# Flexible mixture model approaches that accommodate footprint size variability for robust detection of balancing selection

Xiaoheng Cheng[1,2], Michael DeGiorgio[3]

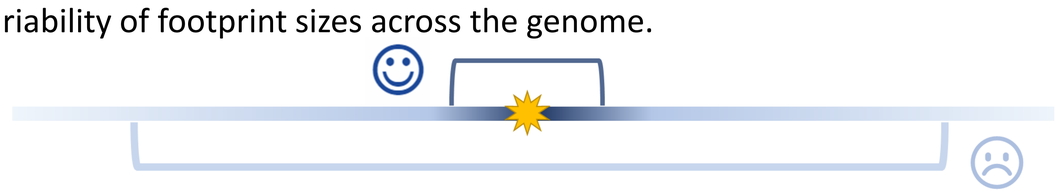
*[1]MCIBS, Huck Institutes of the Life Sciences, Pennsylvania State University; [2]Department of Biology, Pennsylvania State University; [3]Department of Computer & Electrical Engineering and Computer Science, Florida Atlantic University*

## Abstract

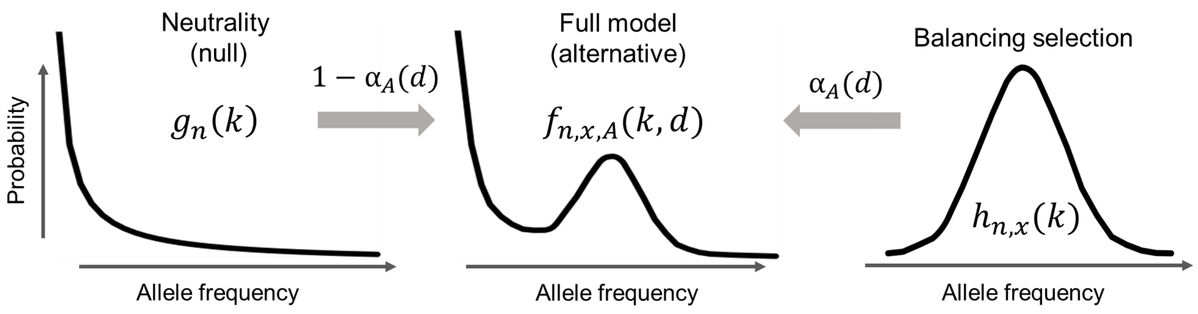
Long-term balancing selection typically leaves narrow footprints of increased genetic diversity, and therefore most detection approaches only achieve optimal performances when sufficiently small genomic regions (*i.e.*, windows) are examined. Such methods are sensitive to window sizes and suffer substantial losses in power when windows are large. Here, we employ mixture models to construct a set of five composite likelihood ratio test statistics—$B_0$, $B_{0,MAF}$, $B_1$, $B_2$, $B_{2,MAF}$—which we collectively term $B$ statistics. These statistics are agnostic to window sizes and can operate on diverse forms of input data. Through simulations, we showed that they exhibit comparable power to the best-performing current methods, and retain substantially high power regardless of window sizes. They also display considerable robustness to high mutation rates and uneven recombination landscapes, as well as an array of other common confounding scenarios. Moreover, we applied $B_2$ on genomic data of two human populations and recovered many top candidates from prior studies, including the then-uncharacterized *STPG2* and *CCDC169-SOHLH2*, both related to gamete functions. We further applied $B_2$ on a bonobo population-genomic dataset. In addition to the *MHC-DQ* and *MHC-DP* genes, we uncovered several novel candidate genes, such as *KLRD1*, involved in viral defense, and *SCN9A*, associated with pain perception. Finally, we show that our methods can be extended to account for multi-allelic balancing selection, and integrated the set of statistics into open-source software named `BalLeRMix` for future applications by the scientific community.

## Composite likelihood ratios based on a nested model

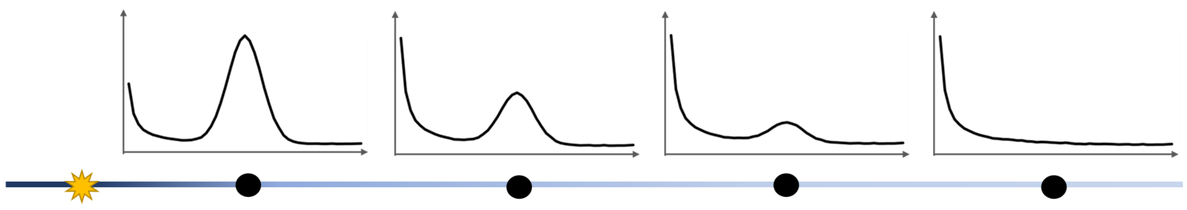
- Methods for detecting balancing selection usually performs best when adopting scanning window similar in size to the footprints of selection, which are narrow.
- All previous methods are sensitive to window sizes and cannot accommodate the variability of footprint sizes across the genome.



- A nested model can be constructed by a mixture of probability distributions describing the influence of neutrality (genome-wide) and balancing selection (binomial).




- Under the mixture model, the probability distribution for allele frequencies converges to neutrality when the distance $d$ is large.

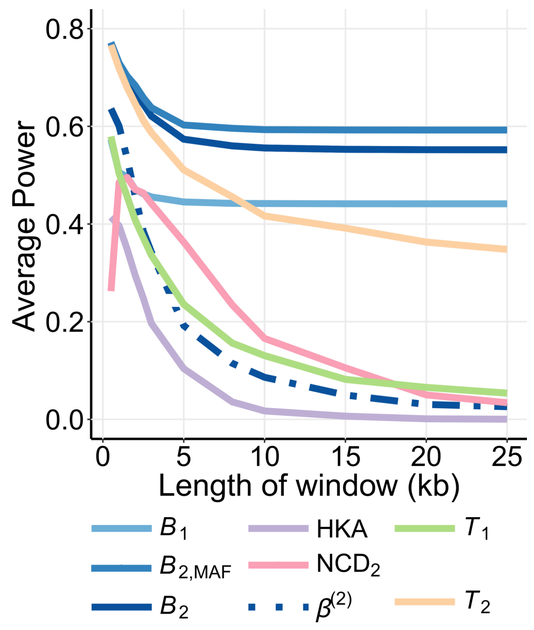


- The resulting composite likelihood ratios, $B$ statistics, are not influenced by sites far from the test site.

$$B = \underset{(\hat{x},\hat{A})}{\operatorname{argmax}} \, 2\ln \frac{\prod_i f_{n,\hat{x},\hat{A}}(k_i,d_i)}{\prod_i g_n(k_i)}$$

- We expanded the framework to five variants—$B_2$, $B_{2,MAF}$, $B_1$, $B_0$, $B_{0,MAF}$—each accommodating a certain type of input data.

## Robust and powerful performance in simulations

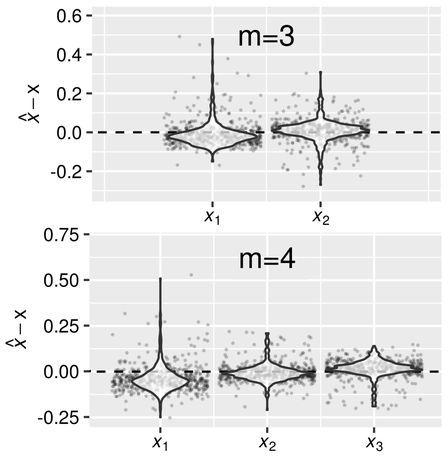



### Robustness to window sizes

- When adopting the optimal window size, $B$ statistics exhibit comparable powers to their analogous model-based and summary statistics.
- Compared with their analogous extant approaches, $B$ statistics experience minor compromise in power with increasing window sizes.
- The $B$ statistics stabilize at a high power under large window sizes, whereas the powers of most current methods are diminished.

### Power on multi-allelic balancing selection

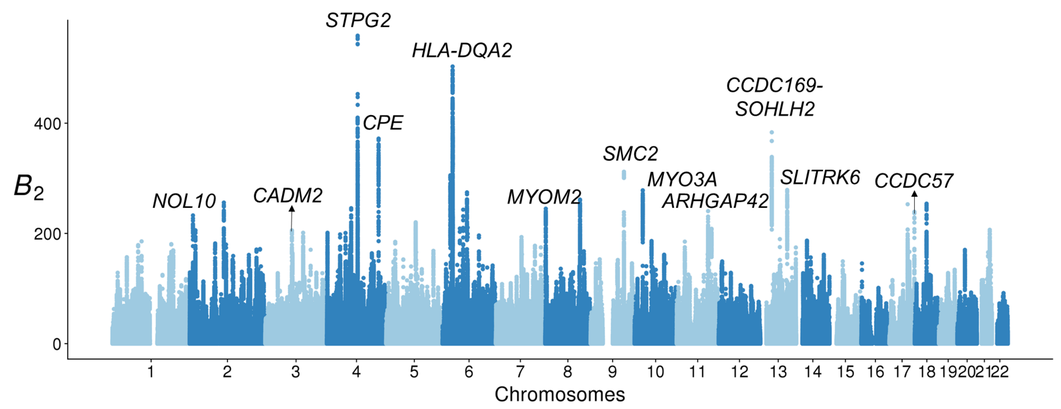
- Our framework can be extended to model multi-allelic balancing selection.
- Multi-allelic $B$ statistics have superior power to detect balancing selection with $m$ balanced allelic classes ($m$>2).
- Multi-allelic $B$ statistics can accurately infer the equilibrium frequencies.




## Uncovering balancing selection on empirical data

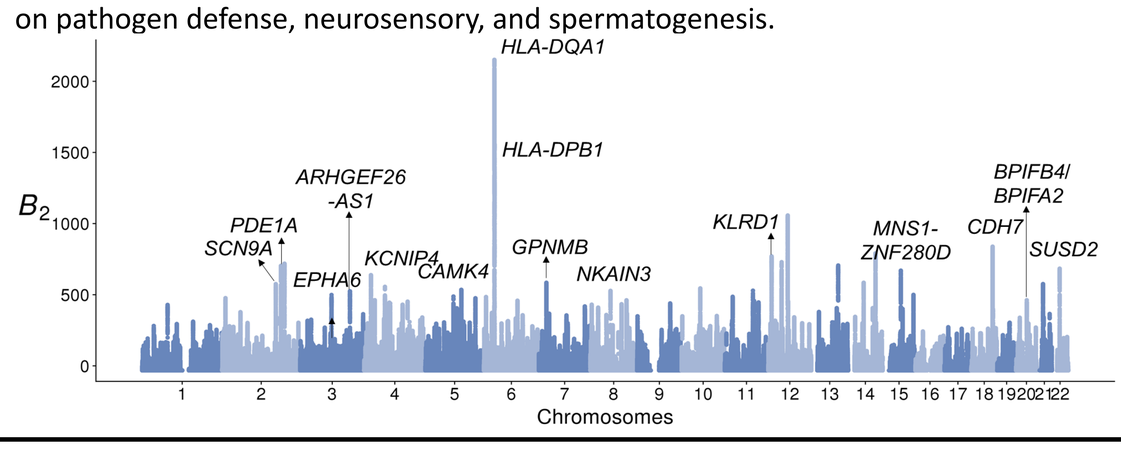
### Balancing selection on European and African humans

- Apply $B_2$ on the genomic data of European (CEU) and African (YRI) populations (The 1000 Genomes Project, 2015).
- Recovered previously-described candidates such as *HLA-DQ* genes and *CADM2*.
- Two top candidates, *STPG2* (*C4orf37*) and *CCDC169-SOHLH2* (*C13orf38*), were reported but not characterized previously. We found both are linked to gamete functionality.




### Balancing selection on bonobos

- Apply $B_2$ on the variant calls from 13 bonobos (Prado-Martinez *et al.*, 2013).
- The MHC region, which is highly integral to adaptive immunity, shows extraordinary evidence of balancing selection, consistent with findings in humans and chimpanzees.
- We found novel candidates, such as *KLRD1*, *SCN9A*, and *PDE1A*, that have implications on pathogen defense, neurosensory, and spermatogenesis.




## References and support

- [1] The 1000 Genomes Project Consortium (2015) *Nature.* 526: 68-74. [2] Prado-Matinez *et al.* (2013) Nature 499:471-475.
- This research was funded by Pennsylvania State University, by National Institutes of Health grant R35-GM128590, by National Science Foundation grants DEB-1753489, DEB-1949268, and BCS-1925825, and by the Alfred P. Sloan Foundation. Portions of this research were conducted with the computing resources from Institute for Computational and Data Science Advanced CyberInfrastructure at Pennsylvania State University.

## Conclusions

- We constructed a novel set likelihood ratio statistics based on mixture models, termed $B$ statistics, which can accommodate the variability in footprint sizes of balancing selection.
- $B$ statistics show minor decay in power when window sizes increases, and perform comparably well for detecting balancing selection.
- Scans for balancing selection on humans and bonobos revealed both previously-characterized and novel candidates.