# Discovery of ongoing selective sweeps within *Anopheles* mosquito populations using deep learning

Short title: Classifying partial sweeps in *Anopheles* populations via deep learning

Alexander T. Xue[1,*], Daniel R. Schrider[2], Andrew D. Kern[3], Ag1000G Consortium

[1] Simons Center for Quantitative Biology, Cold Spring Harbor Laboratory, One Bungtown Road, Cold Spring Harbor, NY 11724

[2] Department of Genetics, University of North Carolina, 120 Mason Farm Road, Chapel Hill, NC 27514

[3] Institute of Ecology and Evolution, 5289 University of Oregon, Onyx Bridge 272, Eugene, OR 97403

[*] Corresponding Author: xanderxue@gmail.com

## Abstract

Identification of partial sweeps, which include both hard and soft sweeps that have not currently reached fixation, provides crucial information about ongoing evolutionary responses. To this end, we introduce *partialS/HIC*, a deep learning method to discover selective sweeps from population genomic data. *partialS/HIC* uses a convolutional neural network for image processing, which is trained with a large suite of summary statistics derived from coalescent simulations incorporating population-specific history, to distinguish between completed versus partial sweeps, hard versus soft sweeps, and regions directly affected by selection versus those merely linked to nearby selective sweeps. We perform several simulation experiments under various demographic scenarios to demonstrate *partialS/HIC*'s performance, which exhibits excellent resolution for detecting partial sweeps. We also apply our classifier to whole genomes from eight mosquito populations sampled across sub-Saharan Africa by the *Anopheles gambiae* 1000 Genomes Consortium, elucidating both continent-wide patterns as well as sweeps unique to specific geographic regions. These populations have experienced intense insecticide exposure over the past two decades, and we observe a strong overrepresentation of sweeps at insecticide resistance loci. Our analysis thus provides a list of candidate adaptive loci that may be relevant to mosquito control efforts. More broadly, our supervised machine learning approach introduces a method to distinguish between completed and partial sweeps, as well as between hard and soft sweeps, under a variety of demographic scenarios. As whole-genome data rapidly accumulate for a greater diversity of organisms, *partialS/HIC* addresses an increasing demand for useful selection scan tools that can track in-progress evolutionary dynamics.

## Introduction

Malaria represents an enormous burden on human health, with an estimated 214 million cases and 438,000 deaths in 2015 [1]. As mosquitos of the *Anopheles gambiae* species complex are the major vector for *Plasmodium* parasites, roughly 70% of global malaria relief budgets have been focused on mosquito control, including insecticide treated bed-nets, indoor residual spraying, and larva control through the direct modification of habitats as well as the application of larvicide. While these vector control efforts have successfully produced major reductions of malaria transmission rates over the past 15 years [1], there has been an alarming increase in mosquitos resistant to insecticides, specifically pyrethroids, observed across nearly all areas of the world covered by anti-malarial efforts [2]. Pyrethroids are the only class of insecticide used in long-lasting insecticidal nets and are applied in many indoor spraying programs, thus the evolutionary innovation of resistance is a well-recognized Achilles heel of anti-malarial efforts.

The increase in insecticide resistant mosquitoes is to be expected from an evolutionary standpoint: anti-malaria control efforts exert a strong selective pressure to which mosquito populations will respond through the differential survivorship and reproduction of those individuals that can best cope with the applied insecticides. Pyrethroid resistance was reported within African malaria vectors first in Sudan during the 1970s, then later in West Africa during the 1990s, most likely stemming from accidental exposure of mosquitos to crop applications of pyrethroids [3,4]. Subsequent analysis showed this earliest resistance to be a result of mutations in the knockdown resistance locus *kdr*, which is known to contribute to pyrethroid resistance in other insect species [5]. Mutations conferring resistance at *kdr* as well as other loci have since spread throughout Africa, and threaten to nullify the gains in malaria control achieved over the past decade [6]. While control efforts are now looking toward non-pyrethroid insecticides [2,7] as well as gene drive technologies [8], it is anticipated that resistance to these control modalities will eventually evolve as well [9,10]. Hence, an important goal in the continued fight against malaria is to identify genomic targets of resistance in *Anopheles*, especially in such a way that might inform vector managers in the field.

Alleles that confer resistance to control efforts should rapidly increase in frequency within *Anopheles* populations in a manner consistent with selective sweeps. When an allele increases in frequency under selection, its linked genetic background comes with it in a process known as genetic hitchhiking. Selective sweeps, through this hitchhiking effect, lead to decreased levels of polymorphism [11–13], skewed allele frequency spectra [14,15], and increases in linkage disequilibrium surrounding the site under selection [16]. Classically, methods for finding sweeps have focused on a particular aspect of genetic variation, for instance observing the site frequency spectrum (SFS) at a locus and comparing it to expectations under neutrality and selective sweeps [17]. More recently, the field has made excellent progress in combining signals across multiple features of genetic variation through supervised machine learning (SML) [18–25], which has substantially improved power, accuracy, and robustness in what have been stubbornly difficult inference problems within population genetics [26]. While much attention has been paid to applying SML for the identification and classification of completed selective sweeps in the genome [22,24], less effort has been made for using SML to identify sweeps that are incomplete within a population, sometimes called partial sweeps (although see Sugden *et al.* [25] for a recent example). In these cases, the beneficial allele is not currently fixed within the population, thereby creating a weaker hitchhiking effect in comparison to a completed sweep, and accordingly a more subtle perturbation of patterns of genetic variation [27]. Although difficult to detect, partial sweeps could be implicated in cases where recently initiated selective forces cause presently ongoing adaptation, directional selection ceases prior to fixation, or an intermediate allele frequency is favored by balancing, polygenic, and/or pleiotropic selection. Therefore, it is important to address the challenge of capturing such genomic signatures that represent a significant facet of evolution, which may also give insight into future dynamics.

## New Approaches

Recent successful efforts to reduce malaria transmission are in danger of collapse due to evolving insecticide resistance in the mosquito vector *Anopheles gambiae*. We aim to

101 understand the genetic basis of current adaptation to vector control efforts by deploying a
102 novel method that can classify multiple categories of selective sweeps, including partial
103 sweep classes, from population genomic data. To this end, we extend a recent SML method
104 to partial sweep inference and apply it to elucidate ongoing selective sweeps from
105 *Anopheles* population genomic samples.

106 Specifically, we introduce here an extension of *S/HIC* [22] and *diploS/HIC* [24],
107 *partialS/HIC*, a SML classifier that uses a deep convolutional neural network (CNN) to
108 classify a genomic window from a set of selection states that includes both hard and soft
109 partial sweeps along with their associated linked classes (*i.e.* regions adjacent to either a
110 partial hard or soft sweep). Importantly, this implementation achieves increased inferential
111 power by utilizing dozens of additional summary statistics, including summaries for the
112 distribution of *iHS* scores [28] as well as derivatives of a recently developed SNP-specific
113 compound statistic called SAFE (selection of allele favored by evolution) [29], under a deep
114 learning framework that involves training convolutional neural networks with coalescent
115 simulations that accommodate demographic history and are converted into spatially-
116 explicit two-dimensional feature vector images. We validate *partialS/HIC*'s performance
117 through extensive simulation-based experiments that were modeled after data from phase
118 I of the *Anopheles gambiae* 1000 genomes project (Ag1000G) [6], with particular emphasis
119 on discovering sweeps currently in progress such as what might result from ongoing vector
120 control efforts (*e.g.* insecticide spraying). Our findings demonstrate that our method is
121 effective for detecting partial sweeps even in the face of complex population size histories
122 such as those found among the Ag1000G samples. Furthermore, for binary classification of
123 selective sweeps when partial sweeps are included, *partialS/HIC* has greater accuracy than
124 *iHS* or SAFE alone as well as two alternative approaches to sweep inference based on the
125 same suite of summary statistics. Subsequently, we apply our method to the empirical
126 Ag1000G data, revealing many partial sweeps as well as completed sweeps from standing
127 genetic variation. Moreover, we find that our sweep candidates are highly enriched for loci
128 that have been previously identified as contributing to insecticide resistance.

129

130 # Results

131 ## Coalescent simulations of feature vector images for *partialS/HIC* training

132 To train *partialS/HIC*, we deployed the program *discoal* [30] to perform coalescent
133 simulations of completed and partial as well as hard and soft selective sweeps, along with
134 simulations without sweeps, in a manner analogous to Schrider and Kern [22] (Figure S1).
135 Individual simulations were converted into two-dimensional matrices, or feature vector
136 images, built from 89 rows corresponding to different summary statistics, and 11 columns
137 corresponding to adjacent sub-windows. The 89 statistics include, along with 14 that were
138 previously implemented in *S/HIC* and/or *diploS/HIC*, 3 genomic region variants of the SNP-
139 specific *iHS* statistic [28] and 72 derivatives of the SAFE score [29]. *partialS/HIC* is trained
140 to classify genomic segments into one of nine states: unaffected by selection (*i.e.* neutral);
141 containing a completed hard, completed soft, partial hard, or partial soft sweep,
142 respectively; or linked to a completed hard, completed soft, partial hard, or partial soft
143 sweep, respectively. To this end, we defined the four completed hard/soft and partial
144 hard/soft selective sweep states as containing a sweep within the central, focal sub-
145 window (*i.e.* the fifth out of the 11 columns of sub-windows). In contrast, the four
146 classification states involving linkage to a selective sweep were defined as having a sweep
147 of the given type within one of the remaining ten sub-windows. Therefore, every linked
148 selection state was trained from simulations that vary in genetic distance to the sweep
149 target site, such that the total set of simulations all had a sweep occurring in any of the ten
150 non-focal sub-windows. This allowed us to accommodate a range of linked classes that
151 differ in spatial patterns within a genomic window, as well as assess how linkage distance
152 affects misclassification bias during our simulation experiments. Simulations were run for
153 each of eight population size histories corresponding to the empirical Ag1000G population
154 datasets, which were previously inferred as part of the initial data release [6]. These
155 *Anopheles* population datasets from Miles *et al.* [6] are labeled here as AOM (*A. coluzzii*
156 from Angola), BFM (*A. coluzzii* from Burkina Faso), BFS (*A. gambiae* from Burkina Faso),
157 CMS (*A. gambiae* from Cameroon), GAS (*A. gambiae* from Gabon), GNS (*A. gambiae* from

Guinea), GWA (*Anopheles* of uncertain species from Guinea-Bissau), and UGS (*A. gambiae* from Uganda).

Heatmaps constructed from median values across simulations reveal expected spatial patterns, such that values immediately flanking a sweep are substantially different than those further from the focal sub-window, while neutral regions display no discernible pattern among sub-windows (Figure S2). Additionally, spatial patterns of statistics differ qualitatively between selection states. These observations are consistent regardless of mosquito population history, suggesting that there is signal within this collection of summary statistics to isolate the location of a sweep to a specific sub-window as well as distinguish among neutral regions and types of selective sweeps.

## Deep learning excels in detecting selective sweeps, including partial hard sweeps

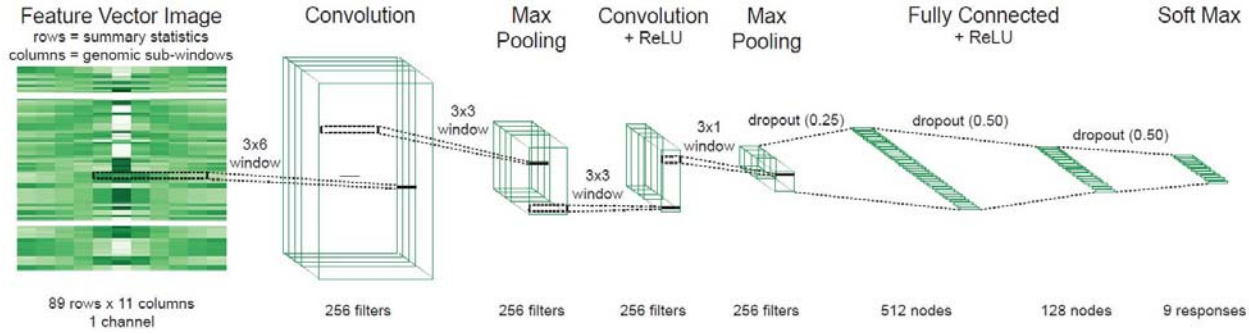


***Figure 1. CNN architecture of neural network layers.*** *Our partialS/HIC classifier utilizes a convolutional neural network whereby the input feature vector image, composed of 89 summary statistics organized into rows and across 11 contiguous genomic sub-windows organized into columns, is passed to a 2D convolutional layer with 256 filters using a 3×6 receptive field. Next is a 2D max pooling layer given a 3×3 receptive field. Then there is a second 2D convolutional layer of 256 filters based also on a 3×3 receptive field and ReLU activation. Afterward is a second 2D max pooling layer with a 3×1 receptive field. The tensor is then flattened after a dropout layer (p=0.25) and passed to two fully-connected layers with ReLU activation, resulting in 512 and 128 nodes with subsequent dropout (p=0.50) respectively. Finally, a softmax activation layer classifies the 9 responses.*

182 We utilized *partialS/HIC* to train and test a CNN for nine-state classification independently
183 on each of the eight demographic histories associated with the Ag1000G population
184 samples (Figures S1, 1). To this end, we produced eight batches of simulations that were
185 split into separate training, validation, and testing sets. During the training process, CNN
186 hyperparameters were tuned on the training set while the validation set allowed mitigation
187 of over-fitting (Figure S3). Each CNN was subsequently assessed for accuracy with the
188 corresponding held-out testing set, which was generated under the same specifications as
189 the training/validation data except linked selection classes were kept discrete versus being
190 grouped together into the four linked selection states. Among the eight test sets, there was
191 moderate overall accuracy for this simulation experiment (median accuracy = 66.4%; Table
192 S1). However, confusion matrix heatmaps provide a more informative view of our
193 classifier's performance, which exhibited reliability in identifying neutral regions,
194 completed hard and soft sweeps, partial hard sweeps, and individual regions linked to
195 completed hard/completed soft/partial hard sweeps (Figures 2, S4). In general (*i.e.* all
196 demographic scenarios save for AOM), completed hard sweep is the class that experienced
197 the highest degree of correct assignment (median accuracy = 96.0%). We also had high
198 accuracy for identification of linked completed hard regions, demonstrating a strong ability
199 to localize completed hard sweeps. The behavior of our *partialS/HIC* classifier is likewise
200 favorable for completed soft sweeps (median accuracy = 84.2%), with completed soft
201 sweeps rarely detected incorrectly as hard sweeps and instead typically mistaken for either
202 the neutral or partial soft sweep state. Moreover, sub-windows linked to completed soft
203 sweeps beyond one sub-window away had low levels of misclassification to any of the non-
204 linked sweep states, again allowing for dependable localization of the sweep.

205 Importantly, the purpose of *partialS/HIC* is to extend our state space to identify
206 ongoing selective sweeps while distinguishing these from completed sweeps. We find that
207 our ability to identify partial hard sweeps was generally strong across population histories
208 (median accuracy = 74.6%) and is often comparable to that of completed soft sweeps.
209 However, localization of partial hard sweeps along the chromosome was more difficult
210 than for completed sweeps, as can be seen from the moderate levels of confusion between
211 partial hard sweep and linked partial hard sweep sub-windows. Undoubtedly, this is due to

212 the limited amount of time recombination has had to whittle down the haplotype carrying
213 the beneficial mutation.

214 Identifying partial soft sweeps was a much more challenging task (median accuracy
215 = 45.3%), with a high false negative rate (median rate of misclassification as neutral =
216 27.6%) as well as a substantial probability of misclassification as a completed soft sweep
217 (median rate = 14.4%). It is encouraging though that partial soft sweeps were almost never
218 misclassified as a completed nor partial hard sweep. Additionally, while our accuracy in
219 classifying partial soft sweeps was poor, false positives were not a major concern (median
220 rate of misclassifying neutral regions as partial soft sweep = 3.2%). Although even fairly
221 low false positive rates can be problematic when the true number of sweeps is low,
222 *partialS/HIC* achieved acceptably low false discovery rates for partial soft sweeps in our
223 application to the *Anopheles gambiae* 1000 Genomes dataset, as we show below.
224 Furthermore, linked partial soft sweeps beyond one sub-window away from the focal sub-
225 window were rarely mistaken for a sweep state, and likewise partial soft sweeps were
226 seldom confused for being linked (median rate of misclassification as linked partial soft
227 sweep = 4.3%), thus demonstrating that accurate localization of partial soft sweeps may be
228 possible.

229 In summary, *partialS/HIC* has excellent ability to distinguish partial from completed
230 sweeps for *de novo* mutations, and lesser yet still substantial power for sweeps from
231 standing variation. Moreover, we demonstrated very strong performance in differentiating
232 between hard and soft sweeps, regardless of whether a sweep was completed or
233 incomplete. Importantly, this is all while maintaining an acceptable false positive rate
234 across each of the population histories tested (median accuracy for neutral regions =
235 85.1%; median rate of misclassifying neutral regions as any one of the four non-linked
236 sweep states = 4.6%).

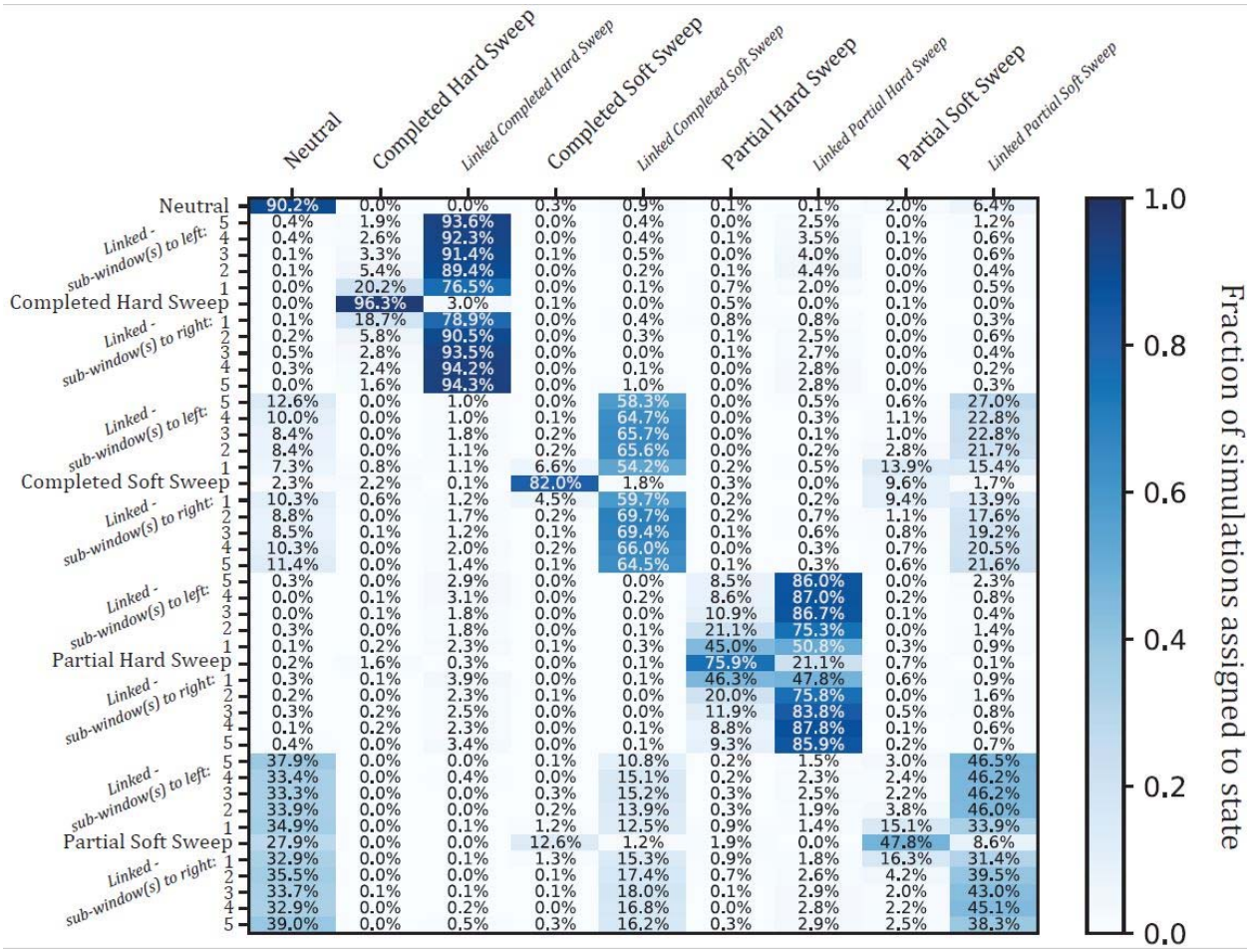


***Figure 2. Confusion matrix heatmap of partialS/HIC simulation experiment.*** *Given the BFS population history. Each row designates a true simulation class, which for linked sweeps is differentiated by distance in genomic sub-windows to the target selective sweep. In total, there are 45 simulated scenarios shown, including 11 for each sweep type (i.e. one case whereby the sweep is within the central sub-window, and 10 whereby a linked sweep is located within one of the flanking sub-windows) and neutrality. Each column indicates one of nine inferred states, with the linked simulation classes collapsed into a single category per selective sweep type for training and classification, allowing partialS/HIC to learn to distinguish between sweeps located in the focal sub-window and those somewhere nearby. Darker blue cells represent a higher proportion of the 1,000 calls for each true class. Importantly, there is generally high accuracy in discriminating between sweeps within the focal sub-window versus linked sub-windows, especially when linkage is beyond one sub-window away and particularly for completed sweeps. However, discovering partial soft sweeps is noticeably a challenging task. Moreover, there is greater sensitivity separating full completion from partial completion for hard sweeps in contrast to soft sweeps.*

## Robustness to demographic model misspecification

To assess robustness to demographic misspecification, we applied a CNN trained on simulations from one population sample to data generated from an alternate demographic history. Specifically, we used training data from the GAS population size history, which was fairly stable over time, and leveraged it against the CMS test dataset, which experienced a dramatic population expansion (overall accuracy = 55.0%; rate of misclassifying neutral regions as any one of the four non-linked sweep states = 1.7%). Despite this misspecification, the confusion matrix (Figure S5) strongly resembles the corresponding matrix that is correctly specified for demography (*i.e.* for CMS in Figure S4). In particular, accuracies for finding neutral regions, completed hard sweeps, and partial hard sweeps are roughly equivalent between the correctly specified model and misspecified model (Figure S5). For soft sweeps, while confusion between completed and partial sweeps is increased for the misspecified model, the overall ability to distinguish sweeps from neutrality is largely preserved. Moreover, the rates at which examples from the linked classes were mistaken for sweeps are seemingly unaffected.

To further examine the impact of increasingly misspecified demography, we produced five additional full testing datasets assuming a simple two-epoch instantaneous contraction model. The five test sets differed only by bottleneck severity, which increased in even intervals from 20× to 100×. We then applied the CNN that was trained given the BFS population history and measured overall accuracies. These accuracy measurements are generally similar to that of our baseline, *i.e.* the corresponding original experiment whereby the testing data were simulated under the inferred demography for BFS, thus the demography was correctly specified with respect to the training set underlying this CNN (Figure S6). However, performance gradually worsens with decreasing bottleneck severity across the test sets, but this may be largely an effect from the number of polymorphisms. To address this, we generated another six test sets that identically replicated the specifications for the BFS training simulations except with $\theta$ fixed in value yet varying between the six datasets. Notably, the range we utilized here for $\theta$ exceeded the $\theta$ prior distribution for the original BFS training and testing simulations, hence further evaluating model misspecification. We find no loss in overall accuracy except in the case with the least genetic diversity simulated, where there is a moderate decrease (Figure S7). This is

285 unsurprising given that low levels of genetic diversity lead to noisy estimates of the
286 statistics employed by *partialS/HIC*. It is encouraging however that sensitivity is
287 maintained throughout most of our $\theta$ prior distribution range and accuracy does not fall
288 below that of our baseline experiment until $\theta$ is at a low value toward the bounds ($\theta$ <
289 5,000 for the full genomic window size of 55,000 base-pairs). To put this into perspective,
290 genetic diversity measures for the corresponding empirical BFS dataset are consistent with
291 simulations generated with $\theta \geq$ 5,000, and likewise are beyond nearly the entirety of the $\theta$
292 = 1,000 simulated values; among the sub-windows that passed our filtering thresholds
293 (Figure S1), there is a median value of $\theta_W$ = 2.633e-2 (central 95% density: 6.484e-3 –
294 3.547e-2) and $\theta_H$ = 6.318e-3 (central 95% density: 2.381e-3 – 1.231e-2), compared to a
295 median value of $\theta_W$ = 1.218e-2 (central 95% density: 1.116e-3 – 1.849e-2) and $\theta_H$ = 3.863e-
296 3 (central 95% density: 8.592e-4 – 9.696e-3) across all sub-windows generated under the $\theta$
297 = 5,000 condition, whereas the simulations given $\theta$ = 1,000 have a median value of $\theta_W$ =
298 2.353e-3 (central 95% density: 9.802e-5 – 3.861e-3) and $\theta_H$ = 7.043e-4 (central 95%
299 density: 5.950e-8 – 2.571e-3).

300 Together, these results indicate that sweep discovery and localization is not strongly
301 impacted by several forms of demographic model misspecification during training.

302

## 303 Partial sweeps are unpredictably misclassified as either completed
## 304 sweep or neutral when not explicitly considered

305 Since the previous versions of *partialS/HIC* (*S/HIC* and *diploS/HIC*) did not allow for partial
306 sweep selection states, we were interested in how such five-state classifiers would behave
307 when confronted with partial sweeps. To explore this, we conducted a simulation
308 experiment that first removed partial hard and soft sweeps as well as their associated
309 linked classes from the CNN training process, thus training on only five states rather than
310 all nine. Next, in an effort to examine the classification behavior for these five-state CNNs,
311 we applied the full test set that included the partial sweep classes. Unsurprisingly, the
312 trend was for partial sweeps to be most often confused for linked selection (Figures 3, S8).

Perhaps more concerning is the false negative rate (*i.e.* rate at which partial sweeps were misclassified as neutral), which was substantial in partial hard sweeps for several populations (median = 8.8%; max = 32.4%; >1% in all populations) and extreme in partial soft sweeps (>50% in three populations, >40% in three more populations, and >24% in all populations). Partial hard sweeps that were discovered were also often misclassified as a completed soft sweep (median rate of misclassification as completed soft sweep = 5.5%). However, when training included partial sweeps, there is universal and dramatic improvement in both finding sweeps and correctly identifying the model of selection (Figures 2, S4). Meanwhile, overall accuracy remains similar among the five-state and nine-state classifiers with respect to simulations of neutral, completed sweep, and linked classes exclusively (Figures 3, S8; Table S2). As a result, accuracy should only stand to benefit from incorporating partial sweeps into training since ignoring such information leads to unacceptably high false negative rates of partial sweeps being called neutral or linked.

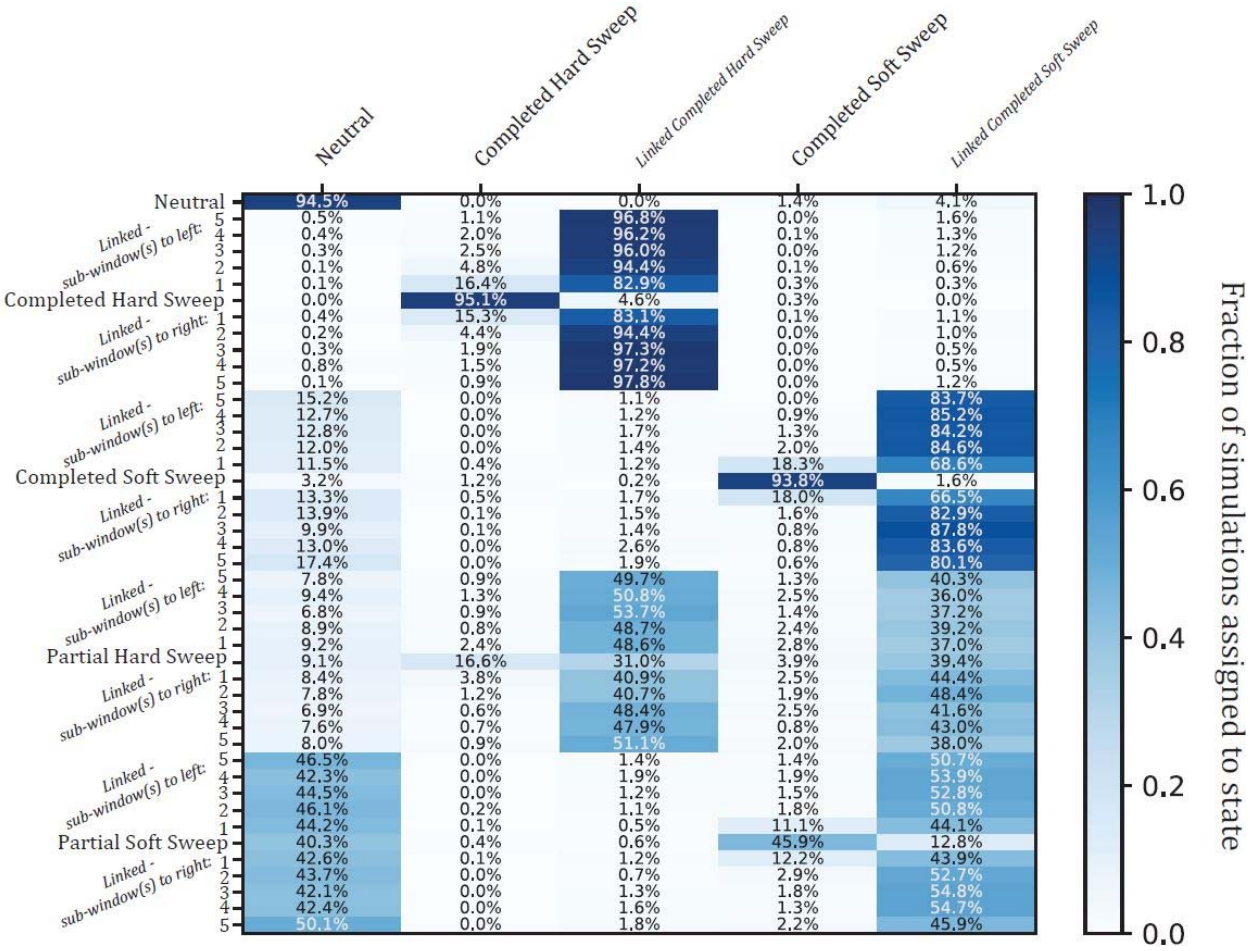

***Figure 3. Confusion matrix heatmap of simulation experiment with partial sweeps ignored during training.*** *Given the BFS population history. Structured in the same manner as Figure 2, each row designates a true simulation class and darker blue cells represent a higher proportion of the 1,000 calls for each true class. In contrast to Figure 2 though, each column indicates one of five inferred states instead of nine, as the two partial sweep and two linked partial sweep classification states were omitted from training to determine misclassification bias when partial sweeps are ignored. There is a substantial decrease in discovery of partial sweeps, especially partial hard sweeps (20.5% total sweep discovery compared to 78.2% in Figure 2; 46.3% total sweep discovery of partial soft sweeps compared to 62.3% in Figure 2). Specifically for partial hard sweeps, a large proportion of the detected sweeps are misclassified as soft sweeps instead of hard sweeps (19.0%;* i.e. *3.9% of calls = soft sweeps : 16.6% of calls = hard sweeps).*

## ***partialS/HIC*** **binary classification outperforms several competing approaches**

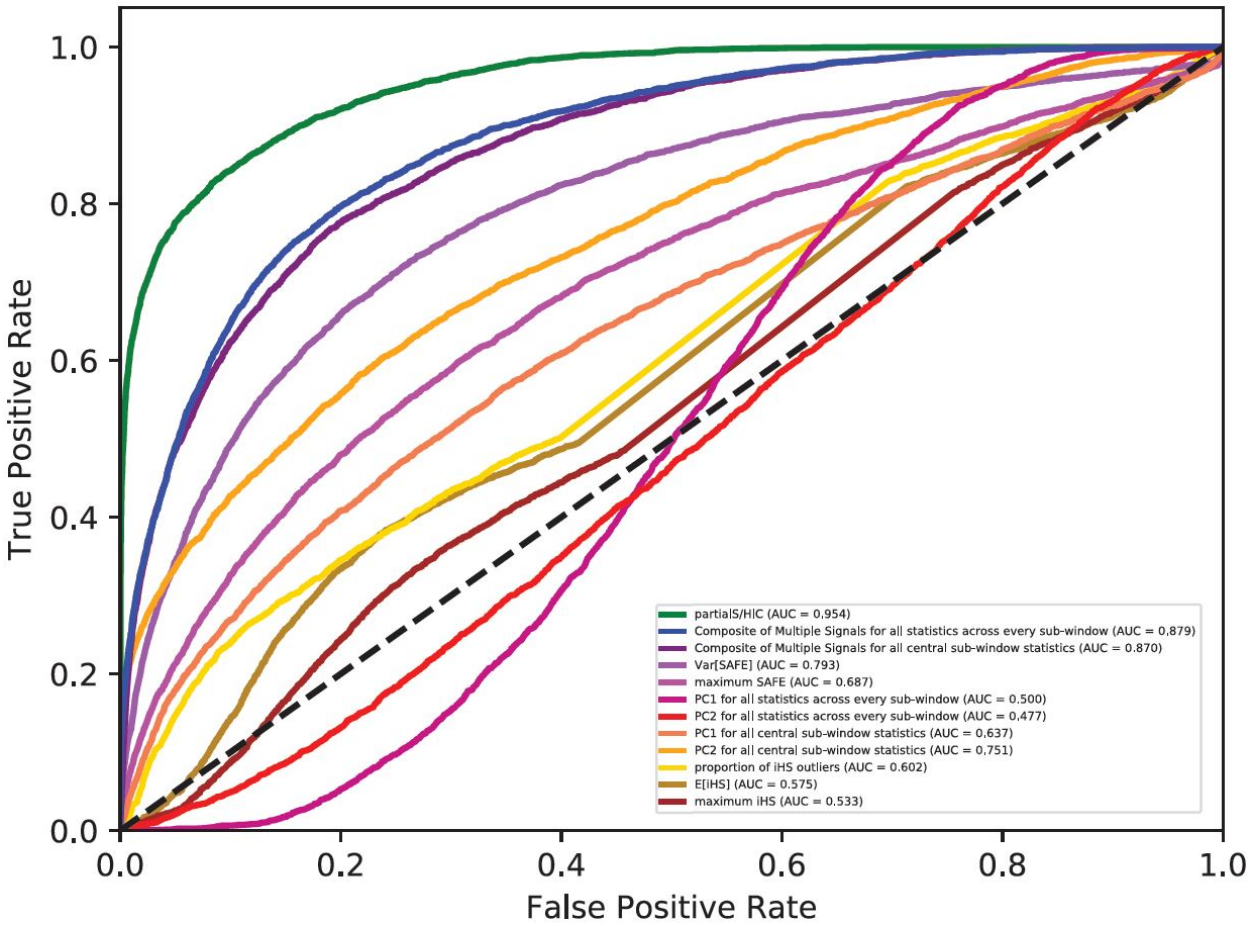


***Figure 4. ROC curves for binary classification of selective sweeps, including partial sweeps, versus regions neutrally evolving or under linked selection.*** *Given the BFS population history. The partialS/HIC deep learning classifier outcompetes two other approaches for managing the same suite of summary statistics: Composite of Multiple Signals and principal component analysis. Furthermore, partialS/HIC excels in performance over several sub-window derivatives of two summary statistics, SAFE and iHS, which were all included within our set of summary statistics used for training. Notably, Composite of Multiple Signals, SAFE, and iHS were all designed to uncover selective sweeps. Additionally, the SAFE score is itself a compound statistic that captures signal from several constituent statistics, and the majority of our training data is derived from either the SAFE score or one of its components.*

To assess whether the collection of summary statistics under our deep learning method extends inferential resolution beyond the signal conferred by *iHS* or SAFE alone, and to determine if our CNN approach leverages these statistics in a more informative manner than other aggregation schemes, we compared receiver operating characteristic (ROC)

curves, which plot true positive against false positive rates given varying thresholds, for the binary classification task of broadly detecting selective sweeps (*i.e.* any of the four selection states involving a sweep within the focal sub-window) vs. neutral regions or linked sweeps. We selected for our comparison *iHS*-derived statistics (mean, maximum, and proportion of outlier *iHS* values within the central sub-window of the testing simulations) because *iHS* was explicitly designed for detecting partial sweeps [28], and SAFE-based statistics (variance and maximum of SAFE scores within the central sub-window of the testing simulations) since SAFE is a compound of many variables, of which the vast majority of our training data are summaries [29]. Additionally, we conducted a principal component analysis (PCA) from the training data, restricted to only the focal sub-window as well as for the entire dataset throughout the full window respectively in two separate PCAs, and projected the testing data onto the first two principal components (PC1, PC2). We then produced a ROC curve from each set of principal component values transformed from the testing data, respectively to both PC1 and PC2 as well as the PCA on solely the central sub-window versus the full-scale window, for a total of four individual ROC curves. These ROC curves allow contrast to a simple dimensionality reduction approach. Lastly, we computed Composite of Multiple Signals scores from the testing data for both the focal sub-window and full window respectively, using the training data to derive probability distributions. Deploying the Composite of Multiple Signals metric provides an alternative method that is intended to exploit multiple summary statistics for inference of positive selection [31]. For a direct comparison, we altered *partialS/HIC* slightly to create a CNN that had two final output responses (*i.e.* sweep in central sub-window vs. no sweep in central sub-window) instead of nine, with the training dataset as well as architecture of network layers (Figure 1) remaining the same. The CNN optimized for binary classification was then applied to the testing simulations, with the output probability of a selective sweep utilized as input to generate the ROC curve.

The *partialS/HIC* binary classifier consistently outperforms other methods in identifying selective sweeps to the focal sub-window when partial sweeps are included (median AUC = 0.943; Figures 4, S9). Interestingly, we demonstrate that the Composite of Multiple Signals also has a strong ability to extract information from our selection of

389 summary statistics, though it always trails behind *partialS/HIC*; in every case, the best
390 performance after *partialS/HIC* is for the Composite of Multiple Signals based on all sub-
391 window statistics (median AUC = 0.876), while the next best ROC curve is from the
392 Composite of Multiple Signals calculated from only the central sub-window (median AUC =
393 0.859). In contrast, with the unexpected exception of PC2 for central sub-window statistics
394 (median AUC = 0.749), PCA appears to be an unfavorable approach for combining summary
395 statistics to infer selective sweeps (*e.g.* median AUC for PC1 of all sub-window statistics =
396 0.500). Furthermore, there is evidence that SAFE scores capture important signal about
397 selective sweeps in a manner that is robust to the final frequency of the selected mutation,
398 particularly the variance of SAFE scores within the focal sub-window (median AUC =
399 0.787). Conversely, there seems to be decreased signal in *iHS*, as all its focal sub-window
400 variants individually performed quite poorly (*e.g.* median AUC for proportion of *iHS*
401 outliers = 0.582).

402

403 **Soft and partial sweeps are commonplace among *A. gambiae* populations**

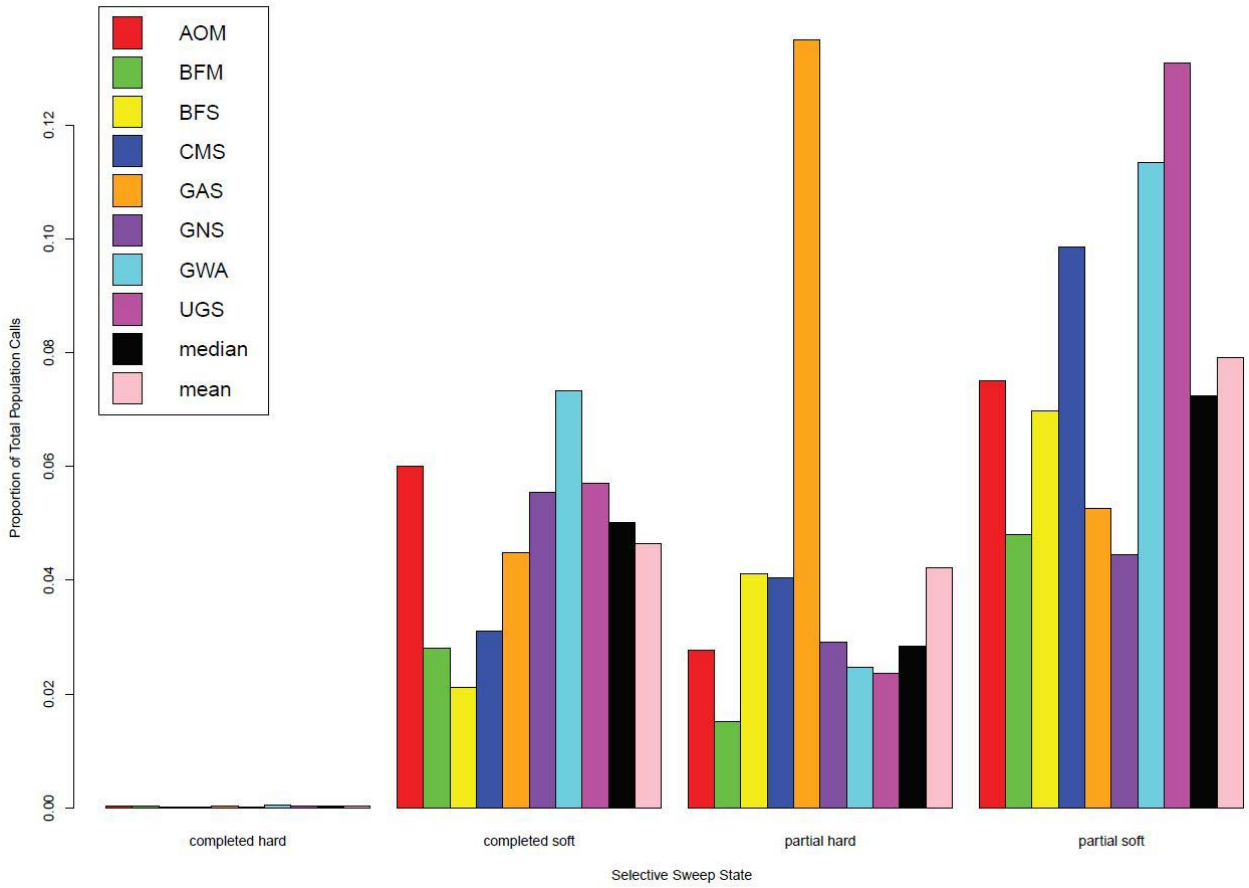


***Figure 5. Genome-wide partialS/HIC sweep calls across empirical mosquito population datasets.*** *For all eight datasets, there is a minimal number of calls for completed hard sweeps, represented here on the Y-axis as the relative proportion to the total set of empirical calls across every genomic window per population dataset. However, there is indeed a substantial proportion of sweep calls for hard sweeps, yet are incomplete or in-progress. In most cases, the number of both completed and partial soft sweeps further exceeds that of partial hard sweeps. Furthermore, the majority of sweeps have not reach fixation across these Anopheles populations.*

Turning our attention to the Ag1000G phase I data, we applied our nine-state CNNs to the corresponding *A. gambiae* population datasets, classifying 5 KB segments using a 55 KB full sliding window throughout the whole genome (Figure S1). Each of the eight mosquito populations contains a large number of sub-windows identified as completed soft sweeps (median fraction of total calls genome-wide = 5.01%) as well as partial sweeps (median fraction of total calls genome-wide for partial hard sweep = 2.84%; median fraction of total calls genome-wide for partial soft sweep = 7.24%), coupled with only a handful of completed hard sweep predictions (median fraction of total calls genome-wide = 0.03%)

421 (Figure 5; Table S3). Partial soft sweeps were typically discovered most often (median
422 proportion of sweep calls = 52.59%), with completed soft sweeps often following (median
423 proportion of sweep calls = 28.80%) and partial hard sweeps usually being the third most
424 numerous class of detected sweep (median proportion of sweep calls = 19.75%). Notably,
425 our estimated false discovery rates (FDRs) are higher for soft sweeps (median FDR for
426 completed soft sweeps = 11.09%; median FDR for partial soft sweeps = 12.20%) compared
427 to hard sweeps (median FDR for completed hard sweeps = 0.00%; median FDR for partial
428 hard sweeps = 0.39%); this implies that individual soft sweep candidates should be viewed
429 with more caution, though we should be able to estimate the genome-wide proportion of
430 these classes well. Specifically, classifications for partial soft sweeps outnumber those for
431 completed soft sweeps in every population besides GNS, as well as partial hard sweep calls
432 in all populations but GAS. Importantly, these findings remain the same, with the exception
433 that calls for partial hard sweeps now slightly exceed partial soft sweep classifications in
434 GNS, after false discovery correction (Table S3). Additionally, had partial sweeps not been
435 accounted for in the training process, our results suggest that we would have both
436 underestimated the total number of sweeps and incorrectly labeled many of our partial
437 sweeps (Figures 3, S8). This would have led to the conclusion that adaptation from
438 standing variation rather than *de novo* mutations dominate selective sweep dynamics in
439 these *A. gambiae* populations. While it is clear that soft sweeps are indeed more common in
440 these data, our results suggest that hard sweeps often occur as well, though with few
441 reaching fixation. Furthermore, the *partialS/HIC* classifications indicate that most selective
442 sweeps in these population samples are incomplete, suggesting that we are capturing a
443 view of selection in progress.

444

## 445 Selective sweeps are significantly enriched in functional regions of the *A.* 446 *gambiae* genome

447 To elucidate broad characteristics underlying the genomic targets of selection, we used
448 permutation tests of sweep call locations to discover enrichment patterns in the following
449 DNA regions of interest: gene, mRNA, exon, CDS, five-prime UTR, and three-prime UTR.

450 Permutation tests were based on the total number of calls for the four selection states with
451 sweeps occurring within the focal sub-window, as well as the individual number of calls for
452 each of these states respectively. Across all eight population datasets and for all six DNA
453 regions under investigation, there is a statistically significant enrichment of total sweep
454 calls along with completed soft sweeps calls in particular, whereas completed hard sweep
455 calls are not significantly enriched in any single case (Figures 6, S10; Table S4). Conversely,
456 partial sweep enrichment varies among populations as well as individual DNA regions.
457 Specifically, partial hard sweeps are significantly enriched in five of the six DNA regions for
458 BFS and CMS, and four of the DNA regions for UGS; while partial soft sweeps are
459 significantly enriched in all six DNA regions for UGS, five of the DNA regions for BFM, and
460 four of the DNA regions for BFS.

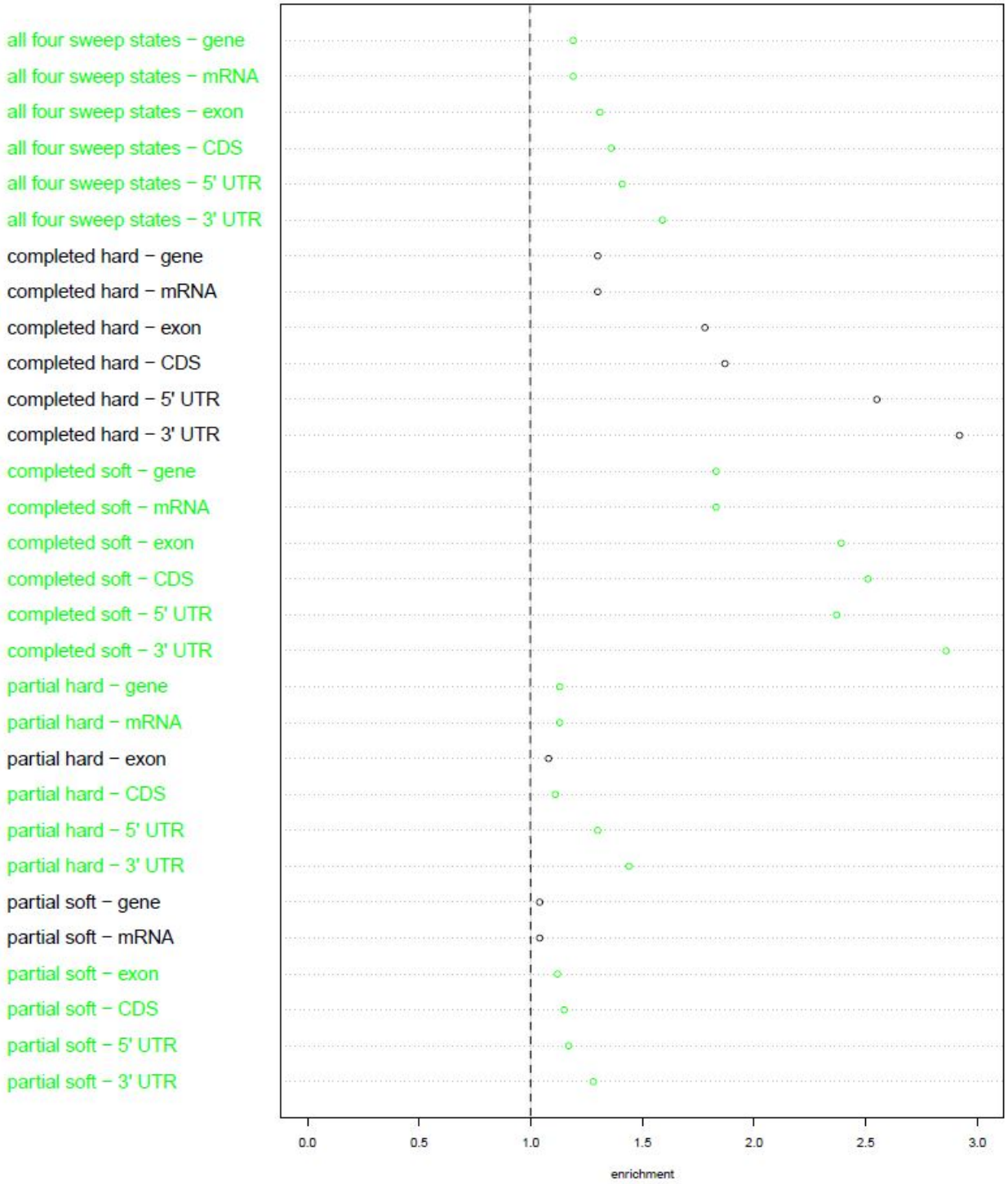


***Figure 6. Enrichments of partialS/HIC empirical selective sweep calls within DNA regions.*** *For the BFS population empirical dataset. Statistically significant (*p *< 0.05) enrichments in green. No enrichment (i.e. single-fold enrichment based on the permuted calls) is represented by the dotted line. Completed soft sweeps are enriched within all six DNA regions that experienced permutation tests. Similarly, all of the DNA regions have an enrichment of partial sweeps, hard and/or soft. In contrast, there is no significant enrichment*

468 *of completed hard sweeps for any of the DNA regions. Overall though, selective sweeps in*
469 *general are enriched for each of the DNA regions.*

470

471 **Insecticide resistance loci, especially related to metabolism, are**
472 **significantly enriched for selective sweeps**

473 We performed a similar permutation analysis for four sets of genes known to confer
474 insecticide resistance (IR), finding at least one set of IR genes to be statistically significant
475 for every population dataset in enrichment of total sweep calls (*i.e.* aggregate of all four
476 sweep classes) and completed soft sweeps, respectively (Figures 7, S11; Table S5). In
477 particular, metabolism-related IR genes are significantly enriched for each of these cases.
478 Furthermore, IR genes corresponding to well-characterized resistance loci (*i.e.* target sites)
479 are significantly enriched for AOM and BFM total sweep calls as well as completed soft
480 sweeps in AOM, BFM, and GNS; IR genes associated with behavior are significantly enriched
481 for BFM and GNS total sweep calls as well as completed soft sweeps in AOM, CMS, and GAS;
482 and IR genes affiliated with cuticular activity are significantly enriched for completed soft
483 sweeps in GWA and UGS. In contrast, completed hard sweeps are only significantly
484 enriched in BFS, GNS, GWA, and UGS for IR genes connected to behavior (as well as
485 metabolism for BFS). For partial sweeps, significant enrichment only occurs within BFM
486 (partial soft sweeps in metabolism as well as behavior IR genes), CMS (partial hard sweeps
487 in metabolism as well as target site IR genes), and UGS (partial hard sweeps in metabolism
488 IR genes).

489

490 **Completed soft sweeps are significantly enriched within the same GO**
491 **term annotations across populations**

492 To uncover further functional traits targeted by selection, we used our permuted datasets
493 to find individual gene ontology (GO) terms that are enriched in our sweep candidates.
494 Partial soft sweeps are not significantly enriched for any GO terms among populations,

495 while completed hard sweeps and partial hard sweeps are each enriched only in UGS for a
496 single GO term belonging to the cellular components GO domain; specifically, "membrane"
497 is enriched of completed hard sweeps and "nuclear cohesion complex" is enriched of partial
498 hard sweeps. Otherwise, GO term enrichment pertains to just completed soft sweeps as
499 well as the total of all four sweep states in combination (Table S6). For the completed soft
500 sweep significant enrichments, we found six cases of the same GO term in all eight
501 populations. Three of these are related to cellular components ("nucleus", "membrane" and
502 "integral component of membrane"), and the other three are connected to molecular
503 function, specifically binding ("nucleic acid binding", "protein binding" and "ATP binding").
504 All six of these terms, especially those conferring binding function, are also enriched for
505 total sweep calls across multiple datasets: "nucleic acid binding" and "ATP binding" in
506 seven populations; "protein binding" and "nucleus" in six populations; "integral component
507 of membrane" in five populations; and "membrane" in three populations, one of which is
508 significantly enriched for this GO term of completed hard sweeps as well, the single
509 example of such among all populations. Other cases involving the same GO term
510 significantly enriched for completed soft sweeps in over half of the populations include:
511 "binding", "cytoplasm", and "zinc ion binding" in seven populations; "RNA binding" in six
512 populations; and "mRNA splicing, via spliceosome" and "ATPase activity" in five
513 populations.

***Figure 7. Enrichments of partialS/HIC empirical selective sweep calls within groupings of insecticide resistance genes.*** *For the BFS population empirical dataset. Statistically significant (p < 0.05) enrichments in green. No enrichment (i.e. single-fold enrichment based on the permuted calls) is represented by the dotted line. Rows without a plotted point are due to the insecticide resistance gene group not having any calls of that sweep type (e.g. there are no completed hard sweeps called within target site genes). In this mosquito population, there*

521 *is evidence for several mechanisms through which insecticide resistance evolved, both in*
522 *sweep types and categories of genes under selection.*

523

524 # Discussion

525 ## *partialS/HIC* elucidates both species-wide and population-specific sweep
526 dynamics within *A. gambiae*

527 The Ag1000G data provided the opportunity to investigate selection at both the continental
528 scale, where wide-reaching impact across the whole species complex could be uncovered,
529 and the regional level, revealing population-specific sweep dynamics. For the former, we
530 observed that *A. gambiae* populations consistently experienced very few completed hard
531 sweeps, with nearly all sweeps being partial and/or soft. In fact, the impact of completed
532 hard sweeps on the adaptive process within mosquitos appears to be even more limited
533 than what was observed previously in humans [32]. This is likely a result of the much
534 larger population sizes and concordant levels of genetic variation that are maintained
535 within *Anopheles* populations. Importantly, we find a large number of ongoing selective
536 sweeps within these populations, particularly in comparison to the number of completed
537 sweeps. There are multiple reasons why this might be the case. A trivial explanation may
538 simply be that we only have power to detect sweeps that have completed in the past few
539 hundred generations, though this seems unlikely. More plausibly, a large number of
540 ongoing sweeps might be expected given the recent change in environment induced by
541 vector control efforts. Another possible explanation is that the frequency dynamics of
542 beneficial alleles within a population is often more complex than assumed and may indeed
543 contain an overdominant component [33]. This would mean that some portion of the
544 partial sweeps that we are observing in *Anopheles* is actually balanced, or transiently
545 balanced, polymorphisms. A fourth class of explanation is that beneficial mutations may
546 not be able to fix in populations due to competition with beneficial mutations on other
547 genes that have originated in different parts of the species range [34]. Indeed, each of these
548 factors may play some role in our reported abundance of partial sweeps.

549 Although such genome-wide sweep patterns occur species-wide, enrichment
550 behavior seems much more population-specific. For instance, while every population
551 possesses significant enrichment of completed soft sweeps coupled with no completed
552 hard sweep significant enrichments for the six functional DNA regions studied here, partial
553 sweep enrichments vary widely among datasets. Sweep behavior is even more
554 idiosyncratic for insecticide resistance genes, as the only constant between populations is
555 that metabolism is a recurring target of selection, especially for completed soft sweeps.

556 These findings from the Ag1000G data provide important genomic resources that
557 could inform continental-wide malaria control strategies for the entire *A. gambiae* species
558 complex, as well as have relevance to management efforts specialized to certain
559 populations and localities. Such insight into mosquito vector evolution may also help curb
560 future insecticide resistance adaptation, and in turn prevent impending crises of vector
561 control failure. However, it is important to consider that our partial sweep calls could be
562 capturing more complex selective dynamics at play, for example polygenic and quantitative
563 trait adaptation [35,36], balancing selection [37], and introgression of beneficial alleles.
564 These could lead to different modes of adaptation for the same genomic region across
565 populations, for instance a favorable SNP undergoing a soft sweep at its origin and then
566 carried to neighboring populations (as was suggested in Miles *et al.* [6]) may appear to be
567 experiencing a partial hard sweep in those recipient populations. Such complicated
568 interactions merit further investigation on the Ag1000G data, which would not only
569 continue advancing methodological development for population genetics, but also address
570 interesting questions for a widespread and ecologically important organism that has
571 crucial ramifications on wildlife management and public health.

572

## 573 *partialS/HIC* offers unprecedented detection of partial sweeps

574 Supervised machine learning approaches are rapidly gaining traction among population
575 geneticists, with deep learning in particular beginning to experience increased attention
576 and methodological development due to its exciting potential to unlock classic population

genetics problems. Examples of successful SML implementation in population genomics include demographic model choice [38], demographic parameter inference [39], comparative analysis of independent single-population size changes [40], identification of introgressed regions [41], recombination rate estimation [42–44], and genomic scans of selective sweeps [22]; deep learning specifically has been employed for joint inference of demography and selection [23], discovery of recombination hotspots [45], estimation of demographic and recombination parameters [46,47], discovery of functional variants [48], prediction of geographic origin [49], and differentiating between hard and soft sweeps from neutral regions [24]. These applications especially benefit from the ability to handle high dimensional input data and bypassing the need of a likelihood function, which is due to SML uncovering data patterns from leveraging *a priori* information through a training algorithm [23,26]. CNNs expand this utility to image processing, which has been demonstrated with *diploS/HIC* to be a powerful tool for exploiting the genomic spatial distribution of multiple population-level summary statistics to detect selective sweeps [24].

Here, we demonstrated with *partialS/HIC* that deep learning can be extended to partial sweeps, especially partial hard sweeps, yielding greater accuracy and robustness than has been previously attained. We also showcased consistent performance in the face of several underlying demographic backgrounds. Specifically, *partialS/HIC* is capable of discovering selective sweeps when partial sweeps are considered, simultaneous disambiguation between partial and completed sweeps as well as between hard and soft sweeps, and spatial localization of selection targets in the genome. Moreover, we have shown that partial sweeps remain mostly undetected if ignored from the training process, even though such selection may be commonplace throughout a genome as with the Ag1000G data. As a result, many previous studies scanning for either complete or ongoing selective sweeps solely (*i.e.* not jointly inferring both types of selection) may have overlooked an important subset of evolutionary events [34]. Researchers may then be interested in reexamining datasets with *partialS/HIC* to elucidate the relative contributions of fixed versus incomplete sweeps to adaptive evolution.

Importantly, the efficacy of *partialS/HIC* relies on several factors that are unexplored here, including simulation prior specifications, CNN architecture with respect

607 to construction and parameterization of neural network layers, and data structure. Hence,
608 it is prudent for future implementations to validate performance by testing a range of
609 configurations, given a project's individual intricacies, to assess robustness and inherent
610 assumptions. In particular, future exploration of alternate image constructions could be
611 potentially of great methodological benefit. Such images could be derived from different
612 ordering schemes and/or suites of summary statistics, as well as without summary
613 statistics entirely, instead directly exploiting sequence alignments [45–47,49] or even raw
614 reads. More broadly, CNNs can be further extended to address other long-standing efforts
615 in evolutionary biology, such as parameter inference under complex isolation-migration
616 models or phylogenetic reconstruction [50].

617

618 # Methods

619 ## Simulations for training and testing CNN classifier

620 We used *discoal* [30] to simulate training and test datasets corresponding to each *A.*
621 *gambiae* population under nine different selection states: neutrally evolving, completed
622 hard sweep, completed soft sweep, partial hard sweep, partial soft sweep, and linked
623 region for every one of the four sweep classes (Figure S1; Table S7). For the four sweep
624 types, the target SNP was located in the exact middle position within the central, or sixth in
625 sequence, sub-window of 11 in total; the selected SNP was placed in the middle within one
626 of the other ten sub-windows for linked sweeps. There were 2,000 training examples per
627 selection state (with the specific sub-window under selection, *i.e.* one of ten possible
628 simulation classes, randomized for linked sweeps) and 1,000 test examples per class
629 (including for each of the ten linked sweep locations), resulting in a training dataset of
630 18,000 simulations and a test dataset of 45,000 simulations given each demographic
631 history, thus totaling 144,000 training and 360,000 test simulations (as well as 495,000
632 additional test simulations from 11 test datasets all under the BFS population history to
633 explore demographic misspecification; see below). To conduct single-population
634 simulations with *discoal*, we used the *stairway plot* [51] point estimates from Miles *et al.* [6]

for size change parameters as well as $N_0$ (present-day effective population size), assumed a mutation rate ($\mu$) of $3.5\times10^{-9}$ mutations per base pair per generation, and performed random draws for locus-wide mutation and recombination rates from the following distributions for each independent replicate: $\theta \sim U(\frac{2\times E[\theta]}{11.0}, \frac{2\times E[\theta]}{1.1})$, whereby $E[\theta] = 4N_0\mu L$ and $L$ is the length of the simulated sequence with $L$ = 55,000 base pairs; $\rho \sim TEXP(2\times E[\theta], 6\times E[\theta])$, where $\rho = 4N_0rL$, $r$ is the recombination rate per base pair, and $TEXP(\beta$, maximum value) is a truncated exponential distribution with mean $\beta$; $s \sim U(1.0\times10^{-4}, 1.0\times10^{-2})$; end time of sweep $\sim U(0, 2{,}000)$ generations ago, which represents fixation for completed sweeps and the transition back to neutral evolution for partial sweeps; selected SNP allele frequency at onset of soft sweep $\sim U(\frac{1}{N_0}, 0.2)$; and selected SNP allele frequency at end of partial sweep $\sim U(0.20, 0.99)$.

## Constructing two-dimensional feature vector images of summary statistics

The eight training and test datasets (including an additional 11 BFS datasets for testing demographic misspecification), as well as empirical datasets, were converted into two-dimensional feature vector matrices for downstream deep learning (Figure S1); this was performed within the Python environment and required usage of the module *numpy*. Prior to this two-dimensional transformation, the simulated data were modified to better account for uncertainty within the empirical data, specifically: 1) sites that were missing any individual calls or could not be polarized against the outgroup were excluded; and 2) incorrect identification of the derived allele. For the former, each simulation randomly drew from a distribution of 1,552 masking profiles (with test simulations drawing without replacement per selection class of 1,000 simulations), which determined the exact sites to be omitted from further analysis; a masking profile consisted of the site positions within a single full 55 KB window on the *A. gambiae* genome that had absent at least one sample throughout the entirety of the Ag1000G data and/or ancestral state information, and the total set represented all 1,552 sequential, non-overlapping windows (*e.g.* 2L: 1–55,000; 2L:

663 55,001–110,000; 2L: 110,001–165,000; etc.) where the proportion of masked sites did not
664 exceed 75% in any of the constituent sub-windows (*i.e.* 1,250 sites). To account for
665 mispolarization, estimated rates were obtained from Miles *et al.* [6] and exploited via a
666 binomial distribution to mispolarize a random subset of SNPs to the other allele per
667 simulation.

668 The empirical data similarly underwent processing for compatibility with the
669 simulated data. First, chromosomes were delineated into sequential 5 KB sub-windows
670 (*e.g.* positions 1-5,000 formed the first sub-window, positions 5,001-10,000 formed the
671 second sub-window, etc.), with the aforementioned masking criteria applied across sites
672 and remaining SNPs polarized. Within each population dataset, all polymorphic positions
673 composed of more than two alleles were further removed from analysis, such that only
674 polarized monomorphic and biallelic sites comprising a full data matrix of no missing data
675 were left. Sub-windows containing no SNPs or less than 25% of the original sites were
676 subsequently discarded, and every configuration of 11 contiguous 5 KB sub-windows of
677 those remaining formed a single full window, which would be classified into one of the nine
678 selection states based upon its central sub-window while using spatial information from
679 the neighboring five sub-windows on either side. To clarify, this eliminated any window
680 that did not have a consecutive sequence of 11 sub-windows that survived data filtering,
681 and resulted in a sliding window that progressed a single sub-window at a time, such that
682 succeeding full windows could be overlapping by up to ten sub-windows.

683 Every independent simulation totaling 55 KB in length from 11 sub-windows of 5
684 KB, as well as empirical sequence of 11 adjacent sub-windows per population, was then
685 transformed into 89 separate summary statistic vectors that capture aspects of population-
686 level variation across the sampled individuals, with each vector consisting of 11 elements
687 corresponding to the sub-windows. The first 17 summary statistics were $\pi$ [52], $\theta_W$ [53],
688 Tajima's $D$ [14], $\theta_H$ [15], Fay-Wu's $H$ [15], number of unique haplotypes, $H_1$ [54], $H_{12}$ [54],
689 $H_2/H_1$ [54], $Z_{nS}$ [55], maximum $\omega$ [16], $E[iHS]$ [28], maximum $iHS$ [28], proportion of outlier
690 $iHS$ values [28], variance of pairwise genotype distances [24], skewness of pairwise
691 genotype distances [24], and kurtosis of pairwise genotype distances [24]. These were
692 previously implemented in *S/HIC* [22] and/or *diploS/HIC* [24] except for $H_1$ (though a

693 multilocus genotype equivalent was used in *diploS/HIC*) and the *iHS*-based statistics. We
694 employed the Python package *scikit-allel* to calculate $\pi$, $\theta_W$, Tajima's $D$, $\theta_H$, Fay-Wu's $H$,
695 number of unique haplotypes, $H_1$, $H_{12}$, $H_2/H_1$, and the statistics related to *iHS*. Values for *iHS*
696 were standardized within 50 derived allele frequency bins, following mispolarization in the
697 case of simulated data. Outlier *iHS* values were defined as within either 2.0% tail of the
698 distribution obtained from simulations of neutral evolution under the appropriate
699 demographic history.

700 The remaining 72 summary statistics were distribution summaries of SAFE and its
701 various components [29], specifically: haplotype allele frequency (HAF), which is the sum
702 of derived allele counts across all the derived alleles present within a sequence; unique
703 HAF score (*i.e.* each unique HAF value is counted only once, even if representing multiple
704 individuals); $\varphi$, which is the sum of HAF scores for sequences harboring the derived allele,
705 divided by the total sum of HAF scores across all sequences; κ, which is the proportion of
706 distinct HAF scores that carry the derived allele; derived allele frequency; and the SAFE
707 score itself, which is the difference between $\varphi$ and κ normalized against the derived allele
708 frequency. Notably, HAF is calculated per sequence, whereas $\varphi$, κ, derived allele frequency,
709 and SAFE are calculated per polymorphism. The following distribution summaries were
710 utilized to construct individual values spanning a sub-window: mean, median, mode; 2.5%,
711 25%, 75%, and 97.5% quartiles; maximum, variance, standard deviation, skewness, and
712 kurtosis. Importantly, each summary statistic vector was normalized, in the same manner
713 as the preceding versions to *partialS/HIC* [22,24], to capture signal solely from the relative
714 spatial distribution of the summary statistics across the 11 sub-windows, rather than
715 allowing influence from absolute values. Subsequently, the 89 vectors were vertically
716 collated to form a two-dimensional matrix that could then be exploited for image
717 processing. The arrangement of these vectors were such that the 11 columns corresponded
718 to the series of sub-windows from left to right, and the 89 rows of summary statistics were
719 in the order presented here (with the distribution summaries iterating first for every SAFE
720 component, *e.g.* row 52: skewness of $\varphi$ values; row 53: kurtosis of $\varphi$ values; row 54: $E$[κ]).
721 Importantly, column and row order affects deep learning optimization, which may have
722 consequences on overall efficacy, due to the convolutional and pooling windows employed

723 by the CNN architecture, hence related summary statistics were grouped together (*e.g.*
724 alternative distribution summaries of a SNP-based statistic, various SAFE derivatives).
725 Heatmap images, based on median values per statistic and sub-window, were generated in
726 R for the neutral case and each of the four sweep states under every population history
727 from the training simulations.

728

## 729 Training and testing CNNs for deep learning implementation

730 The architecture of our CNN was composed of the following sequential layers: 1) 2D
731 convolutional layer with 256 filters using 3×6 windows and "same" padding; 2) 2D max
732 pooling layer given a 3×3 window; 3) a second 2D convolutional layer of 256 filters based
733 also on 3×3 windows, "same" padding, and ReLU activation; 4) a second 2D max pooling
734 layer with a 3×1 window; 5) dropout layer with p=0.25; 6) flattening layer; 7) fully-
735 connected layer with ReLU activation to 512 responses; 8) a second dropout layer with
736 p=0.50; 9) a second fully-connected layer with ReLU activation to 128 elements; 10)
737 another dropout layer with p=0.50; and 11) softmax activation layer to 9 states (Figure 1).
738 This architecture was trained using the Python module *Keras* [56] given the *Adam*
739 optimizer [57], with 20 epochs, batch size of 32 simulations per step within an epoch, and
740 10% of the training data (*e.g.* 1800 simulations from the total nine-state training dataset)
741 randomly removed as a validation set during optimization (Figure S1). Training was
742 performed for every population demography under three experimental settings: 1) given
743 the full set of training data distributed across nine selection states; 2) exploiting a subset of
744 the training data from only five of the selection states, specifically those involving neutral
745 regions or completed sweeps; 3) deploying the entire training data, but with binary
746 classification between selective sweeps in the focal sub-window and all unselected classes
747 (*i.e.* neutral class together with every linked class). The three training regimes were then
748 applied to the test dataset that corresponded in underlying simulated history, resulting in
749 predictions of selection state given the default *Keras* threshold parameters for the first two
750 training schemes, whereas the softmax probability of a focal sweep was instead exploited
751 under training on binary classification. In the former case, individual inferences per test

752 simulation were collated and compared against true values to produce overall accuracy
753 measures as well as confusion matrix heatmaps to assess misclassification bias for each of
754 the 45 simulated scenarios (*i.e.* 11 sub-window sweep locations for each of the four sweep
755 types, plus neutrality).

756 Moreover, to explore the effect of demographic misspecification, we conducted a
757 single additional test under the first experimental set-up whereby the CNN trained on the
758 GAS simulations was applied to the CMS test simulations. Furthermore, we engaged in two
759 more misspecification experiments that involved simulating, for the first experiment, five
760 testing datasets that changed the underlying demographic model, and for the second
761 experiment, six testing datasets that differed in fixed $\theta$ value. In the five test sets that
762 altered the demographic history, we used an instantaneous contraction model that
763 experienced a population crash at 0.0001 time units with an intensity of, respectively
764 across the five datasets: 20×, 40×, 60×, 80×, 100×. Priors for $\theta$, $\rho$, $s$, end time of sweep,
765 selected SNP allele frequency at onset of soft sweep, and selected SNP allele frequency at
766 end of partial sweep remained the same as for the initial BFS simulations. In the six test
767 sets that varied $\theta$, we employed the same parameterization from the initial BFS simulations
768 with the exception that $\theta$ was set to, respectively across the six datasets: 1,000; 5,000;
769 10,000; 15,000; 20,000; 25,000. Our motivation for these intervals was to exceed the
770 bounds of the original prior distribution, $\theta \sim U(1{,}750.204699,\ 17{,}502.046985)$. After
771 simulation, we leveraged the nine-state CNN trained under the BFS specifications against
772 each of the 11 testing datasets to obtain overall accuracies for comparison among these test
773 datasets and to the original test dataset whereby the BFS history was correctly modeled.

774 Regarding the binary classification experiment, the Python module *sklearn* [58] was
775 used for building the ROC curves to evaluate accuracy and sensitivity. Additionally, *sklearn*
776 was utilized for conducting PCA on the training simulations, on either the focal sub-window
777 or across the entire full-scale genomic window respectively, with subsequent projection of
778 the testing simulations into independent sets of PC1 and PC2 values, to obtain a total of
779 four individual ROC curves. To calculate the Composite of Multiple Signals [31], training
780 data were separated into the two categories of selection versus neutrality (*i.e.* 4,000
781 simulations from four selection states versus 41,000 simulations from 41 neutral or linked

782 classes), which were respectively converted into a histogram of 200 bins for every
783 summary statistic. Assuming the Composite of Multiple Signals for the central sub-window
784 only, this resulted in a total of 178 distributions given the 89 statistics for both selection
785 and neutral; for all sub-windows, this was instead a total of 1,958 distributions due to the
786 11 total sub-windows of statistics. These distributions were then employed to calculate
787 probabilities both under selection and neutrality for the corresponding testing data
788 summary statistics; for bins with a zero value, 0.1 of the smallest possible increment (*i.e.*
789 selection: $\frac{0.1\times1}{4000}$; neutrality: $\frac{0.1\times1}{41000}$) was assigned as the probability. To determine the prior
790 for each probability, represented in the original Composite of Multiple Signals equation by
791 π, we utilized the proportional composition of the training simulations (*i.e.* selection: $\frac{4}{45}$;
792 neutrality: $\frac{41}{45}$). Finally, probabilities were natural log transformed and summed among
793 summary statistics to compute the Composite of Multiple Signals.

794

795 **Detecting selective sweeps for *A. gambiae* population datasets**

796 To scan the genome for signatures of selective sweeps, the nine-state trained CNNs were
797 applied to the eight empirical mosquito datasets, with the underlying simulated
798 demography matched to the sampled population (Figure S1). Calls were corrected for false
799 discovery by exploiting the accuracy and error rates for neutral regions from the nine-state
800 simulation experiment, such that the amount of neutral calls was assumed to be
801 underestimated while the amount of calls for the remaining eight selection states were
802 assumed to be inflated. Subsequently, we produced sets of 10,000 randomly permuted calls
803 across the genome to derive null expectations of sweep enrichment, following Schrider and
804 Kern [32]. Using the gene annotation file "Anopheles-gambiae-
805 PEST_BASEFEATURES_AgamP4.7.gff3.gz" from VectorBase, we exploited these permuted
806 datasets to assess statistically significant enrichment within certain DNA regions,
807 groupings of known IR genes (N. Harding, *pers. comm.*), and all basic GO term definitions
808 from http://www.geneontology.org (last accessed February 18, 2015). The DNA regions of
809 interest included gene, mRNA, exon, CDS, five-prime UTR, and three-prime UTR; IR genes

were assigned to four functional categories: metabolism, target sites, behavior, and cuticular. To determine significant enrichment, the number of inferred calls for a particular DNA region or IR gene category had to have a $p$-value < 0.05 based on the respective distribution of 10,000 permutations; for the GO terms, we deployed a corrected $q$-value < 0.05 due to concerns of false discovery stemming from the large number of terms tested for enrichment.

## Acknowledgments

We thank Jeff Adrion for comments on the manuscript. This work was supported by NIH award no. R01GM117241 to ADK and NIH award no. K99HG008696 to DRS.

820 # References


821 1. WHO. World Malaria Report 2015 [Internet]. 2015. Available:
822 http://www.who.int/malaria/publications/world-malaria-report-2015/en/

823 2. Hemingway J, Ranson H, Magill A, Kolaczinski J, Fornadel C, Gimnig J, et al. Averting a
824 malaria disaster: Will insecticide resistance derail malaria control? Lancet. 2016;387:
825 1785–1788. doi:10.1016/S0140-6736(15)00417-1

826 3. Brown AW. Insecticide resistance in mosquitoes: a pragmatic review. J Am Mosq
827 Control Assoc. 1986;2: 123—140.

828 4. Elissa N, Mouchet J, Rivière F, Meunier J-Y, Yao K. Resistance of Anopheles gambiae
829 s.s. to pyrethroids in Côte d'Ivoire. Ann Soc Belg Med Trop (1920). 1993;73: 291–
830 294.

831 5. Martinez-Torres D, Chandre F, Williamson MS, Darriet F, Bergé JB, Devonshire AL, et
832 al. Molecular characterization of pyrethroid knockdown resistance (kdr) in the major
833 malaria vector Anopheles gambiae s.s. Insect Mol Biol. 1998;7: 179–184.
834 doi:10.1046/j.1365-2583.1998.72062.x

835 6. Miles A, Harding NJ, Bottà G, Clarkson CS, Antão T, Kozak K, et al. Genetic diversity of
836 the African malaria vector anopheles gambiae. Nature. 2017;552: 96–100.
837 doi:10.1038/nature24995

838 7. Oxborough RM, N'Guessan R, Jones R, Kitau J, Ngufor C, Malone D, et al. The activity of
839 the pyrrole insecticide chlorfenapyr in mosquito bioassay: Towards a more rational
840 testing and screening of non-neurotoxic insecticides for malaria vector control. Malar
841 J. 2015;14: 124. doi:10.1186/s12936-015-0639-x

842 8. Hammond A, Galizi R, Kyrou K, Simoni A, Siniscalchi C, Katsanos D, et al. A CRISPR-
843 Cas9 gene drive system targeting female reproduction in the malaria mosquito
844 vector Anopheles gambiae. Nat Biotechnol. 2016;34: 78–83. doi:10.1038/nbt.3439.A

845 9. Unckless RL, Messer PW, Connallon T, Clark AG. Modeling the manipulation of
846 natural populations by the mutagenic Chain reaction. Genetics. 2015;201: 425–431.
847 doi:10.1534/genetics.115.177592

848 10. Unckless RL, Clark AG, Messer PW. Evolution of resistance against CRISPR/Cas9 gene
849 drive. Genetics. 2017;205: 827–841. doi:10.1534/genetics.116.197285

850 11. Maynard Smith J, Haigh J. The hitch-hiking effect of a favourable gene. Genet Res
851 (Camb). Rutgers University; 1974;23: 23–35. doi:10.1017/S0016672308009579

852 12. Kaplan NL, Hudson RR, Langley CH. The "hitchhiking effect" revisited. Genetics.
853 1989;123: 887–899.

854 13. Stephan W, Wiehe THE, Lenz MW. The effect of strongly selected substitutions on
855 neutral polymorphism: analytical results based on diffusion theory. Theor Popul Biol.

1992;41: 237–254. doi:10.1016/0040-5809(92)90045-U

14. Tajima F. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics. 1989;123: 585–595.

15. Fay JC, Wu C-I. Hitchhiking under positive Darwinian selection. Genetics. 2000;155: 1405–1413.

16. Kim Y, Neilsen R. Linkage disequilibrium as a signature of selective sweeps. Genetics. 2004;167: 1513–1524. doi:10.1534/genetics.103.025387

17. Nielsen R, Williamson S, Kim Y, Hubisz MJ, Clark AG, Bustamante C. Genomic scans for selective sweeps using SNP data. Genome Res. 2005;15: 1566–75. doi:10.1101/gr.4252305

18. Pavlidis P, Jensen JD, Stephan W. Searching for footprints of positive selection in whole-genome SNP data from nonequilibrium populations. Genetics. 2010;185: 907–922. doi:10.1534/genetics.110.116459

19. Lin K, Li H, Schlötterer C, Futschik A. Distinguishing positive selection from neutral evolution: Boosting the performance of summary statistics. Genetics. 2011;187: 229–244. doi:10.1534/genetics.110.122614

20. Ronen R, Udpa N, Halperin E, Bafna V. Learning natural selection from the site frequency spectrum. Genetics. 2013;195: 181–193. doi:10.1534/genetics.113.152587

21. Pybus M, Luisi P, Dall'Olio GM, Uzkudun M, Laayouni H, Bertranpetit J, et al. Hierarchical boosting: A machine-learning framework to detect and classify hard selective sweeps in human populations. Bioinformatics. 2015;31: 3946–3952. doi:10.1093/bioinformatics/btv493

22. Schrider DR, Kern AD. S/HIC: Robust Identification of Soft and Hard Sweeps Using Machine Learning. PLoS Genet. 2016;12: e1005928. doi:10.1371/journal.pgen.1005928

23. Sheehan S, Song YS. Deep Learning for Population Genetic Inference. PLoS Comput Biol. 2016;12: e1004845. doi:10.1371/journal.pcbi.1004845

24. Kern AD, Schrider DR. diploS/HIC: An Updated Approach to Classifying Selective Sweeps. G3 Genes, Genomes, Genet. 2018;8: 1959–1970. doi:10.1534/g3.118.200262

25. Sugden LA, Atkinson EG, Fischer AP, Rong S, Henn BM, Ramachandran S. Localization of adaptive variants in human genomes using averaged one-dependence estimation. Nat Commun. Springer US; 2018;9: 703. doi:10.1038/s41467-018-03100-7

26. Schrider DR, Kern AD. Supervised Machine Learning for Population Genetics: A New Paradigm. Trends Genet. Elsevier Ltd; 2018;34: 301–312. doi:10.1016/j.tig.2017.12.005

892 27. Coop G, Ralph P. Patterns of neutral diversity under general models of selective
893 sweeps. Genetics. 2012;192: 205–224. doi:10.1534/genetics.112.141861

894 28. Voight BF, Kudaravalli S, Wen X, Pritchard JK. A map of recent positive selection in
895 the human genome. PLoS Biol. 2006;4: e72. doi:10.1371/journal.pbio.0040072

896 29. Akbari A, Vitti JJ, Iranmehr A, Bakhtiari M, Sabeti PC, Mirarab S, et al. Identifying the
897 favored mutation in a positive selective sweep. Nat Methods. Nature Publishing
898 Group; 2018;15: 279–282. doi:10.1038/nmeth.4606

899 30. Kern AD, Schrider DR. Discoal: Flexible Coalescent Simulations With Selection.
900 Bioinformatics. 2016;32: 3839–3841. doi:10.1093/bioinformatics/btw556

901 31. Grossman SR, Shylakhter I, Karlsson EK, Byrne EH, Morales S, Frieden G, et al. A
902 Composite of Multiple Signals Distinguishes Causal Variants in Regions of Positive
903 Selection. Science (80- ). 2010;327: 883–886. doi:10.1126/science.1183863

904 32. Schrider DR, Kern AD. Soft sweeps are the dominant mode of adaptation in the
905 human genome. Mol Biol Evol. 2017;34: 1863–1877. doi:10.1093/molbev/msx154

906 33. Sellis D, Callahan BJ, Petrov DA, Messer PW. Heterozygote advantage as a natural
907 consequence of adaptation in diploids. Proc Natl Acad Sci. 2011;108: 20666–20671.
908 doi:10.1073/pnas.1114573108

909 34. Ralph PL, Coop G. Parallel Adaptation: One or Many Waves of Advance of an
910 Advantageous Allele? Genetics. 2010;186: 647–668.
911 doi:10.1534/genetics.110.119594

912 35. Pritchard JK, Pickrell JK, Coop G. The Genetics of Human Adaptation: Hard Sweeps,
913 Soft Sweeps, and Polygenic Adaptation. Curr Biol. Elsevier Ltd; 2010;20: R208–R215.
914 doi:10.1016/j.cub.2009.11.055

915 36. Booker TR, Jackson BC, Keightley PD. Detecting positive selection in the genome.
916 BMC Biol. BMC Biology; 2017;15: 98. doi:10.1186/s12915-017-0434-y

917 37. Connallon T, Clark AG. Antagonistic versus nonantagonistic models of balancing
918 selection: Characterizing the relative timescales and hitchhiking effects of partial
919 selective sweeps. Evolution (N Y). 2013;67: 908–917. doi:10.1111/j.1558-
920 5646.2012.01800.x

921 38. Smith ML, Ruffley M, Espíndola A, Tank DC, Sullivan J, Carstens BC. Demographic
922 Model Selection using Random Forests and the Site Frequency Spectrum. Mol Ecol.
923 2017; doi:10.1111/mec.14223

924 39. Pudlo P, Marin J-M, Estoup A, Cornuet J-M, Gauthier M, Robert CP. Reliable ABC
925 model choice via random forests. Bioinformatics. 2016;32: 859–866.
926 doi:10.1093/bioinformatics/btv684

927 40. Xue AT, Hickerson MJ. Multi-DICE: R package for comparative population genomic

inference under hierarchical co-demographic models of independent single-population size changes. Mol Ecol Resour. 2017; doi:10.1111/1755-0998.12686

41. Schrider DR, Ayroles J, Matute DR, Kern AD. Supervised machine learning reveals introgressed loci in the genomes of Drosophila simulans and D. sechellia. PLOS Genet. 2018;14: e1007341. doi:10.1371/journal.pgen.1007341

42. Lin K, Futschik A, Li H. A fast estimate for the population recombination rate based on regression. Genetics. 2013;194: 473–484. doi:10.1534/genetics.113.150201

43. Adrian AB, Corchado JC, Comeron JM. Predictive models of recombination rate variation across the drosophila melanogaster genome. Genome Biol Evol. 2016;8: 2597–2612. doi:10.1093/gbe/evw181

44. Gao F, Ming C, Hu W, Li H. New Software for the Fast Estimation of Population Recombination Rates (FastEPRR) in the Genomic Era. G3 Genes, Genomes, Genet. 2016;6: 1563–1571. doi:10.1534/g3.116.028233

45. Chan J, Perrone V, Spence JP, Jenkins PA, Mathieson S, Song YS. A Likelihood-Free Inference Framework for Population Genetic Data using Exchangeable Neural Networks. bioRxiv. 2018; 1802.06153. doi:10.1101/267211

46. Flagel L, Brandvain Y, Schrider DR. The Unreasonable Effectiveness of Convolutional Neural Networks in Population Genetic Inference. bioRxiv. 2018; 336073. doi:10.1101/336073

47. Adrion JR, Galloway JG, Kern AD. Predicting the landscape of recombination using deep learning. Mol Biol Evol. 2020; doi:10.1093/molbev/msaa038

48. Zhou J, Troyanskaya OG. Predicting effects of noncoding variants with deep learning-based sequence model. Nat Methods. 2015;12: 931–934. doi:10.1038/nmeth.3547

49. Battey CJ, Ralph PL, Kern AD. Predicting Geographic Location From Genetic Variation with Deep Neural Networks. bioRxiv. 2019; 2019.12.11.872051. doi:10.1101/2019.12.11.872051

50. Suvorov A, Hochuli J, Schrider DR. Accurate Inference of Tree Topologies from Multiple Sequence Alignments Using Deep Learning. Syst Biol. 2020;69: 221–233. doi:10.1093/sysbio/syz060

51. Liu X, Fu Y-X. Exploring population size changes using SNP frequency spectra. Nat Genet. 2015;47: 555–559. doi:10.1038/ng.3254

52. Tajima F. Evolutionary relationship of DNA sequences in finite populations. Genetics. 1983;105: 437–460.

53. Watterson GA. On the number of segregating sites in genetical models without recombination. Theor Popul Biol. 1975;7: 256–276. doi:10.1016/0040-5809(75)90020-9

964 54. Garud NR, Messer PW, Buzbas EO, Petrov DA. Recent Selective Sweeps in North
965 American Drosophila melanogaster Show Signatures of Soft Sweeps. PLoS Genet.
966 2015;11: e1005004. doi:10.1371/journal.pgen.1005004

967 55. Kelly JK. A test of neutrality based on interlocus associations. Genetics. 1997;146:
968 1197–1206.

969 56. Chollet F. Keras [Internet]. 2015. Available: https://keras.io

970 57. Kingma DP, Ba JL. Adam: A Method for Stochastic Optimization. arXiv. 2014;
971 1412.6980. doi:http://doi.acm.org.ezproxy.lib.ucf.edu/10.1145/1830483.1830503

972 58. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-
973 learn: Machine Learning in Python. J Mach Learn Res. 2011;12: 2825–2830.
974 doi:10.1007/s13398-014-0173-7.2

975

976 # Figure Captions

977 **Figure 1. CNN architecture of neural network layers.** Our *partialS/HIC* classifier utilizes
978 a convolutional neural network whereby the input feature vector image, composed of 89
979 summary statistics organized into rows and across 11 contiguous genomic sub-windows
980 organized into columns, is passed to a 2D convolutional layer with 256 filters using a 3×6
981 receptive field. Next is a 2D max pooling layer given a 3×3 receptive field. Then there is a
982 second 2D convolutional layer of 256 filters based also on a 3×3 receptive field and ReLU
983 activation. Afterward is a second 2D max pooling layer with a 3×1 receptive field. The
984 tensor is then flattened after a dropout layer (p=0.25) and passed to two fully-connected
985 layers with ReLU activation, resulting in 512 and 128 nodes with subsequent dropout
986 (p=0.50) respectively. Finally, a softmax activation layer classifies the 9 responses.

987 **Figure 2. Confusion matrix heatmap of *partialS/HIC* simulation experiment.** Given the
988 BFS population history. Each row designates a true simulation class, which for linked
989 sweeps is differentiated by distance in genomic sub-windows to the target selective sweep.
990 In total, there are 45 simulated scenarios shown, including 11 for each sweep type (*i.e.* one
991 case whereby the sweep is within the central sub-window, and 10 whereby a linked sweep
992 is located within one of the flanking sub-windows) and neutrality. Each column indicates
993 one of nine inferred states, with the linked simulation classes collapsed into a single
994 category per selective sweep type for training and classification, allowing *partialS/HIC* to
995 learn to distinguish between sweeps located in the focal sub-window and those somewhere
996 nearby. Darker blue cells represent a higher proportion of the 1,000 calls for each true
997 class. Importantly, there is generally high accuracy in discriminating between sweeps
998 within the focal sub-window versus linked sub-windows, especially when linkage is beyond
999 one sub-window away and particularly for completed sweeps. However, discovering partial
1000 soft sweeps is noticeably a challenging task. Moreover, there is greater sensitivity
1001 separating full completion from partial completion for hard sweeps in contrast to soft
1002 sweeps.

1003 **Figure 3. Confusion matrix heatmap of simulation experiment with partial sweeps**
1004 **ignored during training.** Given the BFS population history. Structured in the same

1005 manner as Figure 2, each row designates a true simulation class and darker blue cells
1006 represent a higher proportion of the 1,000 calls for each true class. In contrast to Figure 2
1007 though, each column indicates one of five inferred states instead of nine, as the two partial
1008 sweep and two linked partial sweep classification states were omitted from training to
1009 determine misclassification bias when partial sweeps are ignored. There is a substantial
1010 decrease in discovery of partial sweeps, especially partial hard sweeps (20.5% total sweep
1011 discovery compared to 78.2% in Figure 2; 46.3% total sweep discovery of partial soft
1012 sweeps compared to 62.3% in Figure 2). Specifically for partial hard sweeps, a large
1013 proportion of the detected sweeps are misclassified as soft sweeps instead of hard sweeps
1014 (19.0%; *i.e.* 3.9% of calls = soft sweeps : 16.6% of calls = hard sweeps).

1015 **Figure 4. ROC curves for binary classification of selective sweeps, including partial**
1016 **sweeps, versus regions neutrally evolving or under linked selection.** Given the BFS
1017 population history. The *partialS/HIC* deep learning classifier outcompetes two other
1018 approaches for managing the same suite of summary statistics: Composite of Multiple
1019 Signals and principal component analysis. Furthermore, *partialS/HIC* excels in performance
1020 over several sub-window derivatives of two summary statistics, SAFE and *iHS*, which were
1021 all included within our set of summary statistics used for training. Notably, Composite of
1022 Multiple Signals, SAFE, and *iHS* were all designed to uncover selective sweeps. Additionally,
1023 the SAFE score is itself a compound statistic that captures signal from several constituent
1024 statistics, and the majority of our training data is derived from either the SAFE score or one
1025 of its components.

1026 **Figure 5. Genome-wide *partialS/HIC* sweep calls across empirical mosquito**
1027 **population datasets.** For all eight datasets, there is a minimal number of calls for
1028 completed hard sweeps, represented here on the Y-axis as the relative proportion to the
1029 total set of empirical calls across every genomic window per population dataset. However,
1030 there is indeed a substantial proportion of sweep calls for hard sweeps, yet are incomplete
1031 or in-progress. In most cases, the number of both completed and partial soft sweeps further
1032 exceeds that of partial hard sweeps. Furthermore, the majority of sweeps have not reach
1033 fixation across these *Anopheles* populations.

1034 **Figure 6. Enrichments of *partialS/HIC* empirical selective sweep calls within DNA**
1035 **regions.** For the BFS population empirical dataset. Statistically significant ($p < 0.05$)
1036 enrichments in green. No enrichment (*i.e.* single-fold enrichment based on the permuted
1037 calls) is represented by the dotted line. Completed soft sweeps are enriched within all six
1038 DNA regions that experienced permutation tests. Similarly, all of the DNA regions have an
1039 enrichment of partial sweeps, hard and/or soft. In contrast, there is no significant
1040 enrichment of completed hard sweeps for any of the DNA regions. Overall though, selective
1041 sweeps in general are enriched for each of the DNA regions.

1042 **Figure 7. Enrichments of *partialS/HIC* empirical selective sweep calls within**
1043 **groupings of insecticide resistance genes.** For the BFS population empirical dataset.
1044 Statistically significant ($p < 0.05$) enrichments in green. No enrichment (*i.e.* single-fold
1045 enrichment based on the permuted calls) is represented by the dotted line. Rows without a
1046 plotted point are due to the insecticide resistance gene group not having any calls of that
1047 sweep type (*e.g.* there are no completed hard sweeps called within target site genes). In
1048 this mosquito population, there is evidence for several mechanisms through which
1049 insecticide resistance evolved, both in sweep types and categories of genes under selection.

1050 **Figure S1. Flowchart of *partialS/HIC* methodological procedure for training, testing,**
1051 **and empirical application given eight *Anopheles* mosquito population datasets.** Major
1052 steps highlighted in green, specifically data simulation, training and testing of CNN
1053 classifier, and genome-wide inference of selection state per focal sub-window within a full
1054 window of 11 sequential sub-windows. Important data processing details that define
1055 genomic sub-windows as well as account for uncertainties within the empirical data from
1056 missing sites and polarization are also included. However, we omit our various
1057 misclassification experiments, which are ancillary to our primary demonstration of
1058 *partialS/HIC* to detect selective sweeps from population genomic datasets while obtaining
1059 baseline measures of accuracy. For organizational and conceptual purposes, our protocol is
1060 presented here mostly contained within two columns, representing simulated versus
1061 empirical data respectively, but the actual order of events oscillates between these two
1062 sides of the diagram. In fact, many of the tasks could be accomplished in parallel, though
1063 certain actions must occur sequentially. For example, training and testing data were

generated after demographic inference in order to inform parameterization of the simulations. Moreover, we filtered the empirical data of missing calls and missing polarization information prior to constructing masking profiles for application to the simulated data, which invoked an intermediate step hence our usage of a dotted arrow. Additionally, the CNN first had to be optimized from the training simulations before conducting empirical analysis.

**Figure S2. Feature vector heatmaps for neutral and four sweep states under each population history.** Heatmaps derived from median values of a single training dataset, each assuming a different population history. Across populations, feature vectors demonstrate distinctive characteristics per classification state, especially in the focal sub-window and for the neutral state in comparison to the four sweep states.

**Figure S3. Training and validation loss over samples and epochs during *partialS/HIC* CNN optimization under each population history.** Black circles represent loss function outputs from training for every 32 samples out of a total 16,200 samples per each of 20 epochs. Green asterisks indicate loss function outputs from the validation set at the conclusion of epochs. The degree to which training loss is less than validation loss likely reflects the magnitude of overfitting the CNN to the training data.

**Figure S4. Confusion matrix heatmaps of *partialS/HIC* simulation experiment under each population history.** Extension of Figure 2 for all eight population histories. Importantly, the patterns highlighted in Figure 2, which displays the confusion matrix given the BFS population history and is reproduced identically here, are largely present throughout the other confusion matrices.

**Figure S5. Confusion matrix heatmap of *partialS/HIC* demographic misspecification test.** Training on simulations given the GAS population history while testing on simulations given the CMS population history. Importantly, the patterns highlighted in Figures 2 and S4 are largely retained here, though there is a modest decrease in soft sweep discovery for both completed and partial.

1091 **Figure S6. Overall *partialS/HIC* accuracies between testing datasets generated with**
1092 **varying demographic misspecification given training on the BFS population history.**
1093 The CNN trained on the simulations assuming the BFS population history was separately
1094 applied against five testing datasets that were produced under a simple two-epoch
1095 instantaneous population crash model, with each dataset differing only in relative
1096 magnitude of the size change. The overall accuracy from the original testing dataset
1097 whereby the demography was accurately specified on the BFS population history is marked
1098 by the green dotted line at 67.76%. The final population size ranges from 5% (*i.e.* 20×) to
1099 1% (*i.e.* 100×) of the original size, with increased accuracy for more severe contractions
1100 (*i.e.* 45.98% accuracy for reduction to 5.00% of original size; 55.27% for reduction to
1101 2.50%; 59.46% for reduction to 1.67%; 61.99% for reduction to 1.25%; and 63.50% for
1102 reduction to 1.00%). Notably, the greater error at lower magnitudes could be largely due to
1103 the decrease in simulated polymorphisms (Figure S7).

1104 **Figure S7. Overall *partialS/HIC* accuracies between testing datasets generated with**
1105 **varying $\theta$ values given training on the BFS population history.** The CNN trained on the
1106 simulations assuming the BFS population history was separately applied against six testing
1107 datasets that were produced under identical conditions as the training data except differing
1108 in fixed $\theta$ values. The overall accuracy from the original testing dataset whereby $\theta$ was
1109 drawn from a prior distribution is marked by the green dotted line at 67.76%. Across
1110 values of 1,000, 5,000, 10,000, 15,000, 20,000, and 25,000, which collectively is inclusive of
1111 and exceeds the original prior distribution ($\theta \sim U(1{,}750.204699,\ 17{,}502.046985)$), there is
1112 comparable accuracy to the original (50.99%, 67.78%, 70.36%, 70.74%, 70.72%, and
1113 70.74% respectively). In fact, accuracy exceeded the original for all tested $\theta$ values except $\theta$
1114 = 1,000.

1115 **Figure S8. Confusion matrix heatmaps of simulation experiment with partial sweeps**
1116 **ignored during training under each population history.** Extension of Figure 3 for all
1117 eight population histories. Importantly, the patterns highlighted in Figure 3, which displays
1118 the confusion matrix given the BFS population history and is reproduced identically here,
1119 are largely present throughout the other confusion matrices.

1120 **Figure S9. ROC curves for binary classification of selective sweeps, including partial**
1121 **sweeps, versus regions neutrally evolving or under linked selection for each**
1122 **population history.** Extension of Figure 4 for all eight population histories. Importantly,
1123 the patterns highlighted in Figure 4, which displays the ROC curves given the BFS
1124 population history and is reproduced identically here, are largely present throughout the
1125 other plots of ROC curves.

1126 **Figure S10. Enrichments of *partialS/HIC* empirical selective sweep calls within DNA**
1127 **regions for each empirical mosquito population dataset.** Extension of Figure 6 for all
1128 eight empirical population datasets. Importantly, the patterns highlighted in Figure 6,
1129 which displays the dot chart for the BFS permutation test and is reproduced here with a
1130 wider X-axis to match the other plots, are largely present throughout the other dot charts,
1131 though partial sweep enrichment is highly variable.

1132 **Figure S11. Enrichments of *partialS/HIC* empirical selective sweep calls within**
1133 **groupings of insecticide resistance genes for each empirical mosquito population**
1134 **dataset.** Extension of Figure 7 for all eight empirical population datasets. Importantly, as in
1135 Figure 7, which displays the dot chart for the BFS permutation test and is reproduced
1136 identically here, there is evidence for several mechanisms through which insecticide
1137 resistance evolved, both in sweep types and categories of genes under selection as well as
1138 both within individual populations and between different populations.