On detecting selective sweeps using single genomes

Priyanka Sinha
Ecole Polytechnique Federale de Lausanne

Let us know how access to this document benefits you.
Follow this and additional works at: https://escholarship.umassmed.edu/gsbs_sp

Part of the Bioinformatics Commons, Genetics and Genomics Commons, and the Medicine and Health Sciences Commons

Repository Citation

This material is brought to you by eScholarship@UMMS. It has been accepted for inclusion in GSBS Student Publications by an authorized administrator of eScholarship@UMMS. For more information, please contact Lisa.Palmer@umassmed.edu.
Identifying the genetic basis of human adaptation has remained a central focal point of modern population genetics. One major area of interest has been the use of polymorphism data to detect so-called “footprints” of selective sweeps – patterns produced as a beneficial mutation arises and rapidly fixes in the population. Based on numerous simulation studies and power analyses, the necessary sample size for achieving appreciable power has been shown to vary from a few individuals to a few dozen, depending on the test statistic. And yet, the sequencing of multiple copies of a single region, or of multiple genomes as is now often the case, incurs considerable cost. Enard et al. (2010) have recently proposed a method to identify patterns of selective sweeps using a single genome – and apply this approach to human and non-human primates (chimpanzee, orangutan, and macaque). They employ essentially a modification of the Hudson, Kreitman, and Aguade test – using heterozygous single nucleotide polymorphisms from single individuals, and divergence data from two closely related species (human–chimpanzee, human–orangutan, and human–macaque). Given the potential importance of this finding, we here investigate the properties of this statistic. We demonstrate through simulation that this approach is neither robust to demography nor background selection; nor is it robust to variable recombination rates.

**Keywords:** selective sweeps, demography, adaptation, statistical inference

The authors use a method inspired by the Hudson, Kreitman, and Aguade (HKA) test (Hudson et al., 1987). This statistic employs intra-specific polymorphism data and inter-specific divergence data to test for deviations from the Standard Neutral model. Enard et al. (2010) derive their polymorphism data from a single heterozygous individual, and the divergence information from comparison with a single chimp, orangutan, or macaque genome. This approach is termed “K—estimation,” where the corrected level of heterozygosity is indicated by the value of a statistic $K (0 \leq K \leq 1)$.

They assess two single human individuals, the Venter and Watson genomes. They use a co-occurrence test to find the difference between the expected and observed values for the $K$-statistic among orthologous genes. Numerous other recent studies have sought to identify and describe genome-wide positive selection in humans using various aspects of polymorphism and divergence data (e.g., Sabeti et al., 2006, 2007; Voight et al., 2006; Williamson et al., 2007) – mostly using single nucleotide polymorphisms (SNP) data from HapMap (Frazer et al., 2007) or Perlegen (Hinds et al., 2005). Enard et al. (2010) find essentially no overlap with the candidate gene sets generated using these other methodologies. Enard et al. (2010) perform forward-time population simulations to validate their method using fregene. However, they only consider two models – an equilibrium model and a single bottleneck model. Here we further evaluate their method and assertions using simulations under several models of positive and negative selection, variable recombination and mutation, and a variety of non-equilibrium models.
A possible complication with the $K$-statistic is its ability to differentiate between patterns produced by a selective sweep, relative to other neutral processes. Given the essence of the statistical design, significant regions will be those that have reduced heterozygosity relative to flanking genomic sequence. Thus, we investigate whether this approach is robust to: (1) non-equilibrium demographic models, which may greatly increase the variance in heterozygosity across the genome, (2) varying levels of selective constraint across the genome, and (3) simple variation in recombination rate.

**MATERIALS AND METHODS**

**SIMULATIONS**

All simulations were performed using the program SFS_CODE (Hernandez, 2008). This is a generalized Wright–Fisher forward-population genetic simulation for finite-site mutation models with selection, recombination, and demography. The program and documentation are available for download at: http://sfscode.sourceforge.net/SFS_CODE/SFS_CODE_home/SFS_CODE_home.html

The parameters used are human specific and rescaled for computational efficiency. The mutation rate ($\mu$) is $2.35 \times 10^{-8}$ per site per generation, considering human–chimp divergence to be $\sim 1.13\%$, divergence time is $\sim 6$ mya, and a generation time of 25 years (Gutenkunst et al., 2009). For expediency, the effective population size, $N_e$, is 500. To account for the estimated $N_e = 10,000$ for humans, a rescaling factor of 20 was used. Thus, $\theta = 0.00094$ for all simulations. Similarly, the scaled recombination rate $4N\rho = \rho = 0.00074$ (Nielsen et al., 2005). The selection coefficient, $s$, is evaluated at 0.1, 0.01, and 0.001. All simulations were conducted both in the presence and absence of recombination. Additionally, models of recurrent positive and negative selection were modeled with a fraction 0.01, 0.001, and 0.0001 of sites under selection.

Population bottlenecks are modeled in the following way: a population of constant size $N$ is reduced to size $N_b$ at time $t_b$ (in units of $4N$ generations) in the past and then exponentially increases back to size $N$. Population bottlenecks are simulated for various times since the reduction ($t_b = 0.1, 0.54$, and 1, in $4N$ generations), and severities ($0.02$, $0.1$, and $0.722$).

Enard et al. (2010) chose the size of test region ($L$) depending on the level of heterozygosity across the genome, and thus it varies by species. We simulate a test region of 100 kb with adjacent 2000 kb genomic flanking regions. For this test region, a ratio $r_l$ is calculated, which is the ratio of polymorphism to divergence for the $L$ region. A similar ratio $r_g$ is calculated for the adjacent genomic region ($G$) and then a final ratio $R_{obs}$ of $r_{l/g}$. If $R_{obs}$ is less than 1, then there is said to be a local reduction in heterozygosity. Similarly, a ratio $R$ is computed for 5,000 additional windows of size $q$, that are randomly sampled within $G$, but at a distance at least five times $q$ from $L$. The $R_{obs}$ for test region is ranked among the $R$ values for the adjacent regions. $K$ is the proportion of random windows with $R$ lower than $R_{obs}$.

$K$ values $< 0.05$ are statistically significant, and thus reject the model (i.e., are consistent with positive selection). Simulations under models of positive selection were performed in order to characterize the true positive rate; while a variety of simulations under alternative models characterize the false positive rate. These two measures thus describe the performance of the $K$-statistic.

**RESULTS**

**THE STANDARD NEUTRAL MODEL WITH VARIABLE RATES OF RECOMBINATION**

To account for the effect of recombination rate on patterns of selective sweeps, we simulated with and without recombination. Under the equilibrium neutral model (for $\theta = 0.00094$), the false positive rate is $0.167$ ($\rho = 0.00074$) and $0.313$ ($\rho = 0$), with and without recombination respectively (Table 1). The performance under this Standard Neutral model (where the standard false positive rate may be expected to be 0.05), should be taken as the baseline performance of this statistic. Considering the known genomic variation in recombination rate (Broman et al., 1998; Lenzi et al., 2009), this result suggests a strong bias toward false positive signatures being detected in low recombination rate regions (Figure 1).

**NON-EQUILIBRIUM NEUTRAL MODELS, THE EFFECT OF POPULATION SIZE CHANGE**

A tremendous amount of literature has focused upon the ability of test statistics to distinguish between patterns produced under selective vs. demographic models – and distinguishing these models has often proven difficult (see review of Kelley et al., 2006; Thornton et al., 2007). Here, we examine the performance of the $K$-statistic under a variety of bottleneck models of varying timing and severity – models of the sort that are commonly estimated for non-African human populations (e.g., Gutenkunst et al., 2009).

For population bottlenecks the false positive rate ranges from 0.12 to 0.153 in presence of recombination and 0.311–0.344 in...
The poor performance of this statistic is consistent with the poor overlap with other human genomic scans (Carlson et al., 2005; Voight et al., 2006; Tang et al., 2007; Williamson et al., 2007; Pickrell et al., 2009), with the fraction of overlapping genes varying between 6 and 15%. Thus, despite the novelty of the initial claim, these results strongly suggest that the single genome approach to sweep detection is not robust to any local diversity-reducing model. Thus, the variety of polymorphism based test statistics (see review of Nielsen, 2005; Thornton et al., 2007) for distinguishing the above models using patterns in linkage disequilibrium and the site frequency spectrum appear to be the most promising avenue forward – continuing to necessitate the sampling and sequencing of consistently sampled population data.
REFERENCES


Sinha et al. Single vs. multiple genome inference 719


**Conflict of Interest Statement:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 20 July 2011; paper pending published: 06 September 2011; accepted: 13 November 2011; published online: 01 December 2011.


This article was submitted to Frontiers in Evolutionary and Population Genetics, a specialty of Frontiers in Genetics.

Copyright © 2011 Sinha, Dincer, Virgil, Xu, Poh and Jensen. This is an open-access article distributed under the terms of the Creative Commons Attribution Non Commercial License, which permits non-commercial use, distribution, and reproduction in other forums, provided the original authors and source are credited.