Flexible Mixture Model Approaches That Accommodate Footprint Size Variability for Robust Detection of Balancing Selection

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

[1]Huck Institutes of Life Sciences, Pennsylvania State University, University Park, PA

[2]Department of Biology, Pennsylvania State University, University Park, PA

[3]Department of Computer and Electrical Engineering and Computer Science, Florida Atlantic University, Boca Raton, FL

***Corresponding author:** E-mail: mdegiorg@fau.edu.

Associate editor: Yoko Satta

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, 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 show 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 a specific version of the *B* statistics, termed B_2, to a human population-genomic data set and recovered many top candidates from prior studies, including the then-uncharacterized *STPG2* and *CCDC169–SOHLH2*, both of which are related to gamete functions. We further applied B_2 on a bonobo population-genomic data set. In addition to the *MHC-DQ* 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 multiallelic balancing selection and integrated the set of statistics into open-source software named BalLeRMix for future applications by the scientific community.

Key words: **balancing selection, mixture model, likelihood ratio statistic, bonobo, multiallelic balancing selection.**

Introduction

Balancing selection maintains polymorphism at selected genetic loci and can operate through a variety of mechanisms (Charlesworth 2006). In addition to overdominance (Charlesworth B and Charlesworth D 2010), other processes such as sexual selection (Cho et al. 2006), periodical environmental shifts (Bergland et al. 2014), pleiotropy (Andrés 2001; Mitchell-Olds et al. 2007), meiotic drive (Ubeda and Haig 2004; Charlesworth B and Charlesworth D 2010), and negative frequency-dependent selection (Charlesworth B and Charlesworth D 2010) can also maintain diversity at underlying loci. Due to the increasing availability of population-level genomic data, in which allele frequencies and genomic density of polymorphisms can be assessed in detail, there is an expanding interest in studying balancing selection and detecting its genomic footprints (Andrés et al. 2009; Leffler et al. 2013; DeGiorgio et al. 2014; Gao et al. 2015; Hunter-Zinck and Clark 2015; Sheehan and Song 2016; Lonn et al. 2017; Sweeney et al. 2017; Guirao-Rico et al. 2017; Siewert and Voight 2017, 2020; Bitarello et al. 2018; Ye et al. 2018; Cheng and DeGiorgio 2019). However, despite multiple efforts to design statistics for identifying balanced loci (DeGiorgio et al. 2014; Siewert and Voight 2017, 2020; Bitarello et al. 2018; Cheng and DeGiorgio 2019), performances of existing methods still leave room for improvement.

Early methods applied to this problem evaluated departures from neutral expectations of genetic diversity at a particular genomic region. For example, the Hudson–Kreitman–Aguadé (HKA) test (Hudson et al. 1987) uses a chi-square statistic to assess whether genomic regions have higher density of polymorphic sites when compared with a putative neutral genomic background. In contrast, Tajima's *D* (Tajima 1989) measures the distortion of allele frequencies from the neutral site frequency spectrum (SFS) under a model with constant population size. However, these early approaches were not tailored for balancing selection and have limited power. Recently, novel and more powerful summary statistics (Siewert and Voight 2017, 2020; Bitarello et al. 2018) and model-based approaches (DeGiorgio et al. 2014;

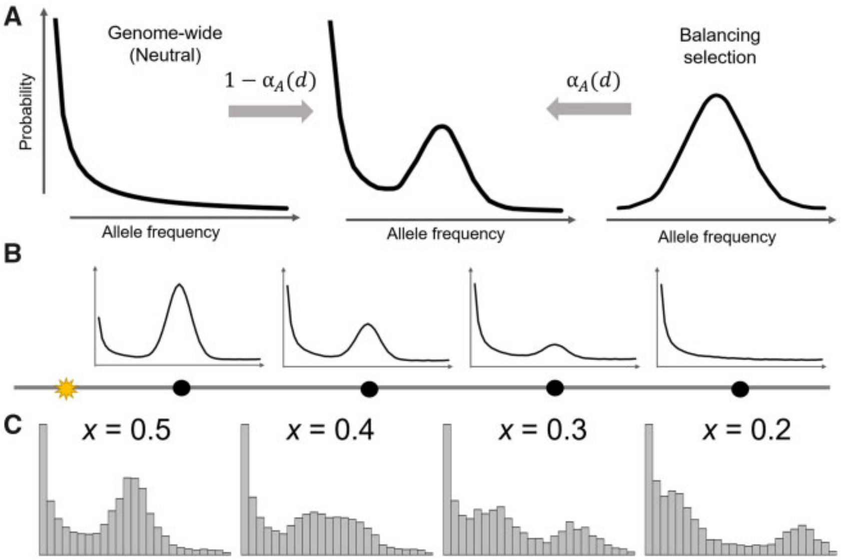

Fig. 1. Schematic of the mixture model underlying the *B* statistics. (*A*) The model for the alternative hypothesis is a mixture of the distribution of allele frequencies under balancing selection at proportion $\alpha_A(d)$, modeled by a binomial distribution, and the distribution under neutrality at proportion $1 - \alpha_A(d)$, modeled by the genome-wide SFS. Here, $\alpha_A(d)$ decays as a function of recombination distance *d*, and so sites close to (i.e., small *d*) the putative selected site will be modeled mostly by the distribution expected under balancing selection, whereas sites far from (i.e., large *d*) the selected site will be modeled mostly by the distribution expected under neutrality. (*B*) Distributions of allele frequencies at neutral sites (black dots) under the mixture model at varying distances *d* from the putative selected site (yellow star). (*C*) Distributions of allele frequencies from the center 10 kb (0.01 cM) of the simulated sequences when balancing selection maintains the equilibrium frequency of $x = 0.2$, 0.3, 0.4, or 0.5.

Cheng and DeGiorgio 2019) have been developed to specifically target regions under balancing selection. In general, the summary statistics capture deviations of allele frequencies from a putative equilibrium frequency of a balanced polymorphism. In particular, the noncentral deviation statistic (Bitarello et al. 2018) adopts an assigned value as this putative equilibrium frequency, whereas the β and $\beta^{(2)}$ statistics of Siewert and Voight (2017, 2020) use the frequency of the central polymorphic site instead. On the other hand, the *T* statistics of DeGiorgio et al. (2014) and Cheng and DeGiorgio (2019) compare the composite likelihood of the data under an explicit coalescent model of long-term balancing selection (Hudson et al. 1987; Hudson and Kaplan, 1988) with the composite likelihood under the genome-wide distribution of variation, which is taken as neutral.

Nevertheless, all extant approaches are limited by their sensitivity to the size of the region that the statistics are computed on (hereafter referred to as the "window size"). Because the footprints of long-term balancing selection are typically narrow (Hudson and Kaplan 1988; Charlesworth 2006), small windows with fixed sizes comparable to that of the theoretical footprint based on a genome-wide recombination rate estimate are commonly used in practice, especially for summary statistics. However, such small fixed window sizes not only lead to increased noise in the estimation of each statistic but also render the statistic incapable of adapting to varying footprint sizes across the genome due to factors such as the uneven recombination landscape (Smukowski and Noor 2011). Though adopting a larger window may reduce noise, true signals will likely be overwhelmed by the surrounding neutral regions, diminishing method power as shown by Cheng and DeGiorgio (2019). Available model-based approaches (DeGiorgio et al. 2014; Cheng and DeGiorgio 2019) could have been made robust to window sizes if they instead adopted the SFS expected under a neutrally evolving population of constant size as the null hypothesis, because their model of balancing selection for the alternative hypothesis converges to this constant-size neutral model for large recombination rates. However, this neutral model does not account for demographic factors that can impact the genome-wide distribution of allele frequencies, such as population size changes. To guard against such demographic influences, the model-based T_1 and T_2 statistics (DeGiorgio et al. 2014; Cheng and DeGiorgio 2019) employ the genome-wide SFS instead, compromising the robustness against large windows. Moreover, Cheng and DeGiorgio (2019) showed that although the power of the T_2 statistic decays much slower than other approaches as window size increases, the loss of power is still substantial.

In this article, we describe a set of composite likelihood ratio test statistics that are based on a mixture model (fig. 1*A* and *B*) that integrates both the genome-wide level of variation and the enrichment of sites with allele frequencies close to the equilibrium allele frequency of long-term balancing selection. Note that the latter has been successfully captured by the summary statistics β (Siewert and Voight 2017, 2020) and non-central deviation (NCD) (Bitarello et al. 2018). Our framework of nested models allows for robust and flexible detection of balancing selection that can augment the size of genomic regions considered in each test to best fit the data. Dependent on the types of data available, we propose a set of five likelihood ratio test statistics termed B_2, $B_{2,MAF}$, B_1, B_0, and $B_{0,MAF}$, which, respectively, accommodate data with substitutions and derived (B_2) or minor ($B_{2,MAF}$) allele frequency polymorphisms, with substitutions and polymorphisms with unknown allele frequency (B_1), and with derived (B_0) or minor ($B_{0,MAF}$) allele frequency polymorphisms only. We comprehensively evaluated their performances under an array of diverse simulated scenarios, including their powers for balancing selection with varying ages, distinct strengths and equilibrium frequencies, robustness against window sizes, and robustness against confounding factors such as demographic history, recombination rate variation, and mutation rate variation. We also compared and discussed their performances with other leading approaches—namely HKA, β, β^*, $\beta^{(2)}$, NCD, T_1, and T_2. To gauge the performance of B statistics on empirical data, we reexamined contemporary human populations in the 1000 Genomes Project data set (1000 Genomes Project Consortium 2015) to uncover previously hypothesized candidates. Furthermore, we performed an exploratory whole-genome scan with B_2 on bonobo genomic data (Prado-Martinez et al. 2013) to probe for long-term balancing selection in the other close relative of humans. We further extended our framework to consider multiallelic balancing selection and examined the performances of extant methods on cases of multilocus balancing selection. Lastly, we developed the software BalLeRMix (balancing selection likelihood ratio mixture models) to implement these novel tests for the convenience of the scientific community.

Model Description

A classical footprint of balancing selection is the increase in the proportion of sites with moderate allele frequencies that are close to the equilibrium frequency at the balanced locus (Kaplan et al. 1988; Siewert and Voight 2017). Previous modeling attempts (Kaplan et al. 1988; Song and Steinrücken 2012; DeGiorgio et al. 2014; Cheng and DeGiorgio 2019) primarily focused on delineating the underlying population-genetic processes, such as through coalescent or diffusion theory. Though these models are able to capture the distortion in the SFS resulting from balancing selection, their intricate mathematical formulations bring challenges to further model extensions to more complicated scenarios as well as the associated computations. As an alternative, it may be appealing to model the effect of balancing selection through statistical approximations of the expected features in the data.

Based on this idea, for a locus under balancing selection that is maintaining a pair of allelic classes, we can approximate the process of observing k_0 copies of the selected allele balanced at equilibrium frequency $x \in (0,1)$ in n samples, as following a binomial sampling process with n trials and a success rate x. For a biallelic neutral site that is linked to this selected locus, we assume that the k derived alleles observed from the n samples at this neutral site are all on the same haplotype with the k_0 selected alleles balanced at frequency x. That is, we assume $k = k_0$ and consider the k derived alleles on the neutral site as surrogates for the balanced alleles of the allelic class with which they are fixed. Therefore, when these two sites are in complete linkage, k can also be considered as binomially distributed with n trials and a success rate x. Meanwhile, for a neutral site not linked to this selected locus, we assume that k follows the distribution expected by the genome-wide SFS. Taken together, the probability $P_n(k)$ of observing k derived alleles out of n sampled alleles at a neutral site can be written as

$$\begin{aligned} P_n(k) = {} & \mathbb{P}[\text{Completely linked to the selected locus}] \\ & \cdot \mathbb{P}[k = k_0 \text{ out of } n \text{ binomially sampled with rate x}] \\ & + \mathbb{P}[\text{Not completely linked to the selected locus}] \\ & \cdot \mathbb{P}[k \text{ out of } n \text{ observed in the genome}]. \end{aligned}$$

Alternatively, this integration of two conditional probabilities can also be viewed as a mixture model, in which the two mixing components represent probabilities under balancing selection and neutrality (based on the genome-wide empirical distribution), with their respective mixing proportions α and $1-\alpha$ representing the probabilities of being completely linked to the selected locus or not, respectively. To approximate α, we chose to consider the exponential decay function, which has been adopted as a proxy for linkage disequilibrium (Nielsen et al. 2005; Moorjani et al. 2011; Loh et al. 2013). To accommodate the varying rates of linkage decay, we introduce a free parameter $A > 0$ for the statistic to optimize over, which essentially determines the size of the footprint of balancing selection, with smaller values of A having wider footprints than larger values. Hence, for a neutral site d recombination units away from the selected locus, the probability that is linked to the selected locus can be approximated by

$$\mathbb{P}[\text{Completely linked to the selected locus}] = \alpha_A(d) = e^{-Ad}.$$

Therefore, for a neutral site d recombination units away from the selected locus, we approximate the probability mass function for sampling k derived alleles out of n sampled alleles as

$$f_{n,x,A}(k,d) = \alpha_A(d) \cdot h_{n,x}(k) + [1 - \alpha_A(d)] \cdot g_n(k),$$

where $h_{n,x}(k)$ denotes the normalized binomial probability of sampling k successes out of n trials with success rate x, and $g_n(k)$ is the normalized genome-wide SFS denoting the proportion of sites with k derived alleles observed out of n sampled alleles. This formulation also applies when k represents the number of minor allele copies, for situations in which the ancestral allele cannot be polarized with an outgroup. See subsequent subsection for precise definitions of normalized $h_{n,x}(k)$ and $g_n(k)$.

Note that although we constructed this mixture model framework by combining conditional probabilities of the derived alleles at a neutral site to be on the same haplotype with one of the two balanced allele classes, the interpretation of the mixing weight $\alpha_A(d)$ is in effect not constrained to linkage and recombination. Other factors that can affect the local SFS, such as the accumulation of low-frequency mutations, can be accounted for by incorporating the genome-wide SFS as well. Although these factors can also vary by recombination distance, we formulate $\alpha_A(d)$ based on the properties of linkage decay alone to simplify our model.

In the following subsections, we describe a set of composite likelihood ratio statistics (B_2, $B_{2,\mathrm{MAF}}$, B_1, B_0, and $B_{0,\mathrm{MAF}}$) constructed based on this mixture model approach for identifying loci undergoing biallelic balancing selection. We also extended this framework to consider multiallelic balancing selection and describe these models in supplementary note 1, Supplementary Material online. Note that all the composite likelihood ratio statistics considered here assume that balancing selection is acting on a single locus. This set of composite likelihood ratio statistics has been implemented in the open-source software package BalLeRMix, which is available at https://github.com/bioXiaoheng/BalLeRMix/tree/master/software.

Probability Distributions Given Derived Allele Polymorphisms and Substitutions

For n sampled alleles at an informative site (i.e., polymorphism or substitution), when the ancestral state to each site can be confidently assigned, denote the number of derived alleles as k, $k = 1, 2, \ldots, n$. Let $\xi_n(k)$ be the total number of informative sites across the whole genome with k derived alleles observed out of n sampled alleles. The probability of observing such a site is therefore

$$g_n^{(2)}(k) = \frac{\xi_n(k)}{\sum_{j=1}^{n} \xi_n(j)}.$$

When balancing selection maintains an equilibrium frequency of x on the site under selection, the outcomes of observing derived alleles on this site (out of n lineages) can be approximated by a binomial distribution of n trials with a success probability of x. Following this binomial model, the probability of observing the selected site with k observed derived alleles is

$$h_{n,x}^{(2)}(k) = \frac{\mathrm{Bin}(k; n, x)}{\sum_{j=1}^{n} \mathrm{Bin}(j; n, x)}.$$

Note the values of $g_n(k)$ and $h_{n,x}(k)$ are conditional on the number of sampled alleles n, and therefore our model requires that the sample size is made explicit at each informative site. Permitting the sample size to differ across sites is important, as missing genotype calls are often common in empirical studies, with sample sizes naturally varying across the genome.

For an informative site d recombination units away from the presumed site under selection, it can either be linked to the derived (with equilibrium frequency x) or ancestral (with equilibrium frequency $1 - x$) haplotype under balancing selection, resulting in a bimodal distribution (fig. 1C). Therefore, the probability of observing k derived alleles out of n sampled alleles is

$$f_{n,x,A}^{(2)}(k, d) = \alpha_A(d)\left[\frac{1}{2}h_{n,x}^{(2)}(k) + \frac{1}{2}h_{n,1-x}^{(2)}(k)\right] + [1 - \alpha_A(d)]g_n^{(2)}(k),$$

where $\alpha_A(d) = \exp(-Ad)$ and where A is a model parameter that determines the size of the genomic footprint of balancing selection. When allele frequency information is unavailable at polymorphic sites, the probability of observing a polymorphic site ($k \neq n$) or substitution ($k = n$) would be

$$f_{n,x,A}^{(1)}(k, d) = f_{n,x,A}^{(2)}(n, d)\mathbb{1}_{\{\mathrm{k=n}\}} + [1 - f_{n,x,A}^{(2)}(n, d)]\mathbb{1}_{\{\mathrm{k\neq n}\}},$$

where $\mathbb{1}_{\{\mathbb{E}\}}$ is a dummy variable that takes the value 1 if the expression E is true, and 0 otherwise.

Similarly, when substitutions are not considered or are missing in the data (i.e., only observe derived allele counts $k = 1, 2, \ldots, n - 1$), the two mixing components can be normalized as

$$g_n^{(0)}(k) = \frac{\xi_n(k)}{\sum_{j=1}^{n-1} \xi_n(j)},$$

and

$$h_{n,x}^{(0)}(k) = \frac{\mathrm{Bin}(k; n, x)}{\sum_{j=1}^{n-1} \mathrm{Bin}(j; n, x)}.$$

The probability of observing a polymorphic site with k derived alleles out of n sampled alleles is then

$$f_{n,x,A}^{(0)}(k, d) = \alpha_A(d)\left[\frac{1}{2}h_{n,x}^{(0)}(k) + \frac{1}{2}h_{n,1-x}^{(0)}(k)\right] + [1 - \alpha_A(d)]g_n^{(0)}(k).$$

Probability Distributions Given Minor Allele Polymorphisms and Substitutions

When alleles cannot be confidently polarized, minor allele frequencies are often used instead. For informative sites with n sampled alleles, denote the minor allele count as k, $k = 0, 1, \ldots, \lfloor n/2 \rfloor$, and the total number of such sites in the genome as $\eta_n(k)$. Substitutions are assigned to $\eta_n(0)$, as the minor allele count is zero. The probability of observing a site with k minor alleles out of n sampled alleles in the genome is

$$g_n^{(2,\mathrm{MAF})}(k) = \frac{\eta_n(k)}{\sum_{j=0}^{\lfloor n/2 \rfloor} \eta_n(j)}.$$

Assume the equilibrium minor allele frequency at the locus undergoing long-term balancing selection is $x \in (0, 0.5]$. The probability of observing k minor alleles out of n sampled alleles is then

$$h_{n,x}^{(2,\mathrm{MAF})}(k) = \frac{\mathrm{Bin}(k;n,x) + \mathrm{Bin}(n-k;n,x)\mathbb{1}_{\{k \neq n/2\}}}{\sum_{j=1}^{n} \mathrm{Bin}(j;n,x)}.$$

Hence, for an informative site d recombination units away from the presumed site under selection, the probability of observing k minor alleles out of n sampled alleles is

$$f_{n,x,A}^{(2,\mathrm{MAF})}(k,d) = \alpha_A(d)h_{n,x}^{(2,\mathrm{MAF})}(k) + [1-\alpha_A(d)]\ g_n^{(2,\mathrm{MAF})}(k).$$

Similarly, when substitutions are not considered or are missing in the data (i.e., only observed minor alleles counts $k = 1, 2, \ldots, \lfloor n/2 \rfloor$), the two mixing components can be normalized as

$$g_n^{(0,\mathrm{MAF})}(k) = \frac{\eta_n(k)}{\sum_{j=1}^{\lfloor n/2 \rfloor} \eta_n(j)}$$

and

$$h_{n,x}^{(0,\mathrm{MAF})}(k) = \frac{\mathrm{Bin}(k;n,x) + \mathrm{Bin}(n-k;n,x)\mathbb{1}_{\{k \neq n/2\}}}{\sum_{j=1}^{n-1} \mathrm{Bin}(j;n,x)}.$$

The probability of observing a polymorphic site with k minor alleles out of n sampled alleles is then

$$f_{n,x,A}^{(0,\mathrm{MAF})}(k,d) = \alpha_A(d)h_{n,x}^{(0,\mathrm{MAF})}(k) + [1-\alpha_A(d)]\ g_n^{(0,\mathrm{MAF})}(k).$$

Composite Likelihood Ratio Tests Based on the Mixture Models

In the preceding subsection, we have provided the marginal probability distributions for the number of observed copies of either a derived or a minor allele at an informative site that is a certain distance from a locus undergoing biallelic balancing selection. Because we cannot obtain the full likelihood that accounts for the joint distribution of allele frequencies across all informative sites that are in high linkage disequilibrium, we instead make the simplifying assumption that neighboring informative sites are independent. This assumption, albeit invalid, allows us to gain insight from the composite likelihood, which is computed by multiplying the marginal probability distributions for all informative sites. By maximizing the resulting composite likelihood from the full model across our parameter space, we can also obtain estimates of the optimal parameter values (i.e., $\widehat{x}$ and $\widehat{A}$), which confer information about the features of the footprints consistent with balancing selection.

Based on the probability distributions described for the five models, for each model $X \in \{$"2", "2,MAF", "1", "0", "0,MAF"$\}$, the composite likelihood of a genomic region with L informative sites under the null hypothesis of neutrality is

$$\mathcal{L}_0^{(X)}(\mathbf{n},\mathbf{k}) = \prod_{i=1}^{L} g_{n_i}^{(X)}(k_i),$$

where $\mathbf{n} = [n_1, n_2, \ldots, n_L]$ and $\mathbf{k} = [k_1, k_2, \ldots, k_L]$ are the vectors of sample sizes and derived or minor allele counts, respectively, at the L informative sites in the genomic region. Recall that the probabilities of sampling a certain number of derived or minor alleles under our model depend on the sample sizes at informative sites, and because sample sizes often vary across the genome due to missing data in empirical studies, we make explicit the sample sizes across all informative sites in the vector $\mathbf{n}$. Similarly, the composite likelihood under the alternative hypothesis of model X would be

$$\mathcal{L}_a^{(X)}(x,A;\mathbf{n},\mathbf{k},\mathbf{d}) = \prod_{i=1}^{L} f_{n_i,x,A}^{(X)}(k_i,d_i),$$

where $\mathbf{d} = [d_1, d_2, \ldots, d_L]$ is the vector of recombination distances between the test site and each of the L informative sites. This likelihood is maximized at

$$(\widehat{x},\widehat{A}) = \underset{(x,A)}{\arg\max}\ \mathcal{L}_a^{(X)}(x,A;\mathbf{n},\mathbf{k},\mathbf{d}).$$

Hence, under model $X \in \{$"2", "2,MAF", "1", "0", "0,MAF"$\}$, the log composite likelihood ratio test statistic for the test site is

$$B_X = 2[\ln\mathcal{L}_a^{(X)}(\widehat{x},\widehat{A};\mathbf{n},\mathbf{k},\mathbf{d}) - \ln\mathcal{L}_0^{(X)}(\mathbf{n},\mathbf{k})].$$

Note that although log-likelihood ratio test statistics can be considered as following χ^2 distributions (of which the degree of freedom is the number of free parameters, e.g., two in the full models described above), B statistics are a set of composite log-likelihood ratio (CLR) statistics, which do not follow regular χ^2 distributions (Pace et al. 2011; Varin et al. 2011). In order for a CLR statistic to approximately follow an asymptotic χ^2 distribution, it needs to undergo adjustment (Pace et al. 2011) that also yields the effective degree of freedom of the asymptotic distribution the adjusted CLR statistic conforms to. This adjustment process is based on the set of observations used to compute the CLR, which is different for every test site. Because for B statistics, the size of the genomic region considered by each test varies across the genome and because the informative sites included in the region are highly correlated, the effective degree of freedom also varies across test sites. Therefore, we cannot infer significance from the values of B statistic alone by referencing the χ^2 distribution.

Moreover, and probably even more importantly, because the model under the null hypothesis only accounts for mean demographic effects based on the genome-wide SFS and not its higher moments (e.g., variance), the resulting P-value obtained from a χ^2 distribution after the statistical adjustment would still deviate from what is commonly expected when the test rejects neutrality (i.e., neutral evolution under an explicit demographic model). We therefore would recommend mass simulation under an appropriate demographic model to generate the "null" distribution of B statistics in order to accurately infer the significance of each test, with the caveat that such an endeavor would require extensive computational resources due to the millions of simulations needed, the lengths of the simulated segments, and the optimization of the B statistics on each of these simulated segments. Lastly, in order to infer genome-wide significance, P-values need to be corrected for multiple testing, for example, through Bonferroni correction, Simes method (Simes 1986), or Benjamini–Hochberg procedures (Benjamini and Hochberg 1995).

Interpretation of Estimated a and x Parameters

The likelihood for the alternative model is maximized over the parameters A and x, where, in our formulation for biallelic balancing selection in the previous subsections, x represents the presumed equilibrium minor allele frequency, and A decides the rate of exponential decay for the probability of two sites being linked, which essentially describes the influence of balancing selection on neutral sites of varying distance away from the test site. After optimizing over this parameter space, the parameter values under the optimal likelihood, $\widehat{A}$ and $\widehat{x}$, provide information on the nature of detected genomic footprints. The value of $\widehat{x}$ should reflect the enriched minor allele frequency across the region. Note that not all mechanisms for balancing selection will maintain the balanced alleles at fixed frequencies (Asmussen and Basnayake 1990; Bergland et al. 2014), so $\widehat{x}$ rather represents the value around which our model presumes the allele frequencies across the region are enriched. Therefore, we advise that caution be used when interpreting $\widehat{x}$ as the equilibrium frequencies without further information about the potential mechanisms that may have acted to maintain the polymorphisms.

Meanwhile, $\widehat{A}$ describes the rate of the exponential decay of the probability $\alpha_A(d) = \exp(-Ad)$ of the two loci being linked and should intuitively be informative of the impact of balancing selection on nearby neutral sites. The smaller the $\widehat{A}$, the wider the footprint would be, and likely the younger the balanced polymorphism. However, multilocus balancing selection can also give rise to wide footprints (Barton and Navarro 2002; Navarro and Barton 2002; Tennessen 2018), which could induce small $\widehat{A}$ values. Furthermore, a large A reduces the number of informative sites that yield meaningful likelihood ratios and can thus also occur when data in the examined area fit the alternative model poorly. Therefore, we advise only comparing the $\widehat{A}$ values among regions with reasonably high composite likelihood ratios, and that caution is used when making inferences from these values as they do not map to an explicit evolutionary model.

Results

Performances on Simulated Data

We simulated 50-kilobase (kb)-long sequences using SLiM3.2 (Haller and Messer 2019), under the three-species demographic model (supplementary fig. S1, Supplementary Material online) inspired by the demographic history of great apes (see Materials and Methods), and extensively evaluated the performances of all five B statistic variants. We also compared the B statistics to the summary statistics β, β^*, HKA, NCD2, and $\beta^{(2)}$, which are, respectively, analogs to B_0, $B_{0,MAF}$, B_1, $B_{2,MAF}$, and B_2, and to the likelihood statistics T_1 and T_2, which are, respectively, analogues to B_1 and B_2.

Robust High Power under Varying Window Sizes

We first examined the robustness of the B statistics to overly large window sizes, under a scenario of strong heterozygote advantage (selective coefficient $s = 0.01$ with dominance coefficient $h = 20$) acting on a mutation that arose 7.5×10^4 generations prior to sampling, with all sites flanking the selected locus evolving neutrally. Because BetaScan (Siewert and Voight 2017, 2020) (which implements the standardized and nonstandardized β, β^*, and $\beta^{(2)}$ statistics, among which we only consider the standardized) operates on windows of fixed physical length, we adopted window sizes of 1, 1.5, 2.5, 3, 5, 10, 15, 20, and 25 kb for all summary statistics and B statistics. The T statistics were applied on windows with matching expected numbers of informative sites. Supplementary note 2, Supplementary Material online, details the calculation for matching the number of informative sites to physical length of a genomic region.

To reduce potential stochastic fluctuations in the number of true positives when the false positive rate is controlled at a low level, we examined the area under a partial curve with no greater than a 5% false positive rate (hereafter referred to as "partial AUC"). As shown in figure 2A (see split views for separate groups of statistics in supplementary fig. S2, Supplementary Material online), under optimal window sizes for most other statistics, all variants of B statistics display substantial partial AUCs comparable to that of the respective T statistic variant, which has outperformed other equivalent summary statistics in most previous simulation studies (DeGiorgio et al. 2014; Siewert and Voight 2017, 2020; Bitarello et al. 2018; Cheng and DeGiorgio 2019). Most remarkably, as the window size increases, while all other statistics exhibit drastic decays in power, the powers of all variants of the B statistic only show minor decreases. In fact, when comparing the powers under 25-kb windows against those under optimal window sizes for each statistic, the powers of all statistics drop more than twice as much as B_1 and B_2 (fig. 2B). In comparison with each method's optimal performance, most statistics (except all B statistics and T_2, the model-based analog of B_2) lose more than 80% of their optimal power under the largest window size examined (fig. 2C). Although T_2 still retains considerably higher partial AUC compared with all other extant methods, it still decreases to a value substantially lower than that of B_2. Such robustness of B statistics to large windows is reasonable and expected, because the probability distribution of allele frequencies at sites far enough from the test site will match the genome-wide SFS, thereby contributing little to the overall likelihood ratio.

Among all statistics evaluated, we found that those considering polymorphism data only (i.e., B_0 variants and β variants) demonstrated relatively poor robustness to increases in window size. This result indicates that the detectable footprint of balancing selection in polymorphism data by itself

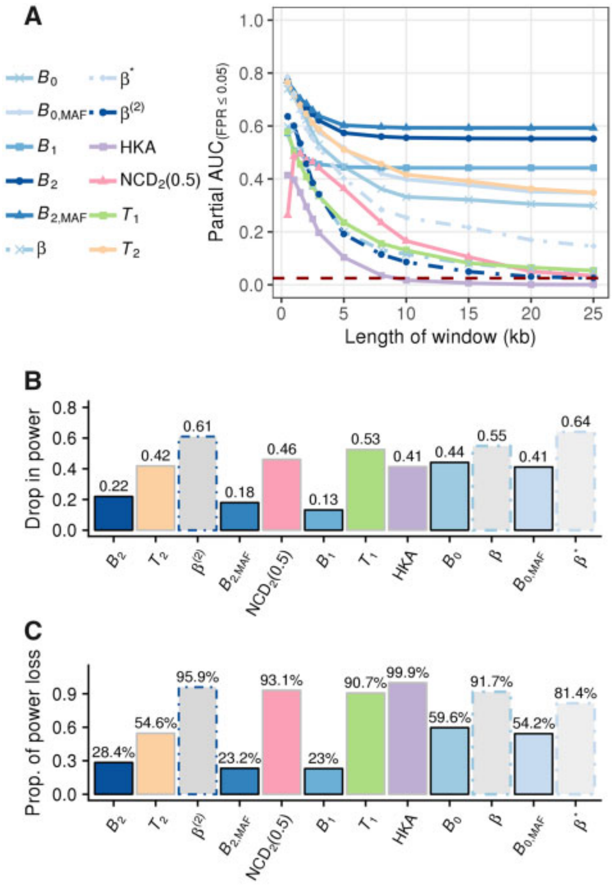

Fig. 2. (*A*) Partial AUC conditioned on false positive rates (FPRs) $\leq$5% (defined such that the maximal value is 1) as a function of window size measured in kilobases (kb) for *B* statistics (varying shades of blue), β statistics (dotted line with varying shades of blue), T_2 (orange), T_1 (green), HKA (purple), and NCD_2 (0.5) (pink), under a scenario in which a mutation undergoing ancient balancing selection (selective coefficient $s = 0.01$ and dominance coefficient $h = 20$) arose 15 Ma (assuming a generation time of 20 years). Statistics that consider the same input type share the same point shape. The dark red-dashed line marks the level of partial AUC expected at the $y = x$ line, or the baseline of randomly choosing between balancing selection and neutrality. (*B*) The amount of partial AUC lost, and (*C*) the proportion of the AUC loss as compared with the optimal value for each statistic when the window size increased from the optimum to 25 kb (e.g., largest evaluated).

may decay faster than other types of information, and that incorporating substitution data may help improve robustness to large windows.

Considering that the powers of all *B* statistics stabilize at a fixed level as the window size increases (fig. 2), we permit the *B* statistics to employ all informative sites on a chromosome. However, to reduce computational load, we only consider sites with mixing proportion $\alpha_A(d) \geq 10^{-8}$ for each value of *A* considered during optimization, which does not create discernible differences in performance from when all data are considered (supplementary fig. S3, Supplementary Material online). However, to ensure that other methods still display considerable power for their comparisons, we applied the summary statistics with their optimal window sizes of 1 kb, and *T* statistics with numbers of informative sites expected in a 1-kb window (see Materials and Methods), unless otherwise stated.

High Power for Detecting Balancing Selection of Varying Age and Selective Strength

Next, we explored the powers of *B* statistics when the selective strength *s*, equilibrium frequency (controlled by the dominance parameter *h*), and the age of balancing selection vary. Specifically, we examined scenarios where the selective coefficients were moderate ($s = 0.01$, fig. 3*A* and *C–E*) or weak ($s = 10^{-3}$, fig. 3*B*), and when the equilibrium frequency of the minor allele is ~0.5 ($h = 20$, fig. 3*A* and *B*), 0.4 ($h = 3$, fig. 3*C*), 0.3 ($h = 1.75$, fig. 3*D*), or 0.2 ($h = 1.33$, fig. 3*E*). Across all scenarios considered, T_2 and β^* show the highest power for old balancing selection. The best-performing *B* variants, B_2 and $B_{2,MAF}$, display high power as well, and are often comparable to that of the $\beta^{(2)}$ statistic. The power of B_1 is also similar to HKA, which is its summary statistic analog. Furthermore, we noticed that *B* statistics exhibit superior power for younger balanced alleles, particularly when balancing selection is more recent than 2 × 10^5 generations, and when the equilibrium frequency does not equal to 0.5 (supplementary fig. S4, Supplementary Material online). For older selected polymorphisms, although several statistics outperform *B* statistics, it is important to point out that all previous methods were provided optimal window sizes, whereas *B* statistics were set to use all sites with considerable $\alpha_A(d)$, under which they show lower power than when window sizes are optimized (fig. 2*A* and supplementary fig. S2*C*, Supplementary Material online). This difference in performance between previous methods applied with their optimal window sizes and *B* statistics can also explain the seemingly inferior performance of the two B_0 variants when compared with the analogous β statistics, as the B_0 variants lose more power than other *B* variants when computed on extended windows. When applied with the same window size, however, B_0 outperforms β by a large margin (fig. 2*A* and supplementary fig. S2*C*, Supplementary Material online). Nevertheless, these results give us confidence that *B* statistics have generally high power to detect young and old balancing selection, even when adopting large windows.

Robustness to Recombination Rate Variation and Elevated Mutation Rates

Despite their flexibility in window size and high power for detecting balancing selection, model-based methods, such as the *T* and *B* statistics, incorporate recombination distances in their inference framework, and can therefore be especially susceptible to potential inaccuracies in input recombination maps. Additionally, because many approaches for detecting balancing selection aim to identify genomic regions with increased genetic diversity, the elevation of mutation rates is also a common and potent confounding factor for detecting balancing selection (Charlesworth 2006; Cheng and DeGiorgio 2019; Siewert and Voight 2020).

To test their robustness to inaccurate recombination rates, we applied *B* and *T* statistics on simulated sequences with

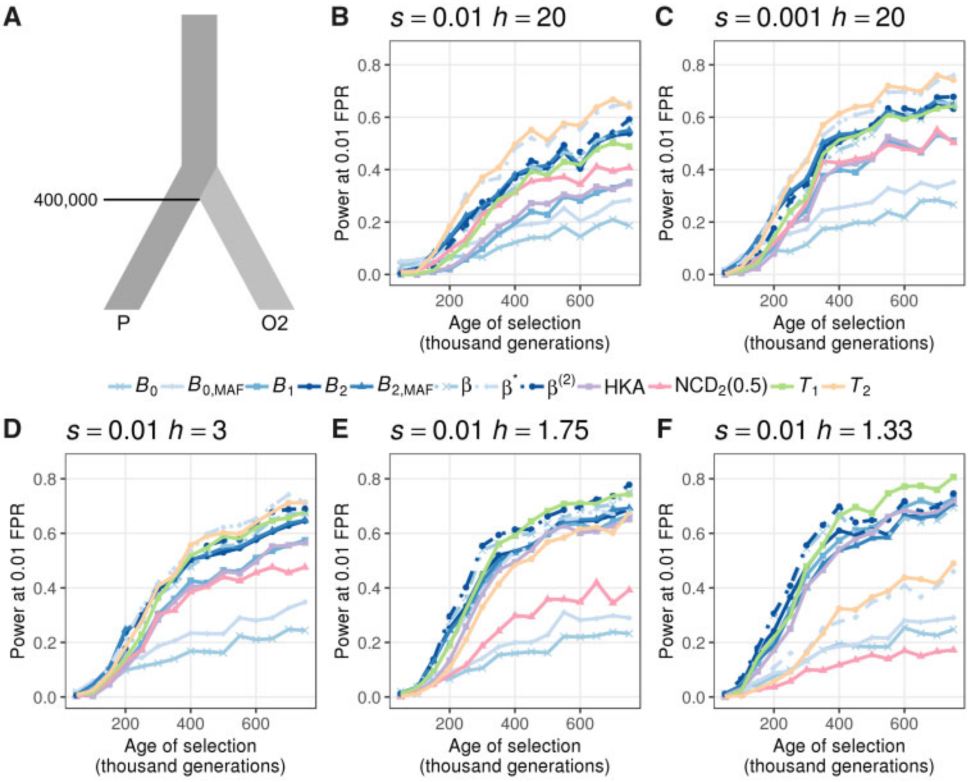

Fig. 3. Ability to detect balancing selection for different heterozygote advantage scenarios. (*A*) Demographic model relating the ingroup (P) and outgroup (O2) populations, with one sample from O2 used as the outgroup sequence. (*B*–*F*) Powers at a 1% false positive rate (FPR) for each statistic as a function of age of the allele undergoing balancing selection for different selection (s) and dominance (h) coefficients. The scenarios considered are (*B*) $s = 0.01$ with $h = 20$, (*C*) $s = 0.001$ with $h = 20$, (*D*) $s = 0.01$ with $h = 3$, (*E*) $s = 0.01$ with $h = 1.75$, and (*F*) $s = 0.01$ with $h = 1.33$. Note that the equilibrium frequencies for panels (*D*), (*E*), and (*F*) are 0.4, 0.3, and 0.2, respectively.

uneven recombination maps (10^2-fold fluctuations in recombination rates; see Materials and Methods). When the sequences evolve neutrally, neither approach is misled (supplementary figs. S5 and S6, Supplementary Material online). When the fluctuation in recombination rate is even more drastic (e.g., 10^4-fold instead of 10^2), all methods tend to report fewer false signals than they would under a uniform map (supplementary figs. S7 and S8, Supplementary Material online). This result suggests that the misleading effects of inaccurate recombination maps are limited.

To examine their robustness against unexpected mutation rate variation, we next simulated a 10-kb mutational hotspot at the center of the 50-kb sequence with a mutation rate five times higher than original and surrounding rate μ and applied each statistic with parameters derived from the original neutral replicates with constant mutation rate μ across the entire sequence. All methods exhibit considerable robustness against this regional increase of mutation rate (supplementary figs. S9 and S10, Supplementary Material online).

We further considered an elevated mutation rate of 5μ across the entire 50-kb sequence and reexamined the robustness of each method. As expected, most statistics display substantially inflated proportions of false signals (i.e., reported signals of balancing selection from sequences neutrally evolving with 5μ mutation rate; supplementary figs. S11A, S11D, and S12, Supplementary Material online). Among them, the B_2 statistic reports the least proportion of false signals, followed by the B_1 statistic. Meanwhile, at low false positive rates, B_2 and $B_{2,MAF}$ statistics report higher proportions of false signals than T_2, their coalescence model-based analog, whereas B_1 outperformed T_1. Additionally, all statistics that consider only polymorphism data, namely the B_0, $B_{0,MAF}$, β, and β^* statistics, are substantially misled. The $\beta^{(2)}$ statistic, albeit taking substitutions into account, also displays surprisingly high proportions of false signals.

We next explored how the regional mutation rate elevation in the genome could affect the detection of balancing selection. To this end, we mixed neutral sequences evolving with 5μ mutation rates with those with the original μ mutation rate at varying proportions (5%, 10%, 25%, or 50%) and used these mixed pools of neutral sequences as the “whole genome” to compute their SFSs, interspecies coalescent times, and polymorphism-substitution ratios to inform T, B, β, and HKA statistics of the neutral variation levels. We then scanned these sequences with summary statistics adopting 1-kb windows, T statistics adopting 12-site windows, and B statistics using the whole sequence. We found that as the proportion of fast-mutating neutral sequences increases, most methods show substantially compromised powers (supplementary fig. S13, Supplementary Material online). Among them, however, T_2 and NCD consistently exhibit considerable power throughout all scenarios examined, followed by T_1, $B_{2,MAF}$, B_1, and B_2, which still retain some power despite substantial drops. Meanwhile, the methods that do not effectively utilize substitutions, that is, B_0, $B_{0,\mathrm{MAF}}$, and β statistics, almost lose all the power. This is consistent with previous results, suggesting that the absence of substitution renders methods for detecting balancing selection susceptible to the confounding effects of unexpected mutation rates.

With knowledge of their robustness against unexpected mutation rate elevation, we further examined the powers of each method to detect balancing selection within sequences evolving with high mutation rates when they are correctly informed. That is, T and β statistics are provided the correct population-scaled mutation rate and interspecies coalescent time, and all except for B statistics adopt their optimal window sizes of 1 kb (60 informative sites for T statistics). We simulated sequences undergoing balancing selection that initiated 250,000 generations ago with a neutral mutation rate of 5μ across the simulated segment and applied summary and T statistics on the sequences mutating at a rate of 5μ, with their optimal window sizes under the correct mutation rate. Supplementary figure S14, Supplementary Material online, demonstrates that the powers of all methods are substantially higher than for the identical scenario with sequences evolving under the original neutral mutation rate μ (compared with fig. 3C and supplementary fig. S4C, Supplementary Material online). This improved detection ability likely results from the roughly 5-fold increase in the number of informative sites included within each window. The T statistics display lower areas under their receiver operating characteristic curves than their equivalent B statistics (supplementary fig. S14A, Supplementary Material online), and the $B_{0,\mathrm{MAF}}$ and $B_{2,\mathrm{MAF}}$ statistics perform substantially worse than their respective derived allele frequency counterparts B_0 and B_2. Moreover, as with other simulated scenarios, we find that the power of $B_{0,\mathrm{MAF}}$ is lower than others (supplementary fig. S14B, Supplementary Material online). However, when the window size for all summary statistics is expanded from the optimal 1 kb to a suboptimal 5 kb, their powers substantially decrease to levels similar to $B_{0,\mathrm{MAF}}$.

Robust Power under Realistic Demographic Models

The influence of demographic history was the major motivation for T statistics to adopt the genome-wide SFS instead of the coalescence-based constant-size neutral model as the null hypothesis, despite that the latter being nested under the alternative model for balancing selection used by the T statistics. This trade-off has endowed T statistics with considerable robustness to population size changes (DeGiorgio et al. 2014; Cheng and DeGiorgio 2019) but has also potentially compromised their robustness to large windows, as shown in "Robust High Power under Varying Window Sizes" subsection of the Results. For B statistics, however, because their null models both reflect the genome-wide SFS and are nested under the alternative models, they should exhibit considerable robustness to both oversized windows and demographic changes.

To evaluate their performances under recent population expansions and bottlenecks, we considered the demographic histories of contemporary European humans (Terhorst et al. 2017, CEU; supplementary fig. S15A, Supplementary Material online) and bonobos (Prado-Martinez et al. 2013, supplementary fig. S16A, Supplementary Material online; see details in Materials and Methods), respectively. The former have been extensively characterized (Lohmueller et al. 2009; Gravel et al. 2011; Terhorst et al. 2017) and therefore can reliably reflect the performance of each method under realistic scenarios. On the other hand, because we intend to apply the B statistics on bonobo genomic data, we are also interested in evaluating their performance under an inferred bonobo demographic model.

As previously described, we applied the B statistics with unlimited window sizes, whereas the other statistics were provided with smaller window sizes matching the theoretical size for a footprint of long-term balancing selection (see supplementary note 2, Supplementary Material online). Despite being provided disadvantageous window sizes, B statistics still demonstrate comparable to, and often higher power than, current summary statistic approaches, both under the human (supplementary fig. S15, Supplementary Material online) and the bonobo (supplementary fig. S16, Supplementary Material online) demographic models. Although T_2 has higher power than the B statistics, we note that the T statistics were operating with optimal window sizes, whereas the window sizes for B statistics are identified across a parameter range. When B_1 and B_2 are applied with identical window sizes as T_1 and T_2 (supplementary figs. S17 and S18, Supplementary Material online), the margins between their performances are no longer substantial. Additionally, we noticed that most statistics tend to have higher power for sequences evolving under the bonobo demographic history than under that of the Europeans (notice that the y-axes in supplementary figs. S15 and S16, Supplementary Material online, have different scales).

Robust Power under Varying Mutation Rates across Target and Outgroup Species

In addition to temporally varying population sizes, differing mutation rates between closely related species may also affect the performance of the coalescence-based T statistics, as they assume a uniform neutral mutation rate along the genealogy relating the lineages from the ingroup and outgroup species. Among great apes, for example, accumulating evidence suggests that humans have substantially lower mutation rates than other great apes (as reviewed by Scally and Durbin 2012).

To examine the behavior of each method when mutation rates of the target and outgroup species differ, we simulated a two-species demographic history, with the target and outgroup species, respectively, evolving at neutral rates $\mu = 1.2 \times 10^{-8}$ and $\mu = 2.5 \times 10^{-8}$ mutations per site per generation (see Materials and Methods for details). We introduced an adaptive mutation evolving under balancing selection at varying time points prior to sampling along this demographic history and examined the power of each statistic to detect balancing selection across a diverse array of selection parameters (supplementary fig. S19, Supplementary Material online).

Across all six combinations of selection parameters considered, we observe similar trends for each statistic when compared with simulations under the constant population size (fig. 3) and CEU (supplementary fig. S15, Supplementary Material online) demographic histories evolving with a constant neutral mutation rate. The T_2 statistic performs the best when $s = 0.01$ with $h = 20$ (supplementary fig. S19*A*, Supplementary Material online), under which the equilibrium frequency is closest to 0.5 and when heterozygotes are most advantageous. As the selective advantage *hs* and equilibrium frequency decrease, the margin between the powers of T_2 and B_2 shrinks and even reverses for all scenarios with small dominance *h* (supplementary fig. S19*C*–*F*, Supplementary Material online). Furthermore, methods based solely on polymorphism and substitution calls (i.e., T_1, B_1, and HKA) show improvements in power as the equilibrium frequency decreases, and some even outperform most of the other statistics (supplementary fig. S19*D* and *E*, Supplementary Material online). Statistics that ignore substitutions (i.e., B_0, $B_{0,\text{MAF}}$, β, and β^*), on the other hand, perform especially well for recent balancing selection with high heterozygote advantage (large *hs*; supplementary fig. S19*A* and *B*, Supplementary Material online). As the balanced alleles reach their equilibrium frequencies sooner when the selective advantage of heterozygotes (i.e., *hs*) is high, sequences with mutations of higher *hs* would have older footprints than those with mutations introduced at the same time but with lower *hs*. In this respect, it is understandable that B_0 and β variants outperform others only for selection with large *hs* that are introduced within 150,000 generations prior to sampling.

Based on this two-species model with diverging mutation rates, we further integrated changes in population size of the target species in accordance with the demographic history of the CEU (Terhorst et al. 2017, supplementary fig. S20, Supplementary Material online). From the four sets of selection parameters tested, we found that most methods exhibit lower power compared with those under constant population sizes (supplementary fig. S19, Supplementary Material online). This result is consistent with the lower powers under simulations with a constant mutation rate when the target population size evolves under the CEU demographic history (supplementary fig. S15, Supplementary Material online) compared with the setting in which the target evolves with constant size (fig. 3). Despite their lower powers in general, we still observe similar relative performances across statistics, with T_1 and B_1 exhibiting higher powers when the heterozygote advantage *hs* is small. Moreover, we found that $B_{2,\text{MAF}}$ shows superior power to B_2.

Reexamining Long-Term Balancing Selection in Human Populations

We applied B_2 on contemporary European (Europeans in Utah; CEU, supplementary fig. S22, Supplementary Material online) and west African (Yoruban; YRI, supplementary fig. S21, Supplementary Material online) human populations from the 1000 Genomes Project data set (1000 Genomes Project Consortium 2015) (see Materials and Methods) to reexamine the footprints of long-term balancing selection, which previous studies (DeGiorgio et al. 2014; Siewert and Voight 2017) have provided cases for reference. The most outstanding candidates in both scans localize in the HLA-D region (human leukocyte antigen, also known as major histocompatibility [MHC] class II region) (supplementary figs. S23 and S24), agreeing with previous findings (Sanchez-Mazas 2007; Leffler et al. 2013; DeGiorgio et al. 2014; Teixeira et al. 2015; Siewert and Voight 2017; Meyer et al. 2018; Bitarello et al. 2018). Within the HLA-D region, the B_2 scores computed for both populations show extraordinary peaks around *HLA-DQ* and *HLA-DP* gene clusters, although CEU (supplementary fig. S23, Supplementary Material online) scores remarkably higher on *HLA-DP* genes than YRI (supplementary fig. S24, Supplementary Material online). Echoing the critical roles of HLA-D genes in adaptive immunity, the gene *ERAP2* exhibits extraordinary scores in both populations (supplementary figs. S25 and S26, Supplementary Material online). This gene has been reported in past studies of balancing selection in humans (Andrés et al. 2009, 2010; Bitarello et al. 2018), and Andrés et al. (2010) demonstrated that its splicing variants can alter the level of MHC-I presentation on B cells. Additionally, we also observed high B_2 scores on *CADM2* (supplementary figs. S27 and S28, Supplementary Material online) and *WFS1* (supplementary figs. S29 and S30, Supplementary Material online), on which Siewert and Voight (2017) characterized potential nonsynonymous mutations segregating in both populations.

In addition to these previously characterized candidates, both scans display extreme B_2 scores on another two top-ranking regions in the T_2 scans by DeGiorgio et al. (2014): the *STPG2* gene (formerly named *C4orf37*; supplementary figs. S31 and S32, Supplementary Material online) and the *CCDC169–SOHLH2* (formerly named *C13orf38-SOHLH2*; supplementary figs. S33 and S34, Supplementary Material online) region, with *STPG2* particularly more outstanding in the scan of YRI than in CEU. Intriguingly, both these genes are associated with gametes. The *STPG2* gene encodes sperm-tail PG-rich repeat-containing protein 2, which, despite the paucity of literature that describes its function, is found in sperm (Uhlén et al. 2015). The high-scoring region on this gene harbors a number of tissue-specific expression quantitative trait loci (eQTLs) for its expression, especially in brain and reproductive tissues (GTEx Consortium 2017). The *SOHLH2* gene, on the other hand, encodes the transcription factor "spermatogenesis- and oogenesis-specific basic helix–loop–helix-containing protein 2," which plays important roles in both spermatogenesis and oogenesis (Toyoda et al. 2009; Suzuki et al. 2012). We observed drastically elevated B_2 scores (supplementary fig. S33, Supplementary Material online) across an extended region upstream of *SOHLH2* that covers the naturally occurring *CCDC169–SOHLH2* read-through transcript (as introduced in RefSeq database; O'Leary et al. 2016). Similar to *STPG2*, this region also features numerous eQTLs for the expression of *SOHLH2*, especially in endocrine glands, brain, and reproductive tissues (GTEx Consortium 2017).

Other regions with outstanding peaks shared by both scans include the genes *CPE* (supplementary figs. S35 and S36, Supplementary Material online) and *MYOM2* (supplementary figs. S37 and S38). *CPE* encodes carboxypeptidase E, a key enzyme for synthesizing peptide hormones such as insulin and oxytocin, and its mutant mice strain (*Cpe*^{fat}) exhibits endocrinic disorders such as obesity and infertility (Naggert et al. 1995). *MYOM2* encodes endosacromeric cytoskeleton M-protein 2, which serves as a structural component of muscle tissues (van der Ven et al. 1999). Both genes harbor eQTLs reported by GTEx Consortium (2017) around the high-scoring regions.

Probing for Footprints of Balancing Selection in Bonobo Genomes

We further applied the B_2 statistic on the variant calls of 13 bonobos (Prado-Martinez et al. 2013) lifted over to human genome assembly GRCh38/hg38. Only biallelic single nucleotide polymorphisms (SNPs) were considered, and substitutions were called using bonobo panPan2 reference sequence (Prüfer et al. 2012), with the human sequence as the ancestral state. Stringent filters were applied to remove repetitive regions and regions with poor mappability (see Materials and Methods). We observed many genomic regions with outstanding B_2 scores (fig. 4), which include both the *MHC-DQ* and *MHC-DP* genes and a few novel candidates.

Among the outstanding peaks, the top two cover the *MHC-DQA1*, *MHC-DQB1*, *MHC-DPA1*, *MHC-DPB1*, and *MHC-DPB2* genes, which harbor all the top 0.01% B_2 scores (fig. 5A). Such high scores can be explained by both the elevated proportion of polymorphic sites, 0.299 as compared with the genome-wide proportion of 0.237 (fig. 5B; note that genes are labeled based on human hg38 genome annotations), as well as the enrichment of polymorphic sites with moderate minor allele frequencies (fig. 5C). Furthermore, the region exhibits a multimodal SFS, which may correlate to the multiple B_2 peaks observed in the region.

In addition to the *MHC-DQ* and *MHC-DP* genes, *KLRD1* also presents prominent B_2 scores (supplementary fig. S39, Supplementary Material online) on its first intron. This gene expresses a natural killer (NK) cell-surface antigen, also known as CD94, and plays a pivotal role in viral defense. Unlike the region covering *MHC-DQ* genes, the minor allele frequencies at polymorphic sites around the *KLRD1* region are clearly enriched near a frequency of 0.45, instead of the multimodal distribution observed around the *MHC-DQ* genes. We also found other high-scoring regions associated with innate immunity, such as the gene *GPNMB* (supplementary fig. S40, Supplementary Material online), gene *LY86* (supplementary fig. S41, Supplementary Material online), and the intergenic region between *BPIFB4* and *BPIFA2* (supplementary fig. S42, Supplementary Material online).

Another interesting candidate is the pain perception gene *SCN9A* (supplementary fig. S43, Supplementary Material online), on which the highest scores overlap with the transcript of its antisense RNA gene that regulates its expression. Instead of enriching toward a single value, the minor allele frequencies at the polymorphic sites across the region are dispersed, with at least two modes (approximate minor allele frequencies of 0.25 and 0.40). This finding may correlate with the multiple peaks identified around this region, which may be sensible given the large number of exons covered. Similarly, the antisense RNA gene *ARHGEF26-AS1* (supplementary fig. S44, Supplementary Material online) harbors high B_2 scores, with allele frequencies enriched around 0.15 and 0.45. Other notable candidates include *PDE1A* (supplementary fig. S45, Supplementary Material online), which encodes a pivotal enzyme in cellular Ca^{2+}- and cyclic nucleotide signaling (Michibata et al. 2001). This gene has multiple splicing variants and plays roles in both neurodevelopment (Pekcec et al. 2018) and sperm functionality (Lefièvre et al. 2012). A few other genes scoring in the top 0.05% are also involved in spermatogenesis or gamete functionality while serving other important functions, such as a Ca^{2+}/calmodulin-dependent protein kinase gene *CAMK4* (supplementary fig. S46, Supplementary Material online; Sikela et al. 1990) and a otherwise cancer-related gene *SUSD2* (supplementary fig. S47,

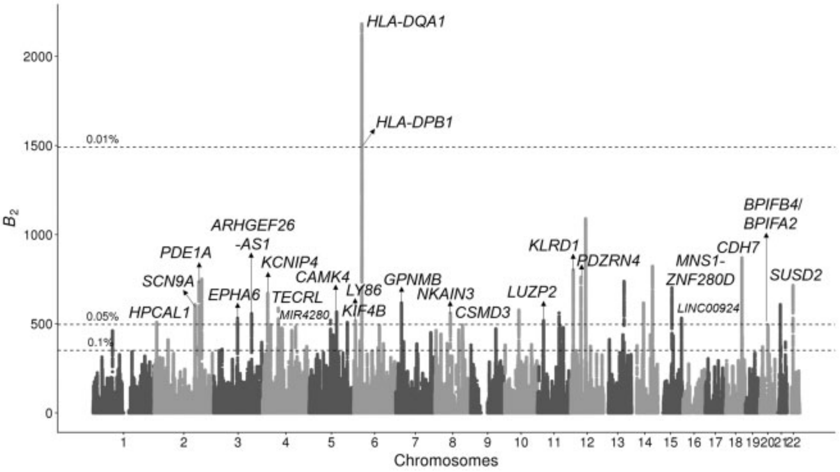

Fig. 4. Manhattan plot displaying B_2 scores across the 22 human autosomes for which the bonobo genomic data were mapped, with the candidates scoring in the top 0.05% annotated. RNA genes are annotated with smaller fonts. Horizontal dotted lines represent cutoff scores for the top 0.1%, 0.05%, and 0.01% across the genome. Peaks higher than 0.05% cutoff but without annotations do not have neighboring protein-coding regions within a 250-kb radius.

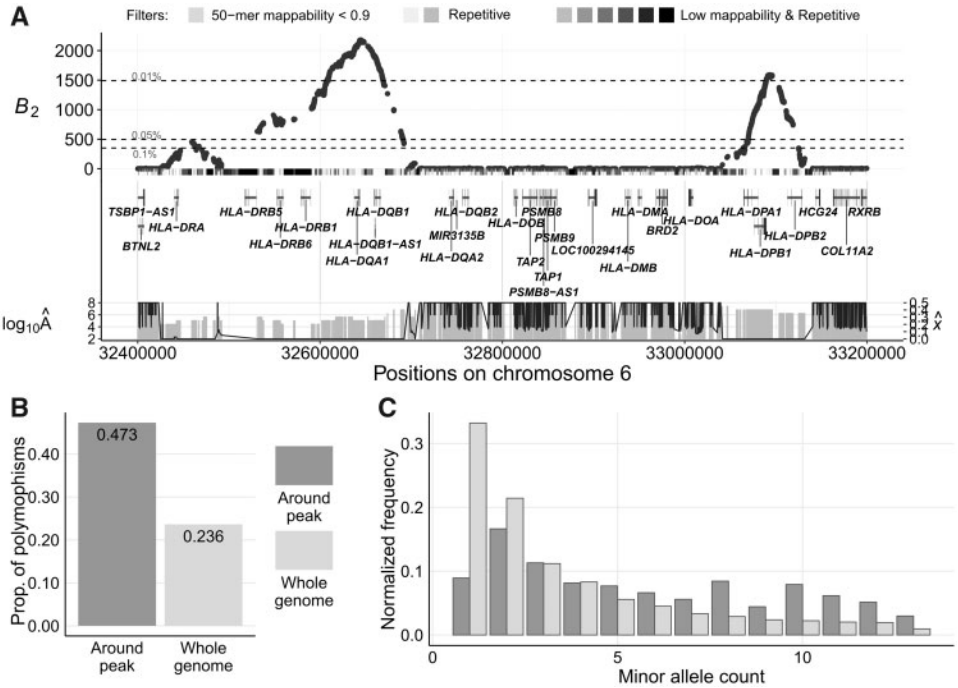

Fig. 5. Evidence for balancing selection on *MHC-DQ* and *MHC-DP* genes in bonobos. Note that the plotted gene names are based on the annotations of human hg38 reference genome. (*A*) B_2 scores across the genomic region on chromosome 6 surrounding the *MHC-DQ* and *MHC-DP* genes. The gray bars directly under the B_2 scores represent the masked regions, as well as the features in these regions. The darker the shade, the greater number of types of repetitive sequences (e.g., RepeatMasker mask, segmental duplication, simple repeats, or interrupted repeats) overlapping the region. Vertical gray bars below display the estimated equilibrium minor allele frequency $\hat{x}$ for each maximum likelihood ratio B_2, and the black line traces the value for the respective inferred footprint size $\log_{10}(\hat{A})$ (*B*) Proportion of informative sites that are polymorphic in the 800-kb region centered on the peak compared with the whole-genome average. (*C*) Minor allele frequency distribution in the 500-kb region centered on the peak compared with the whole-genome average.

Supplementary Material online; Harichandan et al. 2013; Watson et al. 2013).

Discussion

In this study, we introduced a novel set of composite likelihood ratio statistics—B_2, $B_{2,\mathrm{MAF}}$, B_1, B_0, and $B_{0,\mathrm{MAF}}$ —to robustly detect footprints of balancing selection with high power and flexibility. The *B* statistics are based on a mixture model creating a proper nested likelihood ratio test, which helps them overcome the common susceptibility to oversized windows held by current methods. We have extensively evaluated their performances on simulated data compared with current state-of-the-art methods and have demonstrated the superior properties of the *B* statistics under various scenarios. We reexamined balancing selection in human populations (1000 Genomes Project Consortium 2015) and recovered well-established candidates, including the HLA-D genes and *ERAP2*. We further applied B_2 onto the genomic data of bonobos (Prado-Martinez et al. 2013) and uncovered not only the *MHC-DQ* and *MHC-DP* gene cluster but also intriguing candidates that are involved in innate immunity, neurosensory development, and gamete functionality.

Evaluating the Performance of *B* Statistics through Simulations

In our simulation study, the *B* statistics showed remarkable robustness to large window sizes, with only minor decays in power under oversized windows, whereas other methods exhibited large declines in power. Moreover, even when considering all data available as input (i.e., the most disadvantageous window size) all variants of *B* statistics still exhibit comparable power to extant methods and displayed satisfactory performance across varying types and strengths of balancing selection. Under scenarios with confounding factors, such as high mutation rate and nonequilibrium demographic history, the *B* statistics demonstrated satisfactory robustness as well.

The robustness against varying window sizes is of particular interest in this study, not only because it ensures high power under large windows, but it also allows the statistics to augment the size of genomic regions from which they make meaningful inferences. This flexibility grants a key advantage over previous methods that require the window size to be fixed throughout the scan in order to yield comparable results across the genome. In particular, because many factors (such as recombination rates) can influence the footprint size of balancing selection, it is not ideal to adopt a fixed window size for a whole-genome scan based on a uniform population-scaled recombination rate, and *B* statistics naturally accommodate such variability across the genome.

Admittedly, in practice, as the genomic region considered in the tests expands, nonneutral sites will inevitably be included. This indeed violates our assumption that the test locus is surrounded by neutral sites only. Nonetheless, because both positive and purifying selection reduce the presentation of sites with intermediate frequencies (Tajima 1989; Braverman et al. 1995; Fay and Wu 2000; Bamshad and Wooding 2003), their effect on the SFS is in general opposite to the features expected from balancing selection. This suggests that including such sites in the window is unlikely to hamper the power to detect balancing selection. Meanwhile, when multiple sites in the considered region undergo balancing selection, the pattern of polymorphisms across the region will indeed differ from that in regions with a single selected locus. We will discuss the effects of such multilocus balancing selection in the subsequent subsection Performance of Single-Locus Methods on Multilocus Balancing Selection.

One important consideration is that, so far our simulation study (as well as previous ones by DeGiorgio et al. 2014; Bitarello et al. 2018; Siewert and Voight 2020) only evaluates the method performance in the context of single-locus heterozygote advantage. For many other balancing selection mechanisms, such as negative frequency-dependent selection (Asmussen and Basnayake 1990) and periodic environmental fluctuations (Bergland et al. 2014), a stable equilibrium cannot be guaranteed (Cockerham et al. 1972; Asmussen and Feldman 1977; Ginzburg 1977). In nonoverdominance settings for which particular equilibrium frequencies indeed exist, the balanced alleles are still maintained near these fixed frequencies, thereby satisfying the general assumptions of the statistical models underlying our B statistics. Moreover, when such intrinsic equilibrium frequencies do not exist, allele frequencies may still fluctuate around some mean values. Even if such mean values are unattainable, there will still persist an enrichment of sites with intermediate frequencies, thereby presenting characteristic footprints of balancing selection. We therefore believe that our mixture model framework should still have high power to detect footprints of nonoverdominance balancing selection, and that overall our results have comprehensively characterized the promising performance of the B statistics.

Confounding Effects of Mutation Rate or Recombination Rate Variation

In our simulation study, sequences with a central 10-kb mutational hotspot did not mislead methods as much as those with the mutation rate elevated across the entire sequence (supplementary fig. S9, Supplementary Material online). This result may seem counterintuitive at first, as a smaller region of increased mutation rate may better resemble the footprints of long-term balancing selection. However, upon a closer examination of the site frequency spectra and proportions of polymorphic sites (supplementary fig. S48, Supplementary Material online), sequences with an extended region of high mutation rate exhibit a greater departure in these features under scenarios with no elevated mutation rate than for scenarios with a central mutational hotspot. Specifically, these sequences have more sites with high derived allele frequencies and a higher proportion of polymorphic sites overall (supplementary fig. S48*B*, Supplementary Material online), likely resulting from the recurrent mutation on sites that were originally substitutions. The increase is also more profound on sites with high derived allele frequency. For example, the proportions of sites with derived allele frequency of 0.96 increased by almost 2-fold from ~0.00104 to 0.00190, and the proportions of sites with derived allele frequency of 0.98 increased by almost 3-fold from 0.00105 to 0.00273. By contrast, the difference in scale between the proportions of polymorphisms (0.182 vs. 0.189) is minor. The larger fold-change in the proportions of high-frequency polymorphisms (i.e., sites with $k = n - 1,\ \ n - 2$, and $n - 3$ derived alleles) relative to that of substitutions ($k = n$ derived alleles) could explain the more profound inflation in power for the statistics relying only on information at polymorphic sites. Similarly, after folding the SFS, the large changes in the proportions of low-frequency alleles were substantially mitigated, echoing the superior performance of $B_{2,\mathrm{MAF}}$ and β relative to their unfolded counterparts.

Another unexpected result from the simulations of elevated mutation rate is the drastic inflation of false signals reported by β statistics (supplementary fig. S11, Supplementary Material online), which can also be observed in the nonstandardized β statistics (supplementary fig. S49, Supplementary Material online). Although Siewert and Voight (2020) tested their power to detect balancing selection under high mutation rate, it was unexplored whether their β statistics would misclassify highly mutable neutral sequences as those undergoing balancing selection, and our results show that they could be easily misled. However, we further found that the performances of the standardized β statistics largely improve when provided with the correct mutation rate and divergence time (supplementary fig. S49*B*, Supplementary Material online). This result partly confirms the superiority of standardized β statistics over the unstandardized ones. It also suggests that β statistics are considerably susceptible to the confounding effect of mutation rate elevation, and that their performance relies highly on the accuracy of the provided mutation rate. Instead of using a constant mutation rate for the entire scan, we propose that providing locally inferred population-scaled mutation rates θ may help improve the robustness of β statistics. Indeed, when we instead estimate θ using the mean pairwise sequence difference $\widehat{\theta}_\pi$ (Tajima 1983) for each replicate and provided BetaScan the respective inferred value as the θ parameter, the standardized statistics no longer report as many false signals (supplementary fig. S49*C*, Supplementary Material online). However, we also observed that providing a locally inferred θ estimate compromises the power of standardized β statistics to detect balancing selection, both under normal (i.e., μ) and elevated (i.e., 5μ) mutation rates (supplementary figs. S50 and S51, Supplementary Material online, respectively), especially for the unfolded β^* and $\beta^{(2)}$ statistics. This result is probably because, in addition to an elevation in mutation rate, the locally inferred θ can also be inflated by footprints of balancing selection, thereby decreasing the β statistic's sensitivity.

In contrast to mutation rate variation, all statistics are robust to recombination rate variation, with B_0 and $B_{0,\mathrm{MAF}}$ reporting substantially fewer false signals than the others (supplementary fig. S5, Supplementary Material online).

This robustness to recombination rate variation may be explained by the high similarity in the SFS and proportion of polymorphic sites to sequences evolving under a uniform recombination rate (supplementary fig. S52, Supplementary Material online).

Effect of Multiple Testing on Sequences with High Mutation Rates

Because B, T, and β statistics are computed on every informative site, as suggested by Cheng and DeGiorgio (2019), multiple testing can account for some inflation in their powers because sequences with a higher mutation rate will have a greater number of informative sites. To evaluate the effect of multiple testing for sequences with high mutation rates, we downsampled the test sites (see Materials and Methods) such that the number of test scores being computed approximately matches that under the original mutation rate μ. Although all statistics show varying levels of improvements in performance (supplementary figs. S11*B*, *C*, *E*, and *F*, Supplementary Material online), some still report high proportions of false signals, especially all β statistics and $B_{0,\text{MAF}}$. That is, multiple testing cannot account for all the factors that drive these statistics to misidentify features of elevated mutation rates as footprints of balancing selection. This result corresponds to the fact that both the SFS and the density of polymorphic sites are altered under scenarios with extended regions of elevated mutation rate (supplementary fig. S48, Supplementary Material online), likely due to recurrent mutation.

Furthermore, we observed that both before and after downsampling, the T statistics report fewer false signals than their respective B statistic analogs. One potential factor behind their marginally superior performance may be that T statistics perform tests on fixed numbers of informative sites, instead of genomic regions measured by physical lengths (as did B statistics and the summary statistics). For T statistics, the size of the genomic region covered by the same number of informative sites would be much narrower under rapidly mutating sequences than in sequences with the original mutation rate. This means that the resulting T scores in either scenario are reflective of the levels of variation for sequences with drastically different lengths. To account for this factor, we provide B_1 and B_2 with informative site-based windows identical to that of T statistics and reexamined their performances (supplementary figs. S53 and S54, Supplementary Material online). After matching the windows, B_1 and B_2 variants in turn display higher robustness than T_1 and T_2 to elevated mutation rates, suggesting that B statistics are at least comparably robust to T statistics. Meanwhile, we also matched the window size for B_0 variants and β to gauge the effect of adopting large windows on the proportions of false signals from B_0 variants. When B_0 scans the sequences with 1-kb windows, though there is an increase in the resulting number of false signals (supplementary figs. S53*A* and S54*C*, Supplementary Material online), at a 1% false positive rate the proportions of false signals for the two B_0 variants only increase by $<$0.1 and are still substantially lower than that of β and β^* (supplementary figs. S53*B* and S54*C* and *D*, Supplementary Material online).

Comparing the B Statistics with the T Statistics

Because the T statistics of DeGiorgio et al. (2014) have previously been the only model-based approach for the detection of long-term balancing selection from polymorphism data in a single species, the comparison between the model-based B and T approaches is particularly intriguing for researchers with empirical data suitable for the application of either. The T statistics are based on an explicit coalescent model (Hudson and Kaplan 1988; Kaplan et al. 1988) and have been shown to have superior power to a number of other methods in previous studies (DeGiorgio et al. 2014; Siewert and Voight 2017, 2020; Bitarello et al. 2018; Cheng and DeGiorgio 2019), consistent with our simulation results. The B statistics, on the other hand, employ a mixture model, where the component modeling balancing selection is not based on an explicit evolutionary model, but nevertheless shows impressive performance on simulated data, as the shape of the distribution of allele frequencies is similar to what might be expected under balancing selection. The often superior performances of both approaches over summary statistics are understandable, as both utilize the genomic spatial distribution of genetic diversity in their inferences.

However, within the T statistic framework, the model for the null hypothesis (neutrality) is not nested in the alternative hypothesis (balancing selection). Although the T_1 and T_2 statistics could have adopted nested models by employing the standard neutral coalescent as the model for the null hypothesis, doing so would increase susceptibility to demographic factors, which can also alter the genome-wide SFS. To better account for these factors, DeGiorgio et al. (2014) instead employed the genome-wide distribution of genetic variation to compute probabilities under the null hypothesis of neutrality. This explains the substantial decay in power for both T statistics as the window size increases (fig. 2 and supplementary fig. S2*A* and *B*, Supplementary Material online), as well as its robust performance under varying sized demographic models (DeGiorgio et al. 2014; Cheng and DeGiorgio 2019, supplementary figs. S15 and S16, Supplementary Material online). In contrast to the T statistics, the null model for B statistics (which also employs the genome-wide SFS) is nested within the alternative, due to their mixture model framework. This feature mitigates the biases introduced by sites far from the test site, while simultaneously accounting for demographic factors. Consequently, the B statistics display robust performance under oversized windows and realistic demographic models in our simulations (fig. 2 and supplementary figs. S2, S15, and S16, Supplementary Material online).

Another advantage of the B statistics over the T statistic approach, especially for B_2 compared with T_2, is the computational load. Because the probability distribution of allele frequencies under the Kaplan–Darden–Hudson (Kaplan et al. 1988) model is difficult to compute, the T_2 statistic relies on previously generated sets of simulated site frequency spectra over a grid of equilibrium frequencies $x \in \{0.05, 0.10, \ldots, 0.95\}$ for each distinct sample size n and recombination distance *d*. Generation of such frequency spectra is computationally intensive, and the load increases substantially with the increase in sample size, thereby limiting the application of T_2 to data sets with larger sample sizes. However, this is not a limitation of B_2, as the SFS under balancing selection is determined simply as a mixture of the given genome-wide distribution of allele frequencies and a statistical distribution with closed-form solutions whose computational cost is minor, and only increases linearly with the sample size. Moreover, the rapid computation of this spectrum permits a finer grid of equilibrium frequencies *x* to be interrogated.

Considering Multiallelic or Multilocus Balancing Selection

Both model-based approaches employed by the *T* and *B* statistics assume that balancing selection acts on a single biallelic locus. Whereas this case may be the most intuitive and simplistic scenario to model and simulate, many well-established empirical examples of balancing selection—such as the MHC locus in animals (Wills 1991; Hedrick 2002), the ABO blood group in primates (Saitou and Yamamoto 1997; Fumagalli et al. 2008; Ségurel et al. 2012; Leffler et al. 2013), and the plant self-incompatibility locus (Charlesworth et al. 2000)—feature multiple alleles balanced across an extended genomic region. It therefore brings into question how these methods perform on genomic regions evolving under multiallelic or multilocus balancing selection, and whether current frameworks can be extended to consider these more complicated cases of balancing selection.

Extending Mixture Models to account for Multiallelic Balancing Selection

There exist theoretical models of multiallelic balancing selection based on the coalescent (Hey 1991; Muirhead and Wakeley 2009). However, possibly due to computational constraints, such models have not been implemented within a likelihood-ratio framework for detecting the footprints they characterize. Here, instead of following DeGiorgio et al. (2014) to compute the densities of polymorphisms and substitutions or to approximate the SFS using simulations under an explicit coalescent model, our mixture models can be readily extended to account for multiallelic balancing selection at a single locus without the extensive computational burden of coalescent-based approaches that integrate selection. Specifically, we consider samples with multiple balanced alleles as following multinomial distributions (see supplementary note 1, Supplementary Material online), and henceforth use the mixture models to approximate the SFS at biallelic neutral sites that are linked to a selected locus with $m \in \{2, 3, 4, \ldots\}$ balanced allelic classes. This extension is also implemented in our BalLeRMix software, with the special case of $m = 2$ reducing to the model introduced in the Model Description section.

To simulate single-locus multiallelic balancing selection, we employed SLiM version 3.3, which can simultaneously incorporate the four standard nucleotides of DNA and thus allows these distinct nucleotides to coexist at the same site. We introduced two, three, or four distinct mutations with fitness parameters $s = 0.001$ and $h = 20$ in each simulated replicate 500,000 generations in the past to examine the relative performances of *T*, biallelic *B*, and multiallelic *B* statistics. Under this fitness scheme, the equilibrium frequencies when two, three, or four alleles are balanced in the population are approximately $(1/2, 1/2)$, $(1/3, 1/3, 1/3)$, or $(1/4, 1/4, 1/4, 1/4)$, respectively (see Materials and Methods for details). As the number of balanced alleles assumed by *B* statistics (i.e., parameter *m*) increases, the powers of *B* statistics barely change when two (supplementary fig. S55*A*–*C*, Supplementary Material online) overdominant mutations are introduced. When more than two overdominant alleles are balanced in the population, it is remarkable that *B* statistics with *m* set to three or four (supplementary fig. S55*E* and *F*, Supplementary Material online, respectively) outperform those with $m = 2$ (supplementary fig. S55*D*, Supplementary Material online). Furthermore, we also observe that the optimal equilibrium minor allele frequencies reported by the *B* statistics match well with the true equilibrium frequencies in the simulated replicates (supplementary fig. S56, Supplementary Material online).

To further dissect the relative performances of *B* statistics (with $m = 4$), we also applied other statistics with their optimal window sizes on these simulated sequences (supplementary fig. S57, Supplementary Material online). As the number of balanced alleles increases, each statistic demonstrated improvements in their power. Furthermore, the B_1, B_2, and $B_{2,\text{MAF}}$ statistics outperform their respective *T*- or summary-statistic analogs under all three scenarios considered.

Taken together, these results suggest that the multiallelic *B* statistics can substantially improve the detection power for balancing selection with more than two balanced alleles. Moreover, *B* statistics with larger *m* parameters, the presumed number of balanced alleles, are downward compatible with population samples carrying fewer than *m* balanced alleles, as the presumed equilibrium minor allele frequencies of the extra allelic classes would be optimized close to zero (see supplementary fig. S56, Supplementary Material online).

Performance of Single-Locus Methods on Multilocus Balancing Selection

Similar to multiallelic balancing selection, despite previous theoretical work to model or simulate multilocus balancing selection (Barton and Navarro 2002; Navarro and Barton 2002; Tennessen 2018), no detection approach has yet been developed accordingly. Meanwhile, neither model-based detection framework underlying the *T* statistics nor the *B* statistics can address these cases without jointly accounting for allelic combinations at multiple informative sites as the target of selection. Therefore, without shifting the paradigm to consider such site-to-site combinations so as to accurately locate the set of neighboring selected loci, one can still examine the performance of extant balancing selection approaches for locating genomic regions containing more than one locus under balancing selection.

To this end, we tested the simplest case with two nearby loci carrying independent overdominant alleles (see Materials and Methods). To ensure that individuals heterozygous at both loci are as advantageous as in the single-locus balancing selection simulations with $s = 0.001$ and $h = 20$ (supplementary fig. S58*A* and *B*, Supplementary Material online), we set the selective coefficients of both overdominant mutations to $s = 0.0005$. That is, a two-locus genotype that is heterozygous at each of the loci would have fitness approximately equal to $1 + 2hs = 1.2$. Despite this adjustment, we observed that all statistics show drastic improvements in their powers (supplementary fig. S58*C* and *D*, Supplementary Material online), with the lowest power among them of 0.8 (supplementary fig. S58*D*, Supplementary Material online). This result suggests that multilocus balancing selection can potentially create more-prominent footprints compared with single-locus balancing selection. To further gauge the extent to which the additional selected locus can boost detection power, we simulated sequences with two nearby loci each evolving under $s = 10^{-5}$ and $h = 20$, such that the selective coefficient s is 2 orders of magnitude smaller than that of the mutations introduced in the sequences evolving under single-locus balancing selection (supplementary fig. S58*A* and *B*, Supplementary Material online). Remarkably, all methods still exhibit substantially higher powers for sequences with two nearby loci with weakly advantageous ($s = 10^{-5}$) alleles undergoing balancing selection (supplementary fig. S58*E* and *F*, Supplementary Material online).

The higher powers observed for simulated multilocus balancing selection scenarios is understandable, as Tennessen (2018) demonstrated that two noninteracting neighboring loci tend to reinforce the maintenance of polymorphisms when both are independently subjected to balancing selection. However, multilocus balancing selection can also be achieved by epistasis (Barton and Navarro 2002; Navarro and Barton 2002), whereby the fitness effect of one locus is contingent on the allelic state of another locus, and has been shown by a growing body of empirical studies to be pervasive in the genome (as reviewed by Shao et al. 2008; Lehner 2011; Mackay 2014). Though we did not simulate such scenarios in this study, because two interacting loci would better maintain polymorphisms at the selected loci than two noninteracting ones (Barton and Navarro 2002; Navarro and Barton 2002; Tennessen 2018), it would not be surprising that they would produce even stronger footprints than what we observe here.

Furthermore, genomic sequences with multiple nearby balanced loci will have more extended footprints of balancing selection. With the capability to optimize over window sizes, *B* statistics should be more sensitive to such regions than other approaches applied with small fixed windows. Indeed, B_2 substantially outperforms T_2 (applied with 12 informative sites on either side of a test site) when the two neighboring loci under selection are weakly advantageous themselves (supplementary fig. S58*E* and *F*, Supplementary Material online). The margins between their powers still persist even when *T* statistics adopt windows with 122 informative sites on either side of the test site (supplementary fig. S59*E* and *F*, Supplementary Material online), despite the marginal increases in their powers for two-locus balancing selection.

Our exploratory results imply not only that extant approaches for detecting balancing selection have high power when applied to genomic regions carrying multiple balanced loci but also that such power is also likely much higher than they would have for single-locus regions. For *B* statistics in particular, because they optimize over window sizes, the gap between their sensitivity for multilocus balancing selection and that for single-locus settings may be more profound than other methods when applied with small windows. Our results also support the speculation that top candidates identified in previous scans for balancing selection may be more likely to carry more than one functional polymorphic site, as is the case for the MHC locus, considering that all methods we evaluated show higher powers for multilocus balancing selection than for the single-locus process.

Application of B_2 to Empirical Data

In this study, we applied the B_2 statistic on both human and bonobo genomic data and identified sensible candidate targets in each species. We first reexamined the CEU and YRI human populations in the 1000 Genomes Project data set (1000 Genomes Project Consortium 2015) with B_2, which have been previously probed for long-term balancing selection in multiple studies (DeGiorgio et al. 2014; Siewert and Voight 2017; Bitarello et al. 2018). We found that top candidates reported by B_2 overlap largely with previous scans, lending confidence in the power of *B* statistics to make replicable discoveries. Next, we performed the first model-based scan for footprints of balancing selection on bonobo polymorphism data. In addition to the genomic regions previously reported to be under ancient balancing selection in humans and chimpanzees (e.g., the *MHC-DQ* genes at the MHC locus; Leffler et al. 2013; Teixeira et al. 2015; Cheng and DeGiorgio 2019), we have also uncovered novel candidates such as *KLRD1* and *SCN9A*, which play roles in pathogen defense and pain perception, respectively. Our results may correspond to the unique features and evolutionary history of bonobos, as suggested by accumulating evidence (de Waal 1990; Hare et al. 2012; de Groot et al. 2017; Wroblewski et al. 2017; Maibach and Vigilant 2019) on bonobo behavior and physiology.

Potential Balancing Selection on Gamete-Associated Genes in Humans

In the scans of human populations, we recovered previously reported candidates *STPG2* (formerly *C4orf13*) and *CCDC169* (formerly *C13orf38*), in addition to the HLA-D locus and *ERAP2*. Neither of the two former genes was discussed in previous studies after reporting them as top candidates, probably due to their late characterization. Intriguingly, both genes are related to gametogenesis, with recent association and clinical studies underscoring their functional importance. In particular, the expression of *STPG2* has been detected in male tissues, endocrine tissues, as well as the brain (Uhlén et al. 2015). Structural mutations deleting this gene have been linked to azoospermia (Yakut et al. 2013) and velocardiofacial syndrome (Wu et al. 2019), and association studies of SNPs in this have correlated it with autism (Connolly et al. 2017) and preclampsia (Johnson et al. 2012). A recent study even reported footprints of ongoing positive selection on a segregating preclampsia-associated SNP in this gene (Arthur 2018). Note that these authors only analyzed the disease-associating variants and applied haplotype-based selection tests, which tend to reveal regions with at least one dominant haplotype. The footprints reported by Arthur (2018) can result from either recent partial sweeps or balancing selection, with only the latter matching the kilobase-scale size of the increased diversity surrounding the region (supplementary figs. S31 and S32, Supplementary Material online).

The conjoined gene *CCDC169*–*SOHLH2* encodes a read-through transcript of the gene *CCDC169* and its immediate downstream *SOHLH2*, a crucial gene for gametogenesis. In addition to its potential to initiate the transcription of *SOHLH2* on occasions of read-through, *CCDC169* has also been found to have specific expression in prenatal brain tissues (Pletikos et al. 2014). More interestingly, the B_2 scores across this gene do not form a typical peak as seen in many other candidate regions (supplementary figs. S33 and S34, Supplementary Material online). Instead, we observed a plateau of elevated B_2 scores above the region joining the two genes. Furthermore, both the mean pairwise sequence difference (π) and T_2 with a 22-informative-site-radius window show two minor peaks across this region. Considering our results for multilocus balancing selection (supplementary fig. S58, Supplementary Material online), such footprints may be reflective of multiple loci undergoing balancing selection, probably interactively via epistasis, which can create footprints of extended tracks of elevated genetic diversity (Barton and Navarro 2002; Navarro and Barton 2002).

Lastly, despite the intriguing functional implications behind our candidates, we are aware that some of our candidate regions show worrying signs for artifacts. For example, *STPG2* (also a top candidate in the scan by DeGiorgio et al. 2014) has low 35-mer sequence uniqueness scores across the whole 40-kb region examined, despite surviving the 50-mer mappability filter. The peak linking *CCDC169* and *SOHLH2* shows overall higher sequence uniqueness than *STPG2*, but the few regions with relatively lower uniqueness colocalize with the peaks reported by π and T_2. This colocalization is also observed in the gene *CPE*, where peak regions with a drop in sequence uniqueness also display lower sequencing depths than other regions. Though not all regions with low mappability necessarily yield outstanding scores for balancing selection, these signs could still be indicative of erroneous mapping and warrant further investigation and caution in interpretation.

Footprints of Balancing Selection in Bonobos and Their Implications

As one of the two sister species to humans, bonobos (initially known as the pygmy chimpanzees; Prüfer et al. 2012) have been drawing increasing attention from the genomics community (Prüfer et al. 2012; Prado-Martinez et al. 2013; de Manuel et al. 2016). However, compared with chimpanzees (the other sister species), bonobos are relatively understudied, despite their close relationship to humans and unique social behaviors. For bonobos, one of their most idiosyncratic traits is their high prevalence of sociosexual activities (de Waal 1990; Kano 1992; Wrangham 1993), which serve important nonreproductive functions and include frequent same-sex encounters. As a close relative to humans, their female-dominance, low-aggression, and hypersexual social behaviors contrast fiercely with those of humans and chimpanzees (Kano 1992; Wrangham 1993). A growing number of recent studies have also characterized the differences in physiological responses between bonobos and chimpanzees behind their social behaviors (Heilbronner et al. 2008; Hohmann et al. 2009; Wobber et al. 2010; Deschner et al. 2012; Surbeck et al. 2012), yet the genetic component underlying their unique behaviors, however, remains largely elusive. From the B_2 scan of bonobo genomes, we identified a number of interesting top candidates involved in pathogen defense. Despite that most of the MHC region was removed by a mappability filter (see Materials and Methods), we still observed extraordinary signals from the remainder of this region. More specifically, the *MHC-DQ* and *MHC-DP* genes harbor the highest peaks across the genome (figs. 4 and 5). These genes encode the component proteins of MHC-DQ and MHC-DP molecules, which are cell-surface receptors on antigen-presenting cells (Ball and Stastny 1984) and has long been known to be highly polymorphic in great apes (Takahata et al. 1992; Prüfer et al. 2012; Teixeira et al. 2015).

Another immune-related gene, *KLRD1*, which encodes the cell-surface antigen CD94, also exhibited outstanding B_2 scores. The interaction between KLRD1 (CD94) and NKG2 family proteins can either inhibit or activate the cytotoxic activity of NK cells (Pende et al. 1997; Cantoni et al. 1998; Masilamani et al. 2006), as well as pivot the generation of cell memory in NK cells (Cerwenka and Lanier 2016). Furthermore, KLRD1 (CD94) has been shown to play an important role in combating viral infections such as cytomegalovirus (Cerwenka and Lanier 2016) and influenza (Bongen et al. 2018) in humans, as well as the mousepox virus in mice (Fang et al. 2011). In humans and chimpanzees, *KLRD1* is highly conserved (Khakoo et al. 2000; Shum et al. 2002). Here, the involvement in viral defense of *KLRD1* presents an especially intriguing case for bonobos. Bonobos have been recently shown to harbor reduced levels of polymorphism in MHC class I genes (Maibach et al. 2017; Wroblewski et al. 2017), which were further predicted to have lower ability to bind with viral peptides when compared with chimpanzees (Maibach and Vigilant 2019). The genes encoding another regulator of MHC class I molecules, the killer cell immunoglobulin-like receptors (KIR), were also found to have contracted haplotypes in bonobos (Rajalingam et al. 2001; Walter 2014; Wroblewski et al. 2019), with the lineage III *KIR* genes serving reduced functions (Wroblewski et al. 2019). In fact, many studies have pointed out that these reduced features are unlikely the natural consequences of demographic factors—even after considering the harsher bottlenecks bonobos have undergone compared with chimpanzees—and speculate that selective sweeps in bonobos on these regions (Prüfer et al. 2012; Walter 2014; Maibach et al. 2017; Wroblewski et al. 2017, 2019) may have eliminated the diversity in these critical immunity genes. In this light, the polymorphisms on *KLRD1* may be compensating the reduced diversity in their binding partners in bonobos.

Several other genes in high-scoring regions are also found to be involved in immunity. For one, the highest peak on chromosome 7 encompasses the entire gene *GPNMB* (supplementary fig. S40, Supplementary Material online), with elevated B_2 scores particularly on exons. This gene encodes osteoactivin, a transmembrane glycoprotein found on osteoclast cells, macrophages, and melanoblast (Loftus et al. 2009; Yu et al. 2016), and is shown to regulate proinflammatory responses (Ripoll et al. 2007). Aside from its heavy involvement in cancer (Zhou et al. 2012), the protein GPNMB has also been shown to facilitate tissue repair (Li et al. 2010; Rose et al. 2010; Hu et al. 2013) as well as influence iris pigmentation (Bächner et al. 2002; Maric et al. 2013). Other potential evidence for balancing selection operating on innate immunity-related genes includes the high B_2 scores observed around the intergenic region between *BPIFB4* and *BPIFA2* (supplementary fig. S42, Supplementary Material online), which encode two bacterialcidal permeability-increasing fold-containing (BPIF) family proteins (Levy 2000). The *BPIFA2* genic region is recently shown to harbor many SNPs significantly associated with enteropathy (Fujimori et al. 2019), whereas the *BPIFB4* gene is better known by its association with longevity (Villa, Malovini, et al. 2015; Spinetti et al. 2017; Villa et al. 2018), speculated to partly result from its protection of vascular functions (Villa, Carrizzo, et al. 2015; Puca et al. 2016; Spinelli et al. 2017).

In addition to pathogen defense, we also found other interesting candidates relating to neurosensory and neurodevelopment. One such gene is *SCN9A* (supplementary fig. S43, Supplementary Material online), which encodes Na_V 1.7, a voltage-gated sodium channel, with mutations on the gene associated with various pain disorders (Yang et al. 2004; Cox et al. 2006; Reimann et al. 2010). The peak we observe covers the overlapping RNA gene encoding its antisense transcript, *SCN1A-AS1*, which regulates the expression of *SCN9A* (Koenig et al. 2015), suggestive of diversified regulation of pain perception in bonobos. A few other candidate genes are also involved in neurodevelopment, such as *EPHA6* (Das et al. 2016), *SUSD2* (supplementary fig. S47, Supplementary Material online; Nadjar et al. 2015), and *HPCAL1* (Tam 2015).

Lastly, we noticed that some candidate genes carry multiple distinct functions and may have been undergoing balancing selection due to potential evolutionary conflicts between some of their functions. For example, the gene *GPNMB* plays roles not only in tissue repair (Li et al. 2010) but also in iris pigmentation (Bächner et al. 2002). Another candidate, *PDE1A* gene (supplementary fig. S45, Supplementary Material online), encodes a phosphodiesterase that is pivotal to Ca^{2+}- and cyclic nucleotide signaling (Lefièvre et al. 2012). It is expressed in brain, endocrine tissues, kidneys, and gonads (Uhlén et al. 2015) and has multiple splicing variants. In fact, the high-scoring peak we observed on this gene happens to locate around the exons that are spliced out in some variants (supplementary fig. S45, Supplementary Material online). Studies have demonstrated the relation of this gene to brain development (Yan et al. 1994), mood and cognitive disorders (Xu et al. 2011; Martinez and Gil 2013; Pekcec et al. 2018; Betolngar et al. 2019), and hypertension (Kimura et al. 2017). Meanwhile, the PDE1A protein is also a conserved component of mammalian spermatozoa (Lefièvre et al. 2012; Vasta et al. 2005) and is involved in the movement of its flagella. Similarly, the gene *CAMK4* encodes Ca^{2+}- and calmodulin-dependent kinase 4, which also plays important roles in both immunity (Koga and Kawakami 2018) and spermatogenesis (Wu et al. 2000). The cancer-related protein Sushi-domain containing 2, encoded by *SUSD2* (Watson et al. 2013), not only regulates neurite growth in the brain (Nadjar et al. 2015) but can also be used as a marker molecule for human spermatogonial progenitors (Harichandan et al. 2013). Though it is difficult to judge for these genes which functions may be subject to selective pressures, they nonetheless indicate that pleiotropy can be an important driver of balancing selection.

Concluding Remarks

Extant methods for detecting long-term balancing selections are constrained by the pliability of their inferences as a function of genomic window size. In this study, we presented *B* statistics, a set of composite likelihood ratio statistics based on nested mixture models. We have comprehensively evaluated their performances through simulations and demonstrated their robust high performances over varying window sizes in uncovering genomic loci undergoing balancing selection. Moreover, we showed that even when applied with the least optimal window sizes, the *B* statistics still exhibit high power comparable to current methods, which operated under optimal window sizes, in uncovering balancing selection of varying age and selection parameters, as well as robust performance under confounding scenarios such as elevated mutation rates, variable recombination rates, and population size changes. We reexamined the 1000 Genomes Project YRI and CEU populations with B_2 statistics and have recovered well-characterized genes previously hypothesized to be undergoing long-term balancing selection in humans, such as the HLA-D genes, *ERAP2*, and *CSMD2*. We also characterized previously reported top candidates *STPG2* and *CCDC169–SOHLH2*, both of which are related to gametogenesis. We further applied the B_2 statistic on the whole-genome polymorphism data of bonobos and discovered not only the well-established *MHC-DQ* and *MHC-DP* genes but also novel candidates such as *KLRD1*, *PDE1A*, *SCN9A*, and *SUSD2*, with functional implications in pathogen defense, neurodevelopment, as well as gamete functions. Moreover, we have extended the *B* statistics to consider multiallelic balancing selection, with these extensions demonstrating superior properties to all previous methods for detecting selected loci with more than two balanced alleles. Further, we show that all current methods tend to have higher powers for two-locus balancing selection than for single-locus processes. Lastly, we have implemented these statistics in the open source software BalLeRMix, which, along with other key scripts used in this study, can be accessed at https://github.com/bioXiaoheng/BalLeRMix/. We have also released the empirical scan results for balancing selection in both humans and bonobos, which can be downloaded at http://degiorgiogroup.fau.edu/ballermix.html.

Materials and Methods

In this section, we discuss sets of simulations used to evaluate the performances of the B statistics relative to previously published state-of-the-art approaches (Hudson et al. 1987; DeGiorgio et al. 2014; Siewert and Voight 2017, 2020; Bitarello et al. 2018). Finally, we describe the application of our B statistics to an empirical bonobo data set (Prado-Martinez et al. 2013).

Evaluating Methods through Simulations

We employed the forward-time genetic simulator SLiM (version 3.2; Haller and Messer 2019) to generate sequences of 50 kb in length evolving with or without balancing selection. Based on the respective levels in humans and other great apes, we assumed a mutation rate of $\mu = 2.5 \times 10^{-8}$ per base per generation (Nachman and Crowell 2000) and a recombination rate of $r = 10^{-8}$ per base per generation (Payseur and Nachman 2000). In scenarios with constant population sizes, we set the diploid effective population size as $N = 10^4$. To create baseline genetic variation, each replicate simulation was initiated with a burn-in period of $10N = 10^5$ generations. To speed up simulations, we applied the scaling parameter λ to the number of simulated generations, population size, mutation rate, recombination rate, and selection coefficient, which allows for the generation of the same levels of variation with a speed up in computational time by a factor λ^2. For scenarios based on a model of constant population size, we used $\lambda = 100$. For the demographic models of European humans and bonobos, we used $\lambda = 20$. We simulated 500 replicates for each scenario considered and sampled 50 haploid lineages from the target population and one lineage from the outgroup in each simulation for downstream analyses.

We simulated data from two other diverged species, under the demographic history inspired by that of humans, chimpanzees (Kumar et al. 2005), and gorillas (Scally et al. 2012). Specifically, the closer and farther outgroups diverged 2.5×10^5 and 4×10^5 generations ago, respectively, which correspond to 5 million and 8 Ma, assuming a generation time of 20 years.

To evaluate the power of each method to detect balancing selection with varying selective coefficient s, dominance coefficient h, and age, for each combination of s and h, we considered 15 time points at which the selected allele was introduced, ranging from 5×10^4 to 6.5×10^5 generations prior to sampling with time points separated by intervals of 5×10^4 generations. Assuming a generation time of 20 years, these time points are equivalent to $1, 2, 3, \ldots, 15$ My before sampling. In each scenario, a single selected mutation was introduced at the center of each sequence at the assigned time point, and we only considered simulations where the introduced allele was not lost.

Accelerated Mutation Rate

To evaluate whether the B statistics are robust to high mutation rates, we applied the methods on simulated sequences evolving neutrally along the same demographic history (supplementary fig. S1, Supplementary Material online), but instead with a 5-fold higher mutation rate of $5\mu = 1.25 \times 10^{-7}$ per site per generation. To generate sequences with regional increases in mutation rate, we simulated 50-kb sequences with a 5-fold higher mutation rate of $5\mu = 1.25 \times 10^{-7}$ per site per generation at the central 10 kb of the sequence and the surrounding region with the original rate μ.

Recombination Rate Estimation Error

For evaluating the robustness to erroneous estimation of recombination rates, we simulated sequences with uneven recombination maps and applied the model-based methods with the assumption that the recombination rate is uniform. In particular, we divided the 50-kb sequence into 50 regions of 1 kb each, and in turns inflate or deflate the recombination rate of each region by m fold, such that the recombination rates of every pair of neighboring regions have a m^2-fold difference. We tested $m = 10$ and $m = 100$ in this study.

Demographic History

To examine the performance of methods under realistic demographic parameters, we considered the demographic histories of a European human population (CEU; Terhorst et al. 2017) and of bonobos (Prado-Martinez et al. 2013). For the human population, we adopted the history of population size changes inferred by SMC++ (Terhorst et al. 2017) that spans 10^5 generations, assuming a mutation rate of $\mu = 1.25 \times 10^{-8}$ per site per generation (assumed when estimating the CEU demographic history in Terhorst et al. 2017), a generation time of 20 years, and a scaling effective size of 10^4 diploids. To account for recombination rate variation, we allowed each simulated replicate to have a uniform recombination rate drawn uniformly at random between $r = 5 \times 10^{-9}$ and $r = 1.5 \times 10^{-8}$ per site per generation. We also simulated an additional population that split from the human population 2.5×10^5 generations ago, which is identical to the outgroup (named O1) in the demographic model depicted in figure 3A, with an effective size of $N = 10^4$ diploid individuals.

For the bonobo population history, we scaled the pairwise sequentially Markovian coalescent (PSMC) history inferred from the genome of individual A917 (Dzeeta; sample SRS396202) by Prado-Martinez et al. (2013) with a mutation rate of $\mu = 2.5 \times 10^{-8}$ per site per generation, identical to the simulations on the three-population demographic history (fig. 3A). Because the inferred PSMC model provides a specific ratio of the mutation and recombination rates, we set the recombination rate to $r = 2.84 \times 10^{-9}$ per site per generation. To be consistent with the three-population demographic history, we set the population size prior to 71,640 generations ago, which is the maximum time covered by the PSMC inference, to $N = 10^4$ diploid individuals, and had the outgroup split 2.5×10^5 generations ago with the same diploid population size, identical to the outgroup O1 in the three-population demographic history (fig. 2A).

To simulate species with distinct mutation rates, we split the simulation into two stages, with the first stage concerning the sequences in the ancestral species up until the two populations diverge 5 Ma. Upon divergence, two separate SLiM simulations are used to distinguish the mutation rates in the target and outgroup populations, and samples are output separately before being integrated in subsequent analyses. We set the target species to mutate at a rate of $\mu = 1.2 \times 10^{-8}$ per site per generation (Scally and Durbin 2012) after divergence, and the other species (including the ancestral species) evolving with the mutation rate of $\mu = 2.5 \times 10^{-8}$ per site per generation (Nachman and Crowell 2000). The recombination rate across all simulations is $r = 10^{-8}$ per site per generation (Payseur and Nachman 2000). For the simulations with constant population sizes, we set the effective size of all populations as $N = 10^4$ diploid individuals and adopted the scaling parameter $\lambda = 100$. For simulations employing realistic demographic histories, we used $\lambda = 20$, set the effective population size of the ancestral and the outgroup species as $N = 10^4$ diploids (Takahata et al. 1995), and the target species following the demographic history inferred from the CEU human population (Terhorst et al. 2017) for 10^5 generations prior to sampling. Additionally, we set the generation time of the target species to be 25 years (akin to humans; Scally and Durbin 2012), whereas for the outgroup and ancestral species we used 20 years (akin to nonhuman great apes; Prado-Martinez et al. 2013). Consequently, the species divergence occurred 200,000 generations ago for the target species and 250,000 generations ago for the outgroup.

Three- and Four-Allelic Balancing Selection at a Single Site

To simulate balancing selection on a single site with more than two balanced alleles, we used SLiM3.3 (Haller and Messer 2019) so that all four nucleotides, instead of binary representations of 0s and 1 s, can be incorporated into the simulations. We adopted the same three-species demographic history as illustrated in supplementary figure S1, Supplementary Material online, and simulated sequences of length 50 kb consisting of randomly generated strings of four nucleotides at the beginning of each replicate, with equal chance of occurrence for each nucleotide. We considered the Jukes–Cantor substitution model and set the between-nucleotide mutation rate as $\mu = 8.3 \times 10^{-9}$ per site per generation, such that the total mutation rate (three times the between-nucleotide mutation rate) is $\mu = 2.49 \times 10^{-8}$ per site per generation—roughly the same as adopted in the biallelic balancing selection simulations. We also assumed a uniform recombination rate of $r = 10^{-8}$ per site per generation (Payseur and Nachman 2000). At 500,000 generations before sampling, we introduced two, three, or four mutations of distinct nucleotides that have selective coefficient $s = 0.001$ and dominance coefficient $h = 20$. Note that SLiM considers colocalized mutations of distinct types as if they were at different positions and computes fitness for the individual by multiplying fitness values of each mutation. That is, a diploid individual who is heterozygous at a site harboring two distinct selectively advantageous mutant alleles with parameters $s = 0.001$ and $h = 20$ would have fitness $(1 + hs)(1 + hs) = 1.44$, whereas a homozygote for either selectively advantageous mutation would have fitness $1 + s = 1.001$. At the completion of the simulation, we sampled 25 diploid individuals uniformly at random from each of the sister species (P and O1), and one diploid individual was sampled uniformly at random from species O2, with only one haplotype of this individual being considered as the reference sequence. Only biallelic sites were considered in the downstream analysis.

Application to Empirical Data

Human Genomic Data from the 1000 Genomes Project

We obtained variant calls from the 1000 Genomes Project data set (1000 Genomes Project Consortium 2015), which were mapped to human reference genome hg19, and extracted the haplotypes for the CEU and YRI populations. We used the chimpanzee reference genome panTro5 downloaded from the UCSC Genome Browser (Kent et al. 2002; Haeussler et al. 2019) to call ancestral alleles and only retained mappable monomorphic or biallelic polymorphic sites based on the variation in the CEU (or YRI) population together with the chimpanzee reference genome. For mappable sites not included in the variant call data set, we assumed that the site is monomorphic for the hg19 reference genome and called substitutions accordingly.

To avoid making inference on potentially problematic regions, we applied the RepeatMasker filter and removed segmental duplications, both of which were downloaded from the UCSC Genome Browser (Kent et al. 2002; Haeussler et al. 2019). Genomic regions with mappability 50-mer score (Derrien et al. 2012) lower than 0.9 were discarded as well. Moreover, we performed one-tailed Fisher's exact tests for Hardy–Weinberg equilibrium (Wigginton et al. 2005) on each polymorphic site and removed those with a significant ($P < 10^{-4}$) excess of heterozygous genotypes.

We applied B_2 to each CEU and YRI data set separately, assuming the human recombination map of the hg19 reference genome (International HapMap Consortium 2007). We did not fix the window size of these scans, and instead permitted B_2 to optimize over both free parameters A and x. To better compare our results with previous studies, we also applied the T_2 statistic (DeGiorgio et al. 2014) to the same input data sets, adopting window sizes of 22 or 100 informative sites on either side of a test informative site. We also computed sequence diversity π averaged across each 5-kb window as a reference.

For downstream examination of the mappability of candidate regions, we consulted the 35-mer uniqueness score (UCSC hg19 database; Kent et al. 2002; Haeussler et al. 2019) averaged across each 1-kb region. Furthermore, we also downloaded the BAM files for each individual in the CEU or YRI population and generated per-base read depths with BEDTools 2.26 (Quinlan 2014). We then computed sample-wide mean read depths, their standard deviations, and the number of individuals without coverage for each population after merging read depths of all samples with BEDTools. These references further aided in flagging potentially problematic regions that survived initial filters, as they typically feature lower mappability (mean 35-mer uniqueness) or abnormally low or high read depths.

Bonobo Genomic Data from the Great Ape Project

We obtained the genotype calls of 13 bonobos from the Great Ape Project (Prado-Martinez et al. 2013), which were originally mapped to human genome assembly NCBI36/hg18. We lifted over the variant calls to human genome assembly GRCh38/hg38, so that the bonobo genome assembly panPan2 can be used for polarizing the allele frequencies, with the sequence in hg38 considered as the ancestral allele. Only genomic regions mappable across hg38 and panPan2 were considered for further analyses. For mappable polymorphic sites, we only considered biallelic SNPs. For mappable sites without variant calls in bonobo, we assumed that these sites were monomorphic for the panPan2 reference genome sequence and called substitutions based on whether the panPan2 reference allele was different from the hg38 reference allele.

To circumvent potential artifacts, we performed one-tailed Hardy–Weinberg equilibrium tests on each site and removed sites with an excess of heterozygotes ($P < 0.01$). This *P*-value was determined by the distribution of the *P*-values of all polymorphic sites across the genome, such that 0.035% of such sites are outliers. We also applied the RepeatMasker, segmental duplication, simple repeat, and interrupted repeat filters (all downloaded from UCSC Genome Browser) to remove repetitive regions. To assess the mappability of each genomic region, we employed the mappability scores (obtained by setting the maximum mismatches tolerated to zero; Derrien et al. 2012) of 50-mers. Regions with 50-mer mappability scores lower than 0.9 were removed. Because BalLeRMix employs a prespecified grid of *A* values to accompany the distances *d* in centimorgans, when applying the method, we assumed a uniform recombination rate of 10^{-6} cM per site, which is the approximate recombination rate in humans (Payseur and Nachman 2000).

Supplementary Material

Supplementary data are available at *Molecular Biology and Evolution* online.

Acknowledgments

We thank J. Terhorst and J. Prado-Martinez for sharing the inferred parameters for the demographic history of CEU and great apes, respectively. We also thank P. Ribeca and T. Derrien for providing the up-to-date GEMtools software and the ENCODE mappability tracks for human hg38 genome assembly after adjusting for indels. We appreciate the assistance from B. Haller on the fitness scheme of multiallelic balancing selection with SLiM. We thank Hillary Koch for helpful discussions on the properties of composite likelihood ratio statistics. We are also grateful to the editor and anonymous reviewers for their comments that helped improve this manuscript. This work was funded by Pennsylvania State University, by National Institutes of Health (Grant No. R35-GM128590), National Science Foundation (Grant Nos. DEB-1753489, DEB-1949268, and BCS-2001063), and the Alfred P. Sloan Foundation. Portions of this research were conducted with Advanced CyberInfrastructure computational resources provided by the Institute for Computational and Data Sciences at Pennsylvania State University.

References

1000 Genomes Project Consortium. A global reference for human genetic variation 2015. Nature 526(7571):68–74.

Andrés AM. 2001. Balancing Selection in the Human Genome. Chichester, UK: John Wiley and Sons, Ltd.

Andrés AM, Hubisz MJ, Indap A, Torgerson DG, Degenhardt JD, Boyko AR, Gutenkunst RN, White TJ, Green ED, Bustamante CD, et al. 2009. Targets of balancing selection in the human genome. *Mol Biol Evol.* 26(12):2755–2764.

Andrés MY, Dennis WW, Kretzschmar JL, Cannons S-Q, Lee-Lin B, Hurle PL, Schwartzberg SH, Williamson CD, Bustamante R, Nielsen AM, et al. 2010. Balancing selection maintains a form of ERAP2 that undergoes nonsense-mediated decay and affects antigen presentation. *PLoS Genet.* 6(10):e1001157.

Arthur A. 2018. Molecular evolution of genes associated with preeclampsia: genetic conflict, antagonistic coevolution and signals of selection. *J Evol Med.* 6(1):1–9.

Asmussen M, Basnayake E. 1990. Frequency-dependent selection: the high potential for permanent genetic variation in the diallelic, pairwise interaction model. *Genetics* 125(1):215–230.

Asmussen MA, Feldman MW. 1977. Density dependent selection 1: a stable feasible equilibrium may not be attainable. *J Theor Biol.* 64(4):603–618.

Bächner D, Schröder D, Gross G. 2002. mRNA expression of the murine glycoprotein (transmembrane) NMB (*GPNMB*) gene is linked to the developing retinal pigment epithelium and iris. *Gene Expr Patterns.* 1(3–4):159–165.

Ball E, Stastny P. 1984. Antigen-specific HLA-restricted human T-cell lines. II. A GAT-specific T-cell line restricted by a determinant carried by an HLA-DQ molecule. *Immunogenetics* 20(5):547–564.

Bamshad M, Wooding SP. 2003. Signatures of natural selection in the human genome. *Nat Rev Genet.* 4(2):99–110.

Barton NH, Navarro A. 2002. Extending the coalescent to multilocus systems: the case of balancing selection. *Genet Res.* 79(2):129–140.

Benjamini Y, Hochberg Y. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc B.* 57(1):289–300.

Bergland AO, Behrman EL, O'Brien KR, Schmidt PS, Petrov DA. 2014. Genomic evidence of rapid and stable adaptive oscillations over seasonal time scales in *Drosophila*. *PLoS Genet.* 10(11):e1004775.

Betolngar DB, Mota É, Fabritius A, Nielsen J, Hougaard C, Christoffersen CT, Yang J, Kehler J, Griesbeck O, Castro LRV, et al. 2019. Phosphodiesterase 1 bridges glutamate inputs with no-and dopamine-induced cyclic nucleotide signals in the striatum. *Cereb Cortex.* 29(12):5022–5036.

Bitarello BD, de Filippo C, Teixeira JC, Schmidt JM, Kleinert P, Meyer D, Andrés AM. 2018. Signatures of long-term balancing selection in human genomes. *Genome Biol Evol.* 10(3):939–955.

Bongen E, Vallania F, Utz PJ, Khatri P. 2018. KLRD1-expressing natural killer cells predict influenza susceptibility. *Genome Med.* 10(1):45.

Braverman JM, Hudson RR, Kaplan NL, Langley CH, Stephan W. 1995. The hitchhiking effect on the site frequency spectrum of DNA polymorphisms. *Genetics* 140(2):783–796.

Cantoni C, Biassoni R, Pende D, Sivori S, Accame L, Pareti L, Semenzato G, Moretta L, Moretta A, Bottino C. 1998. The activating form of CD94 receptor complex: CD94 covalently associated with the Kp39 protein that represents the product of the *NKG2-C* gene. *Eur J Immunol*. 28(1):327–338.

Cerwenka A, Lanier LL. 2016. Natural killer cell memory in infection, inflammation and cancer. *Nat Rev Immunol*. 16(2):112–123.

Charlesworth B, Charlesworth D. 2010. Elements of evolutionary genetics. Greenwood Village (CO): Roberts and Company Publishers.

Charlesworth D. 2006. Balancing selection and its effects on sequences in nearby genome regions. *PLoS Genet*. 2(4):e64–6.

Charlesworth D, Awadalla P, Mable B, Schierup M. 2000. Population-level studies of multiallelic self-incompatibility loci, with particular reference to brassicaceae. *Ann Bot*. 85:227–239.

Cheng X, DeGiorgio M. 2019. Detection of shared balancing selection in the absence of trans-species polymorphism. *Mol Biol Evol*. 36(1):177–199.

Cho S, Huang ZY, Green DR, Smith DR, Zhang J. 2006. Evolution of the complementary sex-determination gene of honey bees: balancing selection and trans-species polymorphisms. *Genome Res*. 16(11):1366–1375.

Cockerham CC, Burrows P, Young S, Prout T. 1972. Frequency-dependent selection in randomly mating populations. *Am Nat*. 106(950):493–515.

Connolly S, Anney R, Gallagher L, Heron EA. 2017. A genome-wide investigation into parent-of-origin effects in autism spectrum disorder identifies previously associated genes including shank3. *Eur J Hum Genet*. 25(2):234–239.

Cox JJ, Reimann F, Nicholas AK, Thornton G, Roberts E, Springell K, Karbani G, Jafri H, Mannan J, Raashid Y, et al. 2006. An SCN9A channelopathy causes congenital inability to experience pain. *Nature* 444(7121):894–898.

Das G, Yu Q, Hui R, Reuhl K, Gale NW, Zhou R. 2016. Epha5 and epha6: regulation of neuronal and spine morphology. *Cell Biosci*. 6(1):48.

de Groot NG, Heijmans CM, Helsen P, Otting N, Pereboom Z, Stevens JM, Bontrop RE. 2017. Limited MHC class I intron 2 repertoire variation in bonobos. *Immunogenetics* 69(10):677–688.

de Manuel M, Kuhlwilm M, Frandsen P, Sousa VC, Desai T, Prado-Martinez J, Hernandez-Rodriguez J, Dupanloup I, Lao O, Hallast P, et al. 2016. Chimpanzee genomic diversity reveals ancient admixture with bonobos. *Science* 354(6311):477–481.

de Waal FB. 1990. Sociosexual behavior used for tension regulation in all age and sex combinations among bonobos. In: Feierman JR, editor. Pedophilia. New York: Springer-Verlag. p. 378–393.

DeGiorgio M, Lohmueller KE, Nielsen R. 2014. A model-based approach for identifying signatures of ancient balancing selection in genetic data. *PLoS Genet*. 10(8):e1004561.

Derrien T, Estellé J, Sola SM, Knowles DG, Raineri E, Guigó R, Ribeca P. 2012. Fast computation and applications of genome mappability. *PLoS One* 7(1):e30377.

Deschner T, Fuller BT, Oelze VM, Boesch C, Hublin J-J, Mundry R, Richards MP, Ortmann S, Hohmann G. 2012. Identification of energy consumption and nutritional stress by isotopic and elemental analysis of urine in bonobos (*Pan paniscus*). *Rapid Commun Mass Spectrom*. 26(1):69–77.

Fang M, Orr MT, Spee P, Egebjerg T, Lanier LL, Sigal LJ. 2011. CD94 is essential for NK cell-mediated resistance to a lethal viral disease. *Immunity* 34(4):579–589.

Fay JC, Wu C-I. 2000. Hitchhiking under positive Darwinian selection. *Genetics* 155(3):1405–1413.

Frazer KA, Ballinger DG, Cox DR, Hinds DA, Stuve LL, Gibbs RA, Belmont JW, Boudreau A, Hardenbol P, Leal SM, et al.; International HapMap Consortium. 2007. A second generation human haplotype map of over 3.1 million SNPs. *Nature* 449(7164):851–861.

Fujimori S, Fukunaga K, Takahashi A, Mushiroda T, Kubo M, Hanada R, Hayashida M, Sakurai T, Iwakiri K, Sakamoto C. 2019. Bactericidal/permeability-increasing fold-containing family B member 4 may be associated with NSAID-induced enteropathy. *Dig Dis Sci*. 64(2):401–408.

Fumagalli M, Cagliani R, Pozzoli U, Riva S, Comi GP, Menozzi G, Bresolin N, Sironi M. 2008. Widespread balancing selection and pathogen-driven selection at blood group antigen genes. *Genome Res*. 19(2):199–212.

Gao Z, Przeworski M, Sella G. 2015. Footprints of ancient-balanced polymorphisms in genetic variation data from closely related species. *Evolution* 69(2):431–446.

Ginzburg LR. 1977. The equilibrium and stability for n alleles under the density-dependent selection. *J Theor Biol*. 68(4):545–550.

Gravel S, Henn BM, Gutenkunst RN, Indap AR, Marth GT, Clark AG, Yu F, Gibbs RA, The 1000 Genomes Project, Bustamante CD. 2011. Demographic history and rare allele sharing among human populations. *Proc Natl Acad Sci U S A*. 108(29):11983–11988.

GTEx Consortium. 2017. Genetic effects on gene expression across human tissues. *Nature* 550(7675):204.

Guirao-Rico S, Sánchez-Gracia A, Charlesworth D. 2017. Sequence diversity patterns suggesting balancing selection in partially sex-linked genes of the plant *Silene latifolia* are not generated by demographic history or gene flow. *Mol Ecol*. 26(5):1357–1370.

Haeussler M, Zweig AS, Tyner C, Speir ML, Rosenbloom KR, Raney BJ, Lee CM, Lee BT, Hinrichs AS, Gonzalez JN, et al. 2019. The UCSC Genome Browser database: 2019 update. *Nucleic Acids Res*. 47(D1):D853–D858.

Haller BC, Messer PW. 2019. Slim 3: forward genetic simulations beyond the wright-fisher model. *Mol Biol Evol*. 36(3):632–637.

Hare B, Wobber V, Wrangham R. 2012. The self-domestication hypothesis: evolution of bonobo psychology is due to selection against aggression. *Anim Behav*. 83(3):573–585.

Harichandan A, Sivasubramaniyan K, Hennenlotter J, Schwentner C, Stenzl A, Bühring H-J. 2013. Isolation of adult human spermatogonial progenitors using novel markers. *J Mol Cell Biol*. 5(5):351–353.

Hedrick PW. 2002. Pathogen resistance and genetic variation at MHC loci. *Evolution* 56(10):1902–1908.

Heilbronner SR, Rosati AG, Stevens JR, Hare B, Hauser MD. 2008. A fruit in the hand or two in the bush? Divergent risk preferences in chimpanzees and bonobos. *Biol Lett*. 4(3):246–249.

Hey J. 1991. A multi-dimensional coalescent process applied to multi-allelic selection models and migration models. *Theor Popul Biol*. 39(1):30–48.

Hohmann G, Mundry R, Deschner T. 2009. The relationship between socio-sexual behavior and salivary cortisol in bonobos: tests of the tension regulation hypothesis. *Am J Primatol*. 71(3):223–232.

Hu X, Zhang P, Xu Z, Chen H, Xie X. 2013. GPNMB enhances bone regeneration by promoting angiogenesis and osteogenesis: potential role for tissue engineering bone. *J Cell Biochem*. 114(12):2729–2737.

Hudson RR, Kaplan NL. 1988. The coalescent process in models with selection and recombination. *Genetics* 120(3):831–840.

Hudson RR, Kreitman M, Aguadé M. 1987. A test of neutral molecular evolution based on nucleotide data. *Genetics* 116(1):153–159.

Hunter-Zinck H, Clark AG. 2015. Aberrant time to most recent common ancestor as a signature of natural selection. *Mol Biol Evol*. 32(10):2784–2797.

Johnson MP, Brennecke SP, East CE, Göring HHH, Kent JW, Dyer TD, Said JM, Roten LT, Iversen A-C, Abraham LJ, et al.; The FINNPEC Study Group. 2012. Genome-wide association scan identifies a risk locus for preeclampsia on 2q14, near the inhibin, beta B gene. *PLoS One* 7(3):e33666.

Kano T. 1992. The last ape: pygmy chimpanzee behavior and ecology. Vol. 155. Stanford (CA): Stanford University Press Stanford.

Kaplan NL, Darden T, Hudson RR. 1988. The coalescent process in models with selection. *Genetics* 120(3):819–829.

Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler D. 2002. The human genome browser at UCSC. *Genome Res*. 12(6):996–1006.

Khakoo SI, Rajalingam R, Shum BP, Weidenbach K, Flodin L, Muir DG, Canavez F, Cooper SL, Valiante NM, Lanier LL, et al. 2000. Rapid evolution of NK cell receptor systems demonstrated by comparison of chimpanzees and humans. *Immunity* 12(6):687–698.

Kimura M, Tamura Y, Guignabert C, Takei M, Kosaki K, Tanabe N, Tatsumi K, Saji T, Satoh T, Kataoka M, et al. 2017. A genome-wide association analysis identifies PDE1A—DNAJC10 locus on chromosome 2 associated with idiopathic pulmonary arterial hypertension in a Japanese population. *Oncotarget* 8(43):74917.

Koenig J, Werdehausen R, Linley JE, Habib AM, Vernon J, Lolignier S, Eijkelkamp N, Zhao J, Okorokov AL, Woods CG, et al. 2015. Regulation of Na$_v$ 1.7: a conserved SCN9A natural antisense transcript expressed in dorsal root ganglia. *PLoS One* 10(6):e0128830.

Koga T, Kawakami A. 2018. The role of camk4 in immune responses. *Mod Rheumatol*. 28(2):211–214.

Kumar S, Filipski A, Swarna V, Walker A, Hedges SB. 2005. Placing confidence limits on the molecular age of the human-chimpanzee divergence. *Proc Natl Acad Sci U S A*. 102(52):18842–18847.

Leffler EM, Gao Z, Pfeifer S, Ségurel L, Auton A, Venn O, Bowden R, Bontrop R, Wall JD, Sella G, et al. 2013. Multiple instances of ancient balancing selection shared between humans and chimpanzees. *Science* 339(6127):1578–1582.

Lefièvre L, de Lamirande E, Gagnon C. 2002. Presence of cyclic nucleotide phosphodiesterases PDE1A, existing as a stable complex with calmodulin, and PDE3A in human spermatozoa. *Biol Reprod*. 67(2):423–430.

Lefièvre L, Jha KN, de Lamirande E, Visconti PE, Gagnon C. 2012. Activation of protein kinase A during human sperm capacitation and acrosome reaction. *J Androl*. 33(5):1025–1716.

Lehner B. 2011. Molecular mechanisms of epistasis within and between genes. *Trends Genet*. 27(8):323–331.

Levy O. 2000. A neutrophil-derived anti-infective molecule: bactericidal/permeability-increasing protein. *Antimicrob Agents Chemother*. 44(11):2925–2931.

Li B, Castano AP, Hudson TE, Nowlin BT, Lin S-L, Bonventre JV, Swanson KD, Duffield JS. 2010. The melanoma-associated transmembrane glycoprotein Gpnmb controls trafficking of cellular debris for degradation and is essential for tissue repair. *FASEB J*. 24(12):4767–4781.

Loftus SK, Antonellis A, Matera I, Renaud G, Baxter LL, Reid D, Wolfsberg TG, Chen Y, Wang C, NISC Comparative Sequencing Program, et al. 2009. Gpnmb is a melanoblast-expressed, MITF-dependent gene. *Pigment Cell Melanoma Res*. 22(1):99–110.

Loh P-R, Lipson M, Patterson N, Moorjani P, Pickrell JK, Reich D, Berger B. 2013. Inferring admixture histories of human populations using linkage disequilibrium. *Genetics* 193(4):1233–1254.

Lohmueller KE, Bustamante CD, Clark AG. 2009. Methods for human demographic inference using haplotype patterns from genomewide single-nucleotide polymorphism data. *Genetics* 182(1):217–231.

Lonn E, Koskela E, Mappes T, Mokkonen M, Sims AM, Watts PC. 2017. Balancing selection maintains polymorphisms at neurogenetic loci in field experiments. *Proc Natl Acad Sci U S A*. 114(14):3690–3695.

Mackay TF. 2014. Epistasis and quantitative traits: using model organisms to study gene–gene interactions. *Nat Rev Genet*. 15(1):22–33.

Maibach V, Hans JB, Hvilsom C, Marques-Bonet T, Vigilant L. 2017. MHC class I diversity in chimpanzees and bonobos. *Immunogenetics* 69(10):661–676.

Maibach V, Vigilant L. 2019. Reduced bonobo MHC class I diversity predicts a reduced viral peptide binding ability compared to chimpanzees. *BMC Evol Biol*. 19(1):14.

Maric G, Rose AA, Annis MG, Siegel PM. 2013. Glycoprotein non-metastatic b (GPNMB): a metastatic mediator and emerging therapeutic target in cancer. *Onco Targets Ther*. 6:839–852.

Martinez A, Gil C. 2013. Phosphodiesterase inhibitors as a new therapeutic approach for the treatment of Parkinson's. Cambridge: Royal Society of Chemistry.

Masilamani M, Nguyen C, Kabat J, Borrego F, Coligan JE. 2006. CD94/NKG2A inhibits NK cell activation by disrupting the actin network at the immunological synapse. *J Immunol*. 177(6):3590–3596.

Meyer D, Aguiar VR, Bitarello BD, Brandt DY, Nunes K. 2018. A genomic perspective on HLA evolution. *Immunogenetics* 70(1):5–23.

Michibata H, Yanaka N, Kanoh Y, Okumura K, Omori K. 2001. Human Ca2+/calmodulin-dependent phosphodiesterase PDE1A: novel splice variants, their specific expression, genomic organization, and chromosomal localization. *Biochim Biophys Acta*. 1517(2):278–287.

Mitchell-Olds T, Willis JH, Goldstein DB. 2007. Which evolutionary processes influence natural genetic variation for phenotypic traits? *Nat Rev Genet*. 8(11):845–856.

Moorjani P, Patterson N, Hirschhorn JN, Keinan A, Hao L, Atzmon G, Burns E, Ostrer H, Price AL, Reich D. 2011. The history of African gene flow into southern Europeans, Levantines, and Jews. *PLoS Genet*. 7(4):e1001373.

Muirhead CA, Wakeley J. 2009. Modeling multiallelic selection using a Moran model. *Genetics* 182(4):1141–1157.

Nachman MW, Crowell SL. 2000. Estimate of the mutation rate per nucleotide in humans. *Genetics* 156(1):297–304.

Nadjar Y, Triller A, Bessereau J-L, Dumoulin A. 2015. The SUSD2 protein regulates neurite growth and excitatory synaptic density in hippocampal cultures. *Mol Cell Neurosci*. 65:82–91.

Naggert JK, Fricker LD, Varlamov O, Nishina PM, Rouille Y, Steiner DF, Carroll RJ, Paigen BJ, Leiter EH. 1995. Hyperproinsulinaemia in obese fat/fat mice associated with a carboxypeptidase E mutation which reduces enzyme activity. *Nat Genet*. 10(2):135–142.

Navarro A, Barton NH. 2002. The effects of multilocus balancing selection on neutral variability. *Genetics* 161(2):849–863.

Nielsen R, Williamson S, Kim Y, Hubisz MJ, Clark AG, Bustamante C. 2005. Genomic scans for selective sweeps using SNP data. *Genome Res*. 15(11):1566–1575.

O'Leary NA, Wright MW, Brister JR, Ciufo S, Haddad D, McVeigh R, Rajput B, Robbertse B, Smith-White B, Ako-Adjei D, et al. 2016. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res*. 44(D1):D733–D745.

Pace L, Salvan A, Sartori N. 2011. Adjusting composite likelihood ratio statistics. *Stat Sin*. 21(1):129–148.

Payseur BA, Nachman MW. 2000. Microsatellite variation and recombination rate in the human genome. *Genetics* 156(3):1285–1298.

Pekcec A, Schülert N, Stierstorfer B, Deiana S, Dorner-Ciossek C, Rosenbrock H. 2018. Targeting the dopamine D1 receptor or its downstream signalling by inhibiting phosphodiesterase-1 improves cognitive performance. *Br J Pharmacol*. 175(14):3021–3033.

Pende D, Sivori S, Accame L, Pareti L, Falco M, Geraghty D, Bouteiller PL, Moretta L, Moretta A. 1997. HLA-G recognition by human natural killer cells. Involvement of CD94 both as inhibitory and as activating receptor complex. *Eur J Immunol*. 27(8):1875–1880.

Pletikos M, Sousa AM, Sedmak G, Meyer KA, Zhu Y, Cheng F, Li M, Kawasawa YI, Šestan N. 2014. Temporal specification and bilaterality of human neocortical topographic gene expression. *Neuron* 81(2):321–332.

Prado-Martinez J, Sudmant PH, Kidd JM, Li H, Kelley JL, Lorente-Galdos B, Veeramah KR, Woerner AE, O'Connor TD, Santpere G, et al. 2013. Great ape genetic diversity and population history. *Nature* 499(7459):471–475.

Prüfer K, Munch K, Hellmann I, Akagi K, Miller JR, Walenz B, Koren S, Sutton G, Kodira C, Winer R, et al. 2012. The bonobo genome compared with the chimpanzee and human genomes. *Nature* 486(7404):527–531.

Puca AA, Spinetti G, Vono R, Vecchione C, Madeddu P. 2016. The genetics of exceptional longevity identifies new druggable targets for vascular protection and repair. *Pharmacol Res*. 114:169–174.

Quinlan AR. 2014. BEDTools: the Swiss-army tool for genome feature analysis. *Curr Protoc Bioinformatics*. 47(1):11.12.1–11.12.34.

Rajalingam R, Hong M, Adams EJ, Shum BP, Guethlein LA, Parham P. 2001. Short KIR haplotypes in pygmy chimpanzee (Bonobo) resemble the conserved framework of diverse human KIR haplotypes. *J Exp Med*. 193(1):135–146.

Reimann F, Cox JJ, Belfer I, Diatchenko L, Zaykin DV, McHale DP, Drenth JP, Dai F, Wheeler J, Sanders F, et al. 2010. Pain perception is altered by a nucleotide polymorphism in *SCN9A*. *Proc Natl Acad Sci U S A*. 107(11):5148–5153.

Ripoll VM, Irvine KM, Ravasi T, Sweet MJ, Hume DA. 2007. Gpnmb is induced in macrophages by IFN-γ and lipopolysaccharide and acts as a feedback regulator of proinflammatory responses. *J Immunol*. 178(10):6557–6566.

Rose AA, Annis MG, Dong Z, Pepin F, Hallett M, Park M, Siegel PM. 2010. ADAM10 releases a soluble form of the GPNMB/Osteoactivin extracellular domain with angiogenic properties. *PLoS One* 5(8):e12093.

Saitou N, Yamamoto F-I. 1997. Evolution of primate ABO blood group genes and their homologous genes. *Mol Biol Evol*. 14(4):399–411.

Sanchez-Mazas A. 2007. An apportionment of human HLA diversity. *Tissue Antigens* 69(Suppl 1):198–202.

Scally A, Durbin R. 2012. Revising the human mutation rate: implications for understanding human evolution. *Nat Rev Genet*. 13(10):745–753.

Scally A, Dutheil JY, Hillier LW, Jordan GE, Goodhead I, Herrero J, Hobolth A, Lappalainen T, Mailund T, Marques-Bonet T, et al. 2012. Insights into hominid evolution from the gorilla genome sequence. *Nature* 483(7388):169–175.

Ségurel L, Thompson EE, Flutre T, Lovstad J, Venkat A, Margulis SW, Moyse J, Ross S, Gamble K, Sella G, et al. 2012. The ABO blood group is a trans-species polymorphism in primates. *Proc Natl Acad Sci U S A*. 109(45):18493–18498.

Shao H, Burrage LC, Sinasac DS, Hill AE, Ernest SR, O'Brien W, Courtland H-W, Jepsen KJ, Kirby A, Kulbokas EJ, et al. 2008. Genetic architecture of complex traits: large phenotypic effects and pervasive epistasis. *Proc Natl Acad Sci U S A*. 105(50):19910–19914.

Sheehan S, Song YS. 2016. Deep learning for population genetic inference. *PLoS Comput Biol*. 12(3):e1004845.

Shum BP, Flodin LR, Muir DG, Rajalingam R, Khakoo SI, Cleland S, Guethlein LA, Uhrberg M, Parham P. 2002. Conservation and variation in human and common chimpanzee *CD94* and *NKG2* genes. *J Immunol*. 168(1):240–252.

Siewert KM, Voight BF. 2017. Detecting long-term balancing selection using allele frequency correlation. *Mol Biol Evol*. 34(11):2996–3005.

Siewert KM, Voight BF. 2020. Betascan2: standardized statistics to detect balancing selection utilizing substitution data. *Genome Biol Evol*. 12(2):3873–3877.

Sikela J, Adamson M, Wilson-Shaw D, Kozak C. 1990. Genetic mapping of the gene for Ca2+ calmodulin-dependent protein kinase IV (Camk-4) to mouse chromosome 18. *Genomics* 8(3):579–582.

Simes RJ. 1986. An improved Bonferroni procedure for multiple tests of significance. *Biometrika* 73(3):751–754.

Smukowski C, Noor M. 2011. Recombination rate variation in closely related species. *Heredity* 107(6):496–508.

Song YS, Steinrücken M. 2012. A simple method for finding explicit analytic transition densities of diffusion processes with general diploid selection. *Genetics* 190(3):1117–1129.

Spinelli CC, Carrizzo A, Ferrario A, Villa F, Damato A, Ambrosio M, Madonna M, Frati G, Fucile S, Sciaccaluga M, et al. 2017. LAV-BPIFB4 isoform modulates eNOS signalling through Ca^{2+}/PKC-α-dependent mechanism. *Cardiovasc Res*. 113(7):795–804.

Spinetti G, Sangalli E, Specchia C, Villa F, Spinelli C, Pipolo R, Carrizzo A, Greco S, Voellenkle C, Vecchione C, et al. 2017. The expression of the BPIFB4 and CXCR4 associates with sustained health in long-living individuals from Cilento-Italy. *Aging* 9(2):370–380.

Surbeck M, Deschner T, Schubert G, Weltring A, Hohmann G. 2012. Mate competition, testosterone and intersexual relationships in bonobos, *Pan paniscus*. *Anim Behav*. 83(3):659–669.

Suzuki H, Ahn HW, Chu T, Bowden W, Gassei K, Orwig K, Rajkovic A. 2012. SOHLH1 and SOHLH2 coordinate spermatogonial differentiation. *Dev Biol*. 361(2):301–312.

Sweeney CG, Rando JM, Panas HN, Miller GM, Platt DM, Vallender EJ. 2017. Convergent balancing selection on the mu-opioid receptor in primates. *Mol Biol Evol*. 34(7):1629–1643.

Tajima F. 1983. Evolutionary relationship of DNA sequences in finite populations. *Genetics* 105(2):437–460.

Tajima F. 1989. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* 123(3):585–595.

Takahata N, Satta Y, Klein J. 1992. Polymorphism and balancing selection at major histocompatibility complex loci. *Genetics* 130(4):925–938.

Takahata N, Satta Y, Klein J. 1995. Divergence time and population size in the lineage leading to modern humans. *Theor Popul Biol*. 48(2):198–221.

Tam AHT. 2015. Characterization of hippocalcin-like protein 1 (HPCAL1), a neuronal calcium sensor protein in the retina [PhD thesis]. Vancouver: University of British Columbia.

Teixeira JC, de Filippo C, Weihmann A, Meneu JR, Racimo F, Dannemann M, Nickel B, Fischer A, Halbwax M, Andre 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(5):1186–1196.

Tennessen JA. 2018. Gene buddies: linked balanced polymorphisms reinforce each other even in the absence of epistasis. *PeerJ* 6:e5110.

Terhorst J, Kamm JA, Song YS. 2017. Robust and scalable inference of population history from hundreds of unphased whole genomes. *Nat Genet*. 49(2):303–309.

Toyoda S, Miyazaki T, Miyazaki S, Yoshimura T, Yamamoto M, Tashiro F, Yamato E, Miyazaki J-I. 2009. Sohlh2 affects differentiation of KIT positive oocytes and spermatogonia. *Dev Biol*. 325(1):238–248.

Ubeda F, Haig D. 2004. Sex-specific meiotic drive and selection at an imprinted locus. *Genetics* 167(4):2083–2095.

Uhlén M, Fagerberg L, Hallström BM, Lindskog C, Oksvold P, Mardinoglu A, Sivertsson Å, Kampf C, Sjöstedt E, Asplund A, et al. 2015. Tissue-based map of the human proteome. *Science* 347(6220):1260419.

van der Ven PF, Speel EJ, Albrechts JC, Ramaekers FC, Hopman AH, Fürst DO. 1999. Assignment of the human gene for endosarcomeric cytoskeletal M-protein (MYOM2) to 8p23. 3. *Genomics* 55(2):253–255.

Varin C, Reid N, Firth D. 2011. An overview of composite likelihood methods. *Stat Sin*. 21(1):5–42.

Vasta V, Sonnenburg WK, Yan C, Soderling SH, Shimizu-Albergine M, Beavo JA. 2005. Identification of a new variant of pde1a calmodulin-stimulated cyclic nucleotide phosphodiesterase expressed in mouse sperm. *Biol Reprod*. 73(4):598–609.

Villa F, Carrizzo A, Spinelli CC, Ferrario A, Malovini A, Macią g A, Damato A, Auricchio A, Spinetti G, Sangalli E, et al. 2015. Genetic analysis reveals a longevity-associated protein modulating endothelial function and angiogenesis. *Circ Res*. 117(4):333–345.

Villa F, Ciaglia E, Maciag A, Montella F, Ferrario A, Cattaneo M, Puca A. 2018. Longevity associated variant of *BPIFB4* mitigates monocyte mediated acquired immune response. *Innov Aging*. 2(Suppl 1):884.

Villa F, Malovini A, Carrizzo A, Spinelli CC, Ferrario A, Macią g A, Madonna M, Bellazzi R, Milanesi L, Vecchione C, et al. 2015. Serum BPIFB4 levels classify health status in long-living individuals. *Immun Ageing*. 12(1):27.

Walter L. 2014. Immunogenetics of NK cell receptors and MHC Class I ligands in non-human primates. In: Ansari AA, Silvestri G, editors. Natural hosts of SIV. Amsterdam: Elsevier. p. 269–285.

Watson AP, Evans RL, Egland KA. 2013. Multiple functions of sushi domain containing 2 (Susd2) in breast tumorigenesis. *Mol Cancer Res*. 11(1):74–85.

Wigginton JE, Cutler DJ, Abecasis GR. 2005. A note on exact tests of Hardy-Weinberg equilibrium. *Am J Hum Genet*. 76(5):887–893.

Wills C. 1991. Maintenance of multiallelic polymorphism at the MHC region. *Immunol Rev*. 124(1):165–220.

Wobber V, Hare B, Maboto J, Lipson S, Wrangham R, Ellison PT. 2010. Differential changes in steroid hormones before competition in bonobos and chimpanzees. *Proc Natl Acad Sci U S A*. 107(28):12457–12462.

Wrangham RW. 1993. The evolution of sexuality in chimpanzees and bonobos. *Hum Nat*. 4(1):47–79.

Wroblewski EE, Guethlein LA, Norman PJ, Li Y, Shaw CM, Han AS, Ndjango J-BN, Ahuka-Mundeke S, Georgiev AV, Peeters M, et al. 2017. Bonobos maintain immune system diversity with three functional types of MHC-B. *J Immunol*. 198(9):3480–3493.

Wroblewski EE, Parham P, Guethlein LA. 2019. Two to tango: co-evolution of hominid natural killer cell receptors and MHC. *Front Immunol*. 10: 177.

Wu D, Chen Y, Chen Q, Wang G, Xu X, Peng A, Hao J, He J, Huang L, Dai J. 2019. Clinical presentation and genetic profiles of Chinese patients with velocardiofacial syndrome in a large referral centre. *J Genet*. 98(2):42.

Wu JY, Ribar TJ, Cummings DE, Burton KA, McKnight GS, Means AR. 2000. Spermiogenesis and exchange of basic nuclear proteins are impaired in male germ cells lacking camk4. *Nat Genet*. 25(4):448–452.

Xu Y, Zhang H-T, O'Donnell JM. 2011. Phosphodiesterases in the central nervous system: implications in mood and cognitive disorders. In: Francis SH, Conti M, Houslay MD, editors. Phosphodiesterases as drug targets. Berlin, Heidelberg: Springer. p. 447–485.

Yakut S, Cetin Z, Clark OA, Usta MF, Berker S, Luleci G. 2013. Exceptional complex chromosomal rearrangement and microdeletions at the 4q22. 3q23 and 14q31. 1q31. 3 regions in a patient with azoospermia. *Gene* 512(1):157–160.

Yan C, Bentley JK, Sonnenburg WK, Beavo JA. 1994. Differential expression of the 61 kDa and 63 kDa calmodulin-dependent phosphodiesterases in the mouse brain. *J Neurosci*. 14(3):973–984.

Yang Y, Wang Y, Li S, Xu Z, Li H, Ma L, Fan J, Bu D, Liu B, Fan Z, et al. 2004. Mutations in *SCN9A*, encoding a sodium channel alpha subunit, in patients with primary erythermalgia. *J. Med. Genet.* 41(3):171–174.

Ye CJ, Chen J, Villani A-C, Gate RE, Subramaniam M, Bhangale T, Lee MN, Raj T, Raychowdhury R, Li W, et al. 2018. Genetic analysis of isoform usage in the human anti-viral response reveals influenza-specific regulation of ERAP2 transcripts under balancing selection. *Genome Res.* 28(12):1812–1825.

Yu B, Sondag GR, Malcuit C, Kim M-H, Safadi FF. 2016. Macrophage-associated Osteoactivin/GPNMB mediates mesenchymal stem cell survival, proliferation, and migration via a cd44-dependent mechanism. *J Cell Biochem*. 117(7):1511–1521.

Zhou L, Liu F, Li Y, Peng Y, Liu Y, Li J. 2012. Gpnmb/osteoactivin, an attractive target in cancer immunotherapy. *Neoplasma* 59(01):1–5.