978-3-319-58347-1_10.
Gazi NN et al. Genetic structure of Tibeto-Burman populations of Bangladesh: evaluating the gene flow along the sides of Bay-of-Bengal. PLoS One. 2013:8:e75064. https://doi.org/10.1371/journal.pone.0075064.
Gonzalez R. Digital image processing. Pearson Education India; 2009.
Gower G, Picazo P, Fumagalli M, Racimo F. Detecting adaptive introgression in human evolution using convolutional neural networks. eLife. 2021:10:e64669. https://doi.org/10.7554/eLife.64669.
Granka JM et al. Limited evidence for classic selective sweeps in African populations. Genetics. 2012:192:1049–1064. https://doi.org/10.1534/genetics.112.144071.
Günther T, Coop G. Robust identification of local adaptation from allele frequencies. Genetics. 2013:195:205–220. https://doi.org/10.1534/genetics.113.152462.
Haller B, Messer P. SLiM 3: forward genetic simulations beyond the Wright–Fisher model. Mol Biol Evol. 2019:36:632–637. https://doi.org/10.1093/molbev/msy228.
Haque C. Hazards in a fickle environment: Bangladesh. Vol. 10. Springer Science & Business Media; 1997.
Hein M, Andriushchenko M, Bitterwolf J. Why ReLU networks yield high-confidence predictions far away from the training data and how to mitigate the problem. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Institute of Electrical and Electronics Engineers (IEEE); 2019. p. 41–50.
Hejase H, Mo Z, Campagna L, Siepel A. A deep-learning approach for inference of selective sweeps from the ancestral recombination graph. Mol Biol Evol. 2022:39:msab332. https://doi.org/10.1093/molbev/msab332.
Hermisson J, Pennings P. Soft sweeps: molecular population genetics of adaptation from standing genetic variation. Genetics. 2005:169:2335–2352. https://doi.org/10.1534/genetics.104.036947.
Hernandez RD et al. Classic selective sweeps were rare in recent human evolution. Science. 2011:331:920–924. https://doi.org/10.1126/science.1198878.
Hudson R, Kaplan N. Deleterious background selection with recombination. Genetics. 1995:141:1605–1617. https://doi.org/10.1093/genetics/141.4.1605.
Hudson RR. Generating samples under a Wright–Fisher neutral model of genetic variation. Bioinformatics. 2002:18:337–338. https://doi.org/10.1093/bioinformatics/18.2.337.
Islam B et al. Emerging trends in cognitive impairment and dementia among older populations in Asia: a systematic review. J Glob Health. 2024:14:04233. https://doi.org/10.7189/jogh.14.04233.
Jensen JD et al. The importance of the neutral theory in 1968 and 50 years on: a response to Kern and Hahn 2018. Evolution. 2019:73:111–114. https://doi.org/10.1111/evo.13650.
Jensen MØ et al. Mechanism of voltage gating in potassium channels. Science. 2012:336:229–233. https://doi.org/10.1126/science.1216533.
Johri P, Eyre-Walker A, Gutenkunst R, Lohmueller K, Jensen J. On the prospect of achieving accurate joint estimation of selection with population history. Genome Biol Evol. 2022:14:evac088. https://doi.org/10.1093/gbe/evac088.
Karlsson E et al. Natural selection in a bangladeshi population from the cholera-endemic ganges river delta. Sci Transl Med. 2013:5:192ra86. https://doi.org/10.1126/scitranslmed.3006338.
Keinan A, Reich D. Human population differentiation is strongly correlated with local recombination rate. PLoS Genet. 2010:6:e1000886. https://doi.org/10.1371/journal.pgen.1000886.
Kerdoncuff E et al. 50,000 years of evolutionary history of India: impact on health and disease variation. Cell. 2025:188:3389–3404. https://doi.org/10.1016/j.cell.2025.04.027.
Kern A, Schrider D. Discoal: flexible coalescent simulations with selection. Bioinformatics. 2016:32:3839–3841. https://doi.org/10.1093/bioinformatics/btw556.
Kern A, Schrider D. diploS/HIC: an updated approach to classifying selective sweeps. G3 (Bethesda). 2018:8:1959–1970. https://doi.org/10.1534/g3.118.200262.
Kimura M. Evolutionary rate at the molecular level. Nature. 1968:217:624–626. https://doi.org/10.1038/217624a0.
Kimura M. The neutral theory of molecular evolution. Sci Am. 1979:241:98–126. https://doi.org/10.1038/scientificamerican1179-98.

Kristiansen H et al. Extracellular 2–5 oligoadenylate synthetase stimulates RNase L-independent antiviral activity: a novel mechanism of virus-induced innate immunity. J Virol. 2010:84:11898–11904. https://doi.org/10.1128/JVI.01003-10.
Kumar P, Lambert C. Positive unlabeled learning selected not at random (PULSNAR): class proportion estimation without the selected completely at random assumption. PeerJ Comput Sci. 2024:10:e2451. https://doi.org/10.7717/peerj-cs.2451.
Lamason RL et al. SLC24A5, a putative cation exchanger, affects pigmentation in Zebrafish and humans. Science. 2005:310:1782–1786. https://doi.org/10.1126/science.1116238.
Lauterbur M, Munch K, Enard D. Versatile detection of diverse selective sweeps with flex-sweep. Mol Biol Evol. 2023:40:msad139. https://doi.org/10.1093/molbev/msad139.
Lea AJ et al. Applying an evolutionary mismatch framework to understand disease susceptibility. PLoS Biol. 2023:21:e3002311. https://doi.org/10.1371/journal.pbio.3002311.
Lewontin R, Krakauer J. Distribution of gene frequency as a test of the theory of the selective neutrality of polymorphisms. Genetics. 1973:74:175–195. https://doi.org/10.1093/genetics/74.1.175.
Liu X et al. Characterising private and shared signatures of positive selection in 37 Asian populations. Eur J Hum Genet. 2017:25:499–508. https://doi.org/10.1038/ejhg.2016.181.
Lohmueller KE et al. Natural selection affects multiple aspects of genetic variation at putatively neutral sites across the human genome. PLoS Genet. 2011:7:e1002326. https://doi.org/10.1371/journal.pgen.1002326.
Lowe D. Distinctive image features from scale-invariant keypoints. Int J Comput Vis. 2004:60:91–110. https://doi.org/10.1023/B:VISI.0000029664.99615.94.
Lowe G. Object recognition from local scale-invariant features. Proc 7th IEEE Int Conf Comput Vis. 1999:2:1150–1157. https://doi.org/10.1109/ICCV.1999.790410.
Luo Y, Zheng L, Guan T, Yu J, Yang Y. Taking a closer look at domain shift: category-level adversaries for semantics consistent domain adaptation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Institute of Electrical and Electronics Engineers (IEEE); 2019. p. 2507–2516.
Lynch M. Rate, molecular spectrum, and consequences of human mutation. Proc Natl Acad Sci. 2010:107:961–968. https://doi.org/10.1073/pnas.0912629107.
Mallik S. The British East India company and the Great Bengal Famine of 1770: towards a corporate colonial biopolitics. Geogr Rev. 2024:114:464–488. https://doi.org/10.1080/00167428.2024.2325977.
McVicker G, Gordon D, Davis C, Green P. Widespread genomic signatures of natural selection in hominid evolution. PLoS Genet. 2009:5:e1000471. https://doi.org/10.1371/journal.pgen.1000471.
Melchjorsen J et al. Differential regulation of the OASL and OAS1 genes in response to viral infections. J Interferon Cytokine Res. 2009:29:199–208. https://doi.org/10.1089/jir.2008.0050.
Metspalu M et al. Shared and unique components of human population structure and genome-wide signals of positive selection in South Asia. Am J Hum Genet. 2011:89:731–744. https://doi.org/10.1016/j.ajhg.2011.11.010.
Mignone F, Gissi C, Liuni S, Pesole G. Untranslated regions of MRNAS. Genome Biol. 2002:3:1–10. https://doi.org/10.1186/gb-2002-3-3-reviews0004.
Mirza M. Global warming and changes in the probability of occurrence of floods in Bangladesh and implications. Global Environ Change. 2002:12:127–138. https://doi.org/10.1016/S0959-3780(02)00002-X.
Mo Z, Siepel A. Domain-adaptive neural networks improve supervised machine learning based on simulated population genetic data. PLoS Genet. 2023:19:e1011032. https://doi.org/10.1371/journal.pgen.1011032.
Molotkov I, Artomov M. Detecting biased validation of predictive models in the positive-unlabeled setting: disease gene prioritization case study. Bioinform Adv. 2023:3:vbad128. https://doi.org/10.1093/bioadv/vbad128.
Mozzi A et al. OASes and STING: adaptive evolution in concert. Genome Biol Evol. 2015:7:1016–1032. https://doi.org/10.1093/gbe/evv046.
Mughal M, Koch H, Huang J, Chiaromonte F, DeGiorgio M. Learning the properties of adaptive regions with functional data analysis. PLoS Genet. 2020:16:e1008896. https://doi.org/10.1371/journal.pgen.1008896.
Narasimhan V et al. The formation of human populations in South and Central Asia. Science. 2019:365:eaat7487. https://doi.org/10.1126/science.aat7487.
Nicolaisen L, Desai M. Distortions in genealogies due to purifying selection and recombination. Genetics. 2013:195:221–230. https://doi.org/10.1534/genetics.113.152983.
Niculescu-Mizil A, Caruana R. Predicting good probabilities with supervised learning. In: Proceedings of the 22nd International Conference on Machine Learning. Association for Computing Machinery (ACM); 2005. p. 625–632.
Oleksyk T, Smith M, O'Brien S. Genome-wide scans for footprints of natural selection. Philos Trans R Soc B Biol Sci. 2010:365:185–205. https://doi.org/10.1098/rstb.2009.0219.
Ovadia Y et al. Can you trust your model's uncertainty? Evaluating predictive uncertainty under dataset shift. Adv Neural Inf Process Syst. 2019:32.
Pardo-Diaz C, Salazar C, Jiggins C. Towards the identification of the loci of adaptive evolution. Methods Ecol Evol. 2015:6:445–464. https://doi.org/10.1111/2041-210X.12324.
Payseur B, Nachman M. Microsatelllite variation and recombination rate in the human genome. Genetics. 2000:156:1285–1298. https://doi.org/10.1093/genetics/156.3.1285.
Pigllucci M. How organisms respond to environmental changes: from phenotypes to molecules (and vice versa). Trends Ecol Evol. 1996:11:168–173. https://doi.org/10.1016/0169-5347(96)10008-2.
Platt J. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. Adv Large-Margin Class. 1999:10:61–eaat74.
Pouyet F, Aeschbacher S, Thiéry A, Excoffier L. Background selection and biased gene conversion affect more than 95% of the human genome and bias demographic inferences. eLife. 2018:7:e36317. https://doi.org/10.7554/eLife.36317.
Prakash O. The Dutch East India company and the economy of Bengal, 1630–1720. Princeton University Press; 2014.
Prechelt L. Early stopping-but when? In: Neural networks: tricks of the trade. Springer; 2002. p. 55–69.
Przeworski M. The signature of positive selection at randomly chosen loci. Genetics. 2002:160:1179–1189. https://doi.org/10.1093/genetics/160.3.1179.
Ranjan A. Bangladesh liberation war of 1971: narratives, impacts and the actors. India Q. 2016:72:132–145. https://doi.org/10.1177/0974928416637921.
Reich D, Thangaraj K, Patterson N, Price A, Singh L. Reconstructing Indian population history. Nature. 2009:461:489–494. https://doi.org/10.1038/nature08365.
Rexach J et al. 2023. Disease-specific selective vulnerability and neuroimmune pathways in dementia revealed by single cell

genomics [preprint]. bioRxiv. https://doi.org/10.1101/2023.09.29.560245.
Rublee E, Rabaud V, Konolige K, Bradski G. ORB: an efficient alternative to SIFT or SURF. In: 2011 International Conference on Computer Vision. IEEE; 2011. p. 2564–2571.
Sabeti PC et al. Positive natural selection in the human lineage. Science. 2006:312:1614–1620. https://doi.org/10.1126/science.1124309.
Saccheri I, Hanski I. Natural selection and population dynamics. Trends Ecol Evol. 2006:21:341–347. https://doi.org/10.1016/j.tree.2006.03.018.
Saito T, Rehmsmeier M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PLoS One. 2015:10:e0118432. https://doi.org/10.1371/journal.pone.0118432.
Sakharkar M, Chow VT, Kangueane P. Distributions of exons and introns in the human genome. In Silico Biol. 2004:4:387–393. https://doi.org/10.3233/ISB-00142.
Sams AJ et al. Adaptively introgressed Neandertal haplotype at the OAS locus functionally impacts innate immune responses in humans. Genome Biol. 2016:17:246. https://doi.org/10.1186/s13059-016-1098-6.
Scally A, Durbin R. Revising the human mutation rate: implications for understanding human evolution. Nat Rev Genet. 2012:13:745–753. https://doi.org/10.1038/nrg3295.
Schrider D. Background selection does not mimic the patterns of genetic diversity produced by selective sweeps. Genetics. 2020:216:499–519. https://doi.org/10.1534/genetics.120.303469.
Schrider D, Kern A. Soft sweeps are the dominant mode of adaptation in the human genome. Mol Biol Evol. 2017:34:1863–1877. https://doi.org/10.1093/molbev/msx154.
Seger J et al. Gene genealogies strongly distorted by weakly interfering mutations in constant environments. Genetics. 2010:184:529–545. https://doi.org/10.1534/genetics.109.103556.
Sen A. Poverty and famines: an essay on entitlement and deprivation. Oxford University Press; 1982.
Sheehan S, Song Y. Deep learning for population genetic inference. PLoS Comput Biol. 2016:12:e1004845. https://doi.org/10.1371/journal.pcbi.1004845.
Silva M et al. A genetic chronology for the indian subcontinent points to heavily sex-biased dispersals. BMC Evol Biol. 2017:17:1–18. https://doi.org/10.1186/s12862-016-0855-1.
Singh V, Auerbach D. Neurocardiac pathologies associated with potassium channelopathies. Epilepsia. 2024:65:2537–2552. https://doi.org/10.1111/epi.18066.
Slatkin M. Gene flow and the geographic structure of natural populations. Science. 1987:236:787–792. https://doi.org/10.1126/science.3576198.
Stacke K, Eilertsen G, Unger J, Lundström C. Measuring domain shift for deep learning in histopathology. IEEE J Biomed Health Inform. 2020:25:325–336. https://doi.org/10.1109/JBHI.6221020.
Stephan W. Signatures of positive selection: from selective sweeps at individual loci to subtle allele frequency changes in polygenic adaptation. Mol Ecol. 2016:25:79–88. https://doi.org/10.1111/mec.13288.
Stephan W. Selective sweeps. Genetics. 2019:211:5–13. https://doi.org/10.1534/genetics.118.301319.
Strauss KA et al. A population-based study of KCNH7 p.Arg394His and bipolar spectrum disorder. Hum Mol Genet. 2014:23:6395–6406. https://doi.org/10.1093/hmg/ddu335.
Su G, Chen W, Xu M. Positive-unlabeled learning from imbalanced data. In: IJCAI. International Joint Conferences on Artificial Intelligence Organization (IJCAI); 2021. p. 2995–3001.
Sugden LA et al. Localization of adaptive variants in human genomes using averaged one-dependence estimation. Nat Commun. 2018:9:703. https://doi.org/10.1038/s41467-018-03100-7.
Szpiech Z, Hernandez R. selscan: an efficient multithreaded program to perform EHH-based scans for positive selection. Mol Biol Evol. 2014:31:2824–2827. https://doi.org/10.1093/molbev/msu211.
Talkowski ME et al. Next-generation sequencing strategies enable routine detection of balanced chromosome rearrangements for clinical diagnostics and genetic research. Am J Hum Genet. 2011:88:469–481. https://doi.org/10.1016/j.ajhg.2011.03.013.
Tennessen JA et al. Evolution and functional impact of rare coding variation from deep sequencing of human exomes. Science. 2012:337:64–69. https://doi.org/10.1126/science.1219240.
Torada L et al. ImaGene: a convolutional neural network to quantify natural selection from genomic data. BMC Bioinform. 2019:20:337. https://doi.org/10.1186/s12859-019-2927-x.
Trowsdale J. The MHC, disease and selection. Immunol Lett. 2011:137:1–8. https://doi.org/10.1016/j.imlet.2011.01.002.
Vidyasagaran AL et al. Prevalence of mental disorders in South Asia: a systematic review of reviews. Glob Ment Health. 2023:10:e78. https://doi.org/10.1017/gmh.2023.72.
Vinay M, Yuan S, Wu X. Fraud detection via contrastive positive unlabeled learning. In: 2022 IEEE International Conference on Big Data (Big Data). IEEE. 2022. p. 1475–1484.
Voight B, Kudaravalli S, Wen X, Pritchard J. A map of recent positive selection in the human genome. PLoS Biol. 2006:4:e72. https://doi.org/10.1371/journal.pbio.0040072.
Wang X et al. Association study of KCNH7 polymorphisms and individual responses to risperidone treatment in schizophrenia. Front Psychiatry. 2019:10:633. https://doi.org/10.3389/fpsyt.2019.00633.
Wang Z et al. Automatic inference of demographic parameters using generative adversarial networks. Mol Ecol Resour. 2021:21:2689–2705. https://doi.org/10.1111/1755-0998.13386.
Wei H et al. Mitigating neural network overconfidence with logit normalization. In: International Conference on Machine Learning. PMLR; 2022. p. 23631–23644.
Whitehouse L, Schrider D. Timesweeper: accurately identifying selective sweeps using population genomic time series. Genetics. 2023:224:iyad084. https://doi.org/10.1093/genetics/iyad084.
Xu J. An extended one-versus-rest support vector machine for multi-label classification. Neurocomputing (Amst). 2011:74:3114–3124. https://doi.org/10.1016/j.neucom.2011.04.024.
Yang Y, Quan L, Ling Y. RBMS3 inhibits the proliferation and metastasis of breast cancer cells. Oncol Res. 2018:26:9–15. https://doi.org/10.3727/096504017X14871200709504.
Zadrozny B, Elkan C. Transforming classifier scores into accurate multiclass probability estimates. In: Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Association for Computing Machinery (ACM); 2002. p. 694–699.
Zhang M et al. Adaptive risk minimization: learning to adapt to domain shift. Adv Neural Inf Process Syst. 2021:34:23664–23678.
Zhao S, Zhou B, Luo F, Mao X, Lu Y. The structure and function of NKAIN2-a candidate tumor suppressor. Int J Clin Exp Med. 2015:8:17072.
Zügner D, Borchert O, Akbarnejad A, Günnemann S. Adversarial attacks on graph neural networks: perturbations and their patterns. ACM Trans Knowl Discov Data. 2020:14:1–31. https://doi.org/10.1145/3394520.

Associate editor: Sara Mathieson