

SUPPLEMENTARY INFORMATION LEGENDS

Figure S1: Example and average Timesweeper inputs for AFT and HFT formats. The top two rows are individual replicates randomly selected for AFT (left column) and HFT (right column). The bottom two rows are averaged inputs across 10,000 replicates for each class: the third row is zoomed in to show the vicinity of the central polymorphism in closer detail for the AFT method, as well as the HFT method's input zoomed in to focus on the haplotype with the highest frequency increase and those most similar two it; the bottom row shows the entirety of these inputs. For the scenarios with selection (i.e. SSV and SDN) initial sampling time was drawn from a uniform distribution with bounds $[-50, 50]$ and selection coefficient was drawn from a uniform distribution with bounds $[0.00025, 0.25]$. 10 diploid individuals were sampled every 10 generations for 200 generations (see Methods).

Figure S2: Neural network architecture diagrams. Full diagrams with layer types and dimensions for the (A) 1DCNN, (B) Fully Connected Network (FCN), (C) 2DCNN, (D) 1DCNN with an increased number of parameters, and (E) RNN.

Figure S3: Resolution of the AFT and HFT methods assessed on simulated 500 kb chromosomes using varying window sizes. The fraction of polymorphisms classified as a sweep at varying locations along the 500 kb chromosome binned into 501 adjacent windows. When a sweep is present (SSV and SDN columns), it occurs in the center of the chromosome. The three left columns show the classification results of the AFT method, the right three columns show the

classification results of the HFT method, the window size of the feature vector increases from the top to bottom row.

Figure S4: Resolution of the AFT and HFT methods assessed on central 501 polymorphisms using varying window sizes. The fraction of polymorphisms classified as a sweep at each polymorphism along the centralmost 501 polymorphisms. When a sweep is present (SSV and SDN columns), it occurs in the center of the window. The three left columns show the classification results of the AFT method, the right three columns show the classification results of the HFT method, and the window size of the feature vector increases from the top to bottom row.

Figure S5: Timesweeper's classification performance on datasets of varying selection coefficient. `Timesweeper` was trained and tested on 10,000 replicates of each scenario (neutrality, SSV, SDN) with a sampling generation drawn from a uniform distribution with bounds $[-50, 50]$ for selection coefficients of 0.005, 0.01, 0.05, 0.1, and 0.5. Precision-recall (PR) curve, ROC curve, Normalized and non-normalized confusion matrices, and training losses for AFT (left half) and HFT (right half) for s values of 0.005, 0.01, 0.05, 0.1, and 0.5.

Figure S6: ROC Curves for varying selection coefficient classification. ROC curves for varying selection coefficients as described in Figure S5 for classifying sweeps versus neutrality (top row) and the SDN scenario versus the SSV scenario (bottom row) for both AFT (left column) and HFT (right column).

Figure S7: PR Curves for varying selection coefficient classification. PR curves for varying selection coefficients as described in Figure S5 for classifying sweeps versus neutrality (top row) and the SDN scenario versus the SSV scenario (bottom row) for both AFT (left column) and HFT (right column).

Figure S8: Timesweeper's classification performance using different neural network architectures. Timesweeper was trained and tested on the dataset described in Figure S1 using different neural network architectures. PR curves, ROC curves, normalized and non-normalized confusion matrices, and training losses for the classification task (neutrality versus sweep, SDN versus SSV classification) for 1DCNN (top row), 1DCNN with increased parameterization (second row), 2DCNN (third row), and RNN (bottom row) for both AFT (left half) and HFT (right half).

Figure S9: Timesweeper's selection coefficient-inference performance using different neural network architectures. Timesweeper was benchmarked on a 20-timepoint dataset with selection coefficients drawn from a uniform distribution with bounds $[0.00025, 0.25)$ and a starting samppoint generation drawn from a uniform distribution with bounds $[-50, 50)$ post-selection onset. True s versus estimated s values are plotted for both SDN and SSV scenarios with training losses for each network for 1DCNN (top row), 1DCNN with increased parameterization (second row), 2DCNN (third row), and RNN (bottom row) for both AFT (left half) and HFT (right half).

Figure S10: Saliency maps of a trained 2DCNN implementation of Timesweeper. Saliency maps for the 2DCNN architecture trained on the same benchmark data as described in Figure S8, saliency was calculated using examples from the neutral (left column), SSV (central column), and SDN scenarios (right column) for (A) AFT and (B) HFT data formats.

Figure S11: Timesweeper's classification performance on a benchmark dataset using different window sizes. Timesweeper was trained and tested on the dataset described in Figure S1 using window sizes of 1, 3, 11, 51, 101, and 201 SNPs. Precision Recall (PR) curve, ROC curve, normalized and non-normalized confusion matrices, and training losses for AFT (left half) and HFT (right half) are displayed for each window size.

Figure S12: ROC curves for classification performance on datasets with various window sizes. Same as Figure S6 but for the data described in Figure S11.

Figure S13: PR curves for classification performance on datasets with various window sizes. Same as Figure S7 but for the data described in Figure S11.

Figure S14: Timesweeper's regression performance on datasets with various window sizes. Same as Figure S9 but for the data described in Figure S11.

Figure S15: Timesweeper's classification performance on datasets with various sample sizes.

For all simulations 20 timepoints were taken using the same selection coefficient and sampling time parameterization as described in Figure S1. Sample sizes of 1, 2, 5, 10, and 20 individuals were taken at each timepoint. Precision Recall (PR) curve, ROC curve, Normalized and non-normalized confusion matrices, and training losses for AFT (left half) and HFT (right half) are displayed for each sample size.

Figure S16: ROC curves for classification performance on datasets with various sample sizes.

Same as Figure S6 but for the data described in Figure S15.

Figure S17: PR curves for classification performance on datasets with various sample sizes.

Same as Figure S7 but for the data described in Figure S15.

Figure S18: Timesweeper's regression performance on datasets with various sample sizes.

Same as Figure S9 but for the data described in Figure S15.

Figure S19: Timesweeper's classification performance on datasets with various numbers of sampled timepoints.

For all simulations a total of 200 diploid individuals were sampled evenly across the specified number of timepoints (i.e., 200 for the 1-timepoint scenario, 100 per timepoint for the 2-timepoint scenario, etc). Simulation parameterizations for selection coefficient and sampling start time are the same as described in Figure S1. Precision-recall (PR) curve, ROC curve,

normalized and non-normalized confusion matrices, and training losses for AFT (left half) and HFT (right half) are displayed for 1, 2, 5, 10, 20, and 40 timepoints.

Figure S20: ROC curves for classification performance on datasets with various numbers of sampled timepoints. Same as Figure S6 but for the data described in Figure S19.

Figure S21: PR curves for classification performance on datasets with various numbers of sampled timepoints. Same as Figure S7 but for the data described in Figure S19.

Figure S22: Timesweeper's regression performance on datasets with various numbers of sampled timepoints. Same as Figure S9 but for the data described in Figure S19.

Figure S23: Timesweeper's classification performance on datasets with various sampling start times. Simulation parameterizations for selection coefficient are the same as described in Figure S1. Sampling start time was -100 , -50 , 0 , 25 , 100 , or 200 generations post-onset of selection. Precision Recall (PR) curve, ROC curve, normalized and non-normalized confusion matrices, and training losses for AFT (left half) and HFT (right half) are displayed.

Figure S24: ROC curves for classification performance on datasets with various sampling start times. Same as Figure S6 but for the data described in Figure S23.

Figure S25: PR curves for classification performance on datasets with various sampling start times. Same as Figure S7 but for the data described in Figure S23.

Figure S26: Timesweeper's regression performance on datasets with various sampling start times. Same as Figure S9 but for the data described in Figure S23.

Figure S27: Timesweeper's classification performance under various training set sizes. Simulation parameterizations for selection coefficient and sampling start times are the same as described in Figure S1. Data was subsampled to sizes 30,000, 20,000, 10,000, 5,000, 2,000, and 1,000 for each class (neutral, SSV, SDN). Each dataset was partitioned into training, validation, and test data as described in the Methods. Precision Recall (PR) curve, ROC curve, normalized and non-normalized confusion matrices, and training losses for AFT (left half) and HFT (right half) are displayed.

Figure S28: ROC curves for classification performance under various training set sizes. Same as Figure S6 but for the data described in Figure S27.

Figure S29: PR curves for classification performance under various training set sizes. Same as Figure S7 but for the data described in Figure S27.

Figure S30: Timesweeper's regression performance on datasets with varyious training set sizes. Same as Figure S9 but for the data described in Figure S27.

Figure S31: Proportion of positive sweep calls for Timesweeper and competing methods, zoomed in towards the center of the window. The same as Figure S4 but for the data described in Figure 5.

Figure S32: Misspecification classification ROC and PR curves. Pairwise ROC (left) and PR (right) curves for neutral versus sweep (top row) and SDN versus SSV (middle row) tasks. The legend specifies the demographic models used for training and testing (in that order) for each curve. FET's ROC (bottom left) and PR (bottom right) curves are shown for test data simulated under the Bottleneck and OoA models.

Figure S33: Misspecification regression accuracies. Pairwise plots of true versus estimated s for each model trained (rows) and tested on (columns) for both SSV (left) and SDN (right) scenario selection coefficients.

Figure S34: Example and average Timesweeper inputs from simulated training data for the *D. simulans* evolve-and-resequence dataset. (A) – (C) show individual example inputs and (D) shows average inputs of the Neutral and SSV classes for the AFT method where data were simulated for the *D. simulans* evolve-and-resequence dataset as described in the Methods.

Figure S35: Timesweeper's classification performance on simulated test data for the *D. simulans* evolve-and-resequence dataset. (A) – (D) Confusion matrix, ROC curve, training trajectory, and precision-recall (PR) curve showing the performance of the AFT method on simulated *D. simulans* data as described in the Methods.

Figure S36: Replication of top hits for Timesweeper and FET in the *D. simulans* evolve-and-resequence dataset. A datapoint was counted as replicated sweep if it A) was in the top 100 most confidently-detected sweeps (ranked by Timesweeper sweep probability or by FET 1-pvalue) and B) it occurred in a 100kb window containing at least one outlier SNP (i.e. top 1% of all tested SNPs for a replicate) detected in the other replicate.

Table S1: Summaries of Timesweeper's classification results on *D. simulans* evolve-and-resequence dataset binned by Fisher's Exact Test p -value.

Table S2: Summaries of Fisher's Exact Test results on *D. simulans* evolve-and-resequence dataset binned by Timesweeper's sweep probability score.