The search for sexually antagonistic genes: Practical insights from studies of local adaptation and statistical genomics

Filip Ruzicka,1,2 Ludovic Dutoit,3,∗ Peter Czuppon,4,5 Crispin Y. Jordan,6 Xiang-Yi Li,7 Colin Olito,8 Anna Runemark,8 Erik I. Svensson,8 Homa Papoli Yazdi,8 and Tim Connallon1,9

1School of Biological Sciences, Monash University, Clayton VIC 3800, Australia
2Department of Zoology, University of Otago, Dunedin 9054, New Zealand
3Institute of Ecology and Environmental Sciences, UPEC, CNRS, IRD, INRA, Sorbonne Université, Paris 75252, France
4Center for Interdisciplinary Research in Biology, CNRS, Collège de France, PSL Research University, Paris 75231, France
5School of Biomedical Sciences, University of Edinburgh, Edinburgh EH8 9XD, United Kingdom
6Institute of Biology, University of Neuchâtel, Neuchâtel CH-2000, Switzerland
7Department of Biology, Lund University, Lund SE-22362, Sweden
8E-mail: Filip.Ruzicka@monash.edu
9E-mail: tim.connallon@monash.edu

Received April 28, 2020
Accepted July 28, 2020

Sexually antagonistic (SA) genetic variation—in which alleles favored in one sex are disfavored in the other—is predicted to be common and has been documented in several animal and plant populations, yet we currently know little about its pervasiveness among species or its population genetic basis. Recent applications of genomics in studies of SA genetic variation have highlighted considerable methodological challenges to the identification and characterization of SA genes, raising questions about the feasibility of genomic approaches for inferring SA selection. The related fields of local adaptation and statistical genomics have previously dealt with similar challenges, and lessons from these disciplines can therefore help overcome current difficulties in applying genomics to study SA genetic variation. Here, we integrate theoretical and analytical concepts from local adaptation and statistical genomics research—including FST and FIS statistics, genome-wide association studies, pedigree analyses, reciprocal transplant studies, and evolve-and-resequence experiments—to evaluate methods for identifying SA genes and genome-wide signals of SA genetic variation. We begin by developing theoretical models for between-sex FST and FIS, including explicit null distributions for each statistic, and using them to critically evaluate putative multilocus signals of sex-specific selection in previously published datasets. We then highlight new statistics that address some of the limitations of FST and FIS, along with applications of more direct approaches for characterizing SA genetic variation, which incorporate explicit fitness measurements. We finish by presenting practical guidelines for the validation and evolutionary analysis of candidate SA genes and discussing promising empirical systems for future work.

KEY WORDS: Evolutionary genomics, fitness variation, FST, FIS, genomic inference of natural selection, intralocus sexual conflict, local adaptation.

* Ludovic Dutoit is a co-first author, along with Filip Ruzicka
Impact Summary

Genome sequences carry a record of the evolutionary and demographic histories of natural populations. Research over the last two decades has dramatically improved our ability to detect genomic signals of adaptation by natural selection, including several widely-used methods for identifying genes underlying local adaptation and quantitative trait variation. Yet the application of these methods to identify sexually antagonistic (SA) genes—wherein variants that are adaptive for one sex are maladaptive for the other—remains under-explored, despite the potential importance of SA selection as a mechanism for maintaining genetic variation. Indeed, several lines of evidence suggest that SA genetic variation is common within animal and plant populations, underscoring the need for analytical methods that can reliably identify SA genes and genomic signals of SA genetic variation. Here, we integrate statistics and experimental designs that were originally developed within the fields of local adaptation and statistical genomics and apply them to contexts of sex-specific adaptation and SA genetic variation. First, we evaluate and extend statistical methods for identifying signals of SA variation from genome sequence data alone. We then apply these methods to reanalyze previously published data on allele frequency differences between sexes—a putative signal of SA selection. Second, we highlight more direct approaches for identifying SA genetic variation, which use experimental evolution and statistical associations between individual genetic variants and fitness. Third, we provide guidelines for the biological validation, evolutionary analysis, and interpretation of candidate SA polymorphisms. By building upon the strong methodological foundations of local adaptation and statistical genomics research, we provide a roadmap for rigorous analyses of genetic data in the context of sex-specific adaptation, thereby facilitating insights into the role and pervasiveness of SA variation in adaptive evolution.

A population’s evolutionary capacity for adaptation hinges upon the nature and extent of the genetic variation it harbors (Fisher 1930). In simple environments where selection is uniform over time, across space, and among different classes of individuals within the population, adaptation may proceed by fixing unconditionally beneficial mutations and eliminating deleterious ones. Yet species exist in complex environments, where opportunities for adaptation can be limited by genetic trade-offs among traits and fitness components (Otto 2004; Gomulkiewicz and Houle 2009; Chevin 2013; Connallon and Hall 2018) or by gene flow and conflicting directional selection among habitats within a species’ range (Kirkpatrick and Barton 1997; Lenormand 2002; Duputić et al. 2012). Such contexts allow maladaptation to persist in spite of abundant genetic variation within the population (Walsh and Blows 2009).

“Sexually antagonistic” (SA) genetic variation—wherein alleles that are beneficial when expressed in one sex are harmful when expressed in the other—represents a particularly common form of genetic trade-off (Rice and Chippindale 2001; Bonduriansky and Chenoweth 2009; Van Doorn 2009). SA genetic variation arises from sex differences in selection (a.k.a. sex-specific selection) on traits that are genetically correlated between the sexes (Connallon and Clark 2014b), and may contribute substantially to fitness variation (Kidwell et al. 1977; Abbott 2011; Connallon and Clark 2014a; Oliito et al. 2018) and maladaptation (Lande 1980; Matthews et al. 2019). Estimates of phenotypic selection suggest that sex differences in directional selection are common (Cox and Calsbeek 2009; Lewis et al. 2011; Gosden et al. 2012; Stearns et al. 2012; Morrissey 2016; De Lisle et al. 2018; Singh and Punzalan 2018), implying that many genetic variants affecting quantitative traits have SA effects on fitness. Likewise, estimates of genetic variation for fitness suggest that the genetic basis of female and male fitness components is partially discordant, with some multilocus genotypes conferring high fitness in one sex and low fitness in the other (Chippindale et al. 2001; reviewed in Connallon and Matthews 2019).

Although studies of sex-specific selection indicate that SA alleles contribute to fitness variation in several animal and plant populations (e.g., Chippindale et al. 2001; Fedorka and Mousseau 2004; Svensson et al. 2009; Delph et al. 2011; Mokkonen et al. 2011; Berger et al. 2014), the population genetic basis of this fitness variation is largely unknown, leaving several important questions unanswered. For example, what fraction of genetic variance for fitness is attributable to SA alleles versus other classes of genetic variation (e.g., deleterious mutations)? Is SA genetic variation attributable to many small-effect loci or to a few large-effect loci? Are SA polymorphisms maintained under balancing selection, or are they transient and primarily evolving via mutation, directional selection, and drift? Are SA alleles randomly distributed across the genome or are they enriched on certain chromosome types (e.g., sex chromosomes)? These questions are part of the broader debate about the genetic basis of fitness variation and the evolutionary forces that maintain it (Lewontin 1975; Charlesworth and Hughes 1999), which is one of the oldest in this field and perhaps the most difficult to resolve (e.g., Lewontin 1975, p. 23).

As in most areas of evolutionary biology, research on SA selection is increasingly drawing upon genomics. A few studies have identified candidate SA polymorphisms with large effects on traits related to fitness (Roberts et al. 2009; Barson et al. 2015; Rostant et al. 2015; VanKuren and Long 2018; Pearse et al. 2019), and this handful of SA loci almost certainly represents the tip of the iceberg (Ruzicka et al. 2019). Many other studies have
highlighted genomic patterns of expression or sequence diversity that could be indicative of sex-specific selection (e.g., Innocenti and Morrow 2010; reviewed in Kasimatis et al. 2017; Mank 2017; Rowe et al. 2018).

Although there is optimism that genomics will facilitate the study of sex-specific selection, we still face several challenges in applying genomic data to identify and characterize SA genetic variation. For example, some putative genomic signals of sex-specific selection, such as sex-biased gene expression, are ambiguous: at best, they may serve as indirect proxies of sex-specific selection, or at worst provide no information about contemporary selection on each sex (Kasimatis et al. 2017; Rowe et al. 2018). Allele frequency differences between sexes (e.g., between-sex $F_{ST}$) may represent more straightforward genomic signatures of sex-specific selection (e.g., Cheng and Kirkpatrick 2016; Lucotte et al. 2016; but see Bissegger et al. 2019; Kasimatis et al. 2020). Yet ambiguous null hypotheses for empirical estimates of between-sex $F_{ST}$, along with high statistical noise relative to biological signal in these estimates, raise questions about statistical power and the prevalence of false positives within such data (Kasimatis et al. 2019). In addition, we need to better understand the extent to which common pitfalls of genome sequence datasets—for example, mismapped reads (Tsai et al. 2019), sampling biases, hidden population structure, and effects of linkage and hitchhiking—yield artifactual signals of sex-specific processes, and thus questionable candidate SA genes.

The challenges in applying genomics to sex-specific selection have strong parallels in the fields of local adaptation and statistical genomics. Although the study of SA loci is still in its infancy, the fields of local adaptation and statistical genomics have already grappled intensively with many of the conceptual and methodological challenges that research on sex-specific selection now faces (Hoban et al. 2016; Visscher et al. 2017). For example, local adaptation research has long emphasized the importance of clear null models for distinguishing genes involved in local adaptation from false positives that simply reside in the tails of neutral null distributions (Lewontin and Krakauer 1973; Günther and Coop 2013; Whitlock and Lotterhos 2015; Lohse 2017). Similarly, statistical genomics researchers have repeatedly warned that hidden population structure in genome-wide association studies (GWAS) can lead to spurious conclusions about the genetic basis of quantitative traits, complex diseases, and the role of adaptation in population differentiation (Lander and Schork 1994; Price et al. 2010; Barton et al. 2019; Berg et al. 2019; Sohail et al. 2019). Lessons from local adaptation and statistical genomics research can therefore sharpen hypothesis framing, guide statistical methodology, and inform best practices for disentangling signal, noise, and artifacts in studies of sex-specific selection.

Here, by drawing insights from local adaptation and statistical genomics research, we present practical guidelines for population genomic analyses of sex-specific fitness variation. We first outline two statistics that can, in principle, provide indirect evidence of sex-specific fitness effects of genetic variation: between-sex $F_{ST}$, which is sensitive to sex differences in viability selection and some components of reproductive success, and $F_{IS}$, a measure of Hardy-Weinberg deviations in diploids, which is sensitive to sex differences in overall selection (i.e., cumulative effects of viability, fertility, fecundity, and mating competition). We develop theoretical null models for each metric, provide an overview of their sampling distributions and statistical power, and present a reanalysis of published $F_{ST}$ data in light of our models. We also highlight complementary methods adapted from case-control GWAS to overcome some of the limitations of these metrics. Second, we evaluate several direct approaches for characterizing sex-specific genetic variation for fitness, which combine elements from quantitative genetics, reciprocal transplant studies, and experimental evolution. These direct approaches have been extensively employed to study the genetic basis of locally adapted phenotypes and quantitative traits, but rarely to identify SA loci. Third, we outline approaches for validating candidate genes and discuss best practices for the analysis and interpretation of their evolutionary histories.

**Indirect Approaches for Identifying SA Genes**

Estimating fitness under natural conditions is difficult, rendering approaches for identifying SA genes that rely on fitness measurements (i.e., direct approaches; see section “Direct Approaches for Identifying SA Genes”) unfeasible for many populations. Any widely applicable approach must therefore make use of indirect empirical signals of SA selection in genome sequence data, which can now be collected for virtually any species.

Two specific patterns of genome sequence variation could be indicative of contemporary SA selection, as emphasized by several recent studies (e.g., Cheng and Kirkpatrick 2016; Lucotte et al. 2016; Eyer et al. 2019; Kasimatis et al. 2019). First, because sex differences in selection during the life cycle are expected to generate allele frequency differences between breeding females and males (i.e., the members of each sex that contribute to offspring of the next generation; see Box 1), allele frequency differences between samples of females and males of a population could be indicative of sex differences in selection—including SA selection, sex-limited selection, or ongoing sexually concordant selection that differs in magnitude between the sexes. Ideally, inferences of sex-specific selection from allele frequency estimates should be based on samples of breeding adults that have
passed the filter of viability selection, sexual selection, and fertility/fecundity selection, although in practice genome sequences of random samples of adults are more readily obtainable and will reflect viability selection (Cheng and Kirkpatrick 2016; Kasimatis et al. 2019). Second, allele frequency differences between breeding females and males of a given generation elevate heterozygosity in offspring of the next generation relative to Hardy-Weinberg expectations (Kasimatis et al. 2019). Inflated heterozygosity in a large random sample of offspring could therefore reflect sex differences in viability selection, sexual selection, and/or fertility and fecundity selection during the previous generation.

**SINGLE-LOCUS SIGNALS OF SEX-SPECIFIC SELECTION IN FIXATION INDICES**

Fixation indices, which are widely applied in studies of population structure (Wright 1951), can be used to quantify allele frequency differences between sexes ($F_{ST}$) and elevations in heterozygosity in the offspring of a given generation ($F_{IS}$), each of which are predicted consequences of sex-specific selection (Boxes 1-2). Several studies have estimated $F_{ST}$ between sexes using gene sequences sampled from adults (Cheng and Kirkpatrick 2016; Lucotte et al. 2016; Flanagan and Jones 2017; Wright et al., 2018, 2019; Bissegger et al. 2019; Vaux et al. 2019) or from breeding individuals (Dutoit et al. 2018), yet it remains unclear how much information about SA selection is contained within these estimates. A major problem, as recently emphasized by Kasimatis et al. (2019), is that the contribution of sex-specific selection to allele frequency differentiation between the sexes may often be weak compared to effects of sampling error in allele frequency estimates. Indeed, simulations presented in several studies (e.g., Cheng and Kirkpatrick 2016; Lucotte et al. 2016; Connallon and Hall 2018; Kasimatis et al. 2019) confirm that signals of SA selection in between-sex $F_{ST}$ are swamped by sampling error in small population genomic datasets. Nevertheless, without clearly defined probability distributions for between-sex $F_{ST}$ and related metrics of allele frequency differentiation, it remains difficult to evaluate whether sampling error, by itself, is sufficient to account for the empirical distributions of population genomic metrics that are putatively associated with sex-specific selection.

As we show in Appendix A (Supporting Information) (see Box 2), a null distribution for between-sex $F_{ST}$ estimates at loci with no sex differences in selection conforms to a special case of Lewontin and Krakauer’s (1973) classic null model for $F_{ST}$ estimated between populations. An appealing feature of our two-sex null model is its insensitivity to some of the simplifying assumptions inherent in Lewontin and Krakauer’s original model (i.e., that subpopulations are independent; see Nei and Maruyama 1975; Robertson 1975; Charlesworth 1998; Beaumont 2005; Whitlock and Lotterhos 2015), or issues arising from genetic linkage (Charlesworth 1998), which do not affect the two-sex null distribution when $F_{ST}$ is independently estimated per single nucleotide polymorphism (SNP), but can strongly impact the null for concatenated sequences (e.g., gene-wide $F_{ST}$ estimates; see Booker et al. 2020; Appendix A [Supporting Information]). Another appealing feature of the null model for $F_{ST}$ is its insensitivity to the distribution of allele frequencies in the population, provided SNPs with very low minor allele frequencies (MAFs) are excluded from the analysis (e.g., MAFs < 0.05 for small datasets; MAFs < 0.01 for large datasets; see Appendices A and D [Supporting Information]). This insensitivity to MAF under the null is unique to $F_{ST}$; other metrics of allele frequency differentiation strongly covary with MAF, so that simulations based on the MAF distribution of the study population are required to generate genome-wide null predictions against which data can be compared (Appendix D [Supporting Information]; Figs. S6 and S7; see Berner 2019 and Bissegger et al. 2019 for examples).

By comparing the distribution of between-sex $F_{ST}$ under the null (i.e., no sex differences in selection) with the corresponding distribution under SA selection (Box 2), we can formally evaluate the minimum strength of selection ($s_{min}$, defined as the minimum cost, per sex, of inheriting the “wrong” SA allele) required for SA loci to reliably reside within the upper tail of the null distribution. For example, the 99th percentile for the null distribution is $F_{ST(99\%)} \approx 3.32/n_H$, where $n_H$ is the harmonic mean sample size of female- and male-derived gene sequences, and $F_{ST}$ refers to an estimate of $F_{ST}$ (see Box 2; results are based on Nei’s 1973 estimator for $F_{ST}$, which closely aligns with Wright’s 1951 definition for $F_{ST}$ between a pair of populations; see Appendix A [Supporting Information] for discussion of alternative $F_{ST}$ estimators). Roughly 1% of $F_{ST}$ estimates should fall above this threshold when there are no sex differences in selection. The probability that a SA locus resides within the tail of the null distribution depends on the allele frequencies at the locus, the strength of selection, and the sample size of individuals that are sequenced (Appendix A [Supporting Information]). In studies with large sample sizes (i.e., $n_H = 10^5$ or greater, as in some human genomic datasets: see Fig. 1), $F_{ST}$ for a SA locus with intermediate allele frequencies and a fitness effect of a few percent will reliably fall within the upper tail of the null distribution (Fig. 1; Appendix D [Supporting Information]). In contrast, studies where $n_H < 10^4$ require very strong selection ($s_{min} > 0.05$) for SA loci to reliably reside within the upper tail of the null $F_{ST}$ distribution (Fig. 1), and are unlikely to identify individual SA loci (i.e., significant $F_{ST}$ outliers), even in cases where SA genetic polymorphism is common throughout the genome. Under scenarios of sex-limited selection, or sexually concordant selection that differs in magnitude between the sexes, sample sizes required to reliably identify true $F_{ST}$ outliers must be even larger, as the degree of allele frequency differentiation under sex-limited
estimates of porting Information]). Although only a single study has used signals of elevated heterozygosity are expected to be strongest

Of SA selection from noise (Fig. 1A), let alone distinguishing of SA selection from noise (Fig. 1A), let alone distinguishing

values for the 5% tail) for published between-sex studies that include eight human datasets (Cheng and Kirkpatrick 2016; Lucotte et al. 2016; Kasimatis et al. 2020; Pirastu et al. 2020), four fish datasets (see Flanagan and Jones 2017; Wright et al., 2018, 2019; Bissegger et al. 2019; Vaux et al. 2019), and seven bird datasets (see Dutout et al. 2018; Wright et al. 2019). The harmonic mean sample size, \( n_{\text{min}} \), refers to the number of genes sequenced from diploid female and male samples in studies of between-sex studies that include eight human datasets (Cheng and Kirkpatrick 2016; Lucotte et al. 2016; Kasimatis et al. 2020; Pirastu et al. 2020), four fish datasets (see Flanagan and Jones 2017; Wright et al., 2018, 2019; Bissegger et al. 2019; Vaux et al. 2019), and seven bird datasets (see Dutout et al. 2018; Wright et al. 2019). The harmonic mean sample size, \( n_{\text{min}} \), refers to the number of genes sequenced from diploid female and male samples in studies of between-sex studies that include eight human datasets (Cheng and Kirkpatrick 2016; Lucotte et al. 2016; Kasimatis et al. 2020; Pirastu et al. 2020), four fish datasets (see Flanagan and Jones 2017; Wright et al., 2018, 2019; Bissegger et al. 2019; Vaux et al. 2019), and seven bird datasets (see Dutout et al. 2018; Wright et al. 2019). The harmonic mean sample size, \( n_{\text{min}} \), refers to the number of genes sequenced from diploid female and male samples in studies of between sex IS (see Box 2). (B) Probability that \( F_{ST} \) for an additive SA locus (i.e., \( h_T = h_M = \frac{1}{2} \)) with intermediate equilibrium allele frequencies \((\rho, \eta = 1/2)\) is in the top 5% tail of the null distribution for \( F_{ST} \). \( |s_m| \) is the fitness cost of being homozygous for the “wrong” SA allele (see Appendices A and B [Supporting Information]). See Appendices C and D (Supporting Information) for further analyses of statistical power.

Figure 1. Signal of SA selection relative to sampling error in \( F_{ST} \) and \( F_{IS} \) estimates. (A) Lines show the minimum strength of selection \( (s_{\text{min}}; \text{see Appendices A and B [Supporting Information])} \) that causes the expected values of \( \hat{F}_{IS} \) and \( \hat{F}_{ST} \) at an SA locus to reach the 5% tail (broken line) or the 1% tail (solid line) of the null distributions for \( \hat{F}_{ST} \) (black lines) and \( \hat{F}_{IS} \) (gray lines). Circles show the theoretical \( s_{\text{min}} \) values (for the 5% tail) for published between-sex \( F_{ST} \) studies that include eight human datasets (Cheng and Kirkpatrick 2016; Lucotte et al. 2016; Kasimatis et al. 2020; Pirastu et al. 2020), four fish datasets (see Flanagan and Jones 2017; Wright et al., 2018, 2019; Bissegger et al. 2019; Vaux et al. 2019), and seven bird datasets (see Dutout et al. 2018; Wright et al. 2019). The harmonic mean sample size, \( n_{\text{min}} \), refers to the number of genes sequenced from diploid female and male samples in studies of between sex IS (see Box 2). (B) Probability that \( F_{ST} \) for an additive SA locus (i.e., \( h_T = h_M = \frac{1}{2} \)) with intermediate equilibrium allele frequencies \((\rho, \eta = 1/2)\) is in the top 5% tail of the null distribution for \( F_{ST} \). \( |s_m| \) is the fitness cost of being homozygous for the “wrong” SA allele (see Appendices A and B [Supporting Information]). See Appendices C and D (Supporting Information) for further analyses of statistical power.

selection is roughly half the differentiation expected under SA selection, and differentiation is further muted under sexually concordant selection.

A second putative signal of sex-specific selection is an enrichment of heterozygotes among offspring cohorts, as inferred from high values of \( F_{IS} \) (defined in Box 2; Appendix B [Supporting Information]). Although only a single study has used estimates of \( F_{IS} \) (\( \hat{F}_{IS} \)) to test for sex-specific selection (Eyer et al. 2019; see Boxes 1-2), the potential for future applications warrants evaluation of signals of sex-specific selection using this metric. Hardy-Weinberg deviations in a sample, as captured by \( \hat{F}_{IS} \), may arise from selection, nonrandom mating (e.g., inbreeding or population structure), or random sampling of genotypes from the population (Crow and Kimura 1970; Weir 1997; Lachance 2009). Statistical properties of Hardy-Weinberg deviations in genotype samples are well established (Weir 1997), and easily adaptable for our point of interest: the distribution of \( \hat{F}_{IS} \) in a randomly mating population in which allele frequencies may differ between the female and male parents of a given generation (Box 2). As illustrated in Box 2, the sampling variance for \( \hat{F}_{IS} \) exceeds that of \( \hat{F}_{ST} \) by a substantial margin. Consequently, \( \hat{F}_{IS} \) has far less power than \( \hat{F}_{ST} \) to distinguish signal of SA selection from noise (Fig. 1A), let alone distinguishing sex-limited selection or sexually concordant selection that differs in magnitude between the sexes. An additional issue is that signals of elevated heterozygosity are expected to be strongest among cohorts sampled at birth, yet selection occurring during the life cycle can potentially decrease heterozygosity, further dampening signals of sex-specific selection in \( \hat{F}_{IS} \) estimates from adult samples.

\( F_{ST} \) DISTRIBUTIONS AND MULTI-LOCUS SIGNALS OF SEX-SPECIFIC SELECTION

Genome scans for individually significant SA loci (\( \hat{F}_{IS} \) or \( \hat{F}_{ST} \) outliers) are severely underpowered unless SA loci segregate for intermediate frequency alleles with large fitness effects and sample sizes are very large (see above). Indeed, no empirical \( F_{ST} \) study to date has yielded individually significant autosomal candidate SA SNPs that have survived corrections for multiple-testing and rigorous controls for genotyping error and read-mapping artifacts (see below).

Although \( F_{ST} \) scans for individually significant SA loci are highly conservative, the full empirical distribution of \( F_{ST} \) for autosomal SNPs may nonetheless carry a cumulative signature of sex-specific selection at many loci—even in the absence of individually significant SA genes. For example, SNPs responding to sex-specific selection (i.e., SNPs with sex-specific fitness effects or SNPs in linkage disequilibrium with them) should inflate average \( \hat{F}_{ST} \) and the proportion of observations in upper quantiles predicted by null models (Fig. 2). An excess of observations in the upper quantiles of the theoretical null may imply an enrichment of SNPs responding to sex-specific selection in the tail of the
empirical \( F_{ST} \) distribution, with SNPs in the tail representing interesting candidates for follow-up analyses (see section “Validation and Follow-up Analyses of Candidate SA Genes”). Moreover, the fraction of true versus false positives among candidates can be quantified. For example, if 2% of observed SNPs fall within the top 1% quantile of the theoretical null, this implies a 1:1 ratio of true to false positives within the top 2% of observations (i.e., a false discovery rate of 50%).

To test for elevation of empirical \( F_{ST} \) estimates relative to our theoretical null model, we reanalyzed three representative datasets from previously published studies (collared flycatcher *Ficedula albicollis*: Dutoit et al. 2018; gulf pipefish *Synagnathus scovelli*: Flanagan and Jones 2017; human: The 1000 Genomes Project Consortium 2015 data used by Cheng and Kirkpatrick 2016; Code available at https://github.com/ldutoit/male_female_fst). For human and flycatcher whole-genome re-sequencing datasets, we used autosomal coding variants, excluding any SNP with missing data. For the pipefish RAD-seq dataset, coding and noncoding variants with less than 50% missing data were included; sex-linked regions are unknown in this species and could not be excluded. In all datasets, polymorphic sites with MAFs below 5% were also excluded, as sites with low MAF exhibit inflated sampling variances (see Whitlock and Lotterhos 2015; Appendix A [Supporting Information] and Figs. S1-S7). Analyses were carried out in bedtools (Quinlan and Hall 2010), vcftools (Danecek et al. 2011), and R (R Core Team 2020).

For all three datasets, permuted \( F_{ST} \) distributions (i.e., \( F_{ST} \) calculated after randomly permuting sex labels across individuals) conform well to the theoretical null model for \( F_{ST} \) (Figs. S1-S4). For the 1000 Genomes human dataset, \( F_{ST} \) observations (i.e., nonpermuted estimates) are indistinguishable from both the theoretical null and permuted distributions, with no enrichment of observations within the top quantiles of the null (Fig. 2; \( \chi^2 \) tests; 5% tail: \( P = 0.18 \); 1% tail: \( P = 0.33 \); comparison between the nonpermuted mean \( F_{ST} \) and the \( F_{ST} \) means for 1000 permutations of the data: \( P > 0.05 \)). In contrast, flycatcher and pipefish datasets show elevated \( F_{ST} \) values relative to the 5% and 1% tails of their null distributions (Fig. 2; \( \chi^2 \) tests; \( P = 4.33 \times 10^{-9} \) and \( P < 2.2 \times 10^{-16} \) for flycatcher; \( P < 2.2 \times 10^{-16} \) and \( P < 2.2 \times 10^{-16} \) for pipefish), and exhibit inflated means for observed \( F_{ST} \) relative to the \( F_{ST} \) means of 1000 permutations of the data (\( P < 0.05 \) for both datasets). Such enrichment shows that sampling error is not sufficient to explain empirical distributions of \( F_{ST} \), and instead implies that many loci are responding to sex differences in selection (either directly or indirectly through hitchhiking with selected loci), or that a false signal of elevated \( F_{ST} \) has been generated by population structure and/or data quality issues, as discussed below.
ACCOUNTING FOR SPURIOUS SIGNALS OF SEX-SPECIFIC SELECTION

The analyses presented above suggest that \( F_{ST} \), although severely underpowered for detecting individually significant outlier SNPs, may capture polygenic signals of sex-specific selection. However, sex differences in selection should not be invoked as the cause of such elevations in \( F_{ST} \) without excluding artifacts that may generate similar patterns.

First, incorrect mapping of sex-linked markers to autosomes can potentially lead to artificial inflation of \( F_{ST} \) estimates. For example, Y- or W-linked sequences may be erroneously mapped to sequence paralogs on autosomes (Tsai et al. 2019), resulting in artificially high \( F_{ST} \) inferences at autosomal sites (Bissegger et al. 2019; Kasimatis et al. 2020). This problem can be mitigated in species with high-quality reference genomes, where unmapped reads can be eliminated through quality-filtering steps (i.e., removal of SNPs associated with low MAFs or extreme deviations from Hardy-Weinberg expectations), removal of candidate regions showing high sequence similarity to sex chromosome sequences (Kasimatis et al. 2020), and excluding regions with sex-biases in read coverage. However, mismapping is difficult to control for in species lacking high-quality reference genomes, including those where the sex determination system is unknown (e.g., the pipefish dataset in Fig. 2), or where sex chromosomes are young or rapidly evolving. Moreover, the effects of demographic processes, including recent admixture events or sex-biased migration, play out differently between sex chromosomes and autosomes (Hedrick 2007), so that caution is required in interpreting elevated between-sex \( F_{ST} \) on the X or Z, as has been reported in humans (e.g., Lucotte et al. 2016).

Second, sex differences in population structure—arising from the broad geographic sampling of individuals or recent migration into a single sampled population—can also generate signals of genetic differentiation between females and males in the absence of sex differences in selection (Box 1). Taxa with broad contemporary distributions (e.g., humans and Drosophila) often show significant genetic differentiation among populations. Uneven or unrepresentative sampling of individuals of each sex from a set of different locations can, by chance, inflate allele frequency differences between the sexes beyond expectations for a single, panmictic population. If loci showing high between-sex \( F_{ST} \) also exhibit high between-population \( F_{ST} \), this could be indicative of population structure contributing to allele frequency divergence between the sexes in the empirical sample. Studies that sample individuals from a single population may also show artificially elevated between-sex \( F_{ST} \) if migration is sex biased (Box 1), which is common among animals (Trochet et al. 2016).

Sex-specific population structure can be accounted for by leveraging the statistical framework of case-control GWAS, in which associations between polymorphic variants and binary phenotypic states are quantified (e.g., presence or absence of a disease). The case-control GWAS approach treats sex (female or male) as the binary phenotypic state and scans for loci with the strongest associations, which should exhibit elevated absolute odds ratios (see Appendix C [Supporting Information]; Kasimatis et al. 2020; Pirastu et al. 2020). Although the underlying logic is identical to between-sex \( F_{ST} \), existing analytical methods for case-control GWAS can take population structure and relatedness in the empirical sample into account by including kinships (or the top principal components derived from kinships) as covariates (Astle and Balding 2009; Price et al. 2010). The case-control GWAS framework also permits estimation of SNP-based heritability of the phenotype (i.e., sex; Yang et al. 2011; Speed et al. 2017), which can be used to quantify a genome-wide signal of sex-specific selection.

Despite the advantages of leveraging an existing statistical framework, using case-control GWAS to test for associations between alternative alleles and sex does not sidestep all of the challenges faced by \( F_{ST} \) and \( F_{IS} \) statistics (see Appendices C and D [Supporting Information]). As with \( F_{ST} \) and \( F_{IS} \), large sample sizes remain essential for discriminating between sampling variance and true signal of sex differences in selection (especially when selection is weak), and the methods perform poorly when MAFs are low. Additionally, association tests using odds ratios depend heavily on a normal approximation, and there is a deep and still evolving literature regarding hypothesis testing using these methods that users should be aware of (e.g., Haldane 1956; Wang and Shan 2015).

Direct Approaches for Identifying SA Genes

In exceptional study systems, candidate SA polymorphisms can be identified through explicit statistical associations between genotypes and fitness. Such direct inference approaches present two major advantages over indirect methods. First, the inclusion of fitness measurements can potentially increase power to detect individual SA loci, relative to indirect methods (e.g., Fig. 1). Second, association tests can be conducted across many components of fitness (e.g., viability, fecundity, and mating success), facilitating identification of the life history stages and selective contexts affected by SA loci. We outline two general approaches for direct inference of sex-specific selection—GWAS and evolve-and-resquence (E&R) studies—which have been extensively employed to identify genes associated with human trait variation and/or local adaptation (Long et al. 2015; Visscher et al. 2017), yet rarely to identify SA loci.
GWAS OF SEX-SPECIFIC FITNESS

GWAS quantify statistical associations between phenotypic variation and polymorphic SNPs throughout the genome. Using GWAS to identify SA loci further requires that data on fitness components and genotypes are collected from individuals of each sex. A major advantage of GWAS is the availability of statistically rigorous methods to identify candidate loci, including methods to control for covariates in analyses (Price et al. 2010), and approaches that correct for multiple testing (e.g., family-wise or false discovery rate correction; Benjamini and Hochberg 1995) or that reduce the number of tests through gene-based association analysis (Nagamine et al. 2012; Riggio et al. 2013; Bérénos et al. 2015). We discuss the application of GWAS to three dataset types: (i) datasets in which genotypes and phenotypes are each measured independently in each sex (e.g., humans); (ii) systems amenable to experimental manipulation, in which each genotype can be replicated among female and male carriers (e.g., isogenic or hemiclone fruit fly lines); and (iii) pedigreed populations, in which the genealogical relationships between all individuals are known (e.g., some sedentary vertebrate populations).

Where genotypic and phenotypic measurements are performed among independently sampled individuals of each sex, as in humans, SA loci can be identified by first performing a separate GWAS in each sex (“sex-stratified” GWAS) (Fig. 3A), and then quantifying the difference between male- and female-specific effect sizes (see also Gilks et al. 2014). Illustrating this approach, Winkler et al. (2015) performed a sex-stratified GWAS on several human anthropometric traits, then defined a t-statistic as
\[ t = \frac{\hat{\beta}_M - \hat{\beta}_F}{\sqrt{SE_M^2 + SE_F^2 - 2\rho SE_M SE_F}}, \]
where \( \hat{\beta}_M \) and \( \hat{\beta}_F \) are the sex-specific effect sizes, \( SE_M \) and \( SE_F \) are the sex-specific standard errors, and \( \rho \) is the between-sex rank correlation among genome-wide loci. For each polymorphic site, \( P \)-values were generated by comparing the observed \( t \)-statistics to a null \( t \)-distribution with no sex-specific effects (where \( E[t] = E[\hat{\beta}_M - \hat{\beta}_F] = 0 \) under the null). This approach has been applied to nonfitness traits in humans (Randall et al. 2013; Myers et al. 2014; Winkler et al. 2015; Mitra et al. 2016; reviewed in Khramtsova et al. 2019), but has yet to be applied to fitness components (e.g., “number of children” phenotype in the UK Biobank; Sudlow et al. 2015).

In some experimental systems (e.g., fruit flies; flowering plants), the creation of isogenic or hemiclone lines (Abbott and Morrow 2011; Mackay et al. 2012; Berger et al. 2014) allows the same genotypes to be replicated and phenotypically assayed in carriers of each sex. Here, genotypes are effectively transplanted into male and female bodies or “environments,” analogous to the reciprocal transplantation of individuals sampled from different environments in local adaptation studies (Price et al. 2018). Identifying SA loci can then be achieved by transforming the bivariate coordinate system of male and female fitness values of a set of genotypes through matrix rotation (see Berger et al. 2014), which generates a univariate SA phenotype amenable to GWAS analysis (Fig. 3B). The approach is exemplified by a recent study in D. melanogaster (Ruzicka et al. 2019), which identified \( \sim 230 \) candidate SA polymorphic sites.

In pedigreed vertebrate populations, such as Soay sheep (Ovis aries) or Florida scrub jays (Aphelocoma coerulescens),
the genetic relationships between all individuals are known, and transmission of individual alleles across successive generations can be estimated (MacCluer et al. 1986). Because an individual’s genetic contribution to future generations is a genuine representation of its Darwinian fitness, alleles transmitted more frequently by one sex relative to the other represent candidate SA variants. Analyses of pedigreed populations of Florida scrub jays have been used to identify alleles with above-average transmission rates to descendants irrespective of sex (i.e., unconditionally beneficial alleles; Chen et al. 2019), yet this type of GWAS remains to be used to identify SA loci. It should be noted, however, that many pedigreed populations are necessarily small (given the logistics of monitoring them), which may hinder detection of loci affecting fitness variation.

Although GWAS-based identification of candidate SA loci shows promise, two major drawbacks must be kept in mind. First, measurements that capture total lifetime reproductive success are difficult to obtain, and caution is required in interpreting results based on single fitness component (e.g., reproductive but not viability selection), which may correlate imperfectly with total fitness. Second, effect sizes are typically small for polygenic traits (Visscher et al. 2017), including fitness. Powerful GWAS of sex-specific fitness may therefore be logistically prohibitive, and candidate SA loci will necessarily represent the subset of loci with particularly large fitness effects.

**E&R WITH SEX-LIMITED SELECTION**

Experimental elimination of selection in one sex but not the other (i.e., sex-limited selection) is a powerful way to identify SA selection in action. Various sex-limited selection designs have been implemented, including (i) restricting transmission of the genome to the male line, and thereby removing selection through females (Rice, 1996, 1998; Prasad et al. 2007; Bedhomme et al. 2008; Abbott et al. 2010), or vice versa (Rice 1992), using *Drosophila* hemiclones; (ii) eliminating fitness variance in one sex (e.g., by enforcing random contributions to offspring number, or removing opportunity for mate choice) but not the other (Rundle et al. 2006; Morrow et al. 2008; Maklakov et al. 2009; Hollis et al. 2014; Immonen et al. 2014; Chenoweth et al. 2015; Veltzos et al. 2017); and (iii) applying sex-limited artificial selection on a specific fitness component, such as mating success (Dugand et al. 2019), lifespan (Berg and Maklakov 2012; Chen and Maklakov 2014; Berger et al. 2016), reproductive tactic (Bielak et al. 2014), or mating investment (Pick et al. 2017).

To identify SA loci, sex-limited selection can be combined with genotyping at multiple time points during experimental evolution within each selection regime (E&R), thereby connecting population genetic changes to the phenotypic responses accrued during experimental evolution. By tracking allele frequencies in male-limited and female-limited selection lines, alleles that show a significant time-by-treatment interaction point to candidate SA loci, and their frequency dynamics can be characterized using current analytical tools for E&R experiments (Wiberg et al. 2017; Vlachos et al. 2019). E&R is a powerful and proven approach for identifying the genomic basis of phenotypic variation and local adaptation (Turner et al. 2011; Long et al. 2015; Barghi et al. 2019). Moreover, because it is experimental, the issues of sex-specific population structure that arise in between-sex $F_{ST}$ studies (see section “Indirect Approaches for Identifying SA Genes”) can be minimized. Yet despite these advantages, we are not aware of any published study that has used E&R to identify SA loci (see Chenoweth et al. 2015 for the closest effort to date).

Resequencing can be performed using population samples taken at multiple time points within a single generation (Svensson et al. 2018) or across multiple generations, with the latter approach benefitting from the fact that allele frequency responses to selection are cumulative over multiple generations. E&R, like GWAS, remains best suited for detecting loci with relatively large fitness effects. Selection on complex polygenic traits typically leads to small changes in allele frequencies at large numbers of loci, resulting in genomic signals of selection that are difficult to distinguish from genetic drift (Schlöterer et al. 2015). Consequently, the study organism, the number of replicates, the effective population sizes of selection and control lines, and the duration of experiments must be carefully considered in the design of E&R experiments (Baldwin-Brown et al. 2014; Kofler and Schlötterer 2014; Kessner and Novembre 2015).

**Validation and Follow-Up Analyses of Candidate SA Genes**

Candidate SA genes and SNP sets enriched for SA alleles (i.e., identified using methods outlined above) provide context for addressing long-standing questions about SA variation, including the genomic distribution, biological functions, and population genetic processes shaping SA polymorphisms. We focus on two specific issues in follow-up analyses of putative SA variants. First, we outline approaches for biologically validating SA candidates—a crucial task given that candidate gene sets may include appreciable proportions of false positives and artificial signals of sex-specific selection (see section “Indirect Approaches for Identifying SA Genes”). Second, we discuss population genetic analyses and issues of interpretation with bearing on the evolutionary histories of SA genes.

**BIOLOGICAL VALIDATION OF SA GENES**

Candidate SA loci can be directly validated in laboratory-amenable taxa by experimentally manipulating each allele and measuring its sex-specific fitness effect. A good example of experimental validation of naturally occurring SA genes is a
study by VanKuren and Long (2018), in which RNA interference and CRISPR-Cas9 were used to demonstrate SA effects of tandem duplicate genes *Apollo* and *Artemis* on offspring production in *D. melanogaster*. Similarly, Akagi and Charlesworth (2019) used manipulative molecular experiments to study candidate SA genes in several plant species. As a third example, several studies investigated a P450 transposable element insertion that up-regulates the *Cyp6g1* gene and increases DDT resistance in *D. melanogaster* (Smith et al. 2011; Rostant et al. 2015; Hawkes et al. 2016). Although evidence for SA effects at this particular locus is mixed, the experimental approaches used—including measurements of sex-specific fitness among isogenic lines and tracking the frequencies of each alternative allele in experimental cages—represent validation steps with potential for broad usage.

Direct experimental manipulation of candidate SA genes is not always feasible, and in instances where it is not, their biological validity can be assessed in other ways. In organisms harboring nonfunctional genomic material, it is possible to test whether candidate loci are enriched in genomic regions that are putatively functional (e.g., coding or regulatory) rather than inert (e.g., intergenic). Such "genic enrichment," which is expected for SA polymorphisms with genuine phenotypic effects, has previously been used to strengthen validity of candidate alleles for local adaptation (Barreiro et al. 2008; Coop et al. 2009; Key et al. 2016). Another way to increase confidence in candidate loci is to look for multiple signals of SA selection. For example, candidates identified through elevated *F* _ST_ that are also associated with SA fitness effects in a GWAS represent the best candidates for follow-up evolutionary analyses (see below). Finally, if independent data exist on the sex-specific phenotypic effects of individual mutations (e.g., in RNAi databases), these data can be mined to support the validity of candidate SA genes.

**EVOLUTIONARY DYNAMICS OF SA GENES**

We do not outline the range of population genetic analyses that could be used to describe the evolutionary dynamics of candidate SA loci, as these have been comprehensively reviewed elsewhere (e.g., Vitti et al. 2013; Fijarczyk and Babik 2015). Instead we provide guidance on some common issues that are likely to arise when analyzing patterns of genetic variation at SA loci and interpreting their mode of evolution.

First, we emphasize that the evolutionary dynamics of a contemporary SA gene may have, in the past, been governed by any combination of genetic drift, net directional selection (selection favoring fixation of one SA allele), or balancing selection (selection maintaining SA polymorphism). Theory often focuses on the conditions generating balancing selection at SA loci (e.g., Kidwell et al. 1977; Patten and Haig 2009; Fry 2010), leading some empirical studies to use signals of balancing selection (e.g., elevated Tajima’s *D*) as indirect proxies for SA selection (e.g., Du-toit et al. 2018; Wright et al., 2018, 2019; Sayadi et al. 2019). However, whether contemporary candidate SA alleles evolved under balancing selection hinges upon both the historical pattern of sex-specific selection and dominance at such loci (Kidwell et al. 1977; Connallon and Chenoweth 2019), which can be influenced by spatial and temporally varying selection (Connallon et al. 2019), and effective population size (Connallon and Clark, 2012, 2014a; Mullon et al. 2012). SA loci with large and symmetric selection coefficients or beneficial reversals of dominance (e.g., _h_f_ = 1 and _h_m_ = 0 in Box 1) are most conducive to balancing selection, whereas sufficient asymmetry between the sexes in the strength of selection (e.g., Mallet and Chippindale 2011; Mallet et al. 2011; Sharp and Agrawal 2013) should result in net directional selection that removes SA polymorphism rather than maintaining it (Kidwell et al. 1977; Kasimatis et al. 2019). Even when conditions for long-term balancing selection are met, the efficacy of balancing selection relative to drift may often be weak at SA loci, leading to genetic diversity patterns that are indistinguishable from neutrally evolving loci (Connallon and Clark, 2012, 2013; Mullon et al. 2012). In short, loci under contemporary SA selection can have a broad range of possible evolutionary histories. As such, the typical mode of evolution operating at candidate SA loci cannot be assumed a priori and should instead be viewed as a question that must be resolved empirically.

Second, the detection of elevated polymorphism at SA loci does not necessarily imply balancing selection. For example, SA candidate loci may exhibit significantly elevated MAFs relative to non-SA loci (Ruzicka et al. 2019), yet relaxed directional selection can account for this pattern if non-SA loci encompass a mix of neutral sites and sites evolving under sexually concordant directional selection. To establish that SA loci are evolving under balancing selection, it is necessary to show that SA genetic variation is significantly elevated compared to confirmed neutral sites (e.g., short introns, Parsch et al. 2010) and cannot be accounted for by demographic or mutational processes (Andrés et al. 2009; DeGiorgio et al. 2014; Bitarello et al. 2018). On the other hand, significant reductions in polymorphism at SA loci, relative to neutral sites, do not necessarily rule out balancing selection either. Counterintuitively, when the equilibrium frequency of the minor allele is low (i.e., equilibrium MAF < 0.2, approximately), balanced polymorphisms can be lost more rapidly than neutral polymorphisms, leading to reduced genetic variation relative to neutral expectations (Robertson 1962; Mullon et al. 2012).

A third and final point is that nonrandom patterns of genetic variation at SA loci can be generated by ascertainment bias alone. For example, data filtering steps that remove low-MAF SNPs (see “Indirect Approaches for Identifying SA Genes”) necessarily exclude rare SA variants from all downstream analyses. Elevated power to detect fitness effects among intermediate-frequency sites in a GWAS can also generate a spurious positive relationship.
between candidate SA sites and MAF that might be mistaken for nonneutral evolution. Similarly, SA loci could be nonrandomly distributed across the genome (e.g., enriched in regions with low or high recombination), thereby generating spurious patterns in population genomic data that appear to indicate nonneutral evolution. It is therefore important to correct for such biases where possible by, for example, incorporating external data on recombination rate variation (Comeron et al. 2012; Elyashiv et al. 2016) or assessing evidence of trans-species polymorphisms among SA loci before and after removal of CpG sites (Leffler et al. 2013).

Moving Forward

We have critically assessed a broad range of methods for detecting genomic signatures of SA selection, including indirect methods based on genome sequence analysis (section “Indirect Approaches for Identifying SA Genes”) and direct methods based on associations between genome sequences and fitness measurements (section “Direct Approaches for Identifying SA Genes”). An inescapable conclusion from our indirect inference models is that very strong sex differences in selection or very large sample sizes are required to detect individual SA candidate polymorphisms with high confidence (Fig. 1; Appendices A-D [Supporting Information]), in agreement with previous simulations of between-sex $F_{ST}$ (Lucotte et al. 2016; Connallon and Hall 2018; Kasimatis et al. 2019). Nevertheless, estimates of the full distribution of $F_{ST}$ from previously published flycatcher and pipefish datasets reveal an intriguing elevation of genome-wide $F_{ST}$ relative to our null models, which justifies future empirical studies of allele frequency differences between sexes (see below). Although an elevated signal of between-sex $F_{ST}$ is not present in our reanalysis of human data, the absence of a signal is perhaps unsurprising given the small number of loci analyzed, following the removal of noncoding sequences and loci with rare polymorphisms. Genome-wide analyses of sex-specific selection in humans using larger datasets (Kasimatis et al. 2020), and examinations of sex-biased expression among sites with elevated $F_{ST}$ (Cheng and Kirkpatrick 2016)—which were the focus of previous work, but not of our reanalysis—are therefore encouraged. With regard to direct methods for identifying SA genes, the substantial logistical challenge of accurately measuring fitness must be circumvented, yet the approach is powerful when feasible (see Ružička et al. 2019) and certain to be a key component of future work on the genetics of sex-specific fitness variation.

Although there is little doubt that identifying and characterizing SA genes is challenging, there are several reasons for optimism. First, the low power of indirect metrics to detect selection at an individual locus level does not rule out the detection of a cumulative signal of polygenic sex differences in selection (e.g., Fig. 2). Although such an approach implies that candidate SA genes (e.g., those in the highest $F_{ST}$ quantiles) will include many false positives, elevated false discovery rates are not necessarily problematic if we are interested in the general properties of SA candidates relative to samples of putatively neutral (or non-SA) loci. Nevertheless, in studies with low-to-moderate sample sizes, where many candidate genes will be false positives, researchers should minimally demonstrate that (i) the empirical distribution of the metric of interest differs significantly from its appropriate null (see Kasimatis et al. 2019; our reanalyses in section “Indirect Approaches for Identifying SA Genes”), (ii) putative signals of selection are not driven by sex-specific population structure or other artefacts (see section “Indirect Approaches for Identifying SA Genes”), and (iii) candidate loci are situated in putatively functional genome regions (see section “Validation and Follow-up Analyses of Candidate SA Genes”).

Second, the power to detect SA genes using indirect metrics can often be increased in relatively simple ways. For example, pooled sequencing is a cost-effective method for estimating allele frequencies from samples of many individuals (Schlötterer et al. 2015), and well-suited for genome-wide $F_{ST}$ scans (although not for $F_{IS}$ scans, as estimating $F_{IS}$ requires individual-level genotype data). Researchers could, alternatively, focus attention toward large publicly available genomic datasets that are adequately powered for detecting loci under moderately strong SA selection (see Fig. 1), or toward genomic regions predicted to have relatively high statistical power. For example, studies of pseudoautosomal regions of recombining sex chromosomes have substantially higher power to detect $F_{ST}$ outliers driven by sex differences in selection (Qiu et al. 2013; Kirkpatrick and Guerrero 2014). Targeted sampling strategies may also amplify power to identify SA genes. For example, estimating allele frequencies among breeding adults—which have passed filters of viability selection and components of adult reproductive success—increases the number of episodes of selection that can contribute to allele frequency differentiation between sexes, improving the potential for detecting elevated between-sex $F_{ST}$.

Third, well-chosen study systems can improve prospects for accurately measuring lifetime reproductive success and identifying SA loci through direct methods (GWAS or E&R). For example, difficulties in accurately measuring fitness under field conditions can be mitigated in pedigreed populations, where the genetic contribution of each individual to successive generations is known (provided the population is well monitored), and each genotype can therefore be associated with an accurate estimate of total lifetime reproductive success in each sex. Emerging approaches to infer pedigrees from genomic data alone (Snyder-Mackler et al. 2016) may further facilitate identification of SA loci in the absence of long-term monitoring efforts. In some experimental systems, such as laboratory-adapted hemiclones of...
**Table 1. Sex-specific relative fitness of genotypes of a biallelic locus.**

<table>
<thead>
<tr>
<th>Genotype</th>
<th>AA</th>
<th>Aa</th>
<th>aa</th>
</tr>
</thead>
<tbody>
<tr>
<td>Female relative fitness</td>
<td>$1 + s_f$</td>
<td>$1 + s_fh_f$</td>
<td>1</td>
</tr>
<tr>
<td>Male relative fitness</td>
<td>$1 + s_m$</td>
<td>$1 + s_m h_m$</td>
<td>1</td>
</tr>
</tbody>
</table>

*D. melanogaster* (Rice et al. 2005; Abbott and Morrow 2011), relatively accurate measurements of outbred lifetime reproductive success are also possible. E&R is feasible for experimental organisms with short generation times and where large laboratory populations can be maintained (e.g., *Drosophila*, seed beetle *Callosobruchus maculatus*). Here, there is a relatively untapped opportunity to identify SA loci by combining sex-limited selection (e.g., Rice 1992; Prasad et al. 2007; Morrow et al. 2008; Abbott et al. 2010; Bonel et al. 2018) with genotyping across multiple generations (e.g., Turner et al. 2011; Long et al. 2015; Barghi et al. 2019; Abbott et al. 2020).

Finally, despite notable exceptions (e.g., the dioecious plant *Silene latifolia*; Delph et al. 2011; Muyle et al. 2012), plant systems remain underused in research on SA selection. One advantage of plants is their greater amenability to field measurements of fitness components, as widely used in studies of local adaptation and species’ range limits (Hargreaves et al. 2014). Another advantage is the great diversity of reproductive systems in flowering plant species, the vast majority of which are hermaphroditic and susceptible to SA selection (Jordan and Connallon 2014; Tazzyman and Abbott 2015; Olito 2017; Olito et al. 2018), potentially leading to allele frequency differences between the pollen and ovules contributing to fertilization under haploid selection, and to elevated $F_{IS}$ among offspring. A third advantage of plants is their greater tendency to express genetic variation during the haploid stage of their life cycle (e.g., Immler and Otto 2018). Haploid (relative to diploid) expression is expected to inflate the contribution of genetic polymorphism to fitness variance and magnify evolutionary responses to selection, including within-generation allele frequency divergence between sexes (Connallon and Jordan 2016). Exploiting plant systems may thereby increase statistical power to identify candidate SA genes or genomic signals of SA variation using direct (GWAS and E&R) or indirect inference approaches.

**Box 1. Processes generating sexually divergent allele frequencies**

Several evolutionary scenarios can lead to sex differences in the frequencies with which individual alleles are transmitted to offspring (Hedrick 2007; Úbeda et al. 2011; Connallon et al. 2018). We focus on two processes—sex differences in selection and sex-biased migration—that may each commonly arise and affect estimates of allele frequency differences between sexes.

**Sex differences in selection**. Consider a single biallelic locus in which the focal allele (allele $A$) has a frequency of $p$ at birth within a given generation; the alternative $a$ allele has a frequency of $1 - p$. Selection during the life cycle alters the allele frequencies in the set of adults that contribute offspring to the next generation. The frequency of the $A$ allele in breeding females and males (respectively) is follows:

$$p_f = p + s_f p (1 - p) (h_f + p (1 - 2h_f)) + O(s_f^2, s_m^2, s_fs_m)$$

and

$$p_m = p + s_m p (1 - p) (h_m + p (1 - 2h_m)) + O(s_f^2, s_m^2, s_fs_m),$$

where $s_f$ and $s_m$ are female and male selection coefficients for the $A$ allele, $h_f$ and $h_m$ are the dominance coefficients (Table 1), and $O(s_f^2, s_m^2, s_fs_m)$ refers to second-order terms in the selection coefficients, which are negligible (and can be ignored) when $s_f$ and $s_m$ are small, as expected for most loci (Charlesworth and Charlesworth 2010, p. 97).

We expect allele frequency differences between breeding adults of each sex ($p_f \neq p_m$) when the fitness effects of each allele differ between sexes, that is, (1) the same allele is favored in each sex but the strength of selection differs between sexes (e.g., $p_f > p_m$ when $s_f > s_m > 0$); (2) alleles have sex-limited fitness effects (e.g., $p_f > p_m$ when $s_f > 0 = s_m$), or (3) alleles are sexually antagonistic (e.g., $p_f > p_m$ when $s_f > 0 > s_m$). Allele frequencies are expected to remain equal between sexes ($p_f = p_m$) when genetic variation is neutral ($s_f = s_m = 0$), or selection and dominance coefficients are identical between sexes ($s_f = s_m$ and $h_f = h_m$).

**Sex-biased migration**. Consider an island population receiving new migrants each generation, with migration occurring before reproduction during the life cycle. At birth, the frequencies of $A$ and $a$ alleles in the island population are $p$ and $1 - p$, respectively. Let $m_f$ and $m_m$ represent the proportions of breeding females and breeding males on the island that are migrants. The expected frequency of the $A$ allele in breeding females and males (respectively) will be

$$p_f = (1 - m_f) p + m_f \tilde{p}$$

and

$$p_m = (1 - m_m) p + m_m \tilde{p},$$

where $\tilde{p}$ is the probability that any offspring is a migrant (either female or male).
where $\hat{p}$ is the frequency of the $A$ allele in migrant individuals. The identity, $p_f - p_m = (m_f - m_m)(\hat{p} - p)$, implies that allele frequency differences between breeding females and males ($p_f \neq p_m$) would require sex-biased migration ($m_f \neq m_m$) and allele frequency differences between migrant and resident (nonmigrant) individuals ($\hat{p} \neq p$).

### 1. Box 2. Fixation indices ($F_{ST}$ and $F_{IS}$) applied to sex differences

Allele frequency differences between breeding adults of each sex can be quantified by way of fixation indices, originally devised by Wright (1951) for characterizing genetic differentiation among populations. We again consider a biallelic autosomal locus with the focal allele ($A$) at a frequency of $p_f$ in breeding females and $p_m$ in breeding males.

**Between-sex $F_{ST}$**: $F_{ST}$ is a standardized measure of the allele frequency difference between the sexes:

$$F_{ST} = \frac{(p_f - p_m)^2}{4\hat{p}(1 - \hat{p})},$$

where $\hat{p} = (p_f + p_m)/2$ (Cheng and Kirkpatrick 2016). This definition applies to the entire population, and therefore differs from empirical estimates of $F_{ST}$ that are based upon samples of gene sequences from the population. Although there are several estimators of $F_{ST}$ (see Appendix A [Supporting Information]; Bhatia et al. 2013; Gammerdinger et al. 2020), we focus on the simplest:

$$\hat{F}_{ST} = \frac{(\hat{p}_f - \hat{p}_m)^2}{4\hat{\bar{p}}(1 - \hat{\bar{p}})}$$

(Nei 1973), where $\hat{p}_f$ and $\hat{p}_m$ are the allele frequency estimates from samples of females and males, and $\hat{\bar{p}} = (\hat{p}_f + \hat{p}_m)/2$. As we show in Appendix A (Supporting Information), the ratio

$$4n_f n_m \hat{p} (1 - \hat{p}) \hat{F}_{ST} = \frac{4n_f n_m (1 - p_f) + n_f p_m (1 - p_m)}{n_m p_f (1 - p_f) + n_f p_m (1 - p_m)}$$

has an approximately noncentral chi-squared distribution with one degree of freedom and noncentrality parameter of $\lambda = (p_f - p_m)^2(p_1 + p_m) + \frac{n_1 - n_m}{n_1} - 1$, where $n_f$ and $n_m$ are the numbers of sequences derived from females and males, respectively. The approximation can break down when $n_f$ and $n_m$ are small or the minor (rarer) allele at the locus has a frequency close to zero. Under the statistical null distribution, the true allele frequencies do not differ between the sexes ($p_f = p_m$), and therefore $\hat{F}_{ST} \approx X^2_{1}/2n_{1f}$, where $X^2_{1}$ is a chi-squared random variable with one degree of freedom, and $n_{1f} = 2(1/n_f + 1/n_m)^{-1}$ is the harmonic mean sample size.

**$F_{IS}$ in offspring**: $F_{IS}$ can be used to quantify deviations between the observed heterozygosity in a cohort of individuals before selection (e.g., individuals sampled and sequenced at birth) and the expected heterozygosity under Hardy-Weinberg equilibrium. With random mating among breeding adults with female and male allele frequencies $p_f$ and $p_m$, and ignoring effects of genetic drift or segregation distortion, the frequency of the $A$ allele in offspring of the next generation will be $\hat{p} = (p_f + p_m)/2$, and the proportion that is heterozygous will be $P_{aa} = p_f(1 - p_m) + p_m(1 - p_f)$. Under these conditions, $F_{IS}$ will be

$$F_{IS} = \frac{P_{aa}}{2\hat{p}(1 - \hat{p})} - 1 = \frac{(p_f - p_m)^2}{4\hat{\bar{p}}(1 - \hat{\bar{p}})},$$

where the final expression is equivalent to $F_{ST}$ between breeding females and males of the prior generation (Kasimatis et al. 2019). The above expression for $F_{IS}$ applies to the entire set of offspring in a population, whereas empirical estimates of $F_{IS}$ will be based on the genotypes of offspring sampled from the population. As shown in Appendix B (Supporting Information), estimates of $F_{IS}$ will be approximately normally distributed with a mean and variance of

$$E[\hat{F}_{IS}] = \frac{1}{2n} + \frac{(p_f - p_m)^2}{4\hat{\bar{p}}(1 - \hat{\bar{p}})},$$

and

$$\text{var}[\hat{F}_{IS}] = \frac{1}{n}.$$
AUTHOR CONTRIBUTIONS
All authors contributed to the conceptualization of the manuscript. TC, CO, and PC developed mathematical models. LD conducted data reanalyses, with input from TC and FR. TC and FR coordinated manuscript writing, with initial drafts of "Direct approaches for identifying SA genes" by AR, XYL, ES, and FR, initial drafts of "Validation and follow-up analyses of candidate SA genes" by HPY and CYJ, and remaining sections drafted by TC and FR. All authors contributed to manuscript editing.

DATA ARCHIVING
All code for reproducing reanalyses presented in the manuscript is available at https://github.com/ldutoit/male_female_fst. Scripts and other data can be found on dryad: https://doi.org/10.5061/dryad.b2rbznsscc

LITERATURE CITED


THE SEARCH FOR SEXUALLY ANTAGONISTIC GENES


Associate Editor: J. Mank

Supporting Information

Additional supporting information may be found online in the Supporting Information section at the end of the article.

Figure S1. Permuted versus observed $F_{ST}$ for simulated data.

Figure S2. Permuted versus observed $F_{ST}$ for the flycatcher dataset.

Figure S3. Permuted versus observed $F_{ST}$ for the pipefish dataset.

Figure S4. Permuted vs. observed $F_{ST}$ for the human 1000 Genomes data.

Figure S5. Human 1000 Genomes data with SNPs with low minor allele frequencies (MAFs) included in the analysis.

Appendix A. Distribution of $F_{ST}$ estimates.

Table S1. Sex-specific relative fitness of genotypes of a biallelic locus with additive fitness effects in each sex.

Appendix B. Sex-specific allele frequencies and $F_{ST}$ estimates.

Appendix C. Case-control GWAS and the Log-Odds Ratio.

Figure S6. Probability that $L_{/SE}$ for an additive SA locus.

Figure S7. Theoretical signal of multilocus SA polymorphism: the ratio of observed versus permuted absolute log odds ratio estimates within 100 quantiles of the theoretical null of $L_{/SE}$.

Appendix D. Comparison of null model distributions for different metrics allele frequency differentiation between sexes.
Figure S1. Permuted vs. observed $\hat{F}_{ST}$ for simulated data. Details follow those described in Fig. 2 of the main text.

Simulation (mix of neutral & SA loci)

$n_f = 10^5$, $n_m = 10^5$, $N_{SNPs} = 101,000$
Figure S2. Permuted vs. observed $\hat{F}_{ST}$ for the flycatcher dataset. Details of the analysis follow those described in Fig. 2 of the main text.
Figure S3. Permuted vs. observed $\hat{F}_{ST}$ for the pipefish dataset. Details of the analysis follow those described in Fig. 2 of the main text. Note that there are two null distributions in the right-hand panel. One is the idealized theoretical null distribution (the solid grey curve) which would apply if there was no variation in sample sizes across SNPs (i.e., it requires that $n_H$ is roughly constant across polymorphic sites in the genome). The idealized null model is inappropriate for the pipefish dataset due to missing data leading to sample size variation among sites. The simulated null accounts for missing data by simulating $\hat{F}_{ST}$ per site from a null distribution with site-specific $n_H$. 

Pipefish

$n_f = 114$, $n_m = 334$, $N_{SNPs} = 44,773$
Figure S4. Permuted vs. observed $\hat{F}_{ST}$ for the human 1000 Genomes dataset. Details of the analysis follow those described in Fig. 2 of the main text.
Figure S5: Human 1000 Genomes data with SNPs with low minor allele frequencies (MAFs) included in the analysis. All else is the same as Fig. S4.
Appendix A: Distribution of $F_{ST}$ estimates

Following Cheng and Kirkpatrick (2016), between-sex $F_{ST}$ for a bi-allelic locus is defined as:

$$F_{ST} = \frac{(p_f - p_m)^2}{4\bar{p}(1 - \bar{p})}$$

where $p_f$ and $p_m$ are the female and male population frequencies of the focal allele at the locus, and $\bar{p}$ is the sex-averaged allele frequency, $\bar{p} = (p_f + p_m)/2$. The above expression applies to the entire population. Define $F_{ST}$ estimated from allele frequencies in a sample of gene sequences as:

$$\hat{F}_{ST} = \frac{(\hat{p}_f - \hat{p}_m)^2}{4\bar{p}(1 - \bar{p})}$$

(see Nei 1973), where $\hat{p}_f$ and $\hat{p}_m$ are the female and male frequencies calculated from the sample of sequences, and $\hat{p} = (\hat{p}_f + \hat{p}_m)/2$. The subsequent analysis is based on this simple estimator of between-sex $F_{ST}$. We note that several other $F_{ST}$ estimators are widely used in the empirical literature (Bhatia et al 2013; Gammerdinger et al. 2020), and return to these alternative estimators further below.

Suppose that, at birth, the focal allele at a given locus has an overall population frequency of $p$ in both sexes (the other allele at the locus has a frequency of $1 - p$). Although genetic drift can lead to different allele frequencies between the sexes at birth, we assume that the population size is sufficiently large that such differences are negligible. As long as the sample size of sequenced individuals is large and the strength of selection at the locus is not extreme, then the allele frequencies in the panel of sequenced adults will be similar to the allele frequencies at birth in the population. We can therefore approximate $F_{ST}$ from the sample of sequences using a Taylor Series expansion around $\hat{p}_f = \hat{p}_m = p$:

$$\hat{F}_{ST} = (\hat{p}_f - p) \frac{\partial F_{ST}(\hat{p}_i = p)}{\partial \hat{p}_f} + (\hat{p}_m - p) \frac{\partial F_{ST}(\hat{p}_i = p)}{\partial \hat{p}_m}$$

$$+ \frac{1}{2} \left[ (\hat{p}_f - p)^2 \frac{\partial^2 F_{ST}(\hat{p}_i = p)}{\partial \hat{p}_f^2} + (\hat{p}_m - p)^2 \frac{\partial^2 F_{ST}(\hat{p}_i = p)}{\partial \hat{p}_m^2} \right] + O((\hat{p}_i - p)^3) \approx \frac{(\hat{p}_f - \hat{p}_m)^2}{4p(1 - p)}$$

with the final approximation dropping third-order and higher terms of $\hat{p}_i - p$.

Given the overall allele frequencies of $p_f$ and $p_m$ in adult females and males of the population, then $n_f \hat{p}_f$ and $n_m \hat{p}_m$ from samples of gene sequences will be independent, binomially distributed random variables: $\hat{p}_i \sim B(n_i, p_i)/n_i$, where $n_i$ is the number of sequences derived from members of the $i$th sex. With sufficiently large $n_i$, the distribution of $\hat{p}_i$ will be approximately normal, $\hat{p}_i \sim N(n_i p_i, n_i p_i(1 - p_i))/n_i$, leading to:
Using the normal approximation, \( \hat{F}_{ST} \) simplifies to:

\[
\hat{F}_{ST} \approx \frac{(p_f(1-p_f) + p_m(1-p_m))}{4p(1-p)} \left( \frac{\hat{p}_f - \hat{p}_m}{\frac{p_f(1-p_f)}{n_f} + \frac{p_m(1-p_m)}{n_m}} \right)^2
\]

where \( \hat{p}_f \) and \( \hat{p}_m \) are the sample proportions from each sex, \( p_f \) and \( p_m \) are the expected proportions, and \( n_f \) and \( n_m \) are the sample sizes for each sex.

For a locus with no sex differences in selection (\( p \approx p_f \approx p_m \)), \( \hat{F}_{ST} \) reduces to:

\[
\hat{F}_{ST} \approx \frac{n_m + n_f}{4n_f n_m} X_0 = \frac{1}{2n_H} X_0
\]

where \( X_0 \) is a chi-squared random variable with one degree of freedom, and \( n_H = 2/(1/n_f + 1/n_m) \) is the harmonic mean of the sex-specific sample sizes. The mean and variance for \( \hat{F}_{ST} \) become:

\[
E[\hat{F}_{ST}] \approx \frac{n_m + n_f}{4n_f n_m} = \frac{1}{2n_H}
\]

and
\[
\text{var}[\hat{p}_{ST}] \approx \frac{1}{8} \left( \frac{n_m + n_f}{n_f n_m} \right)^2 = \frac{1}{2n_H^2}
\]

Under the null model, the variance of \(\hat{p}_{ST}\) values divided by the square of the mean is:

\[
\frac{\text{var}[\hat{p}_{ST}]}{E[\hat{p}_{ST}]^2} \approx 2
\]

which is identical to the ratio for a recently subdivided pair of populations (Lewontin and Krakauer 1973; Nei and Chakravarti 1977).

For sufficiently large \(n_H\), the median \(\hat{p}_{ST}\) for neutral loci is approximately:

\[
\text{med}[\hat{p}_{ST}] \approx \frac{1}{2n_H} \left( \frac{7}{9} \right)^3 \approx 0.47 \frac{2}{2n_H}
\]

Since the distribution of \(\hat{p}_{ST}\) for neutral loci is a simple linear function of chi-squared random variables, we can define the cumulative distribution of neutral \(\hat{p}_{ST}\) using known properties of the chi-squared distribution. The cumulative distribution function for the chi-squared distribution with one degree of freedom is:

\[
F(\hat{p}_{ST}) = \text{erf}\left( \sqrt{n_H \hat{p}_{ST}} \right)
\]

where \(\text{erf}(x)\) refers to the error function.

**Effects of selection on allele frequency divergence between sexes**

As in Box 1, supposing that selection is strong relative to genetic drift, then the allele frequency in the \(i\)th sex, following selection, will be:

\[
p_i = p + s_i p(1 - p)(h_i + p(1 - 2h_i)) + O(s_i^2)
\]

where \(s_i\) and \(h_i\) are the selection and dominance coefficients for the focal allele in the \(i\)th sex.

Under weak selection (\(|s| \ll 1\)), the first-order approximation (i.e., neglecting \(O(s_i^2)\)) provides a good approximation for the allele frequencies, per sex, following selection.

The first-order approximation can, however, prove inaccurate in scenarios involving strong selection (i.e., \(|s|, |s_m| > 0.1\)), including cases of SA selection with which we are concerned. For the case of arbitrarily strong SA selection, assume that the fitness effects of SA alleles are additive in each sex (i.e., \(h_f = h_m = \frac{1}{2}\)), and the population is at polymorphic equilibrium \((p_f + p_m = 2p)\). For this scenario, we can calculate exact allele frequencies per sex, following selection.

Assume that the \(A\) allele is female-beneficial and the \(a\) allele is male-beneficial (i.e., \(s_f > 0 > s_m\) in Box 1 of the main text). Following Kidwell et al. (1977), we first re-parameterize the general fitness model by substituting \(s_m = -|s_m|\), and \(t_f = s_f/(1 + s_f)\) so that relative fitnesses in each sex are scaled against the best genotype in that sex (Table S1). With no dominance (\(h_f =\)
$h_m = \frac{1}{2}$, we obtain a symmetric SA parameterization, where $|s_m|$ and $t_f$ represent the fitness costs to females and males (respectively) of carrying the “wrong” allele (Table S1). Both selection parameters are constrained to be between zero and one ($0 < |s_m|$, $t_f < 1$), and $t_f = |s_m|$ represents the special case where the strength of selection is equal between the sexes.

**Table S1.** Sex-specific relative fitness of genotypes of a bi-allelic locus with additive fitness effects in each sex (no dominance; $h_i = \frac{1}{2}$)

<table>
<thead>
<tr>
<th>Genotype</th>
<th>Original parameterization (Box 1)</th>
<th>Re-parameterization for a SA locus ($s_f &gt; 0 &gt; s_m$)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>female relative fitness</td>
<td>female relative fitness</td>
</tr>
<tr>
<td></td>
<td>male relative fitness</td>
<td>male relative fitness</td>
</tr>
<tr>
<td>AA</td>
<td>$1 + s_f$</td>
<td>$1$</td>
</tr>
<tr>
<td>Aa</td>
<td>$1 + s_f/2$</td>
<td>$1 - t_f/2$</td>
</tr>
<tr>
<td>aa</td>
<td>$1$</td>
<td>$1 -</td>
</tr>
<tr>
<td></td>
<td>$1 + s_m$</td>
<td>$1 -</td>
</tr>
</tbody>
</table>

Given the reparameterization in Table S1, previous theory (see Connallon and Jordan 2016) has shown that the equilibrium $A$ allele frequency of a polymorphic SA locus will be:

$$p = \frac{t_f - |s_m| + t_f |s_m|}{2 t_f |s_m|}$$

and the equilibrium frequency difference between breeding females and males is:

$$p_f - p_m = \frac{2(1 - p)|s_m|}{|s_m|} \left(1 + \frac{t_f |s_m| p (1 - p)}{(1 - p |s_m|)(1 - (1 - p)t_f)} - 1\right)$$

With an intermediate equilibrium frequency ($p = \frac{1}{2}$, which requires that $t_f = |s_m|$), the expected $\tilde{F}_{ST}$ for the locus is:

$$E[\tilde{F}_{ST}] = \frac{1}{2n_H} + \frac{(p_f - p_m)^2}{4p(1 - p)} \left(1 - \frac{1}{2n_H}\right)$$

Given the sampling distribution of $\tilde{F}_{ST}$ (Box 2; see above), when should we expect $\tilde{F}_{ST}$ for a SA SNP to reliably reside within the tail of the null distribution for $\tilde{F}_{ST}$? Consider the following approximations for when such outcomes are likely. Assuming that each SA locus has additive fitness effects (i.e., $h_f = h_m = \frac{1}{2}$ and $s_f > 0 > s_m$ in Box 1), and each has evolved to an intermediate equilibrium frequency ($p = \frac{1}{2}$ and $|s_m| = s_f/(1 + s_f)$), where $|s_m|$ represents the fitness cost to males of inheriting $AA$ and $s_f/(1 + s_f)$ is the cost to females of inheriting $aa$; see Connallon and Jordan 2016), $\tilde{F}_{ST}$ for a selected locus will have a relatively high probability of residing within the tail of the null when its expected value, $E[\tilde{F}_{ST} | SA]$, is greater than the percentile that is used to define the tail of the null. Following Box 2, the $k^{th}$ percentile of the null distribution for $\tilde{F}_{ST}$ is:
\[ \hat{F}_{ST(k)} = \frac{[\text{erf}^{-1}\left(\frac{k}{100}\right)]^2}{n_H} \]

where \( \text{erf}^{-1}(x) \) refers to the inverse error function. Given the stated assumptions, the expected \( \hat{F}_{ST} \) for an SA locus will be greater than the \( k^{th} \) percentile of the null (i.e., \( \mathbb{E}[\hat{F}_{ST|SA}] > \hat{F}_{ST(k)} \)), when \( |s_m| \) exceeds the following threshold:

\[
s_{\text{min}} = \frac{4 \sqrt{2 \left[ \text{erf}^{-1}\left(\frac{k}{100}\right) \right]^2 - 1}}{1 - 2 \left[ \text{erf}^{-1}\left(\frac{k}{100}\right) \right]^2 - 1 + 2 \sqrt{2 \left[ \text{erf}^{-1}\left(\frac{k}{100}\right) \right]^2 - 1}}
\]

(see Fig. 1A), where \( s_{\text{min}} \) represents the minimum strength of selection (i.e., the minimum cost of inheriting the “wrong” SA allele, per sex) that satisfies \( \mathbb{E}[\hat{F}_{ST|SA}] > \hat{F}_{ST(k)} \). Simulations over a range of sample sizes show that \( \hat{F}_{ST} \) for a SA locus at equilibrium with \( |s_m| = s_f/(1 + s_f) = s_{\text{min}} \) has a ~40-50% probability of exceeding the 99th percentile of the null. For the exact probability that \( \hat{F}_{ST} \) for a SA SNP exceeds the 95th percentile of the \( F_{ST} \) null distribution, see Fig. 1B of the main manuscript and Appendix D, below. Even stronger selection is required to satisfy the condition \( \mathbb{E}[\hat{F}_{ST|SA}] > \hat{F}_{ST(k)} \) when one SA allele at the locus is rarer than the other (\( p \neq \frac{1}{2} \); see Appendix D) and/or when dominance is reversed between sexes (i.e., \( h_f > \frac{1}{2} > h_m \); see Connallon and Jordan 2016).

**Other \( F_{ST} \) estimators**

The \( F_{ST} \) estimator presented above and in Box 2 is a biased estimator with an average that is greater than zero under the null model in which there are no allele frequency differences between the sexes in the population as whole (i.e., \( \hat{F}_{ST} \), as defined above has an expectation greater than zero when \( p_f = p_m \) in the population). Since alternative estimators are often used in empirical studies of population or sex differentiation, it is worth exploring the distribution of \( \hat{F}_{ST} \) based on alternative estimators.

The most widely used estimator, Weir-Cockerham’s \( F_{ST} \) (Weir and Cockerham 1984; Bhatia et al 2013), is defined as:

\[
\hat{F}_{ST} = 1 - \frac{n_H}{n_f + n_m - 2} \left[ n_f \hat{p}_f (1 - \hat{p}_f) + n_m \hat{p}_m (1 - \hat{p}_m) \right] - \frac{n_H}{n_f + n_m - 2} \left[ n_f \hat{p}_f (1 - \hat{p}_f) + n_m \hat{p}_m (1 - \hat{p}_m) \right]
\]

where \( n_H = 2(1/n_f + 1/n_m)^{-1} \). Assuming large populations and (at most) moderate selection, then the estimator can be approximated using a Taylor series expansion around \( \hat{p}_f = \hat{p}_m = p \), where \( p \) is the allele frequency at birth:
\[ F_{ST} = \frac{1}{1-n_H} + \frac{1}{2} \left[ \left( \hat{p}_f - p \right)^2 \frac{\partial^2 F_{ST}(\hat{p}_i = p)}{\partial \hat{p}_f^2} + \left( \hat{p}_m - p \right)^2 \frac{\partial^2 F_{ST}(\hat{p}_i = p)}{\partial \hat{p}_m^2} \right] + O((\hat{p}_i - p)^3) \]

\[ \approx \frac{1}{1-n_H} + \left( \frac{n_H}{n_H-1} \right)^2 \left( 1 - \frac{2}{n_f + n_m} \right) \left( \hat{p}_f - p \right)^2 + \left( \hat{p}_m - p \right)^2 - 2(\hat{p}_f - p)(\hat{p}_m - p) \]

\[ = \frac{1}{1-n_H} + \left( \frac{n_H}{n_H-1} \right)^2 \left( 1 - \frac{2}{n_f + n_m} \right) \left( \hat{p}_f - \hat{p}_m \right)^2 \]

Under the null model, we have (approximately):

\[ \hat{p}_f - \hat{p}_m \sim N \left( 0, \frac{2p(1-p)}{n_H} \right) \]

and

\[ \hat{F}_{ST} \approx \frac{1}{n_H} (X_0 - 1) \]

where \( X_0 \) is a chi-squared random variable with one degree of freedom (as above). For large \( n_H \), we can approximate the \( F_{ST} \) estimator as:

\[ \hat{F}_{ST} \approx \frac{1}{n_H} (X_0 - 1) \]

which gives a mean and variance of:

\[ \mathbb{E}[\hat{F}_{ST}] \approx 0 \]

\[ \text{var}[\hat{F}_{ST}] \approx \frac{2}{n_H^2} \]

The mean is now centred at zero, which shows that the Weir and Cockerham estimator is asymptotically unbiased. Nevertheless, the sampling variance remains inversely proportional to the square of the sample size and will therefore show a similar magnitude of sampling variance as our focal estimator: \( \hat{F}_{ST} = \left( \hat{p}_f - \hat{p}_m \right)^2 / 4\hat{p}(1-\hat{p}) \).

Another commonly used estimator is that of Hudson et al. (1992), which for our case is expressed as:

\[ F_{ST} = \frac{\left( \hat{p}_f - \hat{p}_m \right)^2 - \hat{p}_f(1-\hat{p}_f) - \hat{p}_m(1-\hat{p}_m)}{\hat{p}_f(1-\hat{p}_m) + \hat{p}_m(1-\hat{p}_f)} \]
(see Bhatia et al. 2013; Gammerdinger et al. 2020). The estimators of Hudson et al. and Weir and Cockerham are identical when the sample sizes of female- and male-derived sequences are equal \((n_f = n_m)\), and in practice is approximately the same as Weir and Cockerham’s estimator under the null model \((p_f = p_m)\) with very large sample sizes of female and male sequences.

**Genic \(F_{ST}\) estimates**

Booker et al. (2020) recently emphasized that the distribution of \(F_{ST}\) estimated from concatenated sequences will strongly depend on the degree of linkage between polymorphic sites in the sequence. Specifically, the chi-square distribution of \(\hat{F}_{ST}\) under the null remains applicable when there is no recombination between sites in the concatenated sequence. With recombination, the expected \(\hat{F}_{ST}\) remains the same, whereas the variance declines. This effect of concatenation applies to the context of \(F_{ST}\) between sexes, as we illustrate below. As such, we advocate using non-concatenated estimates to evaluate fits between observed and theoretical distributions for \(\hat{F}_{ST}\).

Using our original \(F_{ST}\) estimator, between-sex \(F_{ST}\) for a single neutral locus is:

\[
\hat{F}_{ST} = \left( \hat{p}_f - \hat{p}_m \right)^2 / 4 \hat{p}(1 - \hat{p}) = X_0 / 2n_H
\]

where, as before, \(X_0\) is a chi-square random variable with one degree of freedom and \(n_H\) is the harmonic mean sample size of female and male gene sequences.

Now define a concatenated genic \(F_{ST}\) as the average across \(k\) polymorphic sites in a given gene. Assuming the \(k\) sites are neutral, we have:

\[
\hat{F}_{ST} = \frac{1}{k} \sum_{i=1}^{k} \hat{F}_{ST,i} = \frac{1}{k} \sum_{i=1}^{k} \frac{X_i}{2n_{H,i}}
\]

where subscripts refer to the \(i\)th polymorphic site in the gene.

If the \(k\) sites are independently segregating and sequence sample sizes are identical across sites \((n_{H,i} = n_H)\), then genic \(\hat{F}_{ST}\) simplifies to:

\[
\hat{F}_{ST} = \frac{1}{2kn_H} \sum_{i=1}^{k} X_i \sim \chi_k^2 / 2kn_H
\]

where \(\sum_{i=1}^{k} X_i \sim \chi_k^2\) is a chi-square random variable with \(k\) degrees of freedom. The mean and variance for the estimated genic \(\hat{F}_{ST}\) are:

\[
E[\hat{F}_{ST}] = \frac{1}{2n_H}
\]

and
We see that under the null model, the expected $\hat{p}_{ST}$ for a SNP and for a genic region will be the same, whereas the former has a $k$-fold greater variance than the latter. Outliers are, therefore, more likely to arise in genes with lower numbers of polymorphisms or nonindependence (non-zero LD) between polymorphic sites.

$$\text{var}[\hat{p}_{ST}] = \frac{1}{2kn_{H}^{2}}$$
Appendix B: Sex-specific allele frequencies and $F_{IS}$ estimates

We assume an outbred population with random mating between the breeding adults of each sex. Let $p_f$ and $p_m$ represent the allele frequencies in breeding females and males, as in Appendix A. Following Kasimatis et al. (2019), we define $F_{IS}$ as the degree of enrichment in heterozygosity among offspring of the population, which is a consequence of the allele frequency differences between breeding adults of the prior generation. Defining $p$ as the frequency of the focal allele in offspring of the breeding adults (so that $2p = p_f + p_m$), we have:

$$F_{IS} = \frac{p_f(1 - p_m) + p_m(1 - p_f)}{2p(1 - p)} - 1 = \frac{p_f(1 - p_m) + p_m(1 - p_f) - 2p(1 - p)}{2p(1 - p)}$$

$$= \frac{2(p - p_fp_m - p(1 - p))}{2p(1 - p)} = \frac{(p_f - p_m)^2}{4p(1 - p)}$$

Note that the final expression is identical to $F_{ST}$ between sexes for the population of breeding females and males of the prior generation (see above). The deviation from Hardy-Weinberg Equilibrium (HWE) can be quantified using a disequilibrium coefficient, $D_A$ (see Weir 1996, p. 94), in which case $F_{IS}$ becomes:

$$F_{IS} = \frac{P_{AA}}{2p(1 - p)} - 1 = \frac{2p(1 - p) - 2D_A}{2p(1 - p)} - 1 = -\frac{D_A}{p(1 - p)}$$

where $P_{AA}$ is the frequency of heterozygotes in the population, and $D_A = (p(1 - p) - P_{AA}/2)$ quantifies the HWE deviation. Note that $D_A > 0$ implies a deficit of heterozygotes relative to HWE (as expected under inbreeding or population subdivision), whereas $D_A < 0$ implies an excess of heterozygotes (as expected when there are allele frequency differences between the sexes).

The null distribution of $F_{IS}$

With large genotype samples (large $n$, where $n$ refers to the number of diploid offspring that are sequenced), empirical estimates of $D_A$, denoted $\hat{D}_A$, will be approximately normal (Weir 1996, p. 95), with expectation and variance:

$$E[\hat{D}_A] = D_A - \frac{p(1 - p) + D_A}{2n}$$

$$\text{var}[\hat{D}_A] = \frac{p^2(1 - p)^2 + (1 - 2p)^2D_A - D_A^2}{n}$$

Consequently, $(\hat{D}_A - E[\hat{D}_A]) / \sqrt{\text{var}[\hat{D}_A]}$ will follow a standard normal distribution. If the population really is at HWE (as under the null model), and the sample size is sufficiently large so that the variance in the test statistic can be approximated using $\text{var}[\hat{D}_A] = \frac{p^2(1 - p)^2}{n}$, then:

$$\frac{\hat{D}_A}{\sqrt{\text{var}[\hat{D}_A]}} = \frac{\hat{D}_A\sqrt{n}}{\hat{p}(1 - \hat{p})} \sim N\left(-\frac{1}{2\sqrt{n}}, 1\right)$$
From this, we see that the estimate of \( F_{IS} \) (denoted \( \hat{F}_{IS} \)) is approximately normal with mean \( E[\hat{F}_{IS}] = 1/2n \) and variance \( \text{var}[\hat{F}_{IS}] = 1/n \), i.e.:

\[
\hat{F}_{IS} = -\frac{\hat{D}_A}{\hat{p}(1 - \hat{p})} = -\frac{1}{\sqrt{n}} \left( \hat{D}_A \sqrt{n} \right) \sim N \left( \frac{1}{2n}, \frac{1}{n} \right)
\]

Simulations of the mean and variance for \( \hat{F}_{IS} \) under the null model for \( 10^5 \) polymorphic loci with minor allele frequencies (MAF) greater than 0.05, are plotted below (lines show the analytical predictions, which match up well against the simulated data; note that the y-axis is in log scale).

**\( F_{IS} \) under sex-specific selection**

With sex differences in selection, the frequency of heterozygotes at birth will be:

\[
p_f(1 - p_m) + p_m(1 - p_f) = 2p(1 - p) + \frac{1}{2} (p_f - p_m)^2
\]

where \( p = (p_f + p_m)/2 \). The HWE deviation in the population becomes:

\[
D_A = -\frac{1}{4} (p_f - p_m)^2
\]

Assuming that the sample size is large and the HWE deviation is small (\( |D_A| << 1 \)), so that terms of \( O(D_A^2, D_A n^{-1}) \) can be neglected, the distribution of \( \hat{F}_{IS} \) converges, approximately, to:

\[
\hat{F}_{IS} = -\frac{\hat{D}_A}{\hat{p}(1 - \hat{p})} \sim N \left( \frac{1}{2n} - \frac{D_A}{p(1 - p)^{1/2} n} \right)
\]

Substituting \( D_A = -\frac{1}{4} (p_f - p_m)^2 \), we obtain:
\[ f_{IS} = -\frac{\tilde{D}_A}{\tilde{p}(1 - \tilde{p})} \sim N\left( \frac{1}{2n} + \frac{(p_f - p_m)^2}{4p(1 - p)}, \frac{1}{n} \right) \]

The \( k \)th percentile of the null distribution \((p_f = p_m)\) is:

\[ \hat{f}_{IS(k)} = \frac{1}{2n} + \frac{2}{\sqrt{n}} \text{erf}^{-1} \left( \frac{2k}{100} - 1 \right) \]

Using the exact expressions for a polymorphic locus at equilibrium under additive SA selection \((h_i = \frac{1}{2}; \text{see Appendix A})\), the minimum fitness cost that will cause \( E[\hat{f}_{IS}] \) for a SA SNP to reach the \( k \)th quantile for the null model \((i.e., \text{the value of } |s_m|, \text{assuming } p = \frac{1}{2}, \text{that leads to } E[f_{IS}] = \hat{f}_{IS(k)} \) is:

\[ s_{\text{min}} = \frac{4 \sqrt{\frac{2}{n}} \text{erf}^{-1} \left( \frac{2k}{100} - 1 \right)}{1 - \sqrt{\frac{2}{n}} \text{erf}^{-1} \left( \frac{2k}{100} - 1 \right) + 2 \sqrt{\frac{2}{n}} \text{erf}^{-1} \left( \frac{2k}{100} - 1 \right)} \]

(see Fig. 1A).
Appendix C: Case-control GWAS and the Log-Odds Ratio

As described in the main text and highlighted in a recent study of SA selection in humans (Kasimatis et al. 2020 bioRxiv), case-control GWAS methods offer a promising alternative to fixation indices ($F_{ST}$ and $F_{IS}$) for estimating and testing for statistical significance of between-sex allele frequency differences. The advantage of case-control GWAS methods relative to genome-wide scans of between-sex $F_{ST}$ and $F_{IS}$ is that they leverage a well-developed linear modelling framework to explicitly account for confounding factors such as population connectivity when testing for allele frequency differences between sexes. Broadly speaking, case-control GWAS rely on a logistic regression of allele frequency on sex (coded as a categorical variable) and other possible explanatory variables, and hypothesis testing is based on the log-odds ratio. Below, we briefly outline the use of the log-odds ratio to test for between-sex differences in allele frequency, illustrate the limits of statistical power when using this approach, and highlight comparisons with the $\hat{F}_{ST}$ statistic, which is presented in the main text.

We emphasize that our brief summary is not meant to be a comprehensive statistical review of odds-ratios, nor a review of case-control GWAS methods and their use to control for population structure and other confounding variables (see e.g., Pirastu et al. 2020; Astle and Balding 2009; Price et al. 2010). We also note that there is a deep and still-evolving statistical literature on hypothesis testing in case-control methods that readers should be aware of (e.g., Haldane 1956, Wang 2015).

The log-odds ratio ($\mathcal{L}$)

Given genome sequence data for a random sample of adult males and females, one can construct a $2 \times 2$ contingency table of allele counts in males and females for a given biallelic locus as:

<table>
<thead>
<tr>
<th></th>
<th>$A$</th>
<th>$a$</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Males</td>
<td>$n_{m,A}$</td>
<td>$n_{m,a}$</td>
<td>$n_m$</td>
</tr>
<tr>
<td>Females</td>
<td>$n_{f,A}$</td>
<td>$n_{f,a}$</td>
<td>$n_f$</td>
</tr>
<tr>
<td>Total</td>
<td>$n_A$</td>
<td>$n_a$</td>
<td></td>
</tr>
</tbody>
</table>

where $n_{ij}$ is the number of copies of the $j$th allele type in the sample of genes sequenced from the $i$th sex; as in our $F_{ST}$ model $n_f$ and $n_m$ are the total sample sizes of genes sequenced in females and males, respectively. The allele frequency estimates from the samples are:

<table>
<thead>
<tr>
<th></th>
<th>$A$</th>
<th>$a$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Males</td>
<td>$\hat{p}_m$</td>
<td>$1 - \hat{p}_m$</td>
</tr>
<tr>
<td>Females</td>
<td>$\hat{p}_f$</td>
<td>$1 - \hat{p}_f$</td>
</tr>
</tbody>
</table>

where $\hat{p}_i = n_{i,A}/n_i$ with $n_i = n_{i,A} + n_{i,a}$. The sample log-odds ratio can be expressed as:

$$\mathcal{L} = \log\left(\frac{\hat{p}_m(1 - \hat{p}_f)}{\hat{p}_f(1 - \hat{p}_m)}\right)$$
For large-sample sizes, $\hat{\mathcal{L}}$ is approximately normal with mean of $E[\hat{\mathcal{L}}] = \log \left( \frac{p_m(1-p_f)}{p_f(1-p_m)} \right)$, where $p_f$ and $p_m$ refer to the true allele frequencies in the population, and a variance of $\text{var}[\hat{\mathcal{L}}] = \frac{1}{n_{m,A}} + \frac{1}{n_{m,a}} + \frac{1}{n_{f,A}} + \frac{1}{n_{f,a}}$ (Woolf 1955; Section 14.6 in Bishop et al. 1975). We note that hypothesis testing relies heavily on this normal approximation, which breaks down for small sample sizes, and for loci with very low minor allele frequency (MAF) (see e.g., Wang & Shan 2015).

**The null distribution of $\hat{\mathcal{L}}$**

With large genotype samples (large $n_f$ and $n_m$), and assuming that there is no sex difference in selection (and thus no allele frequency difference between sexes in the population), empirical estimates of $\hat{\mathcal{L}}$ can be approximated by the Taylor series expansion:

$$\hat{\mathcal{L}} = \log \left( \frac{\hat{p}_m(1 - \hat{p}_f)}{(1 - \hat{p}_m)\hat{p}_f} \right) = -\frac{\left( \hat{p}_f - \hat{p}_m \right)}{\hat{p}(1 - \hat{p})} + O \left( \left( \hat{p}_f - \hat{p}_m \right)^2 \right)$$

Ignoring terms of $O \left( \left( \hat{p}_f - \hat{p}_m \right)^2 \right)$, which are negligible with large sample sizes, $\hat{\mathcal{L}}$ is approximately normal with expectation and variance (respectively) of:

$$E[\hat{\mathcal{L}}] = 0$$

$$\text{var}[\hat{\mathcal{L}}] = \frac{1}{n_{m,A}} + \frac{1}{n_{m,a}} + \frac{1}{n_{f,A}} + \frac{1}{n_{f,a}} \approx \frac{2}{n_H \hat{p}(1 - \hat{p})}$$

The scaled log-odds ratio, $\hat{\mathcal{L}}/\sqrt{\text{var}[\hat{\mathcal{L}}]}$, follows a standard normal distribution, which provides a useful null distribution in the absence of sex-specific selection. The approximation breaks down when any of the allele counts are small, and reliable hypothesis testing therefore requires that the MAF is not too low, or that the sample size is large enough to ensure reasonably high counts of each allele type.

**$\mathcal{L}$ under sex-specific selection**

From Box 1 in the main text, the population frequency of the $A$ allele within the $i$th sex, following sex-specific selection, will be:

$$p_i = p + s_i p (1 - p) \left( h_i + p (1 - 2h_i) \right) + O(s_i^2)$$

where $s_i$ and $h_i$ are the selection and dominance coefficients for the focal allele in the $i$th sex. Analytic expressions for $\mathcal{L}$ under sex-specific selection quickly become unwieldy. We can, however, briefly illustrate the expected behavior of the distribution under sex-specific selection in the idealized scenario of intermediate initial allele frequency (i.e., $p = \frac{1}{2}$) and weak and additive sex-specific selection ($h_f = h_m = \frac{1}{2}$; second-order terms $O(s_i^2)$ ignored). Under these simplifying assumptions, the distribution of $\hat{\mathcal{L}}$ is normal with mean and standard deviation:

$$E[\hat{\mathcal{L}}] = \log \left( \frac{(4 - s_f)(4 + s_m)}{(4 + s_f)(4 - s_m)} \right)$$
Simulations of the mean, standard error, and scaled log odds ratio for $10^4$ loci conforming to the above assumptions show that the approximation works well for an illustrative scenario of SA selection ($s_m = 0.03, s_f = -0.03$) with equal sample sizes for both sexes (where $i$ indexes sex and $n_i = n_f$), provided sample sizes are sufficiently large. Note that results for the null in the second panel are overlain by those for SA selection.

The expected distribution of the scaled sample log odds ratio will also be asymptotically normal, with

$$\frac{\hat{L}}{SE} \sim N\left(\frac{\log \left(\frac{(4-s_f)(4+s_m)}{(4+s_f)(4-s_m)}\right)}{8 \frac{1}{n_f(16-s_f^2)} + \frac{1}{n_m(16-s_m^2)}}, 1\right)$$

As noted above, the asymptotic properties of the sampling distribution of $\hat{L}/SE$ break down when either MAFs are low, sample sizes are small, or both.

**Statistical power to detect individual outlier loci**

The probability that a SA locus resides within the tail of the null distribution depends on the allele frequencies at the locus, the strength of selection, and the sample size of individuals that are sequenced. The expected difference in allele frequency between the sexes, after selection, is maximized for a locus with intermediate allele frequencies and symmetric SA selection ($p = 1/2$, described above), representing a best-case scenario for detecting individual outlier loci. The robustness of the normal approximation for the scaled log odds ratio greatly simplifies calculation of statistical power to identify outlier loci under sex-specific selection. For our idealized scenario where $p_f = p_m = 1/2$ it is simply the area under $N\left(\frac{\hat{L}}{SE}, 1\right)$ that falls above/below the $\alpha/2$ and $(1-\alpha)/2$ quantiles of a standard normal distribution, $N(0,1)$. Below, we illustrate the probability that $\hat{L}/SE$ reliably resides in the tail of the null distribution under the same parameter conditions shown for $\hat{F}_{ST}$ in Fig. 1B of the main text. Overall, the scaled log odds ratio provides similar statistical power to detect individual outlier loci as $\hat{F}_{ST}$. This final conclusion applies across the spectrum of MAF (see Appendix D, below).
**Figure S6** Probability that $\hat{L}/SE$ for an additive SA locus (i.e., $h_f = h_m = \frac{1}{2}$) with intermediate equilibrium allele frequencies ($p, q = 1/2$) is in the top 2.5% tail of the null distribution $N(0,1)$. Sample sizes are assumed to be equal for the $i$th sex, and $|s_i|$ is the fitness cost of being homozygous for the “wrong” SA allele, where $i \in \{f, m\}$. Blue lines show the analytic approximation based on $N \left( E \left( \frac{\hat{L}}{SE} \right), 1 \right)$, points show the proportion of $10^4$ simulated loci which fall in the upper and lower 2.5% tails of the null distribution. Red and black horizontal lines benchmark the 2.5% and 97.5% power thresholds.

**Multi-locus signals of sex-specific selection using $|\hat{L}/SE|$**

Following identical logic as for $\hat{P}_{ST}$ in the main text, the full empirical distribution of the absolute value for $\hat{L}/SE$ for autosomal SNPs may carry a cumulative signature of SA selection in the form of an inflated total number of observations in the upper quantiles of the observed $\hat{L}/SE$ distribution relative to the number of observations predicted under the null. Below we illustrate the signal of enrichment of observations in the upper theoretical quantiles caused by SA selection using the ratio of observed versus permuted $\hat{L}/SE$ for SNPs within 100 quantiles of the theoretical null of $\hat{L}/SE$ using identical simulation methods as those presented in Fig. 2 in the main text (i.e., the top panel of Figure 2). The scaled log odds ratio appears to provide a similar signal of SA selection to $\hat{P}_{ST}$, but is slightly less sensitive to shifts in the observed distribution due to SA selection (compare to Fig. 2 in the main text)
Figure S7. Theoretical signal of multi-locus SA polymorphism: the ratio of observed versus permuted absolute log odds ratio estimates within 100 quantiles of the theoretical null of $\hat{L}/SE$. **Theoretical data:** $\hat{L}/SE$ values were simulated for $10^5$ neutrally evolving loci with MAF above 5% in the dataset, and for $10^3$ loci responding to SA selection prior to sampling of gene sequences (each SA-responding locus had a SA fitness effect or was in perfect linkage disequilibrium with a SA locus). SA-responding loci had intermediate allele frequencies in the population ($p = \frac{1}{2}$). SA selection coefficients were drawn from an exponential distribution with mean of $s_{avg} = 0.03$, and additive fitness effects ($h = \frac{1}{2}$). The top 1% theoretical quantile is enriched by ~30%, implying that ~2/10 of SNPs above the 99% threshold of the null are true positives.
Appendix D: Comparison of null model distributions for different metrics
allele frequency differentiation between sexes

Three metrics have been used in empirical studies of allele frequency differentiation between the sexes: (1) between-sex $\hat{p}_{ST}$ (e.g., Cheng and Kirkpatrick 2016; see Appendix A), (2) absolute allele frequency divergence between sexes, $|\hat{d}_{fm}|$ (e.g., Berner 2019; Bissegger et al. 2019), and (3) the log odds ratio, $\hat{L}$ (e.g., Kasimatis et al. 2020 bioRxiv; see Appendix C). While each of these metrics has a distinct sampling distribution, and their null distributions differ in sensitivity to the minor allele frequency per SNP (MAF), the metrics share commonality in that all three are functions of estimated allele frequency differences between the sexes. Below, we briefly outline and compare the properties of their null distributions, and evaluate and compare their statistical power to detect SA SNPs with different minor allele frequencies.

**Between-sex $\hat{p}_{ST}$**
For large sample sizes, $\hat{p}_{ST}$ under the null (i.e., no sex differences in selection) follows a chi-square distribution with one degree of freedom (see Appendix A). The expectation and the variance for $\hat{p}_{ST}$ are approximately:

$$E[\hat{p}_{ST}] = \frac{1}{2n_H}$$

and

$$\text{var}[\hat{p}_{ST}] = \frac{1}{2n_H^2}$$

and the $k^{th}$ quantile in the distribution of $\hat{p}_{ST}$ is:

$$\hat{p}_{ST(k)} = \frac{[\text{erf}^{-1}\left(\frac{k}{100}\right)]^2}{n_H}$$

Note that the expectation, the variance, and quantiles of $\hat{p}_{ST}$ are all independent of minor allele frequencies in the population. Thus, the distribution of $\hat{p}_{ST}$ under the null model is constant across the spectrum of MAFs after SNPs with very rare alleles have been excluded from the analysis (i.e., the approximations break down for polymorphic sites with MAF near zero; see Appendix A). Genome-wide predictions under the null model should, therefore, be relatively insensitive to details of the site frequency spectrum of the study population.

**Absolute difference between female and male allele frequencies**
The difference between female and male allele frequency estimates for a locus is:

$$\hat{d}_{fm} = \hat{p}_f - \hat{p}_m$$
which is well-approximated by a normal distribution provided the sample size of the study is large and the MAF at the locus is not too close to zero (see Appendix A). The expectation and variance for $\hat{d}_{fm}$ under the null hypothesis, conditioned on the SNP’s MAF in the population ($p$), will be:

$$E[\hat{d}_{fm}|p] = 0$$

$$\text{var}[\hat{d}_{fm}|p] = p(1-p)\left(\frac{1}{n_f} + \frac{1}{n_m}\right) = \frac{2p(1-p)}{n_H}$$

(see Appendix A). The absolute allele frequency difference under the null will therefore follow a half-normal distribution with mean and variance of:

$$E\left[|\hat{d}_{fm}|\right] = 2\sqrt{\frac{p(1-p)}{\pi n_H}}$$

$$\text{var}\left[|\hat{d}_{fm}|\right] = \frac{2p(1-p)}{n_H}\left(1 - \frac{2}{\pi}\right)$$

The $k^{th}$ quantile of the null distribution is:

$$|\hat{d}_{fm}|(k) = 2\sqrt{\frac{p(1-p)}{n_H}}\text{erf}^{-1}\left(\frac{k}{100}\right)$$

Note that each of these predictions is sensitive to the MAF of the locus and, consequently, the genome-wide null distribution for $\hat{d}_{fm}$ or for $|\hat{d}_{fm}|$ will be contingent on the focal population’s site frequency spectrum, which will be unique to each study population.

**Log Odds Ratio ($\hat{L}$)**

With a large sample size, and very low MAF sites excluded from the analysis (as before), the log odds ratio under the null model is well approximated by:

$$\hat{L} = \log\left(\frac{\hat{p}_m(1-\hat{p}_f)}{1-\hat{p}_m\hat{p}_f}\right) = -\frac{(\hat{p}_f - \hat{p}_m)}{\hat{p}(1-\hat{p})}$$

(see Appendix C). Under the null model, $\hat{L}$ will be approximately normal with mean and variance of:

$$E[\hat{L}|p] = 0$$

$$\text{var}[\hat{L}|p] = \frac{2}{n_H p(1-p)}$$
The distribution of the absolute value of $\mathcal{L}$ ($|\mathcal{L}|$) under the null will follow a half-normal distribution with mean and variance:

$$E[|\mathcal{L}|] = \frac{2}{\sqrt{\pi n_H p(1-p)}}$$

$$\text{var}[|\mathcal{L}|] = \frac{2}{n_H p(1-p)} \left(1 - \frac{2}{\pi}\right)$$

and the $k^{th}$ quantile of the null distribution will be:

$$|\hat{\mathcal{L}}|_k = \frac{2 \text{erf}^{-1}\left(\frac{k}{100}\right)}{\sqrt{n_H p(1-p)}}$$

Each of these predictions is sensitive to the MAF of the locus and, consequently, the genome-wide null distribution for $\mathcal{L}$ or for $|\mathcal{L}|$ will be contingent on the population’s site frequency spectrum.

Evaluation of the metrics

The following figure shows simulated data for 50,000 non-singleton SNPs with minor allele frequencies uniformly distributed between ~0 to 0.5. Allele frequencies in the population were equal between the sexes, and 500 individuals of each sex were sampled and sequenced, resulting in sample sizes of $n_f = n_m = n_H = 1000$ gene sequences, per diploid locus, per sex. Each locus is polymorphic in both female and male samples, which excludes singleton polymorphisms from the analysis. Each of the three metrics of allele frequency differentiation was calculated from the simulated allele frequency estimates, with each grey dot showing a given metric’s value for a single SNP. We compared the distribution of each metric relative to its appropriate null model. For $F_{ST}$ estimates, we calculated the fraction of simulated SNPs where $\hat{F}_{ST}$ was greater than the 99% quantile defined by the null model approximations presented above. For the other two metrics—in which the null distribution either increases ($\hat{d}_{fm}$) or decreases ($|\hat{\mathcal{L}}|$) with MAF—we evaluated the proportion of SNPs that exceeded the 99% quantile defined by the SNP’s MAF.

Across many simulation runs, we found that the proportion of SNPs above the appropriate 99% theoretical quantile was always identical between $\hat{F}_{ST}$ and $|\hat{d}_{fm}|$, which is unsurprising given that the theoretical distributions for both metrics are simple functions of the absolute difference between $\hat{p}_f$ and $\hat{p}_m$ (for the simulations shown in the figure below, 0.912% of SNPs for each metric are above the 99% threshold defined by the theoretical null model; individual simulation runs always yielded values near 1%, as expected). Roughly 1% of simulated $|\hat{\mathcal{L}}|$ values exceeded the 99% theoretical quantile for $|\hat{\mathcal{L}}|$ (0.988% for the simulation shown in the figure). Slight quantitative differences between the proportion of outliers in $|\hat{\mathcal{L}}|$ versus $\hat{F}_{ST}$ and $|\hat{d}_{fm}|$ reflects our use of an additional approximation in deriving the distribution of $|\hat{\mathcal{L}}|$ (simulated data relied on no such approximation).
Statistical power to detect SA loci

The power to detect SA loci is roughly the same between the three metrics of allele frequency divergence between sexes, as shown in the following simulations. Simulations show the probability that $F_{ST}$, $|d_{fm}|$, or $|E|$ for an additive SA locus (i.e., $h_f = h_m = \frac{1}{2}$) is in the top 5% tail of the null distribution for the metric. The average cost to the sexes of inheriting the wrong SA allele is given by $\bar{s} = (t_f + |s_m|) / 2$, where $|s_m|$ is the fitness cost to males of
being homozygous for the female-beneficial allele, and $t_f$ is the cost to females of being homozygous for the male-beneficial allele (see Appendix A). Simulation results show the proportion of $10^5$ simulated SA SNPs (per combination of $n_H$, $p$, and $\bar{s}$) that exhibits a greater between-sex divergence than the 95th percentile of the null distribution for the appropriate metric (95th percentile thresholds are based on the equations presented directly above). For each simulated SNP, the population is assumed to be at equilibrium for an additive SA locus with $p$ representing the sex-averaged frequency of the female-beneficial allele at equilibrium; female and male allele frequencies, after selection, are $p_f = p + (p_f - p_m)/2$ and $p_m = p - (p_f - p_m)/2$, where the exact expression for $p_f - p_m$, at equilibrium, is presented in Appendix A.

Supplementary References


