Genetics and population analysis

**BalLeRMix+: mixture model approaches for robust joint identification of both positive selection and long-term balancing selection**

Xiaoheng Cheng 1,*,† and Michael DeGiorgio2,*

1Department of Biology, Pennsylvania State University, University Park, PA 16802, USA and 2Department of Electrical Engineering and Computer Science, Florida Atlantic University, Boca Raton, FL 33431, USA

*To whom correspondence should be addressed.
†Present address: Department of Ecology and Evolution, University of Chicago, Chicago, IL 60637, USA

**Abstract**

**Summary:** The growing availability of genomewide polymorphism data has fueled interest in detecting diverse selective processes affecting population diversity. However, no model-based approaches exist to jointly detect and distinguish the two complementary processes of balancing and positive selection. We extend the BalLeRMix B-statistic framework described in Cheng and DeGiorgio (2020) for detecting balancing selection and present BalLeRMix+, which implements five B statistic extensions based on mixture models to robustly identify both types of selection. BalLeRMix+ is implemented in Python and computes the composite likelihood ratios and associated model parameters for each genomic test position.

**Availability and implementation:** BalLeRMix+ is freely available at https://github.com/bioXiaoheng/BallerMixPlus.

**Contact:** xcheng@uchicago.edu or mdegiorg@fau.edu

**Supplementary information:** Supplementary data are available at Bioinformatics online.

1 Introduction

Footprints of natural selection provide valuable insights into the evolutionary history of populations. As a result, they have been key features that evolutionary biologists probe for within sequenced genomes. Positive selection increases the prevalence of beneficial genetic variation and can reduce genetic diversity in regions nearby the selected loci, and is one of the most examined modes of natural selection (Booker et al., 2017, offers a good review). Meanwhile, balancing selection maintains polymorphisms at selected loci and sharply increases genetic diversity in regions adjacent to the selected loci. The deluge of polymorphism data available from contemporary sequencing technologies has fueled interest in both method development (e.g. Bitarello et al., 2018; Cheng and DeGiorgio, 2019, 2020; DeGiorgio et al., 2014; Islak et al., 2021; Sheehan and Song, 2016; Siewert and Voight, 2017, 2020) and empirical data analysis (e.g. Croze et al., 2017; Leffler et al., 2013; Teixeira et al., 2015) on balancing selection.

However, despite these methodological advancements, few model-based methods exist to jointly detect and distinguish positive and balancing selection from genomic data. Most approaches suited to this task, such as the summary statistics Tajima’s D (Tajima, 1989) and the HKA test (Hudson et al., 1987), as well as the anomaly detection approach of Tsel (Hunter-Zwick and Clark, 2013), identify genomic regions displaying patterns of variation unexpected under neutrality. Though Tsel showcases improved power and robustness compared with previous summary statistics, it cannot indicate the nature of selection, and none of these statistics provide direct details about selected footprint features at outlier regions, such as balanced polymorphism frequency, width of the footprint and magnitude of distortion of the distribution of allele frequencies in support of either positive or balancing selection. Instead, alternative contemporary machine learning strategies for distinguishing between balancing selection and positive selection have been developed and proven to be powerful (Isildak et al., 2021; Sheehan and Song, 2016), yet these methods often rely on accurate estimates of key population parameters such as demographic histories, extensive training datasets and substantial computational resources to deploy. Hence, it is desirable to have a computationally efficient model-based approach that makes minimal assumptions and that has power to discriminate both positive and balancing selection from neutrality, as well as the ability to classify the mode of selection at genomic regions strongly deviating from neutrality.

Initially aiming at accommodating the variability of footprint sizes of long-term balancing selection, Cheng and DeGiorgio (2020) described a flexible mixture model framework, collectively termed B statistics, that we extend here to consider positive selection as well. The B statistics assume the number of balanced alleles follows a binomial distribution with n trials (sample size) and success rate p.
frequencies; example in Fig. 1B). Therefore, extending the mixture
sweeps (when pending on value of ing model based on the same set of assumptions.
models in this way broadens the applicability of the Cheng and
DeGiorgio (2020) B example in Fig. 1A), but can also fit data generated by selective
observed data under long-term balancing selection (where
beta-binomial into the models, the allele count probability distribution.
statistics to adopt a beta-binomial distribution instead to approximate
librium frequency (e.g.Bergland accounting for by the binomial model can inflate the variance, such
observed allele counts are fixed. However, many factors not
sign of log10
A, B (positive selection) are consistent with expectations based on
selection signal), as well as the number of informative sites included
in the computation. The sign of log10 (a) can be indicative of the mode
of selection, as exemplified in Figure 1A and B.

3 Performance evaluation
To evaluate the performance of BalLeRMix+ compared to
BalLeRMix to detect balancing selection, we simulated sequences
under both neutrality and long-term balancing selection using SLiM3.3 (Haller and Messer, 2019) following the protocol of Cheng and DeGiorgio (2020). Both the original and extended B statistics show comparable power (Fig. 1C), confirming that BalLeRMix+ can powerfully detect long-term balancing selection. For positive selection, we simulated sequences evolving along the inferred demographic history of Europeans (see Supplementary Note), and introduced a de novo mutation with per-generation selective advantage of s = 0.01 at 104 or 2500 generations before sampling. Unlike the original B statistics that show little to no power, the extended B1, B2 and B2,MAF statistics of BalLeRMix+ exhibit high power to identify the selective sweep (Fig. 1D). Moreover, Figure 1A and B displays peaks of high-magnitude B2 statistic at the centers of the simulated sequences, showing that the signal of both balancing and positive selection can be localized. We also simulated scenarios of partial selective sweeps, sweeps on standing variation, adaptive introgression and recent balancing selection (see Supplementary Note), and our results confirm that BalLeRMix+ can powerfully and robustly identify and distinguish diverse modes of selection. Given the overall power and robustness of BalLeRMix+, we believe that it represents a comprehensive suite of statistics and will be a welcome addition to the evolutionary biologists’ toolbox.

Acknowledgements
We thank Dr Matthias Steinru¨ cken for helpful discussions. We also thank Dr Matteo Fumagalli and Dr Martin Hölzer for their careful reviews and helpful suggestions to help improve our manuscript and software repository. Simulation studies were conducted on the high-performance computing cluster maintained by Institute of Computational and Data Sciences at Pennsylvania State University.

Funding
This work was supported by the Pennsylvania State University; the University of Chicago; the National Institutes of Health [R35GM128590]; and the National Science Foundation [DEB-1949268, BCS-2001063, DBI-2130666, IIS-2077339].

Conflict of Interest: none declared.
Data availability

The data underlying this article are available in the online supplementary material.

References

Bergland, A.O. et al. (2014) Genomic evidence of rapid and stable adaptive oscillations over seasonal time scales in drosophila. *PLoS Genet.*, 10, e1004775.

Bitarello, B.D. et al. (2018) Signatures of long-term balancing selection in human genomes. *Genome Biol. Evol.*, 10, 939–955.

Booker, T.R. et al. (2017) Detecting positive selection in the genome. *BMC Biol.*, 15, 1–10.

Cheng, X. and DeGiorgio, M. (2019) Detection of shared balancing selection in the absence of trans-species polymorphism. *Mol. Biol. Evol.*, 36, 177–199.

Cheng, X. and DeGiorgio, M. (2020) Flexible mixture model approaches that accommodate footprint size variability for robust detection of balancing selection. *Mol. Biol. Evol.*, 37, 3267–3291.

Croze, M. et al. (2017) A genome-wide scan for genes under balancing selection in drosophila melanogaster. *BMC Evol. Biol.*, 17, 15.

DeGiorgio, M. et al. (2014) A model-based approach for identifying signatures of ancient balancing selection in genetic data. *PLoS Genet.*, 10, e1004561.

Haller, B.C. and Messer, P.W. (2019) SLiM 3: forward genetic simulations beyond the Wright-Fisher model. *Mol. Biol. Evol.*, 36, 632–637.

Harris, C.R. et al. (2020) Array programming with numpy. *Nature*, 585, 357–362.

Hudson, R.R. et al. (1987) A test of neutral molecular evolution based on nucleotide data. *Genetics*, 116, 153–159.

Hunter-Zinck, H. and Clark, A.G. (2015) Aberrant time to most recent common ancestor as a signature of natural selection. *Mol. Biol. Evol.*, 32, 2784–2797.

Isildak, U. et al. (2021) Distinguishing between recent balancing selection and incomplete sweep using deep neural networks. *Mol. Ecol. Resour.*, 1–13. doi:10.1111/1755-0998.13379.

Leffler, E.M. et al. (2013) Multiple instances of ancient balancing selection shared between humans and chimpanzees. *Science*, 339, 1578–1582.

Sheehan, S. and Song, Y. (2016) Deep learning for population genetic inference. *PLoS Comput. Biol.*, 12, e1004845.

Siewert, K.M. and Voight, B.F. (2017) Detecting long-term balancing selection using allele frequency correlation. *Mol. Biol. Evol.*, 34, 2996–3005.

Siewert, K.M. and Voight, B.F. (2020) BetaScan2: standardized statistics to detect balancing selection utilizing substitution data. *Genome Biol. Evol.*, 12, 3873–3877.

Tajima, F. (1989) Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics*, 123, 585–595.

Teixeira, J.C. et al. (2015) Long-term balancing selection in lad1 maintains a missense trans-species polymorphism in humans, chimpanzees, and bonobos. *Mol. Biol. Evol.*, 32, 1186–1196.