Bioinformatics, 32(12), 2016, 1895–1897
doi: 10.1093/bioinformatics/btw051
Advance Access Publication Date: 15 February 2016
Applications Note

OXFORD

Genetics and population analysis

SweepFinder2: increased sensitivity, robustness and flexibility

Michael DeGiorgio[1,*], Christian D. Huber[2], Melissa J. Hubisz[3], Ines Hellmann[4] and Rasmus Nielsen[5]

[1]Department of Biology and Institute for CyberScience, Pennsylvania State University, University Park, PA, USA, [2]Department of Ecology and Evolutionary Biology, University of California, Los Angeles, CA, USA, [3]Department of Biological Statistics and Computational Biology, Cornell University, Ithaca, NY, USA, [4]Department Biologie II, Ludwig-Maximilians-Universität München, Planegg-Martinsried, Germany and [5]Department of Integrative Biology, University of California, Berkeley, CA, USA

*To whom correspondence should be addressed.

Associate Editor: Janet Kelso

Received on 28 May 2015; revised on 6 January 2016; accepted on 19 January 2016

Abstract

Summary: SweepFinder is a widely used program that implements a powerful likelihood-based method for detecting recent positive selection, or selective sweeps. Here, we present SweepFinder2, an extension of SweepFinder with increased sensitivity and robustness to the confounding effects of mutation rate variation and background selection. Moreover, SweepFinder2 has increased flexibility that enables the user to specify test sites, set the distance between test sites and utilize a recombination map.

Availability and implementation: SweepFinder2 is a freely-available (www.personal.psu.edu/mxd60/sf2.html) software package that is written in C and can be run from a Unix command line.

Contact: mxd60@psu.edu

1 Introduction

Polymorphism frequency spectra provide sensitive statistics for identifying signatures of positive selection. SweepFinder (Nielsen *et al.*, 2005) is a widely used program (Li *et al.*, 2011; Pavlidis *et al.*, 2010; Svetec *et al.*, 2009; Williamson *et al.*, 2007) that uses an empirical background frequency spectrum for identifying genomic sites affected by recent positive selection. Specifically, SweepFinder performs a composite likelihood ratio test for positive selection (Kim and Stephan, 2002), in which the likelihood of the null hypothesis is calculated from the neutral (or genome-wide) frequency spectrum, and the likelihood of the alternative hypothesis is calculated from a model in which the neutral spectrum was altered by a recent selective sweep.

Footprints of positive selection can be confounded by other evolutionary forces. One important confounding factor that is rarely considered in the studies of positive selection is background selection, which is a loss of neutral variation due to purging of linked deleterious alleles by negative selection (Charlesworth, 2012; Charlesworth *et al.*, 1993; Hudson and Kaplan, 1995a). Recent studies have shown that background selection is ubiquitous in humans (Lohmueller *et al.*, 2011; McVicker *et al.*, 2009; Wilson Sayres *et al.*, 2014), with estimates of mean reductions in genetic diversity due to background selection ranging from 19 to 26% and 12 to 40% on autosomes and the X chromosome, respectively (McVicker *et al.*, 2009). Thus, the influence of background selection on genetic diversity has important ramifications for making inferences about past adaptive processes from patterns of diversity. In particular, when a beneficial allele is carried to fixation by positive selection, there is a substantial decrease in diversity locally in the genome and a reduction in diversity relative to divergence with other species, both of which can span megabases in length (Maynard Smith and Haigh, 1974). Background selection can similarly affect diversity levels (Akashi *et al.* 2012; Boyko *et al.*, 2008; Charlesworth, 2012; Charlesworth *et al.*, 1993 1995; Hudson and Kaplan, 1995a,b; McVean and Charlesworth, 2000; Nordborg *et al.*, 1996), particularly in regions of low recombination.

© The Author 2016. Published by Oxford University Press. All rights reserved. For Permissions, please e-mail: journals.permissions@oup.com

Downloaded from http://bioinformatics.oxfordjournals.org/ at University of California, Los Angeles on August 30, 2016

Because patterns of background selection can mimic those of positive selection, methods for identifying signatures of positive selection that are based on diversity reduction alone may be confounded by background selection. These conflicting signals have likely contributed to a current debate of the role of recent positive selection in shaping the landscape of human genetic variation (Akey, 2009; Enard *et al.*, 2014; Granka *et al.*, 2012; Hawks *et al.*, 2007, Hernandez *et al.*, 2011; Lohmueller *et al.*, 2011; Williamson *et al.*, 2007), emphasizing the need for methods that can identify sweeps while accounting for background selection. Further, because the effects of background selection may be pronounced in regions of low recombination, it is important that methods jointly account for background selection and local recombination rate, which is also expected to affect patterns of a selective sweep.

2 SweepFinder2

SweepFinder2, which is based on the statistical framework of SweepFinder (Nielsen *et al.*, 2005), jointly accounts for background selection and local recombination rate by modeling the effect of background selection on genetic diversity. It does this by modifying the neutral derived frequency spectrum with respect to *B*-values and by including invariant sites (specifically substitutions), as introduced by Huber *et al.* (2015). *B*-values range from 0 to 1 and are proportional to local reductions in genetic diversity or effective population size due to background selection. McVicker *et al.* (2009) provide a method for inferring *B*-values using comparative data, thereby providing an opportunity for separating background selection from the effect of selective sweeps inferred from within-population polymorphism data. Because background selection reduces diversity by a factor *B*, we multiply each polymorphic frequency class (i.e. allele counts 1, 2, …, $n-1$ in a sample of n) by *B*, as shown in Figure 1A (Huber *et al.*, 2015). Furthermore, because background selection affects diversity relative to divergence with another species, we scale the fixed difference class (i.e. allele count n), and then renormalize the frequency spectrum to sum to 1 (Fig. 1A). Note that this effect depends on the current and ancestral population sizes, as well as on the divergence time in generations between the pair of species. Further, our correction is a first-order approximation, as background selection can alter the frequency spectrum in other ways (e.g. Charlesworth *et al.*, 1993, 1995; Nicolaisen and Desai, 2013; Seger *et al.*, 2010). This point is exemplified by empirical results indicating that diversity reduction in regions with low recombination rates is less than expected under simple models of background selection (e.g. Kaiser and Charlesworth, 2009), though our approach is conservative under this scenario. Figure 1B illustrates how this procedure modifies the neutral frequency spectrum, such that diversity decreases and the proportion of fixed differences increases with increasing effect of background selection (i.e. decreasing *B*-value).

Our method detects selective sweeps in regions under background selection by scaling the neutral frequency spectrum locally in the genome by estimated *B*-values (Fig. 1), using the scaled spectrum in the null hypothesis, and the spectrum under a model of a selective sweep (accounting for local recombination rate) in the alternative hypothesis (Huber *et al.*, 2015). Regions with reductions in diversity and low *B*-values show little evidence of selective sweeps under this test because frequency spectra under the null and alternative hypotheses are similar (Fig. 1C). However, regions with reductions in diversity and relatively high *B*-values may provide evidence of recent selective sweeps, because frequency spectra under the alternative hypothesis will exhibit lower diversity than those under the null

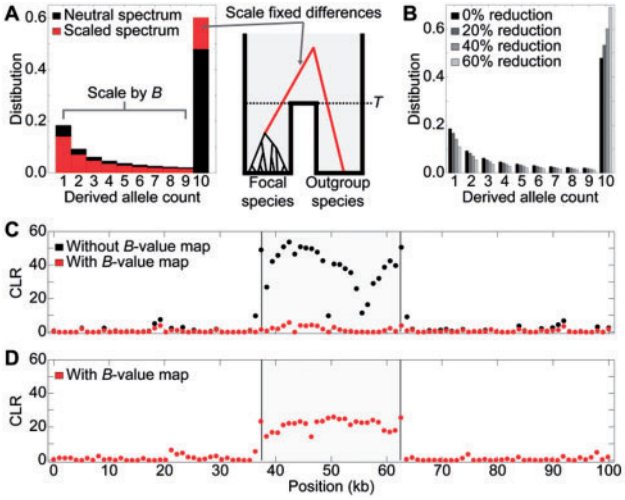

Fig. 1. Generating derived frequency spectra from a neutral frequency spectrum under background selection in a sample of 10 alleles and an outgroup sequence. (A) Polymorphic sites (allele counts 1–9) are scaled by a factor *B*, reducing diversity by 1-*B*. The proportion of fixed sites (allele count 10) is scaled by $(T+2BN/n)/(T+2N/n)$, and the spectrum is then normalized to sum to 1. The scaling factor for the fixed difference class assumes a model in which a pair of species split *T* generations ago, with all populations having effective size *N* (SweepFinder2 implementation permits unequal sizes). (B) Modified frequency spectra for 0, 20, 40 and 60% reductions in diversity due to background selection (*B*-values of 1.0, 0.8, 0.6 and 0.4, respectively). As *B*-value decreases, the level of diversity decreases, and the ratio of diversity to divergence decreases. (C, D) Simulation results (Huber *et al.*, 2015) indicating composite likelihood ratio test statistics as a function of position along a sequence without (C) and with (D) a fixed selective sweep in the center of the sequence. The gray region represents a reduction in recombination rate by two orders of magnitude. Including the *B*-value map decreases false inferences of positive selection (C), yet still can identify positively-selected alleles in regions with background selection (D). Though panel D only displays results when correcting for background selection, it should be noted that the selection signal is overestimated when not controlling for background selection, and could lead to biased estimates of selection coefficients

hypothesis. In addition, recent positively-selected alleles within regions undergoing background selection can still be detected (Fig. 1D). Furthermore, changes in *B*-values across the genome can be incorporated by modifying frequency spectra to preserve the spatial structure in genetic variation leveraged by SweepFinder. While *B*-value maps are currently available only for humans (McVicker *et al.*, 2009) and *Drosophila melanogaster* (Comeron, 2014), the methodology introduced by McVicker *et al.* can be employed to generate maps for other species.

SweepFinder2 is the first method that accounts for the effects of negative selection on diversity when searching for adaptive alleles. In addition, it incorporates novel features that provide the user with increased flexibility, such as the ability to specify a set of test sites, set distances between test sites and employ a recombination map. Thus, our new composite likelihood ratio test generalizes the one implemented in SweepFinder (Nielsen *et al.*, 2005), and provides a substantial improvement in power and flexibility to the popular SweepFinder software.

Conflict of Interest: none declared.

References

Akashi,H. *et al.* (2012) Weak selection and protein evolution. *Genetics*, **192**, 15–31.

Downloaded from http://bioinformatics.oxfordjournals.org/ at University of California, Los Angeles on August 30, 2016

Akey,J.M. (2009) Constructing genomic maps of positive selection in humans: where do we go from here? *Genome Res.*, **19**, 711–722.

Boyko,A.R. *et al.* (2008) Assessing the evolutionary impact of amino acid mutations in the human genome. *PLoS Genet.*, **4**, e1000083.

Charlesworth,B. (2012) The role of background selection in shaping patterns of molecular evolution and variation: evidence from variability on the *Drosophila X* chromosome. *Genetics*, **191**, 233–246.

Charlesworth,B. *et al.* (1993) The effect of deleterious mutations on neutral molecular variation. *Genetics*, **134**, 1289–1303.

Charlesworth,D. *et al.* (1995) The pattern of neutral molecular variation under the background selection model. *Genetics*, **141**, 1619–1632.

Comeron,J.M. (2014) Background selection as baseline for nucleotide variation across the *Drosophila* genome. *PLoS Genet.*, **10**, e1004434.

Enard,D. *et al.* (2014) Genome wide signals of pervasive positive selection in human evolution. *Genome Res.*, **24**, 885–895. doi:10.1101/gr.165822.113.

Granka,J.M. *et al.* (2012) Limited evidence for classic selective sweeps in African populations. *Genetics*, **192**, 1049–1064.

Hawks,J. *et al.* (2007) Recent acceleration of human adaptive evolution. *Proc. Natl Acad. Sci. USA*, **104**, 20753–20758.

Hernandez,R.D. *et al.* (2011) Classic selective sweeps were rare in recent human evolution. *Science*, **331**, 920–924.

Huber,C.D. *et al.* (2015) Detecting recent selective sweeps while controlling for mutation rate and background selection. *Mol. Ecol.*, **25**, 142–156. doi:10.111/mec.13351.

Hudson,R.R. and Kaplan,N.L. (1995a) Deleterious background selection with recombination. *Genetics*, **141**, 1605–1617.

Hudson,R.R. and Kaplan,N.L. (1995b) The coalescent process and background selection. *Philos. Trans. R. Soc. B*, **349**, 19–23.

Kaiser,V.B. and Charlesworth, B. (2009) The effects of deleterious mutations on evolution in non-recombining genomes. *Trends Genet.*, **25**, 9–12.

Kim,Y. and Stephan, W. (2002) Detecting a local signature of genetic hitchhiking along a recombining chromosome. *Genetics*, **160**, 765–777.

Li,J. *et al.* (2011) Global patterns of genetic diversity and signals of natural selection for human *ADME* genes. *Hum. Mol. Genet.*, **20**, 528–540.

Lohmueller,K.E. *et al.* (2011) Natural selection affects multiple aspects of genetic variation at putatively neutral sites across the human genome. *PLoS Genet.*, **7**, e1002326.

Maynard Smith,J. and Haigh, J. (1974) The hitch-hiking effect of a variable gene. *Genet. Res.*, **23**, 23–35.

McVean,G.A. and Charlesworth, B. (2000) The effects of Hill-Robertson interference between weakly selected mutations on patterns of molecular evolution and variation. *Genetics*, **155**, 929–944.

McVicker,G. *et al.* (2009) Widespread genomic signatures of natural selection in hominid evolution. *PLoS Genet.*, **5**, e1000471.

Nicolaisen,L.E. and Desai,M.M. (2013) Distortions in genealogies due to purifying selection and recombination. *Genetics*, **194**, 221–230.

Nielsen,R. *et al.* (2005) Genomic scans for selective sweeps using SNP data. *Genome Res.*, **15**, 1566–1575.

Nordborg,M. *et al.* (1996) The effect of recombination on background selection. *Genet. Res.*, **67**, 159–174.

Pavlidis,P. *et al.* (2010) Searching for footprints of positive selection in whole-genome SNP data from non-equilibrium populations. *Genetics*, **185**, 907–922.

Seger,J. *et al.* (2010) Gene genealogies strongly distorted by weakly interfering mutations in constant environments. *Genetics*, **184**, 529–545.

Svetec,N. *et al.* (2009) Recent strong positive selection on *Drosophila melanogaster HDAC6*, a gene encoding a stress surveillance factor, as revealed by population genomic analysis. *Mol. Biol. Evol.*, **26**, 1549–1556.

Williamson,S.H. *et al.* (2007) Localizing recent adaptive evolution in the human genome. *PLoS Genet.*, **3**, e90.

Wilson Sayres,M.A. *et al.* (2014) Natural selection reduced diversity on human Y chromosomes. *PLoS Genet.*, **10**, e1004064.

Downloaded from http://bioinformatics.oxfordjournals.org/ at University of California, Los Angeles on August 30, 2016