A likelihood approach for uncovering selective sweep signatures from haplotype data

Alexandre M. Harris[1,2] and Michael DeGiorgio[3]

[1]*Department of Biology, Pennsylvania State University, University Park, PA 16802*

[2]*MCIBS Graduate Program at the Huck Institutes of the Life Sciences, Pennsylvania State University , University Park, PA 16802*

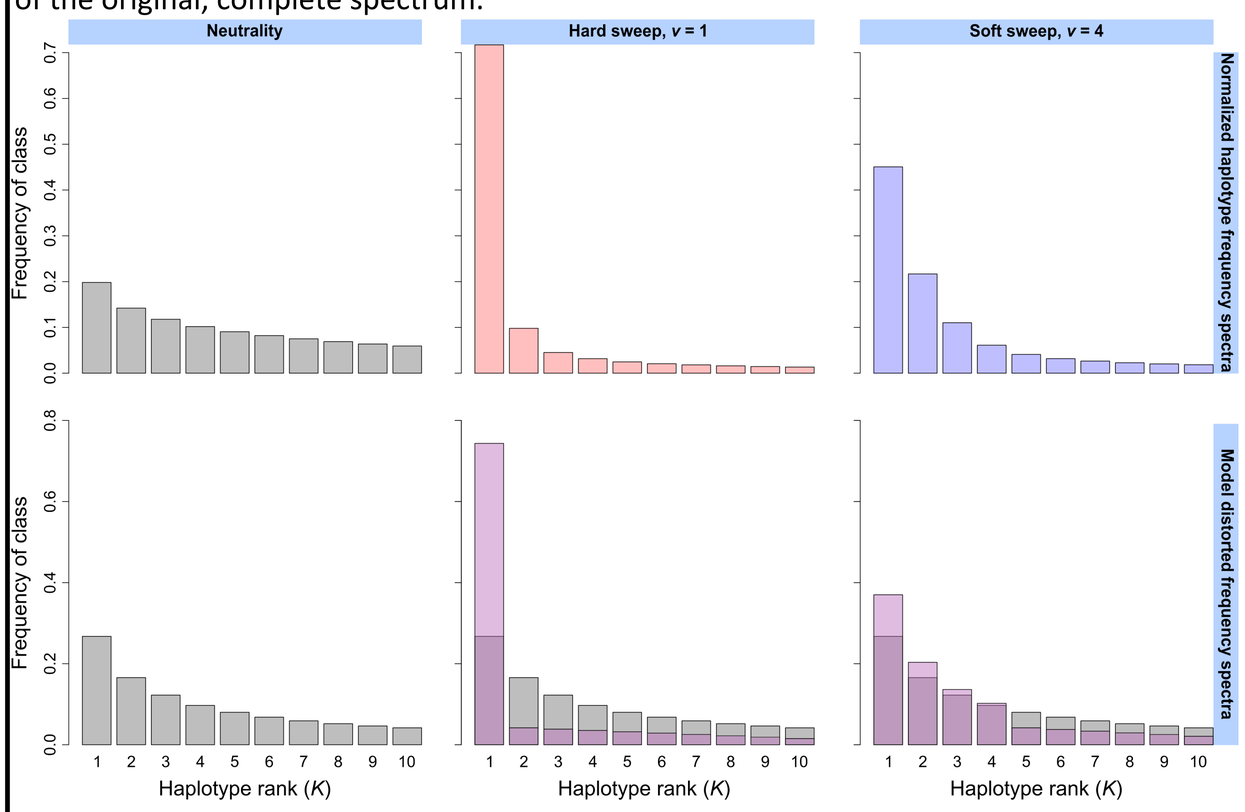
[3]*Department of Computer and Electrical Engineering and Computer Science, Florida Atlantic University, Boca Raton, FL 33431*

Abstract

Selective sweeps are frequent and varied signatures in the genomes of natural populations, and detecting them is consequently important in understanding mechanisms of adaptation by natural selection. Following a selective sweep, haplotypic diversity surrounding the site under selection decreases, and this deviation from the background pattern of variation can be applied to identify sweeps. Multiple methods exist to locate selective sweeps in the genome from haplotype data, but none leverage the power of a model-based approach to make their inference. Here, we propose a likelihood ratio test statistic *T* to probe whole genome polymorphism datasets for selective sweep signatures. Our framework uses a simple but powerful model of haplotype frequency spectrum distortion to find sweeps and additionally make an inference on the number of presently sweeping haplotypes in a population. We found that the *T* statistic is suitable for detecting both hard and soft sweeps across a variety of demographic models, selection strengths, and ages of the beneficial allele. Accordingly, we applied the *T* statistic to variant calls from European and sub-Saharan African human populations, yielding primarily literature-supported candidates, including *LCT*, *RSPH3*, and *ZNF211* in CEU, *SYT1*, *RGS18*, and *NNT* in YRI, and *HLA* genes in both populations. We also searched for sweep signatures in *Drosophila melanogaster*, finding expected candidates at *Ace*, *Uhg1*, and *Pimet*. Finally, we provide open-source software to compute the *T* statistic and the inferred number of presently sweeping haplotypes from whole-genome data.

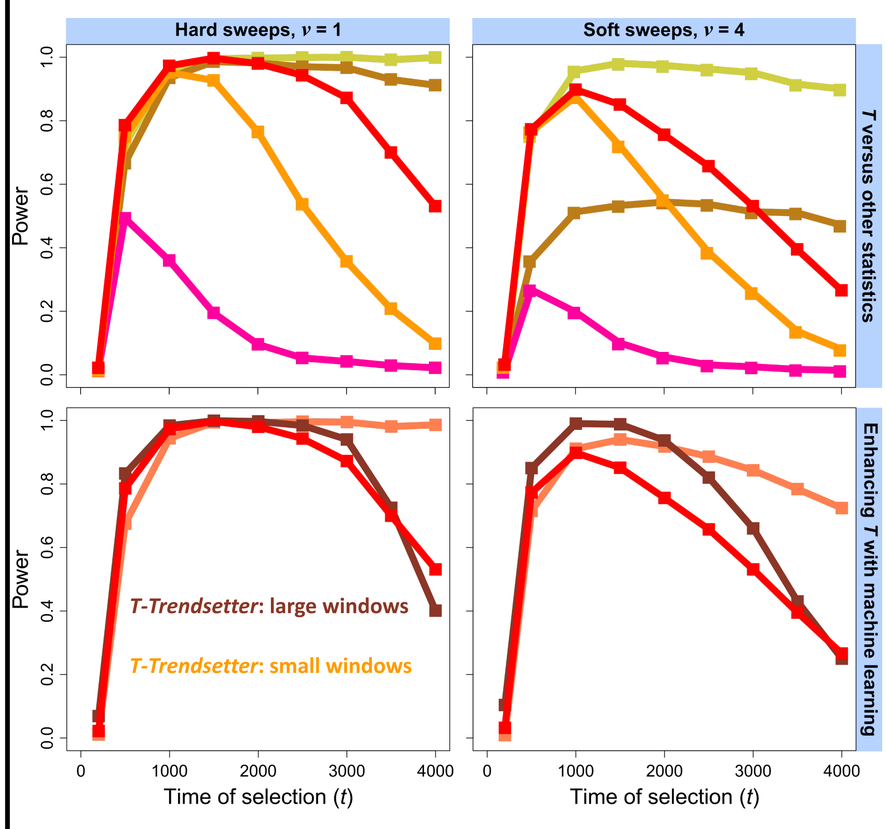
Formulation of the *T* statistic

Selective sweeps (red, blue) cause a distortion in the haplotype frequency spectrum of a population at the site under selection relative to neutrality (gray). We capture this distortion using a truncated, normalized haplotype frequency spectrum, which still preserves the shape of the original, complete spectrum.



The *T* statistic quantifies the likelihood that a haplotype frequency spectrum is consistent with a model of a selective sweep. To generate selective sweep models, we distort the neutral spectrum by adding weight to the sweeping classes at the expense of the non-sweeping classes (purple). Thus, the *T* statistic provides a likelihood and a most likely model.

The *T* statistic has high power to detect recent sweeps



The *T* statistic has high power to detect recent hard and soft sweeps occurring within 2000 generations of sampling. Power is greater for our novel ***T statistic*** than for comparable methods **H12**, ***SweepFinder2***, and ***nS$_L$***. The machine learning method ***Trendsetter*** is more powerful, however.

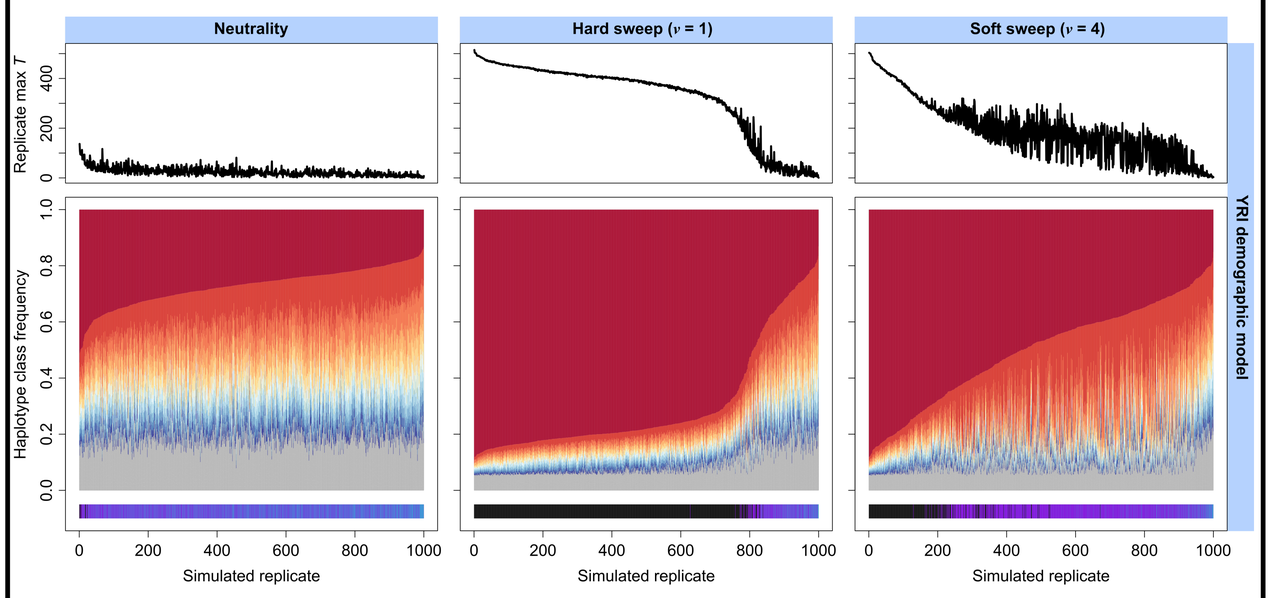
We can enhance the power of the *T* statistic by incorporating it into *Trendsetter*, which uses the spatial distribution of *T* to identify sites under selection. Our ***T-Trendsetter*** approach is more powerful than *T*.

Using the *T* statistic to classify sweeps as hard or soft

Because the *T* statistic optimizes over all possible sweep models, it can be used not only to identify a putative site under selection, but also to classify that site as a hard or soft sweep based on the number of presently-sweeping haplotypes.

We achieve this by assigning a value of $\hat{m}$ to each region under analysis, where larger values of $\hat{m}$ represent a greater number of sweeping haplotypes. Thus, $\hat{m} = 1$ represents a hard sweep, while $\hat{m} \geq 2$ is a soft sweep.

In the figure below, we show results from 1000 replicate simulations of neutrality, hard sweeps, and soft sweeps, arranged in decreasing order of the most frequent haplotype frequency. Each replicate is a slice in the larger plot. In the top row, we include the *T* statistic for the replicate. In the bottom is the haplotype frequency spectrum for 20 haplotypes; the first 10 haplotypes are colored from red (most frequent) to blue (10th-most frequent). The coloration underneath each frequency spectrum indicates the assigned $\hat{m}$ for the replicate. Black is a hard sweep while purples and blues are soft sweeps.



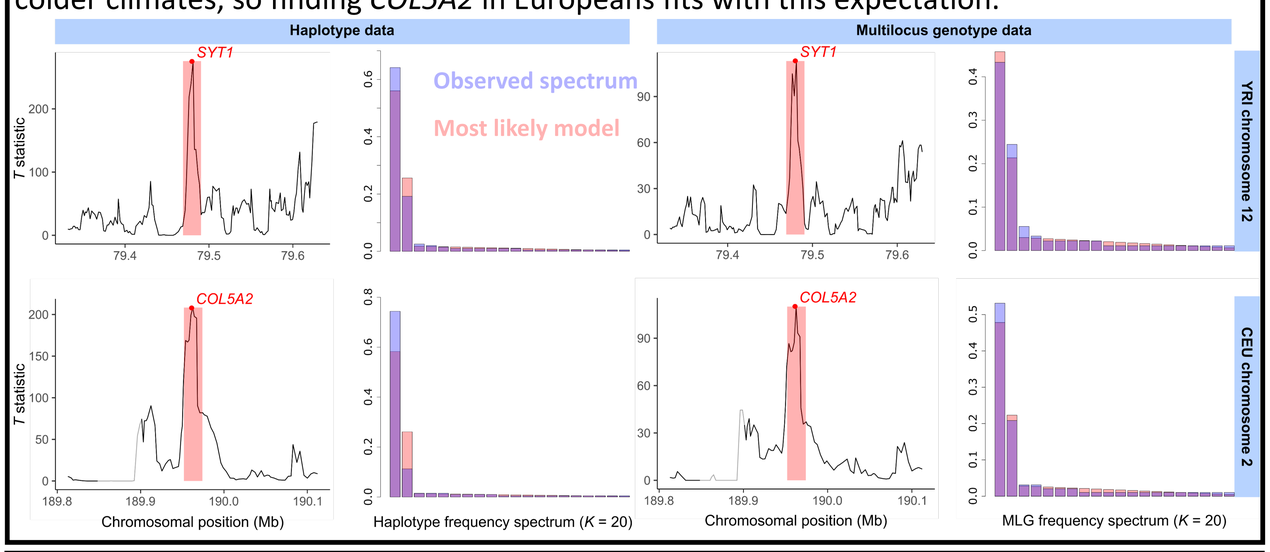
The vast majority of simulated hard sweeps (center) are assigned as hard (black coloration), while the vast majority of simulated soft sweeps are assigned as soft (blue and purple).

Application to human empirical data

We applied the *T* statistic to human whole-genome polymorphism data from the 1000 Genomes Project Phase 3 dataset (2015), identifying and classifying literature-supported and novel sweep candidates. We used not only phased haplotype data (left) for inference, but unphased multilocus genotype data (right) as well, finding that results between the two data types are generally congruent. Thus, phasing is not required to use the *T* statistic.

In the sub-Saharan African YRI population, we identified *SYT1* as a soft sweep ($\hat{m}$) candidate on chromosome 12. This gene has extensive literature support to suggest it is a target of selection in humans, possibly involved in resistance to foodborne bacterial infection.

COL5A2, found as a candidate soft sweep ($\hat{m}$) in the CEU population of western European descent, is a collagen gene. Other genes in this family have been implicated in adaptation to colder climates, so finding *COL5A2* in Europeans fits with this expectation.



Conclusions

The *T* statistic is the first and only method of its kind, using haplotypes in a likelihood framework to identify and classify selective sweeps.

Our approach is unique, as it can classify sweeps as hard or soft without additional analyses.

We can apply the *T* statistic to unphased data, meaning that we can make inferences for non-model organisms, for which phased haplotypes may not be available.

Support and funding

This work was funded by National Institutes of Health grant R35-GM128590, by National Science Foundation grants DEB-1753489, DEB-1949268, and BCS-2001063, and by the Alfred P. Sloan Foundation. Computations for this research were performed on the Pennsylvania State University's Institute for Computational and Data Sciences Advanced CyberInfrastructure (ICS-ACI).