Inferring parameters of selective sweeps through supervised learning

Ian V. Caldas[1], Andrew G. Clark[1,2], Philipp W. Messer[1]

[1]Department of Computational Biology, Cornell University, Ithaca, NY, USA
[2]Department of Molecular Biology and Genetics, Cornell University, Ithaca, NY, USA

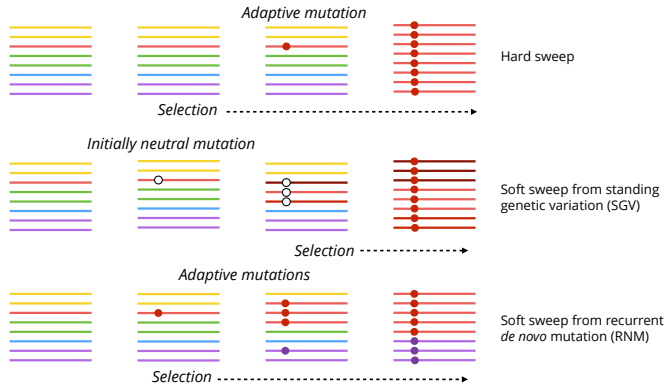
ivc2@cornell.edu @ianvcaldas

Introduction

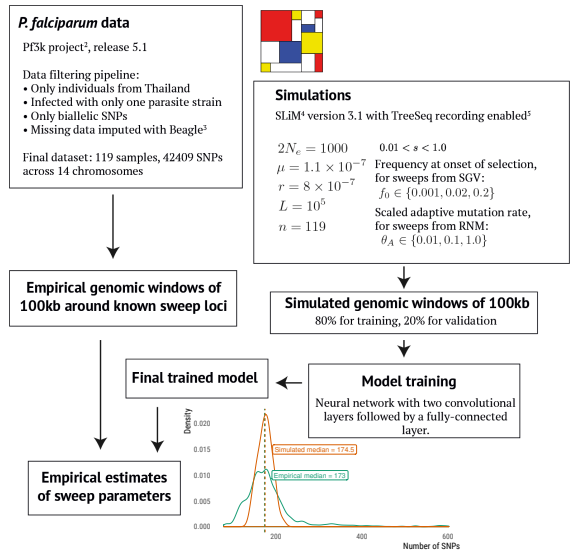
In a selective sweep, an adaptive allele quickly rises in frequency, producing recognizable patterns in surrounding genetic diversity. Deep learning is an efficient way to use this signal to detect sweeps.
Is the pattern of surrounding genetic diversity informative about the evolutionary history of a sweep? Can we estimate relevant sweep parameters from it?

Supervised learning inference methods are useful when the problem is hard to analyze mathematically, like selective sweeps under complex population histories. But only simulations can provide the necessary amount of training data for evolution.
How do we simulate realistic sweep scenarios? How do we create simulations that are relevant to empirical applications?

The parasite *Plasmodium falciparum* causes malaria and Thai populations have evolved resistance to all known classes of antimalarial drugs. The loci responsible for resistance are recently finished hard and soft sweeps, well characterized in the literature[1].
How well does a model to infer parameters of selective sweeps perform on these positive control loci in *P. falciparum*?



Methods

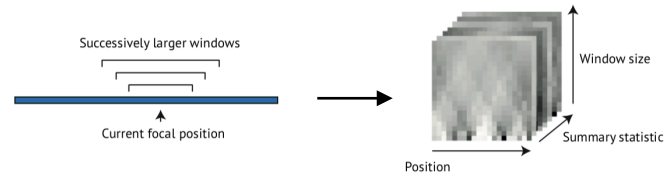


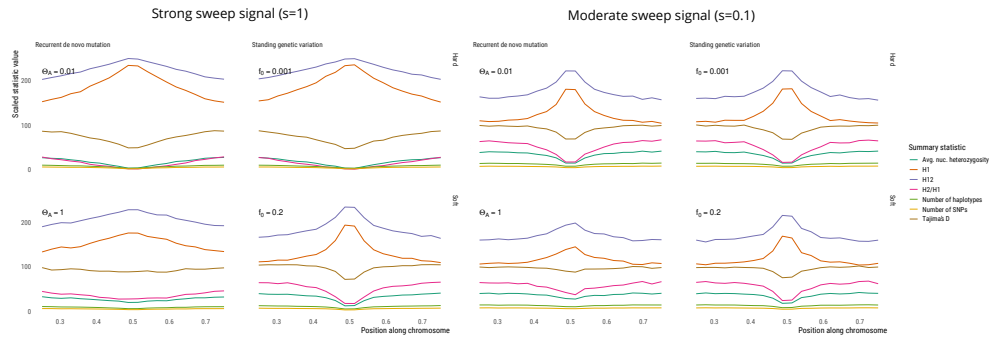
To estimate parameters of sweeps with machine learning, we simulate sweep scenarios, calculate summary statistics on the simulated populations, and train a model to estimate the desired parameter. The model's power is validated on simulated data and can then be applied to empirical data.

SLiM enables forward simulation of arbitrary evolutionary scenarios. We implement simulations of selective sweeps with tree sequence recording and a coalescent burn-in period to allow very fast computing times.

The simulated data must fit the empirical data as much as possible. We take well established mutation and recombination rates from the literature, then adjust population size until the distribution of SNP number per simulated window matches the empirical distribution.

Within each genomic window, we do a moving subwindow analysis, calculating summary statistics across the chromosome for each subwindow size. The result is a "brick" of data. A convolutional neural network is able to leverage the full correlation structure of genomic position, summary statistic and subwindow sizes from this representation.

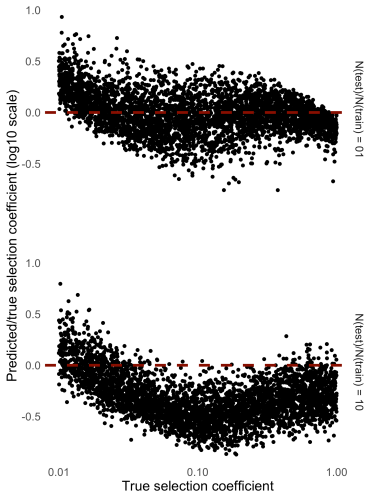




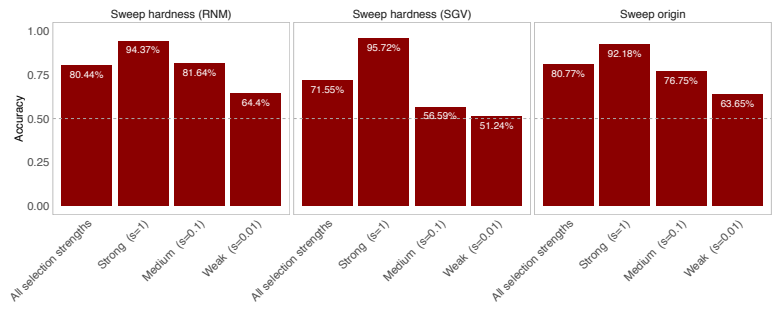
Results

We fit one neural network to estimate each of the following sweep parameters:

How strong was the sweep?



Was it a hard or a soft sweep?



Did it happen from standing genetic variation or recurrent *de novo* mutation?

The model performs well in estimating selection coefficient, and estimates sweep softness and origin with an accuracy above 90% if sweeps are strong.

If the true population size is 10 times larger than the one the model was trained on, the model underestimates the true selection coefficient, consistent with a signature that depends on sweep time.

Estimates of parameters of known sweep loci from *P. falciparum* match what we expect from the literature.

Statistics averaged over all simulations reveal that sweep origin and strength produce distinguishable patterns. However, no single summary statistic drives most of the method's power.

Misspecifying demographic history when training the model can decrease the method's power, depending on how strong the misspecification is.

The method can be applied to a sample of incomplete sweeps. Low sweep frequency reduces power to estimate selection strength, but estimates of softness and sweep origin are robust to incomplete sweeps.

		dhfr	*mdr1*	*crt*	*dhps*	*gch1*
Chromosome; Coordinates (kb)		4; 748-750	5; 958-962	7; 403-406	8; 548-551	12; 974-976
Presumed from literature		Hard	Soft from RNM	Hard	Soft from RNM	Soft from RNM
Softness	P(Hard)	**0.61**	0.03	**0.96**	0.20	0.09
	P(Soft)	0.39	**0.97**	0.04	**0.80**	**0.91**
Sel. coefficient	P(s=1.0)	0.04	0.26	0.10	**1.00**	**1.00**
	P(s=0.1)	**0.79**	0.36	**0.88**	0.00	0.00
	P(s=0.01)	0.17	0.38	0.02	0.00	0.00
Origin	P(RNM)		**0.96**		**0.90**	**0.90**
	P(SGV)		0.04		0.10	0.10

Discussion

Fitting simulated training data to empirical data is the most important step in using supervised learning in population genetics. SLiM is the ideal tool for this, as it allows simulation of sweep scenarios under realistic, biological model, as well as partial sweeps. As a forward simulation, SLiM provides great flexibility in modeling realistic populations, including complex population structure and demography. It will be an important tool for generating training datasets for applications of machine learning in population genetics. Further progress needs to be made on how to formally fit simulations to empirical populations; most current studies validate their performance on human data, assuming a "good-enough" fit with established demographic hypotheses that aren't available for other organisms.

If the model is trained on appropriate simulated data, as discussed, supervised learning model has great power to infer population parameters. Clearer signatures come, predictably, from stronger and harder sweeps. Supervised learning methods, including deep learning methods, have the potential to inform and generate hypotheses about the evolution of drug resistance loci in pathogens, which often undergo selective sweeps. We illustrate that by applying our model to *Plasmodium falciparum,* where it predicts parameters for known sweeps that are in line with the expectations from previous molecular studies of these loci.

References

1. Anderson T., Nkhoma S., Ecker A. and Fidock D. (2011) How can we identify parasite genes that underlie antimalarial drug resistance? *Pharmacogenomics* 12(1), pp.59-85.
2. MalariaGEN *Plasmodium falciparum* Community Project. (2016) Genomic epidemiology of artemisinin resistant malaria. *eLife* 5:eo8714.
3. Browning S.R. and Browning B.L. (2007) Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *The American Journal of Human Genetics* 81(5), pp.1084-1097.
4. Haller B.C. and Messer P.W. (2019) SLiM 3: Forward genetic simulations beyond the Wright-Fisher model. *Molecular Biology and Evolution* 36(3), pp. 632-637.
5. Haller B.C., Galloway J., Kelleher J., Messer P.W. and Ralph P.L. (2019) Tree-sequence recording in SLiM opens new horizons for forward-time simulation of whole genomes. *Molecular Ecology Resources* 19(2), pp.552-566.