1 Introduction

Many individual statistics have been designed to detect genomic signatures of hard selective sweeps, which fall into 3 categories: changes in the site frequency spectrum (SFS), long haplotype blocks, and population differentiation. By combining multiple statistics into one, composite tests can increase the power to detect such sweeps by exploiting multiple genomic signals at once. Existing composite tests include CMS [cite], and the Boosting classifier introduced by Lin *et al.* [cite].

An issue that arises with composite tests is that of missing data, which has been largely ignored until now. Specifically, composite tests are based on a selection of statistics, some of which may be undefined at any given locus. The statistics that are most often undefined in our study are iHS and the related ΔiHH, which have a minor allele frequency threshold of 0.05, since they rely on scores calculated for both derived and ancestral alleles [cite Voight, Grossman]. Furthermore, iHS, ΔiHH, and other long-range haplotype statistics like XP-EHH [cite] rely on the calculation of extended haplotype homozygocity (EHH) at long distances from the locus of interest, which can be interrupted by the ends of chromosomes or sequenced regions. [Something about Fst?] In data from the 1000 Genomes project, we found that XX% of loci had at least one undefined statistic, and YY% had two or more. Existing composite methods either do not address this problem, or apply heuristic compensation mechanisms that can introduce artifacts into the method. The Naive Bayes classification framework that we introduce here for sweep detection naturally accounts for missing data in a way that other methods cannot, simply by computing the probability that a locus has undergone a sweep, given the set of statistics that are available at that locus.

The Naive Bayes framework has many advantages over existing methods in addition to its natural treatment of missing data. One is interpretability; instead of relying on a threshold for classifying loci as under selection or neutrally evolving, it returns a probability for each scenario, which can be used . In addition, the framework naturally extends to classification with more than two scenarios. In this paper, we use this feature to investigate the age of selective sweeps, and to distinguish between sweeps and bottlenecks, which are known to produce similar genomic signatures. The framework is general, however, and could be used to investigate other sweep features such as selective strength, or different types of selection, including background selection and selection from standing variation ("soft sweeps").

- previous methods
  - CMS
  - SweepFinder
  - SweeDetector

2 Methods

2.1 Naive Bayes Framework

- calculation of true posterior:
  $$P(\text{sweep}|s_1, ..., s_n) = \frac{\pi \prod_{i=1}^{n} P(s_i|\text{sweep})}{\pi \prod_{i=1}^{n} P(s_i|\text{sweep}) + (1-\pi) \prod_{i=1}^{n} P(s_i|\text{neutral})}$$
- advantages: missing data, probabilistic interpretation, multi-class classification

3 Results

3.1 Simulations

- ROC curves comparing NB and CMS
- localization plots
- discussion of missing data issue – NB outperforms in particular realms
- Goal: also compare SweepFinder, SweeDetector

3.2 Known targets

- Show that we can identify previously-discovered targets of selection (LCT, etc...)
- Compare to CMS scores
- Genome-wide scan

3.3 Timing sweeps(/Bottlenecks?)

- Goal: show we can distinguish between ancient and recent sweep times
- will need SFS measures for this to work

3.4 AODEs

- Discussion of independence assumption
- Goal: show that AODEs increase the power on simulations (can also look at known targets)

3.5 Robustness to demographic models

- Goal: show that mis-specification of demographic model doesn't change things too much (or alternatively, make a statement about what effect certain kinds of mis-specification might have)

4 Application to San Haplotypes

- Goal: what has to change when considering exome data?
- Goal: infer demographic model of San ($\partial a \partial i$), information from Brenna
- Goal: run NB (with AODE?) on San haplotypes (with YRI as outgroup?), identify genome-wide targets
- Bonus: interesting functional story

5 Discussion

- Possible extensions of the framework: soft sweeps

5.1 Informative statistics

- Goal: draw conclusions about which statistics are informative for which situations (i.e. SFS measures are helpful for distinguishing sweep time)
- Can look at this by making classifiers based on single or joint distributions of the component statistics