# Sweepstakes reproductive success via pervasive and recurrent selective sweeps

Einar Árnason,[1,2*] Jere Koskela,[3] Katrín Halldórsdóttir,[1] and Bjarki Eldon[4]

[1]Institute of Life- and environmental Sciences, University of Iceland, Sturlugata 7, IS102, Reykjavík, Iceland,
[2]Department of Organismal and Evolutionary Biology, Harvard University, Cambridge, MA 02138, USA
[3]Department of Statistics, University of Warwick, Coventry, CV4 7AL, UK
[4]Leibniz Institute for Evolution and Biodiversity Science, Museum für Naturkunde, 10115 Berlin, Germany

*To whom correspondence should be addressed; E-mail: einararn@hi.is.

**Abstract**

Highly fecund natural populations characterized by high early mortality abound, yet our knowledge of such population's recruitment dynamics is rudimentary at best. This knowledge gap has implications for our understanding of genetic variation, population connectivity, local adaptation, and resilience of highly fecund populations. The concept of sweepstakes reproductive success, which posits huge variance in individual reproductive output, is key to understanding recruitment dynamics, the distribution of individual reproductive and recruitment success. However, it is unknown whether highly fecund organisms reproduce by sweepstakes and if they do, the relative roles of neutral and selective sweepstakes. Here we use coalescent-based statistical analysis of genomic population data and show that selective sweepstakes are a strong candidate for explaining recruitment dynamics in the highly fecund Atlantic cod. The sweepstakes result from recurrent and pervasive selective sweeps of new variation generated by mutation. We show that the Kingman coalescent and the Xi-Beta coalescent (modelling random sweepstakes), including complex demography and background selection, are inadequate explanations. Our results show that sweepstakes reproduction processes and multiple-merger coalescent models are relevant and necessary for understanding genetic diversity in highly fecund natural populations. Our findings have fundamental implications for understanding the recruitment variation of fish stocks and general evolutionary genomics of high fecundity.

## Introduction

Individual recruitment success is a fundamental demographic object in ecology and evolution. The distribution of individual recruitment success affects the distribution and abundance of organisms (the subject of ecology) and the genotypic and phenotypic changes resulting from the major forces of evolution. Individual recruitment success determines individual fitness, the currency of natural selection. Many marine organisms are highly fecund, producing huge numbers of juvenile offspring that experience high mortality (type III survivorship) going through numerous developmental stages, fertilization, zygote, larvae, fry, etc. until finally recruiting as adults of the next generation. The concept of sweepstakes reproductive success (Hedgecock, 1994), suggested to have 'a major role in shaping marine biodiversity' (Hedgecock and Pudovkin, 2011, p. 971), is a key to understanding the mechanism behind individual recruitment success. Sweepstakes reproduction has few winners and many losers leading to a very high variance and skew in individual reproductive output. High fecundity by itself does not lead to sweepstakes absent a mechanism for generating high-variance and high-skew offspring distribution. Two main ecological mechanisms turn high fecundity into sweepstakes: a random and a selective mechanism. The first is the chance matching of reproduction to a jackpot of temporally favourable conditions, a case of random sweepstakes (Hedgecock and Pudovkin, 2011). The match/mismatch hypothesis (Cushing, 1969) explains the dynamics of recruitment variation and year-class strength by the timing of reproduction with favourable but erratic environmental conditions, such as weather and climatic conditions. As an example, climatic variability leads to random temporal shifts in planktonic blooms that are food for developing fish larvae, a match means reproductive success, mismatch a reproductive failure (Cushing, 1969). By chance a random individual hits a jackpot of favorable environmental conditions that results in a very large reproductive output of reproducing offspring (Schweinsberg, 2003; Eldon and Wakeley, 2006).

The second mechanism is selective sweepstakes where the genetic constitution of survivors is different from that of non-survivors (Williams, 1975). Under the second scenario, an organism's different developmental stages pass through numerous independently acting selective filters with the cumulative effect of a high-variance high-skew offspring distribution. Here, the winning genotypes are Sisyphean (Williams, 1975) (after Sisyphus from Greek mythology, punished with forever pushing a boulder up a hill). They are ephemeral and must be continuously reassembled. By analogy, the population climbs a selective peak by positive selection, but the environment changes continuously because the sequence of selective filters changes. Only a new or a recombined genotype can climb the selective peak the next time around (Williams, 1975). The population is forever tracking an elusive optimum by climbing an adaptive peak. The selective filters can arise from abiotic factors, and biotic density- and frequency-dependent effects arising from inter and intraspecific competition, and from predation and predator avoidance (Reznick, 2016).

A third alternative is that random survival hits every family, the offspring of every pair, to the same degree. In this case, there is no mechanism turning high fecundity into sweepstakes reproduction. High fecundity by itself does not lead to sweepstakes absent a mechanism turning high fecundity into a high-variance high-skew offspring distribution. Juvenile mortality might even be compensatory and reduce variance of offspring number via density-dependent competition or predation. In this scenario reproduction does not match favourable conditions by chance, no individual hits a jackpot, nor does selective filtering happen. The resulting offspring distribution has a much smaller variance than in the sweepstakes models, the same low and unchanged coefficient of variation in the distribution of zygotes and the distribution of adult offspring (Nunney, 1996). Such reproduction would result in a similar distribution of reproducing offspring as in the assumed mode of reproduction of low fecundity and model organisms (Wright, 1931; Fisher, 1930). A low variance in individual recruitment success modeled through the Wright-Fisher model (or similar models), is nearly universally assumed in population genetics (Wakeley, 2007).

Genomics and coalescent theory offer powerful tools to test our three hypotheses: first of non-sweepstakes versus sweepstakes reproduction and secondly to test the two sweepstakes hypotheses, the random and the selective one. Conducting similar tests would be a daunting task with ecological methods, requiring one to follow the fate of the offspring of different individuals (Grant and Grant, 2014). The first question regards identifying non-sweepstakes versus sweepstakes reproduction in our population genomic data. Our second question regards testing the two sweepstakes hypotheses, the random vs the selective sweepstakes, given evidence of sweepstakes reproduction in the data.

Here we conduct an extensive, simulation-based analysis of site frequency spectra (SM 1.3) and linkage disequilibrium under various coalescent and individual-based models. The resulting predicted patterns of summary statistics allow us to infer the likely mechanisms of individual reproductive and recruitment success in Atlantic cod.

The classical Kingman coalescent (Kingman, 1982; Wakeley, 2007), in essence, models reproduction of low fecundity organisms. Multiple merger coalescents (Donnelly and Kurtz, 1999; Pitman, 1999; Sagitov, 1999; Schweinsberg, 2003, 2000) describe the genealogies for the two kinds of sweepstakes reproduction, the random and the selective sweepstakes. The Xi-Beta coalescent (Schweinsberg, 2000; Birkner et al., 2018) modelling the genealogy of a population with large reproductive events in which a random individual has enormous reproductive success well approximates the random or jackpot sweepstakes hypothesis (Hedgecock and Pudovkin, 2011). The Durrett-Schweinsberg model of recurrent selective sweeps (Durrett and Schweinsberg, 2005), implying a forever changing environment that continuously favors new mutations, well approximates selective sweepstakes (Williams, 1975). The multiple-merger Durrett-Schweinsberg coalescent (SM 1.3) describes the genealogy at a single site, the "neutral" site, that is linked at some recombinational distance to a site hit by a favorable mutation. The population experiences recurrent strongly beneficial mutations at sites linked to a neutral site, and it is assumed that a neutral site never experiences mutation. A beneficial mutation sweeps to fixation in a time measured in $\log N$ time units, where $2N$ is the population size, and the probability of a sweep does not depend on the population size. However, a vital component of the Durrett-Schweinsberg model is an assumption of a high rate of recombination between the neutral and the mutated site, giving ancestral lineages at the neutral site a chance to escape a sweep through recombination.

Several recent studies show evidence of reproductive skew. Many marine organisms have star-like gene genealogies of mitochondrial DNA (e.g. Atlantic cod and Japanese sardines (Árnason, 2004; Niwa et al., 2016)) with an excess (relative to predictions of the Kingman coalescent) of the singleton class of the site frequency spectrum (mutations on the external branches of the genealogy). The nuclear *Ckma* gene of Atlantic cod (Árnason and Halldórsdóttir, 2015) shows such an excess of singletons. An excess of singletons can also result from demographic changes such as post-Pleistocene population expansion. However, the overall effect of population growth and low variance in individual recruitment success on the site-frequency spectrum is different from sweepstakes reproduction. This distinguishes between the models. For example, the site frequency spectrum of the *Ckma* gene also shows an excess of mutations in the right tail of the site frequency spectrum as predicted by multiple-merger coalescent models of sweepstakes (Birkner et al., 2013a; Eldon et al., 2015; Blath et al., 2016) and not by the Kingman coalescent under arbitrary population size history (Sargsyan and Wakeley, 2008). Multiple-merger coalescents occur in models of rapidly adapting populations (Neher and Hallatschek, 2013; Schweinsberg, 2017), under both directional selection (Neher, 2013; Sackman et al., 2019) and possibly strong purifying (background) selection (Irwin et al., 2016; Cvijović et al., 2018). However, background selection is generally not expected to mimic selective sweeps. Sweepstakes reproduction may apply to many different organisms and could be more prevalent than previously thought. There is, therefore, a need for a critical examination of the contrasting hypotheses.

Here we compare our genomic data to predictions of three coalescent models: the Kingman coalescent (Kingman, 1982) with arbitrary demographic histories, the neutral $\Xi$-Beta$(2-\alpha, \alpha)$ coalescent (Schweinsberg, 2000, 2003; Birkner et al., 2018) modelling random jackpot sweepstakes in diploid highly fecund organisms, and the Durrett-Schweinsberg coalescent derived from a population model of recurrent selective sweeps (Durrett and Schweinsberg, 2005) in a population (SM 1.3). Under the Durrett-Schweinsberg model, the environment is forever changing favoring a new mutation each time that every so often sweeps to fixation. Thus this model well approximates the selective sweepstakes of Williams (Williams, 1975). The Durrett-Schweinsberg coalescent assumes the Moran model of reproduction, which assumes a single offspring produced at any time, is at face value a low-fecundity model. However, the underlying scaling can be very high and thus the rate of offspring production (e.g. egg laying) can be practically infinite making it a high-fecundity model. Studying the site frequency spectrum under recurrent selective sweeps Kim (2006) stated that "the excess of high-frequency derived alleles, previously shown to be a signature of single selective sweeps, disappears with recurrent sweeps." This effect is sometimes—incorrectly—taken to mean that the site frequency spectrum is no longer U-shaped under recurrent selective sweeps. However, the excess or deficiency of high-frequency derived alleles is in reference to expectations of the Kingman coalescent (Kim, 2006) and how that affects Fay and Wu's $H$ statistic (Fay and Wu, 2000). The site frequencies of alleles at intermediate allele frequencies (the alleles contributing most to the variance in fitness) still are reduced under recurrent sweeps (Kim, 2006) preserving the U-shaped site frequency spectrum observed under a single selective sweep.

We analyze whole-genome sequences (at $16\times$ and $12\times$ coverage respectively) of the highly fecund marine fish Atlantic cod (*Gadus morhua*) sampled from two separate localities in Iceland, with the localities serving as statistical replicates (SM Fig. **1**). We also consider whether other forces can explain the observed patterns. We consider population expansion, cryptic population structure, balancing and background selection, and the joint action of several forces.

## Results

### Neutrality under no sweepstakes?

The classical Kingman coalescent, derived from the Wright-Fisher (or similar) model of low-variance reproduction, is the no-sweepstakes model. Several tests of a neutral equilibrium under the Wright-Fisher model of reproduction and the Kingman coalescent use a standardized difference of different estimators of $\theta = 4N_e\mu$ the mutation-rate scaled by population size (Tajima, 1989; Fu and Li, 1993; Fay and Wu, 2000; Zeng et al., 2006; Przeworski, 2002). These tests are sensitive to mutations on different parts of the genealogy and thus of different frequency classes of the site frequency spectrum that also may be influenced by demography, background selection, and selective sweeps (Tajima, 1989; Fu and Li, 1993; Fay and Wu, 2000; Zeng et al., 2006; Przeworski, 2002). A negative Tajima's $D$ indicates an excess of low frequency over intermediate frequency alleles, and a negative Fu and Li's $D$, which contrasts mutations on internal and external branches of a genealogy, indicates an excess of singletons. Thus these statistics are sensitive to deviations from neutrality affecting the left tail of the site frequency spectrum, such as population expansion and background selection (Nielsen, 2005). In contrast, negative values of Fay and Wu's $H$ (Fay and Wu, 2000) and Zeng's $E$ (Zeng et al., 2006) statistics, which weigh the frequency of high-frequency derived alleles, are sensitive to deviations from neutrality affecting the right tail of the site frequency spectrum such as positive selection and selective sweeps (Fay and Wu, 2000; Przeworski, 2002; Nielsen, 2005). Jointly viewing Tajima's $D$ and Fay and Wu's $H$ (the $DH$ test Zeng et al., 2006) is relatively robust against demographic changes and background selection and thus indicative of effects of positive selection and selective sweeps. Our genomic scan of these test statistics (Fig. 1 **a** and **b** and SM Fig. **2** **a** and **b**, and SM Tables **1** and **2**) show extensive and genome-wide deviations from expectations of neutral equilibrium of the classical theory, including indications consistent with selective sweeps occurring throughout the genome (Fay and Wu, 2000; Zeng et al., 2006; Przeworski, 2002). The neutrality index $(NI)$ (Rand and Kann, 1996) derived from the McDonald-Kreitman test (McDonald and Kreitman, 1991) is a ratio of ratios: the number of polymorphic non-synonymous to synonymous sites over the number of fixed non-synonymous to synonymous sites. The $\log$ of $NI$ is a log odds ratio. Under neutrality, $NI = 1$, negative values of $-\log(NI)$ indicate negative purifying selection, and a positive values indicate positive selection. Our estimates show both negative and positive selection effects distributed throughout each chromosome (Fig. 1 **c**). The distribution of the neutrality index (Fig. 1 **d**) is heavier on the side of positive selection for all but two chromosomes (10 and 23 for which the median is close to the neutral expectation). Only a minority of individual tests reach a nominal significance (Fig. 1 **e**). Moreover, none is significant, taking multiple testing into account. Overall, however, the cloud of points indicative of positive selection is heavier than the cloud indicative of negative selection, similar to findings in *Drosophila* and different from humans and yeast where negative selection predominates (Li et al., 2008). Thus the results of these classical tests of neutrality (Fig. 1 and SM Fig. **2**) are similar to the deviations from the $\Xi$-Beta$(2-\alpha,\alpha)$ model and are consistent with the operation of pervasive positive selection in the Atlantic cod genome. The classical no-sweepstakes model with population growth (such as post-Pleistocene population expansion, Hewitt, 2004) is known to affect primarily the singleton class and left tail of the site frequency spectrum. We will also show below that plausible demographic scenarios do not materially improve the fit of neutral models without sweepstakes.

## Random vs selective sweepstakes?

The life table of cod (SM 1.2 and SM Table **3**), showing an exponential decay of the number of survivors with age and an exponential increase in fecundity with age, implies that fewer and fewer individuals produce a larger and larger number of eggs. A few females may live 25 years still increasing fecundity with age. The life table by itself thus results in a large variance in offspring number. Old surviving females may be the lucky few to be alive or the very fit that have passed all selective filters. We next compared our observations to predictions of the $\Xi$-Beta$(2-\alpha,\alpha)$ coalescent, which models random jackpot sweepstakes reproduction in a diploid highly fecund population. Here the parameter $\alpha \in (1,2)$ determines the skewness of the offspring distribution, in essence, the size of the jackpot. A lower $\alpha$ essentially means a larger jackpot. We used a range of approximate Bayesian computation (ABC) posterior estimates of the $\alpha$ parameter (SM 1.3.2). The observed site frequency spectra were overall more V-shaped than the U-shape of the expected normalized site-frequency spectrum predicted by this model (SM 1.3.2) (SM Fig. **3a** and **b**). Singletons and low-frequency variants are closest to expectations of an $\alpha = 1.35$ (SM Fig. **3**). However, as the derived allele frequency increases, the observations are closer to a lower and lower $\alpha$ (as low as $\alpha = 1.0$) predictions. The expected site-frequency spectrum of this model shows local peaks at intermediate allele frequencies, which represent the expected simultaneous multiple mergers of two, three, and four groups, corresponding to the four parental chromosomes involved in each large reproduction event. In diploid highly fecund populations a single pair of diploid parents may occasionally produce huge numbers of juvenile offspring (Möhle and Sagitov, 2003; Birkner et al., 2013b, 2018). The observations do not show these peaks (SM Fig. **3a** and **b**). The expectations of this model are also mainly outside of the bootstrap error bars of the observations (Fig. 2). However, comparing the observed site frequency spectra to expectations of the haploid $\Lambda$-Beta$(2-\alpha,\alpha)$ coalescent, a haploid version of random sweepstakes, (SM Fig. **4**) shows a better fit. Low-frequency variants fit reasonably well to an $\alpha = 1.35$, however, as the derived allele frequency increases, a lower and lower $\alpha$ (as low as $\alpha = 1.0$, the Bolthausen-Sznitman coalescent) gives a good fit. This is likely a signal of either positive or negative natural selection. Rare alleles (less than 10–12%) contribute little to the variance in fitness. Once an allele (a site) reaches an appreciable and intermediate frequency it can contribute significantly to the variance in fitness such that selection quickly moves it out of intermediate frequency ranges. Negative selection moves it to a low frequency and positive selection moves it to a high frequency so alleles spend short time at intermediate frequencies (sojourn times are short). The Durrett-Schweinsberg model is haploid and the resulting coalescent is a Lambda-coalescent. The $\Lambda$-Beta$(2-\alpha,\alpha)$ coalescent for small values of $\alpha$ approximates the Durrett-Schweinsberg coalescent and the fact that the $\Lambda$-Beta$(2-\alpha,\alpha)$ fits better to our data than $\Xi$-Beta$(2-\alpha,\alpha)$ coalescent is further indication for selection.

Furthermore, we used approximate Bayesian computation (ABC) to estimate jointly the parameters $\alpha$ and $\beta$, where $\beta$ denotes a population size rescaled rate of exponential growth of the population forward in time, using the $\Xi$-Beta$(2-\alpha,\alpha)$ coalescent (SM 1.3.2). These processes, reproductive skew and population growth, can account for certain features of the site frequency spectrum. Thus, by jointly estimating $\alpha$ and $\beta$ we hope to obtain a more accurate description of the observed data. The resulting posterior distribution show low values of both parameters (Fig. 3) implying high reproductive skew and little growth. That the distribution of the growth parameter spreads more with higher reproductive skew (as $\alpha \to 1$) is not surprising, as population size is known to affect the model only weakly when the reproductive skew is high. Furthermore, the impact of a variable population size vanishes entirely when the reproductive skew is maximum ($\alpha = 1$) (Freund, 2020; Koskela and Wilke Berenguer, 2019).

Earlier work (Matuszewski et al., 2017), using a model in which a single individual reproduces each time and occasionally wins the jackpot whose size is constant over time, also found reproductive skew over demographic expansion in Japanese sardines. We used a more realistic model (Schweinsberg, 2003), in which the whole population reproduces simultaneously, however occasionally, a single random female hits a jackpot, whose size will vary over time.

The $\Xi$-Beta$(2-\alpha,\alpha)$ model of random sweepstakes shows that reproductive skew is a more likely explanation than demographic expansion under the classical Kingman model and the model predicts an upswing, as observed at the right tail of the site frequency spectrum. It nevertheless cannot adequately explain our data. There were systematic deviations from expectations of the model (see residuals in Fig. 4 **a** and **b** and SM Fig. **5 a** and **b**). The deviations were nearly symmetric around a derived allele frequency of 50% (logit of 0), and rare (less than 12%, logit of $-2$) and common alleles (greater than 88%, logit of 2) were too frequent. In contrast, intermediate alleles were too few compared to model expectations. The deviations immediately suggest the action of positive natural selection by selective sweeps. The path of the allele frequency of a new advantageous mutation can be divided into phases (Barton, 1998). Most new mutations with a selective advantage $s$ are lost from the population in the first few generations when they are scarce (lost with a probability $1-2s$). Mutations conditioned to fix will enter a deterministic fate but are at first still rare enough (< 12%; lag phase) that they contribute very little to the variance in fitness, and hence their frequencies change only a little from one generation to the next. When an allele reaches an appreciable frequency it contributes to the variance in fitness, and selection drives it through intermediate frequencies in a short time that is the inverse of the selection coefficient ($1/s$; the exponential phase). In the last (stationary) phase, the variance in fitness is again low, and the mutation lingers in the population at high frequency until fixed (Barton, 1998; Coop and

Ralph, 2012).

Therefore, we investigated the hypothesis of selective sweepstakes by comparing our observations to predictions of the Durrett-Schweinsberg coalescent derived from the Durrett-Schweinsberg model (SM 1.3.3). In the Durrett-Schweinsberg model, a random site on a chromosome is hit by a beneficial mutation that goes to fixation in a time measured in $\log N$ coalescent time units, where $2N$ is the population size. The beneficial mutation sweeps with it neutral sites that are some recombinational distance from the selected site (Durrett and Schweinsberg, 2005; Nielsen, 2005). Distant sites are more likely to escape this hitch-hiking effect than neighbouring sites because of the higher recombination rates. Even though the model is built from a whole chromosome undergoing recurrent selective mutations, the resulting coalescent only describes a single site under the joint effect of hitchhiking and recombination (c.f. Nielsen, 2005). Thus, the model cannot make joint predictions about several sites, such as measures of linkage disequilibrium. In the supplementary material (SM 1.3.3), we propose a two-site extension of the Durrett–Schweinsberg model in the restricted case of two sampled sequences, facilitating predictions of linkage disequilibrium. This model of recurrent selective sweeps explains our results for all subsets of the data (Fig. 2 and Fig. 5), and is also consistent with the decay of linkage disequilibrium observed in the data (SM Fig. **6**) provided that a small fraction of sweeps (on the order of 10%) are taken to affect the whole chromosome regardless of recombination. Such "sweeps" are characteristic of e.g. population bottlenecks. The compound parameter $c = \delta s^2/\gamma$ of the Durrett-Schweinsberg model measures the rate of selective sweeps ($\delta$) times the squared selection coefficient ($s^2$) of the beneficial mutation over the recombination rate ($\gamma$) between the selected site and the site of interest. ABC estimates yield similar values across all replicate data sets, an average of about 10 ( Fig. 5 and SM Fig. **7**). The residuals of the fit to the Durrett-Schweinsberg coalescent (SM Fig. **5 c** and **d**) show deviations that are both smaller and the opposite of the deviations of those of the neutral $\Xi$-Beta($2-\alpha, \alpha$) model (Fig. 4 **a** and **b** and SM Fig. **5 a** and **b**) with intermediate frequency classes too frequent. The compound parameter $c$ ranges from 5 to 18 for different functional groups (Fig. 5). Actual sweeps may also affect nearly neutral or even slightly deleterious sites in addition to the neutral sites that are linked to the selected site in the model. This means that the rate of selective sweeps is 5–18 times more frequent (Fig. 5) than the coalescence rate of a population with a low variance mode of reproduction described by the classical Kingman coalescence. The Durrett-Schweinsberg model is essentially a haploid model, and we suggest that a diploid model, where dominance generates two phenotypes such that selection acts on pairs of chromosomes jointly rather than single chromosomes as in the Durrett-Schweinsberg model would provide an even better fit. However, developing a diploid multi-locus version of the Durrett-Schweinsberg model is outside the scope of the present work. Nevertheless, a comparison of our data with predictions of the Durrett-Schweinsberg model, in particular in comparison with our additional analysis, is perfectly valid. Overall, the selective sweepstakes hypothesis embodied in the Durrett-Schweinsberg coalescent (Durrett and Schweinsberg, 2005) modelling recurrent selective sweeps, in essence, explains our data, whereas the hypothesis of low variance reproduction and the one of random sweepstakes do not.

To further investigate and take into account the effects of selection and recombination on the observed patterns of allele frequencies, we take several steps. We did a principal component based genome-wide scan of selection (using PCangsd Meisner and Albrechtsen, 2018) and detected several peaks (SM Fig. **8**). We used sites that are at least 500 kb away from the selective peaks. We refer to these as non-selection sites. We extracted sites from the genome that are likely under different selective constraints. We thus extracted fourfold degenerate sites (referred to as 4Dsites), intron sites, intergenic sites, promoter sites, $5'$ UTR sites, $3'$ UTR sites, and exon sites. The less constrained sites are not necessarily neutral to selection For example, although silent at the protein level, mutations at fourfold degenerate sites could affect transcriptional and translational efficiency and mRNA stability, thus giving rise to selection for or against such sites. However, the first three classes are generally considered less constrained and the other classes more constrained by selection.

Furthermore, we used OmegaPlus (Alachiotis et al., 2012) and RAiSD (Alachiotis and Pavlidis, 2018) to detect selective sweeps genome wide. Both methods use local linkage disequilibrium to detect sweeps (Nielsen, 2005) and in addition RAiSD uses a local reduction in levels of polymorphism and shifts in the frequencies of low- and high-frequency derived alleles affecting respectively the left and righ tails of the site frequency spectrum. Both methods indicate pervasive selective sweeps on all chromosomes (Fig. 6). We also used SLiM (Haller and Messer, 2019) to simulate positive selection under the no-sweepstakes Wright-Fisher model and under a random sweepstakes model in the domain of attraction of a Xi-Beta coalescent (SM Fig. **9**). We tried various forms of dominance of selection among diploid genotypes (semidominance, $h = 0.5$ and full dominance, $h = 1.0$) with different strength of selection (selection coefficient $s$). The model of successive selective pass or fail filters suggests that lacking a function (a derived allele) is a failing genotype while having a function (derived allele) is a passing genotype as modeled by full dominance. The observation of the heavy mortality of immatures (type III survivorship, SM Table **3**) therefore suggests a model of selection against a recessive lethal and for a dominant. This a a two-phenotype model for a diploid organism. The results of the SLiM simulations of positive selection (SM Fig. **9**) gave site frequency spectra that are qualitatively similar to the observed spectra. Selection for a semidominant produced more U-shaped spectra while selection for a dominant produced more V-shaped spectra similar to the observed. Recurrent hard sweeps interrupting the standard Kingman coalescent (simulated using `msprime`) produce a U-shaped site-frequency spectra (SM Fig. **10**) that is qualitatively similar to our data from the South/south-east coast.

## Can forces other than selective sweeps better explain the patterns?

The effects of demography (changes in population size, population structure, and migration) can be hard to distinguish from various forms of selection (Nielsen, 2005). And different forms of selection can affect the various parts of the site frequency spectrum in similar ways. We now consider whether forces other than selective sweeps can provide better explanations for the observed patterns.

## Historical demography and low variance reproduction

Our estimated demographic history (SM Fig. **11** and SM Fig. **12**) show population expansion in the distant past leading to relative stability of population size in the recent past to modern times. In some cases, an apparent population crash in recent times (SM Fig. **11 c**), which is chromosome-specific, is an exception to this. Demography produces genome-wide effects and, thus, this is likely a peculiarity of runs of homozygosity of some chromosomes (such as centromeric regions, for example) and not reflecting historical size changes of the population. Based on these population growth curves (SM Fig. **11** and SM Fig. **12**) we generated population size change scenarios for simulating site frequency spectra using `msprime` (Kelleher et al., 2016; Baumdicker et al., 2021). The results (SM Fig. **13**) show monotonically decreasing frequency with the size of the mutation or L-shaped site frequency spectra that neither capture the singleton class nor the upswing of the right tail of the observed site frequency spectra (Fig. 2, SM Fig. **3** and SM Fig. **14**). Thus, there is no evidence in our results for a low-variance no-sweepstakes mode of reproduction modelled by the Kingman coalescent, even taking demographic histories of population expansion or collapse into account. Our simulations are in line with the theoretical proof (see Appendix B of (Sargsyan and Wakeley, 2008)), showing that the normalized expected site-frequency spectrum of a Kingman-coalescent under arbitrary population size history is L-shaped.

## Potential confounding due to cryptic population structure

Here we examine alternative explanations for our observations. In particular, are the site frequency spectra influenced by cryptic population structure?

The effect of hidden population structure on the site frequency spectra is expected to look similar to the patterns seen for the inversion chromosomes. These are chromosomes Chr01, Chr02, Chr07, and Chr12 known to carry large inversion (Kirubakaran et al., 2016; Berg et al., 2016). They show two peaks in the site frequency spectrum (SM Figs. **17** and **18**) at the frequency of the variants's haplotype frequency and show a block of values for neutrality statistics (Fig. 1 and SM Fig. **2**). If a sample of size $n$ diploid organisms is composed of two cryptic reproductively isolated populations (sample sizes $n_1$ and $n_2$) we expect to see peaks in the site frequency spectra at the relative frequencies of the two groups. If $n_1 = n_2 = n$ we expect a sharp peak at $n/(2n)$. This peak would include all fixed sites in both populations ($n_1/n$ and $n_2/n$) and spread over neighboring frequency classes ($(n_1 - 1)/n, (n_1 - 2)/n, (n_2 - 1)/n, (n_2 - 2)/n$ and so on). If the frequencies of the two groups differ ($n_1 \neq n_2$) two peaks will appear, but are expected to be narrow. They will always include all sites fixed in either population (because fixed sites in either population will appear to be segregating in the sample as a whole).

To study the potential effects of population structure, we used `msprime` (Kelleher et al., 2016; Baumdicker et al., 2021) to simulate the Kingman coalescent with two isolated populations exchanging a varying number of migrants under population growth as determined by the growth parameter $\beta$. Thus we examined the effects of cryptic structure on the site frequency spectrum by varying the growth rate and the effective number of migrants between subpopulations ($4N_em$), and varied the number of individuals sampled from the population with fewer individuals represented (referred to as the minor population). Parameters of the simulations were the number of individuals from the minor population ($k \in \{4, 3, 2, 1\}$), the migration rate ($m = 10^{-5} \ldots 10^{-3}$), and growth rate ($g = 10^{-4} \ldots 10^{-1}$. The effective size was set at $N_e = 500$ and thus the effective number of migrants per subpopulation per generation was $4N_em = 0.02 \ldots 2$.

We use a two-island model with exponential growth under the Kingman coalescent as a simple tool for assessing the qualitative, joint effect of demography and substructure on the site frequency spectrum (SM Figs. **15** and **16**). Two narrow peaks at opposite allele frequencies are evident (much like the two narrow peaks for the inversion chromosomes, SM Figs. **17** and **18**) becoming smaller with increasing migration. Only if the sample contained a single individual of the minor population is there a remote resemblance to the observed data (SM Fig. **15 g**, **h**, and **j**). Nevertheless, even in this case, doublets are more common than singletons, and it is hard to find combinations of growth and migration rates to mimic the observed data. We used the Xi-Beta coalescent for similar simulations (SM Fig. **16**) and got the same results qualitatively. Therefore, population structure in a population evolving according to the Wright-Fisher (or a similar) low-fecundity model or in a population evolving under a neutral sweepstakes model is an improbable explanation for our results. Both simulations (SM Figs. **15** and **16**) show that only for a minor sample size of one diploid individual do the models show a remote resemblance to our data. To further address this issue, we, therefore, estimated the site frequency spectra with a leave-one-out approach (SM Fig. **19**). The leave-one-out approach is model-free: whichever model is correct, one of the leave-one-out samples should behave differently if a cryptic population structure with a minor sample size of one is present in our data. None of them do. There is no indication that our sample from the South/south-east coast is composed of 67 individuals from one population and a single individual from a divergent population.

To further study the potential effects of cryptic population structure, we note that principal component analysis (PCA) of variation at each of the four chromosomes harbouring large inversions reveal two or three groups that likely represent genotypes of the inversion alleles. There are three groups for Chr01 (which we refer to as Chr01-*AA*, Chr01-*AB*, and Chr01-*BB*), Chr02 (Chr02-*CC*, Chr02-*CD*, and Chr02-*DD*), and Chr07 (Chr07-*EE*, Chr07-*EF*, and Chr07-*FF*), and two groups for Chr12 (Chr12-*GG* and Chr12-*GH*), which has a low frequency of one inversion allele (Fig. **20** and SM Figs. **17** and **18**). If we take these groups as representing the haplotypes of the inversions the genotypic frequencies at each chromosome do not deviate from Hardy-Weinberg equilibrium, and there is thus no evidence for breeding structure (no Wahlund effects, SM Table **4**). However, as the inversions effectively suppress recombination between the inversion alleles, we can also look at the chromosomes of the inversion genotypes as effectively isolated populations with no recombination (migration) between them and estimate the site frequency spectra within genotypes for the inversion chromosomes. Furthermore, we conjecture that the PCA groups observed at inversion chromosomes represent reproductively isolated but cryptic populations. Because demography has genome-wide effects the cryptic structure should be evident in the rest of the genome. We, therefore, estimate the site frequency spectra for the 19 non-inversion chromosomes (chromosome 3–6, 8–11, 13–23) for these groups.

Principal component analysis (PCA) did not show any structure for the non-inversion chromosomes. However, the four inversion chromosomes each showed two narrow peaks at intermediate allele frequencies (SM Figs. **17** and **18**) indicative of either balancing selection or cryptic population breeding structure. If this is breeding structure it should affect the whole genome. To disentangle the effects of balancing selection and potential breeding structure we used the groups defined by PCA at the inversion chromosomes to investigate the inversion chromosomes themselves and the non-inversion chromosomes. We thus conjecture that the PCA groups represent cryptic breeding units.

PCA revealed three (or two) groups on the first principal axis that explains 4–36% of the variation at the inversion chromosomes (SM Fig. **20 a**, **d**, **g**, and **j**). The PCA groups most likely represent genotypes of inversion haplotypes. Taking membership in PCA groups to represent inversion genotype, their frequencies fit the Hardy-Weinberg equilibrium (SM Table **4**) and thus there is no evidence of heterozygote deficiency or Wahlund effect (Wahlund, 1928) indicative of breeding structure. The only exception is chromosome 7 in the Þistilfjörður population, which shows a slight heterozygote excess (SM Table **4**). Furthermore, the site frequency spectra of the PCA groups (SM Fig. **20 b**, **e**, **h**, and **k**) show the same overall V-shape pattern as the site frequency spectra for the overall data (Fig. 2). Additionally, the intermediate PCA group shows a sharp peak at a derived allele frequency of $n/(2n)$ (an equal frequency of two types or 0 on the logit scale) as expected for a group composed of heterozygotes only. Similarly, the site frequency spectra of these PCA groups for the 19 non-inversion chromosomes combined (Fig. **20 c**, **f**, **i**, and **l**) show a pattern characteristic of the site frequency spectra for the overall data. There is not the slightest hint of a Kingman coalescent-like behaviour for any of these PCA groups. Similarly, expectations of the $\Xi$-Beta$(2-\alpha, \alpha)$ coalescent do not explain the data. Overall, the shape of the site frequency spectra for each of the inversion chromosomes (SM Figs. **17** and **18**) and for the PCA groups of each of the inversion chromosomes (SM Fig. **20**) is the same as the shape of the site frequency spectra of the non-inversion chromosomes (Fig. 2). This shape is well explained for all PCA groups by the Durrett-Schweinsberg coalescent (Durrett and Schweinsberg, 2005), for which we estimated the $c$ parameter using ABC for the PCA group of the respective inversion chromosome (SM Fig. **20**).

The observed V-shaped site frequency spectra are inconsistent with an amalgamation of cryptic units reproducing under a Wright-Fisher model. The PCA groups are not cryptic breeding units, and we reject the above conjecture. Instead, we consider them to represent polymorphic inversion genotypes maintained by some form of balancing selection such as frequency-dependent fitnesses arising from the accumulation of deleterious recessives on homokaryotypes (Jay et al., 2021) or other mechanisms of balancing selection (Faria et al., 2019).

### Balancing and background selection and functional constraints

Besides the Durrett-Schweinsberg model, various mechanisms of selection may influence the results. Here we examine the effects of balancing selection, different selective constraints, and background selection.

There are several signs that natural selection affects the observed patterns. Balancing selection retains linked neutral or nearly neutral variants at intermediate frequencies. The tighter the linkage and less the recombination, the longer the coalescent time of the neutral variants (Charlesworth, 2006). The observed site frequency of intermediate frequency alleles is higher among the four inversion chromosomes than the 19 non-inversion chromosomes. All comparisons of the four inversion chromosomes and the 19 non-inversion chromosomes show this effect (SM Figs. **17**, **18** and 2). However, balancing selection does not affect the overall V-shape of the site frequency spectrum of the inversion chromosomes (SM Figs. **17** and **18**).

The selection scan is model-free and is based on finding genes or genomic regions that are outliers relative to the overall genome-wide allele frequencies and taking potential population structure into account. A principal components based genomic scan of selection (SM Fig. **8**) showed many peaks that are likely indicative of recent and strong positive selection. A few peaks were population-specific, but the two populations share most peaks. The region under a peak ranged from a single site to about 2 Megabases (Mb). We extracted sites 500 kb or more away from the peaks (referred to as no-selection) and included with genomic regions under different selective constraints. We extracted fourfold degenerate sites, introns, intergenic sites as less constrained regions, promoter regions, exons, $3'$-UTR, and $5'$-UTR as more selectively constrained regions. The mean, median and mode of the estimated compound parameter $c$ of the Durrett-Schweinsberg model for the different genomic regions ranked from least constrained to most constrained sites (Fig. 5). The ABC-MCMC was well mixed in all cases.

There are two possible explanations for the rank order of the compound parameter $c$ with functional genomic regions. First, the more functionally important a region of the genome is, the stronger the selection coefficient of a new advantageous mutation will be. Such mutations will sweep through the population and carry with them tightly linked neutral mutations in these same regions ($c$ being inversely proportional to the recombination rate $\gamma$). Alternatively, different functional regions are preserved and constrained by purifying (negative) selection. If the sites are tightly linked, a positively selected mutation sweeping through will affect neutral, nearly neutral, and even deleterious sites. A tug of war between the effects of the sweep and purifying selection at a site results in a net effective selection coefficient for that site. The compound parameter $c$ of the Durrett-Schweinsberg model estimates the net effective selection coefficient squared over the recombination rate, which may generate the observed

rank order. Of course, both explanations may apply to different positive mutations. Thus selective sweeps permeate the genome affecting most if not all sites (Pouyet et al., 2018).
To study the effects of background selection, we carried out forward-in-time simulations of the Wright-Fisher model (using SLiM (Haller and Messer, 2019)). Simulations that ran for a relatively short number of generations (on the order of the population size) produced V-shaped site frequency spectra (SM Fig. **13 d**). However, when simulations of the same parameter values ran for a large number of generations (up to ten times the population size of $10^5$ chromosomes) they accumulated more variation (SM Table **5**) and produced monotone L-shaped site frequency spectra. Thus only in a narrow window of non-equilibrium between the input of mutation and its removal by purifying selection or loss by drift can background selection site frequency spectra resemble our observed spectra. In general, however, background selection does not fit our data.

## The joint action of several evolutionary forces

The analysis thus has shown that considered singly, the various factors such as demography and background selection do not provide a good fit, particularly not involving the derived alleles at the right tail of the site frequency spectrum. Analysing the joint action of demography, purifying and background selection with or without random sweepstakes on the genome level is computationally prohibitive. We, therefore, resorted to simulations using SLiM (Haller and Messer, 2019) of a sizeable fragment of a chromosome evolving under the joint action of several forces of evolution (SM Fig. **13**). As is common in complex, multi-component simulations, it may be possible to tweak parameters to obtain results matching the observed data. Nevertheless, a comprehensive model-fitting search is infeasible in our setting. However, the combined effect of negative background selection without selective sweeps did not produce qualitatively accurate, U-shaped site frequency spectra for any parameter combination we tested. Furthermore, a combination of random sweepstakes, randomly occurring bottlenecks, and background selection (SM Fig. **13**, **e** and **f**) did not produce a qualitatively similar U-shaped pattern as the data. Hence, even if best-fit parameters could match the data, we expect the fit would not be robust to small changes in either parameter values or observed data, thus having low predictive and explanatory power. In contrast, scenarios involving selective sweeps routinely produced the right qualitative shape of the site frequency spectra. Hence, we expect a (hypothetical) best-fit analysis to be far more robust.

## Synopsis of results

We have shown that the Durrett-Schweinsberg coalescent modelling recurrent selective sweeps affecting linked sites gives a best fit to our observations (Fig. 2). By extension the hypothesis of reproduction by selective sweepstakes is best supported by our data. The Kingman coalescent and the Wright-Fisher model of reproduction, without strong positive selection of recurrent strongly beneficial mutations (SM Figs. **9** and **10**), cannot explain our data. Similarly, the model of random sweepstakes, the Xi-Beta coalescent, in which a random individual has windfall reproductive success, although fairing better than the Kingman coalescent nevertheless cannot explain the observations. Furthermore, through analysis and forward and backward simulations we study whether other evolutionary forces can adequately explain the data. Historical demographic expansions or contractions do not explain our data (SM Fig. **11**). Analysis of potential cryptic population structure does not provide answers to our patterns (SM Fig. **20**). Similarly, modelling sampling from divergent populations a combination of extreme parameter values can produce patterns similar to the observed patterns (SM Figs. **15** and **16**). However, a leave-one-out analysis of our sample shows that our sample was not produced under such extreme parameter values (SM Fig. **19**). There are clear signals of balancing selection of large inversions at four chromosomes (SM Figs. **17** and **18**). However, balancing selection does not change the overall shape of the site frequency spectrum of these chromsomes, which is the summary statistic that we use for our analysis. Simulations of background selection show that a narrow window of parameter space can resemble observed patterns but in general background selection does not fit our results (SM Figs. **9 d** and **13**). Finally, simulations of the joint action of several evolutionary forces, notably of demography and background selection with or without selective sweeps do not produce qualitatively accurate U-shaped site frequency spectra similar to the observed except in simulations that included selective sweeps (SM Fig. **13**).

## Discussion

Understanding recruitment dynamics and what shapes the distribution of individual reproductive and recruitment success is a fundamental challenge in evolutionary genomics of high fecundity and is key to further understanding metapopulation and community dynamics, predicting response to anthropogenic challenges, for conservation and management, and further development of ecological and evolutionary theory (Eldon, 2020). We show that selective sweepstakes, modeled by a particular example of the Durrett-Schweinsberg multiple-merger coalescent derived from a population model of recurrent selective sweeps (Durrett and Schweinsberg, 2005), essentially explains our data. Even a model of recurrent but incomplete selective sweeps (Coop and Ralph, 2012) similarly leads to U-shaped site frequency spectra generated by a multiple-merger coalescent model similar to the Durrett-Schweinsberg model. We further show that neither no-sweepstakes reproduction nor random-sweepstakes reproduction can explain our data. Our results indicate that strong pervasive positive natural selection is pivotal in reproductive sweepstakes, more so than windfall sweepstakes (Hedgecock and Pudovkin, 2011).

Interpreting the Durrett-Schweinsberg model as approximating selective sweepstakes, we conclude that our findings are strong evidence for selective sweepstakes (Williams, 1975) characterizing the distribution of individual recruitment success of the highly fecund Atlantic cod. Under the Durrett-Schweinsberg coalescent of recurrent selective sweeps, of a new mutation each time, happen very fast compared to the coalescent timescale. The continuous input of new beneficial mutations represent the Sisyphean genotypes that forever climb a selective peak under Williams' concept of selective sweepstakes (Williams, 1975). By extension, selective sweepstakes is the life history of highly fecund organisms with skewed offspring distribution.

Recurrent bottlenecks may mimic the effects of recurrent selective sweeps (Galtier et al., 2000). The duration, depth, and rate of recovery of a bottleneck (Nei et al., 1975) relative to the coalescent $\log N$ timescale of recurrent sweeps under the Durrett-Schweinsberg model is an important issue. A small number of individuals having large numbers of descendants due to a bottleneck and rapid recovery or due a selective sweep will in both cases lead to multiple mergers in the genealogy. Our simulations of random sweepstakes with recurrent bottlenecks yield roughly a U-shaped site frequency spectrum but the fit is not as good as for the selective sweepstakes model. In the Durrett-Schweinsberg model, interpreting a small fraction of sweeps (on the order of 10%) as population bottlenecks resulted in a model which was able to explain the decay of linkage disequilibrium observed in Atlantic cod, without affecting the good fit of the site frequency spectrum. Overall, therefore, the Durrett-Schweinsberg model explains our data although it is formally only applicable to single-locus data from a haploid species (the resulting coalescent process traces the genealogy of a single site), assumes a constant population size, disallows competing, simultaneous sweeps (Kim and Stephan, 2003), and only models hard sweeps.

High fecundity matters in two ways in this process. First, each round of replication results in many new mutations in the genome of a new gamete. Even though the probability of a positive mutation is very low, the millions of gametes produced by each female multiplied by the billions of individuals in a population ensure a steady input of new positive mutations to each generation. Second, high fecundity makes available a high reproductive excess which permits substitutions to occur at high rates by natural selection without the population going extinct (Felsenstein, 1971). Reproduction of a high fecundity organism compares with reproduction of a virus in an epidemic. Each infected individual produces hundreds of billions of viral particles. Even with a tiny proportion of positive mutations the numbers of new mutations are so enormous that it is all but certain that an epidemic produces a steady stream of more contagious and fitter viral variants that sweep to fixation by selection. If the population crashes (Hutchings, 2000) the mutational input of adaptive variation diminishes. The population may run out of fuel for responding to environmental challenges via selective sweeps and go extinct (Felsenstein, 1971). Kimura's neutral theory of molecular evolution and polymorphisms (Kimura, 1968) relied on excessive genetic load based on Haldane's dilemma (Haldane, 1957) that the cost of adaptive substitution would limit the rate of evolution lest the population go extinct (Felsenstein, 1971). Truncation selection of continously distributed characters, where genetic and nongenetic factors independently affect the probability of survival and act cumulatively in each individual (Williams, 1975), mitigates the genetic load (King, 1967; Sved et al., 1967). Our considerations above of full dominance with selection against a lethal homozygote would entail a large genetic load. However, there can be strong selection in one patch and near neutrality in another due to differences in competition and predation. The marginal fitness differences would then be less but such soft selection (Wallace, 1975; Reznick, 2016) would not drive the population extinct (Charlesworth, 2013). Marginal fitnesses would still preserve full dominance and a two phenotype selection scheme and thus behave similarly to the haploid Durrett-Schweinsberg model. The high fecundity and consequent excessive reproductive capacity in our study organism may also alleviate the genetic load problem. However, both loss of mutational input and genetic load (a case of selective extinction) may nevertheless be a factor in the non-recovery of a population following a crash (Hutchings, 2000).

Our estimate of the rate of selective sweeps (SM 1.5) amounts to mergers of ancestral lineages of our sample happening because sweeps occur at 5 to 18 times higher rates than mergers due to ordinary low-variance reproduction (Fig. 5). In the classical model, the coalescence rate is on the order of the population size, or $N$ generations, but the duration of selective sweeps is on the order of $\log N$ generations. If we assume that there are a billion cod in the Icelandic population, this is some 20 generations or about 100 years from when a beneficial mutation arises until fixation. The sigmoid nature of the positive selection curve, with a lag phase followed by an exponential phase and ending in a stationary phase, the main action of selection bringing an allele from a low frequency to a high frequency during the exponential phase may only take a few generations, say 15–20 years. Erratic climatic variability, such as the great salinity anomalies (Cushing, 1969; Dickson et al., 1988) in the North Atlantic, which can greatly affect cod reproduction and ecology, is detectable over decadal time scales, similar to the exponential phase of our estimated selective sweeps.

We estimate that each chromosome in Atlantic cod is affected by a selective sweep every 23 to 50 years on average (SM 1.5). Since we also see evidence of rapid recombination (SM Fig. **6**), we expect that any one sweep will not strongly affect a large region of a chromosome. The rapid recombination will modulate the genomic footprints of sweeps. There is clear evidence that sweeps happen everywhere along the genome (in chromosomal fragments of different sizes, different functional groups, and on all chromosomes (Fig. 5 and SM Figs. **3**, **17**, and **18**). It is, therefore, likely that the true rate of sweeps is even faster than our estimate. For example, if an average sweep were to affect, say, 10% of a chromosome, we would expect to see sweeps every three to four years or roughly once a generation to explain our results. Building a fully quantitative, data-informed picture of the rate of sweeps requires the development of a diploid, genomic version of the Durrett-Schweinsberg model, which is currently absent from the literature, and for which task our results provide strong applied motivation.

Our findings provide a new perspective on coalescent models in population genetics and genomics. For the first time, a test involving genomic data, i.e. using copies of chromosomes from several pairs of homologous chromosomes, is made on the contrasting hypotheses of reproduction using multiple merger coalescents in a diploid organism. It is also the first time multiple merger coalescent models based on neutral evolution and selection are contrasted. Previously, two neutral $\Lambda$-coalescents have been compared to data of outbreaks of the tuberculosis bacterium and used the Bolthausen-Sznitman coalescent ($\alpha = 1$) to model rapid selection (Menardo et al., 2019) Our findings have repercussions for and give impetus to further theoretical development of multiple merger coalescents, particularly for multiple merger coalescent models of strong selection.

We suggest that sweepstakes reproduction is much more common than previously thought. It is essential to understand sweepstakes and the natural and anthropogenic ecological processes conducive to sweepstakes (Hedgecock and Pudovkin, 2011; Williams, 1975). Are selective sweepstakes (Williams, 1975) the rule or is there a role for random sweepstakes (Hedgecock and Pudovkin, 2011; Vendrami et al., 2021)? It is possible that big-bang, the semelparous reproductive strategy of reproducing once and die, is sweepstakes reproduction if there are ecological mechanisms generating a high-variance highly skewed offspring distribution. This mode of reproduction characterizes many annual plants, a myriad of insects, and vertebrates such as the Pacific salmon (*Oncorhynchus*) and the Arctic cod (*Boreogadus saida*), a close relative of the Atlantic cod. We further posit that sweepstakes may be the mode of reproduction of viruses (Timm and Yin, 2012) as inferred from overdispersion of offspring distribution from superspreader individuals and events (Endo et al., 2020), some cancer cells (Kato et al., 2017), and various bacteria (Wright and Vetsigian, 2019; Menardo et al., 2019; Ypma et al., 2013). Fungi and plant pathogens, which cause extensive crop losses of great economic importance (Pimentel et al., 2000), may also reproduce by sweepstakes. Similarly, many repeat-reproducers, the iteroparous reproductive strategy, produce vast numbers of tiny eggs in each reproductive season. It applies to many marine organisms such as oysters (Hedgecock and Pudovkin, 2011), and the Atlantic cod and its Pacific relatives (Árnason and Halldórsdóttir, 2019) that support large fisheries of great economic importance. The dynamics of all these systems can be profitably studied using multiple merger coalescents (Freund et al., 2022), be they generated by random or selective sweepstakes.

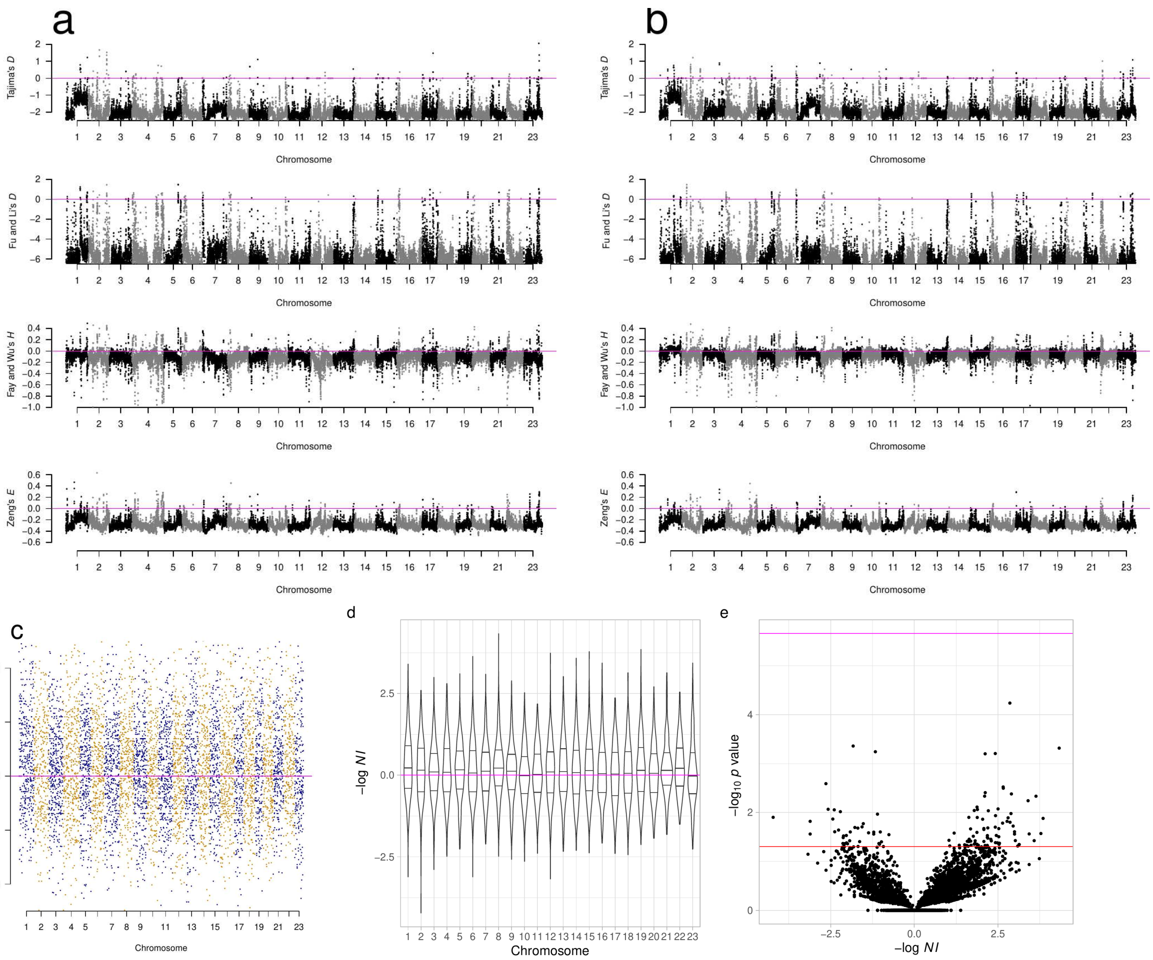


Figure 1: **Neutrality test statistics and neutrality index across all chromosomes. a**, **b**, Manhattan plots of Tajima's $D$ (Tajima, 1989), Fu and Li's $D$ (Fu and Li, 1993), Fay and Wu's $H$ (Fay and Wu, 2000), and Zeng's $E$ (Zeng et al., 2006) show mostly negative values at all chromosomes implying deviations from neutrality. Sliding window estimates (window size 100 kb with 20 kb step size) using GL1 genotype likelihoods for the South/south-east population and the Þistilfjörður population.< Value of statistic under Kingman coalescent neutrality equilibrium indicated with magenta horizontal line. **c**, The neutrality index (Rand and Kann, 1996) associated with the McDonald-Kreitman test (McDonald and Kreitman, 1991) $NI = (P_n \times D_s)/(P_s \times D_n)$ where $P_n$, $P_s$, $D_n$, and $D_s$ are the number of non-synonymous and synonymous polymorphic and fixed sites respectively for all genes of each chromosome. Negative values of $-\log NI$ implying purifying (negative) selection and positive values implying positive selection (selective sweeps) are distributed throughout each chromosome. The outgroup is Pacific cod (Gma) and the magenta horizontal line is at neutral equilibrium. **d**, The distribution of $-\log NI$ per chromosome (violin plots with quartiles) are heavier on the positive side. **e**, The $-\log_{10} p$ value significance of Fisher's exact test for the McDonald-Kreitman test (McDonald and Kreitman, 1991) for all genes in the genome plotted against the $-\log NI$ neutrality index (Rand and Kann, 1996). Overall, the cloud of positive values is denser than the cloud of negative values. The outgroup is Pacific cod (Gma). The red horizontal line is at nominal significance level of 0.05 for individual tests and the magenta line is $0.05/n$ the Bonferroni adjustment for multiple testing. Neutrality index of data from the South/south-east population.

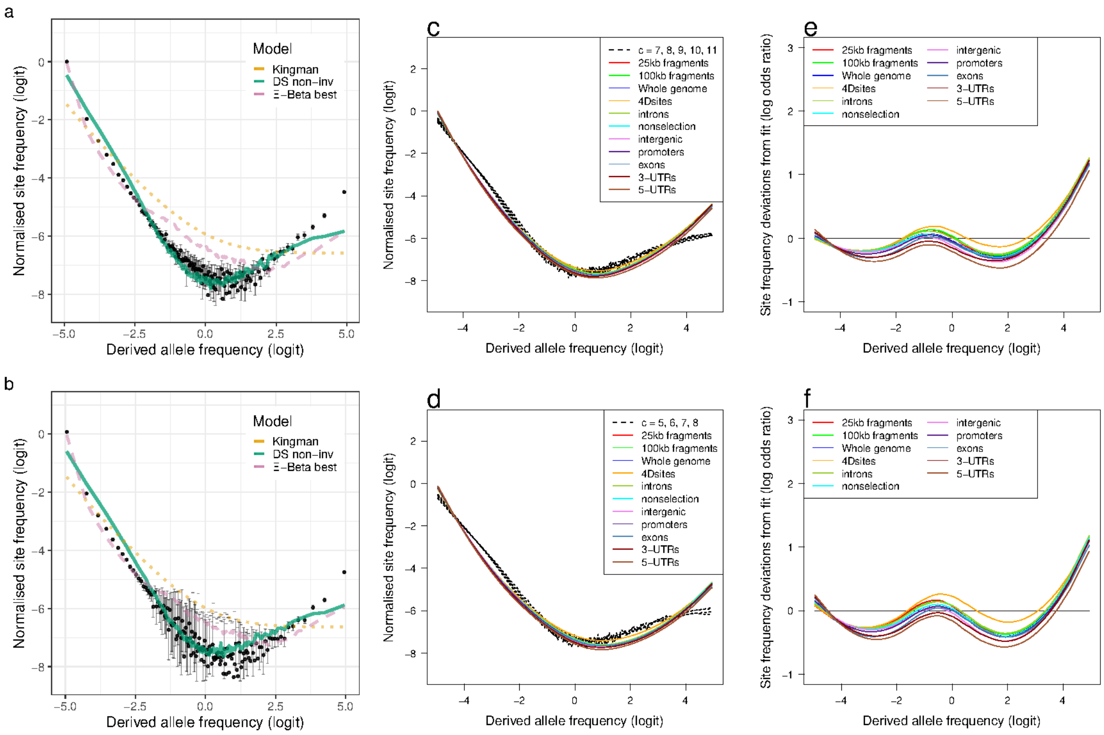


Figure 2: **Fit of observations to models: the no-sweepstakes model, the random sweepstakes model, and the selective sweepstakes model.** **a**, **b**, Mean observed site frequency spectra for the 19 non-inversion chromosomes combined estimated with GL1 likelihood for the South/south-east and Þistilfjörður replicate populations respectively. Error bars of observed data are ±2 standard deviations of the bootstrap distribution. Expected site frequency spectra are the Kingman coalescent modelling non-sweepstakes, the best approximate maximum likelihood estimates (Eldon et al., 2015) of the Ξ-Beta model modelling random sweepstakes, and the ABC estimated Durrett-Schweinsberg coalescent (DS) modelling selective sweepstakes. The observed site frequency spectra of different sized fragments and various functional classes compared to expectations of the Durrett-Schweinsberg coalescent (DS) ABC estimated for the non-inversion chromosomes for the South/south-east population (**c**) and the Þistilfjörður population (**d**). The compound parameter $c$ ranges from 5 to 11. The different functional groups are four-fold degenerate sites (4Dsites), intronic sites, non-selection sites (sites more than 500 kb away from peaks of selection scan, SM Fig. **8**, intergenic sites, promoters, exons, $3'$-UTR sites (3-UTRs), and $5'$-UTR sites in order of selective constraints. **e** and **f** Deviations from expectations of the Durrett-Schweinsberg model of recurrent selective sweeps of different sized fragments and functional groups for the South/south-east population and the Þistilfjörður population respectively.

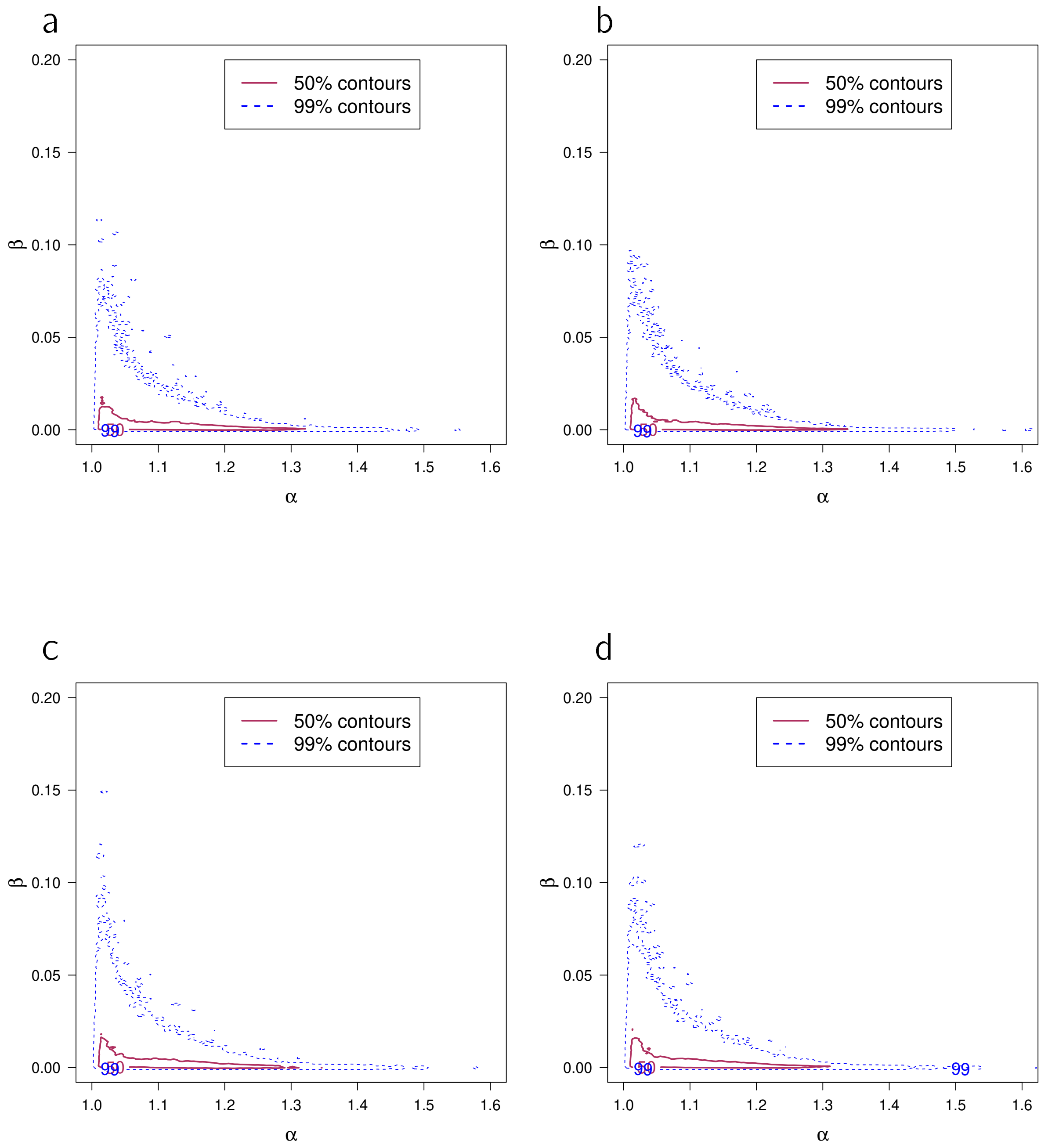


Figure 3: **Approximate Bayesian computation (ABC) joint estimation of parameters of the neutral $\Xi$-Beta$(2-\alpha,\alpha)$ coalescent (random sweepstakes) and of population growth. a, b, c, d** A kernel density estimator for the joint ABC-posterior density of $(\alpha, \beta) \in \Theta_{\mathrm{B}}$. The parameter $\alpha$ determines the skewness of the offspring distribution in the neutral Beta$(2-\alpha,\alpha)$ coalescent model, and the $\beta$ is a population-size rescaled rate of exponential population growth. Estimates are for the GL1 (**a**) and GL2 (**b**) for the South/south-east population and for the GL1 (**c**) and GL2 (**d**) for the Þistilfjörður population. A bivariate model-fitting analysis adding exponential population growth to the $\Xi$-Beta$(2-\alpha,\alpha)$ coalescent does not improve model fit.

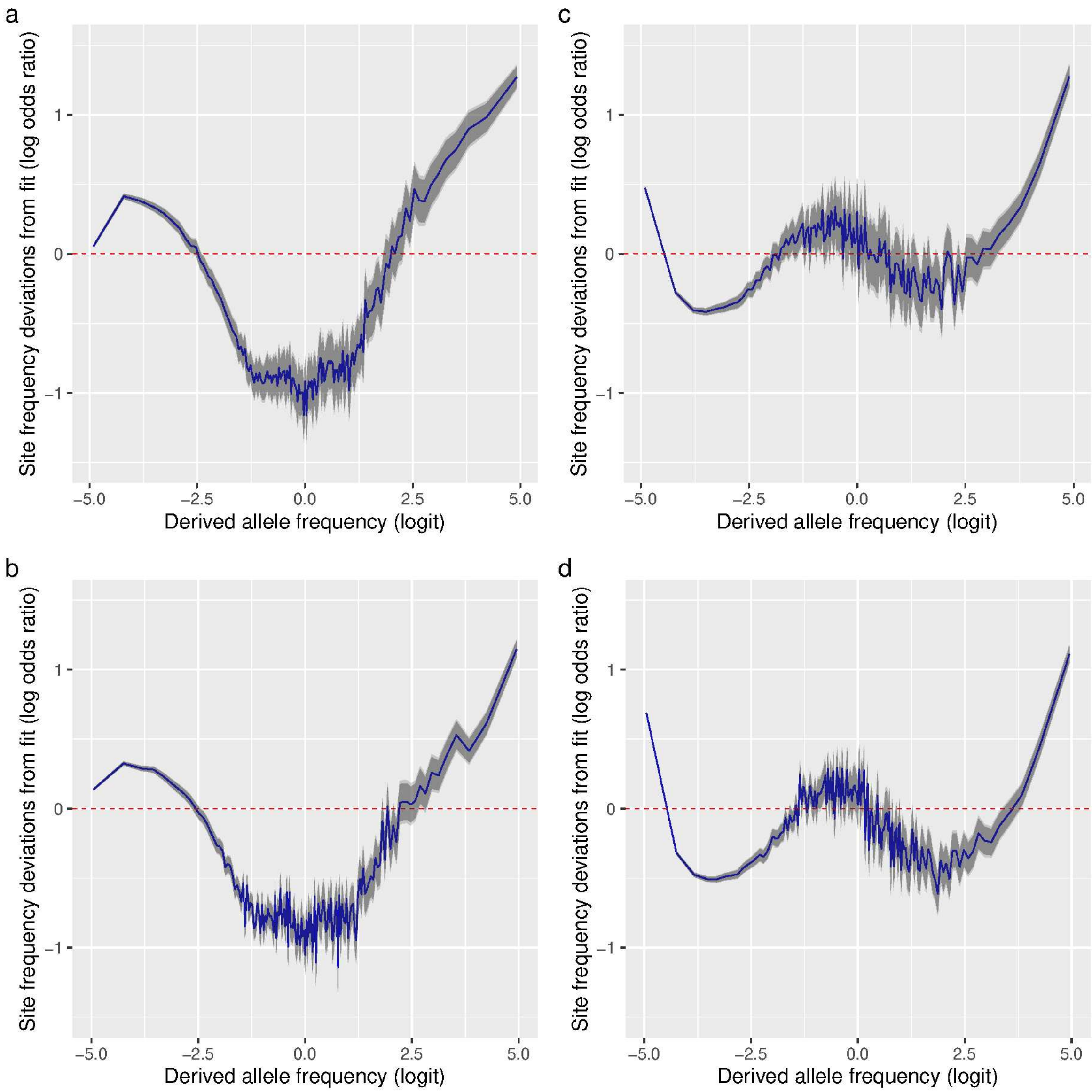


Figure 4: **Deviations from fit to the random sweepstakes model and the selective sweepstakes model.** **a**, **b**, Deviations of site frequencies from approximate maximum likelihood best-fit expectations of the neutral $\Xi$-Beta($2-\alpha$,$\alpha$) coalescent modelling random sweepstakes. Deviations of the mean site frequencies of non-inversion chromosomes 3–6, 8–11, and 13–23 estimated with genotype likelihoods GL1 from best fit expectations of the $\Xi$-Beta($2-\alpha$,$\alpha$) coalescent with $\hat{\alpha} = 1.16$ for the South/south-east population (**a**) and with $\hat{\alpha} = 1.16$ for the Þistilfjörður population (**b**). Deficiency of intermediate allele frequency classes and excess mainly at right tail of site frequency spectrum. **c**, **d**, Deviations of GL1 estimated site frequencies from expectations of the Durrett-Schweinsberg model of recurrent selective sweeps for the South/south-east population with a compound parameter $c = 8.25$ and the Þistilfjörður population with a compound parameter $c = 6.3$ respectively. Better fit than random model but also with excess at right tail of site frequency spectrum. Deviations reported as the log of the odds ratio (in blue), the difference of the observed and expected logit of site frequencies. The dashed red line at zero represents the null hypothesis of no difference. The darker and lighter shaded gray areas represent the 95% and the 99% confidence regions of the approximately normally distributed log odds ratio.

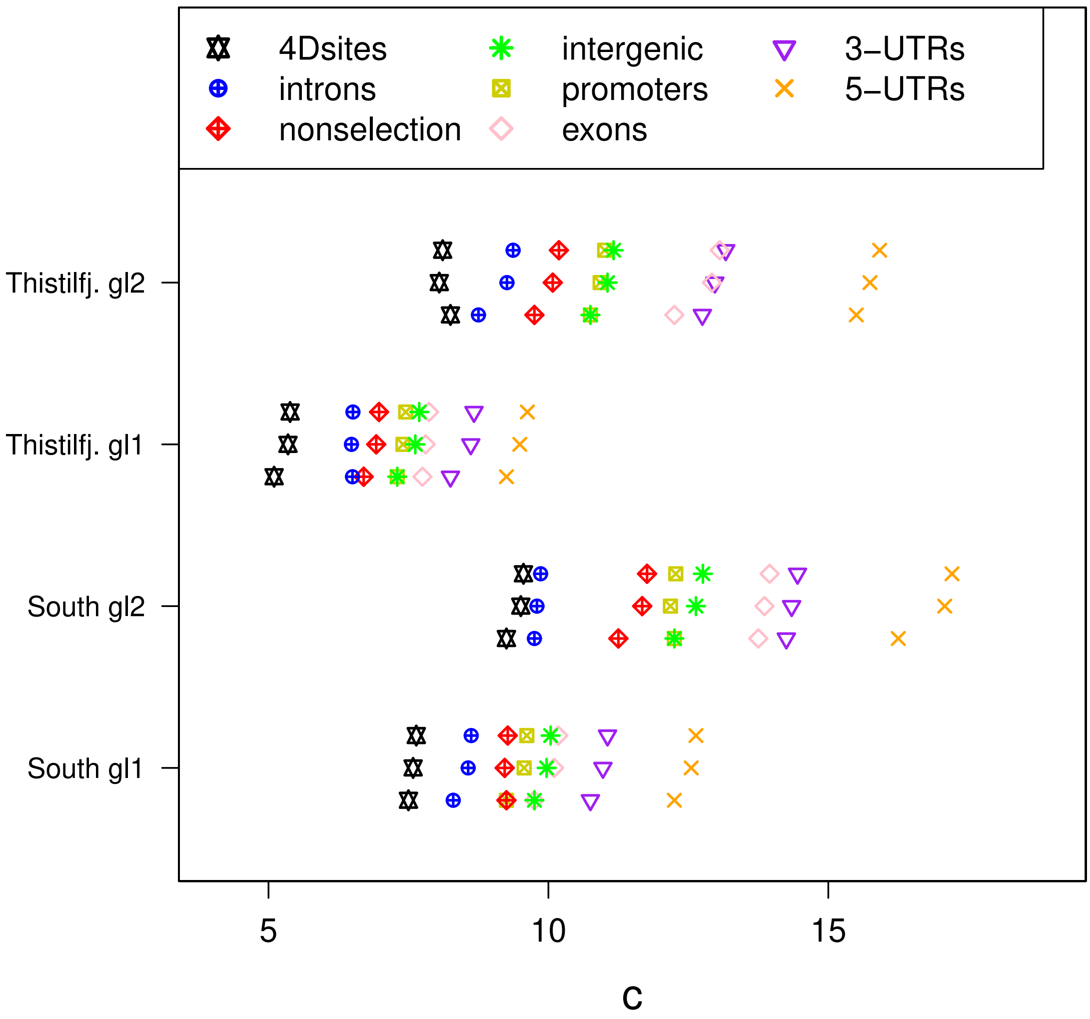


Figure 5: **Approximate Bayesian computation (ABC) estimation of parameters of the Durrett-Schweinsberg coalescent (Durrett and Schweinsberg, 2005) (the selective sweepstakes model) for various functional regions of the genome.** For each category from top to bottom the mean, the median, and the mode of the ABC-posterior distribution of the compound parameter $c \in \Theta_{\mathrm{DS}}$ using SFSs computed from likelihood GL1 and GL2 for the South/south-east and Þistilfjörður populations. The different functional groups are fourfold degenerate sites (4Dsites), intronic sites, non-selection sites (sites more than 500 kb away from peaks of selection scan, SM Fig. **8**), intergenic sites, promoters, exons, $3'$-UTR sites (3-UTRs), and $5'$-UTR sites (5-UTRs), regions ranging from less to more constrained by selection.

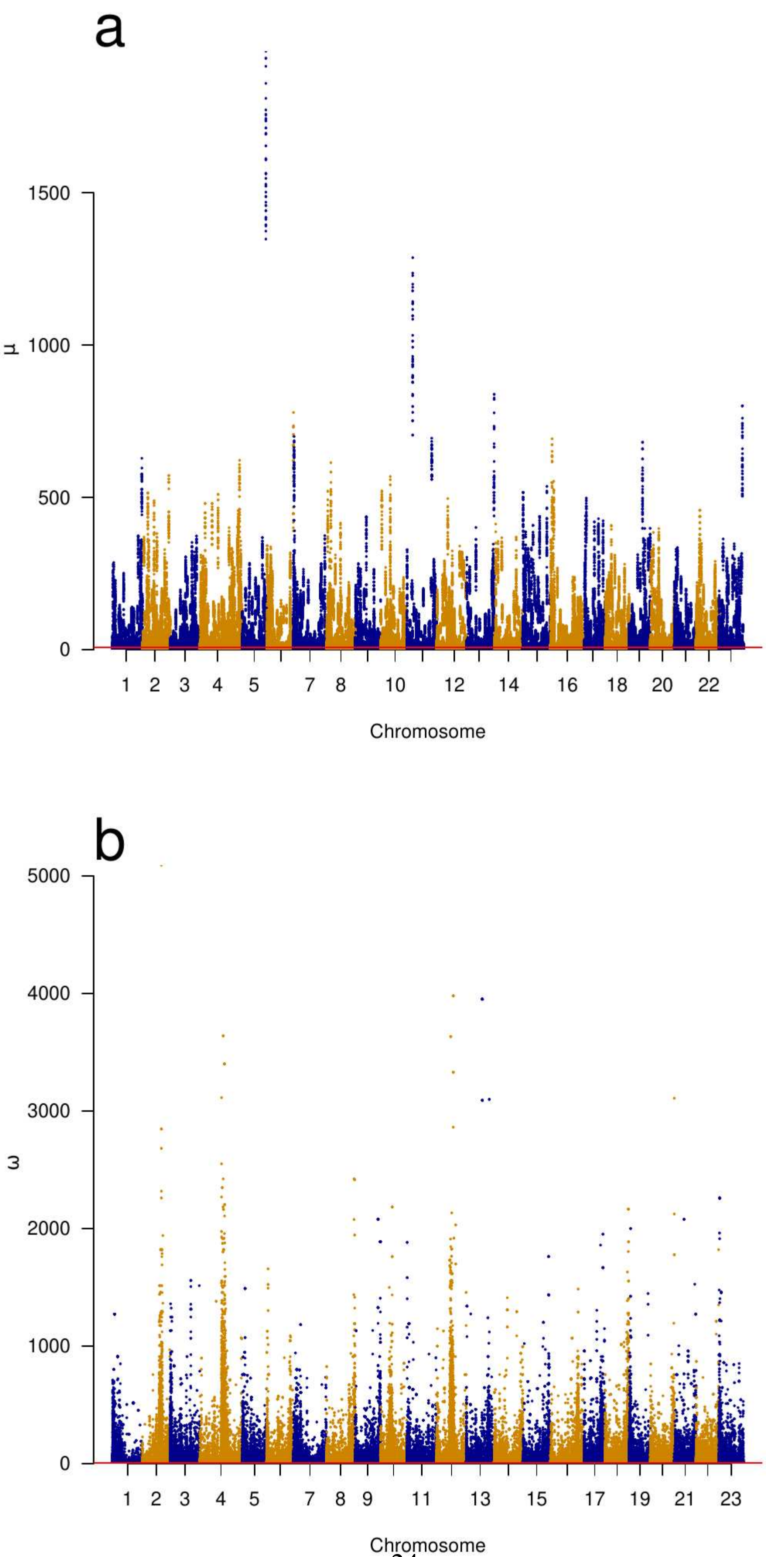




Figure 6: **Genomic scans of selective sweeps by two methods.** **a** Manhattan plots from detection of selective sweeps using RAiSD Alachiotis and Pavlidis (2018) and by using **b** OmegaPlus Alachiotis et al. (2012). Chromosomes with alternating colors. Indications of selective sweeps are found throughout each chromosome.

# Materials and Methods

## Sampling

We randomly sampled adults from our extensive tissue collection (Árnason and Halldórsdóttir, 2015; Halldórsdóttir and Árnason, 2015) from two localities in Iceland, the South/south-east coast ($n = 68$) and Þistilfjörður on the north-east coast ($n = 71$) (SM Fig. **1**). The Icelandic Marine Research collected the fish during spring spawning surveys (Árnason and Halldórsdóttir, 2015). All fish selected here had running gonads (eggs and milt with maturity index 3), indicating that they were spawning at the capture locality.

## Ethics statement

The Icelandic Committee for Welfare of Experimental Animals, Chief Veterinary Officer at the Ministry of Agriculture, Reykjavik, Iceland has determined that the research conducted here is not subject to the laws concerning the Welfare of Experimental Animals (The Icelandic Law on Animal Protection, Law 15/1994, last updated with Law 157/2012). DNA was isolated from tissue taken from dead fish on board research vessels. Fish were collected during the yearly surveys of the Icelandic Marine Research Institute (and other such institutes as already described (Árnason and Halldórsdóttir, 2019)). All research plans and sampling of fish, including the ones for the current project, have been evaluated and approved by the Marine Research Institute Board of Directors, which serves as an ethics board. The Board comprises the Director-General, Deputy Directors for Science and Finance and heads of the Marine Environment Section, the Marine Resources Section, and the Fisheries Advisory Section.

## Molecular analysis

We shipped tissue samples of cod from the South/south-east coast population of Iceland to Omega Bioservices. Omega Bioservices isolated genomic DNA using the E-Z 96 Tissue DNA Kit (Omega Biotek), made picogreen DNA sample quality checks, made sequencing libraries using Kapa Hyper Prep WGS (Kapa Biosystems), used Tapestation (Agilent Technologies) for sizing libraries, and sequenced libraries on a 4000/X Ten Illumina platform with a $2 \times 150$ bp read format, and returned demultiplexed fastq files.

Genomic DNA was isolated from the tissue samples of Þistilfjörður population using the E-Z 96 Tissue DNA Kit (Omega Biotek) according to the manufacturer's recommendation. The DNA was normalized with elution buffer to 10 ng/ul. The normalized DNA was analyzed at the Bauer Core of Harvard University. According to the manufacturer's recommendation, the Bauer Core used the Kapa HyperPrep Plus kit (Kapa Biosystems) for enzymatic DNA fragmentation and adapter ligation, except that the reaction volume was 1/4 of the recommended volume. The target insert size was 350 base pairs (bp) with a resulting average of 487 bp. The libraries were indexed using IDT (Integrated DNA Technologies) unique dual 8 bp indexes for Illumina. The Core uses Tapestation (Agilent Technologies) and Picogreen qPCR for sizing and quality checks. Multiplexed libraries were sequenced on NovaSeq (Illumina) S4 lanes at the Broad Institute with a $2 \times 150$ bp read format, and demultiplexed fastq files were returned.

## Bioinformatic analysis

The sequencing centers returned de-multiplexed `fastq` files for different runs of each individual. Data processing followed the Genome Analysis Toolkit (GATK) best practices (Auwera et al., 2013) as implemented in the `fastq_to_vcf` pipeline of Alison Shultz (`github.com/ajshultz/comp-pop-gen`). Using the pipeline the raw reads were adapter trimmed using `NGmerge` (Gaspar, 2018), the trimmed `fastq` files aligned to the gadMor3.0 chromosome-level reference genome assembly (`NCBIaccessionID:GCF_902167405.1`) using `bwa mem` (Li and Durbin, 2009), and the resulting `bam` files deduplicated, sorted, and indexed with `gatk` (Auwera et al., 2013).

The deduplicated `bam` files were used for population genetic analysis with `ANGSD` (Korneliussen et al., 2014). Outgroup fasta sequences were generated with `-dofasta 3`, which chooses a base using an effective depth algorithm (Wang et al., 2013). A high coverage specimen (Árnason and Halldórsdóttir, 2019) from each of Pacific cod *Gadus macrocephalus* (labeled Gma), walleye pollock, also from the Pacific, *G. chalcogrammus* (labeled Gch), Greenland cod *G. ogac* (labeled Gog), and Arctic cod *Boreogadus saida* (labeled Bsa) were each taken as an outgroup. To estimate site frequency spectra the site allele frequency likelihoods based on genotype likelihoods were estimated using `ANGSD` and polarized with the respective outgroup using the `-anc` flag with `-doSaf 1` and `-doMajorMinor 1` for both genotype likelihoods 1 and 2 (the SAMtools genotype likelihood, `-GL 1` and the GATK genotype likelihood, `-GL 2`). Filtering was done on sequence and mapping quality `-minMapQ 30` `-minQ 20`, indel realignment `-baq 1 -C 50`, quality checks `-remove_bads 1` `-uniqueOnly 1 -only_proper_pairs 1 -skipTriallelic 1`, and finally the minimum number of individuals was set to the sample size (e.g. `-minInd 68`) so that only sites present in all individuals are selected. Errors at very low-coverage sites maybe called as heterozygotes. Similarly, sites with very high-coverage (more than twice or three times the average) may represent alignment issues of duplicated regions such that paralogous sites will be called as heterozygous. We addressed the issues of coverage with two steps. First, we screened out individuals with an average genome-wide coverage less than $10\times$ giving samples sizes of $n = 68$ and $n = 71$ for the South/south-east and the Þistilfjörður populations, respectively. This resulted in an average coverage of $16\times$ and $12\times$ for the South/south-east and the Þistilfjörður populations, respectively. Second, we determined the overall coverage of all sites in the genome that passed the quality filtering. We then used the minimum and maximum of the boxplot statistics ($Q_1 - 1.5\text{IQR}$ and $Q_3 + 1.5\text{IQR}$, which represent roughly $\mu \pm 2.7\sigma$ for a normal distribution) to filter sites using the `ANGSD` flags `-setMinDepth` $Q_1 - 1.5\text{IQR}$ and `-setMaxDepth` $1.5Q_3 + \text{IQR}$ thus removing sites with a boxplot outlier coverage. We did this filtering separately for each chromosome. All our site frequency spectra are estimated using these flags. The site frequency spectra of the full data for each chromosome were then generated with `realSFS` using default flags. Site frequency spectra for genomic regions used the `-sites` flag of `realSFS` with the sample allele frequency files (`saf`) files estimated with the above filtering and was thus based on the same filtering.

We use the logit transformation, the log of the odds ratio $\log(p/(1-p))$, to analyse the site frequency spectra. We transform both the derived allele frequency and the normalized site frequency. Under this transformation, the overall shape of the site frequency spectrum (L-shape, U-shape, V-shape) is invariant.

To investigate divergence among gadid taxa we used `ANGSD` to generate beagle likelihoods (`-GL 1`, `-doGlf 2`) and the quality filtering above. We then used `ngsDist` (Vieira et al., 2015) to estimate the $p$-distance as nucleotide substitutions per nucleotide site between Atlantic cod and walleye pollock. The number of sites (`-n_sites`) was set to the number of variable sites and the total number of sites (`-tot_sites`) was set equal to the number of sites that passed the quality filtering in the estimation of the site frequency spectra above (SM Table **6**). A tree (SM Fig. **21**) was generated with `fastME` (Lefort et al., 2015) and displayed using `ggtree` (Yu et al., 2016).

To evaluate deviations from neutrality, we used `ANGSD` to estimate the neutrality test statistics Tajima's $D$ (Tajima, 1989), Fu and Li's $D$ (Fu and Li, 1993), Fay and Wu's $H$ (Fay and Wu, 2000), and Zeng's $E$ (Zeng et al., 2006) in sliding windows (window size 100 kb with 20 kb step size).

We generated `vcf` files for the South/south-east population using `GATK` (Auwera et al., 2013). We used the genomic features files (`gtf`) of the Gadmor3 assembly to extract sites belonging to different functional groups. We used `ReSeqTools` (He et al., 2013) to extract fourfold degenerate sites, `bedtools` (Quinlan and Hall, 2010) to extract exon and intron sites using genomic feature files (`gtf`), and we used the `GenomicFeatures Bioconductor` package (Lawrence et al., 2013) for extracting other functional regions. We then used the `-sites` flag of `realSFS` to estimate site frequency spectra from the sample allele frequency (`saf`) files of the entire data for each chromosome, thus keeping the quality and coverage filtering applied to the full data (**SM** ). We used `PopLDdecay` (Zhang et al., 2018) to estimate the decay of linkage disequilibrium. To perform the McDonald-Kreitman test of selection (McDonald and Kreitman, 1991) we used `SnpEff` (Cingolani et al., 2012) to estimate the number of polymorphic non-synonymous and synonymous ($P_n$ and $P_s$) sites of protein-coding genes. To estimate the number of fixed non-synonymous and synonymous ($D_n$ and $D_s$) sites, we used a single individual with the highest coverage (32×) from the South/south-east population and a single high coverage (31×) Pacific cod individual and counted homozygous sites. We used the neutrality index $NI = (P_n/P_s)/)D_n/D_s)$ (Rand and Kann, 1996) transformed as $-\log NI$ as an index of selection with negative values implying negative (purifying and background) selection and positive values implying positive selection (selective sweeps).

We did a principal components (PC) based scan of selection using `PCangsd` (Meisner and

Albrechtsen, 2018) (`python pcangsd.py -selection`). We then removed regions of 500 kb on either side of selective peaks that exceeded $\log_{10} p \geq 4$ (SM Fig. **8**) to define regions of no selection that we compared with other genomic regions (e.g. Fig. 5).

We used `OmegaPlus` (Alachiotis et al., 2012) and `RAiSD` (Alachiotis and Pavlidis, 2018) scanning for selective sweeps genome-wide. Both methods use local linkage disequilibrium to detect sweeps (Nielsen, 2005) and in addition `RAiSD` uses a local reduction in levels of polymorphism and shifts in the frequencies of low- and high-frequency derived alleles affecting respectively the left and righ tails of the site frequency spectrum.

## Methods for analyzing coalescent models

This section describes the model fitting procedure we used for each family of models discussed in SM 1.3. Where possible, we have resorted to documented state-of-the-art simulators and inference packages, though that was not possible in all cases, particularly for the Durrett-Schweinsberg model. A description of various terms is given in SM Table **7**. All custom code has been made available via GitHub, with links below.

### Kingman coalescent

There are numerous, well-documented packages for inferring population size profiles from whole-genome data under the Kingman coalescent, typically relying on the sequentially Markovian coalescent approximation (McVean and Cardin, 2005). We used `scm++` (`https://github.com/popgenmethods/smcpp`) (Terhorst et al., 2016) to produce best-fit profiles. We also used the stairway plot (`https://github.com/xiaoming-liu/stairway-plot-v2`) (Liu and Fu, 2015, 2020) that uses the site frequency spectra for a model-flexible demographic inference. Both packages were installed according to their respective documentations, and run using default settings. To treat runs of homozygosity, which may represent centromeric regions, as missing, we set the flag `-missing-cutoff 10` in `smc++` runs.

### $\Xi$-Beta$(2-\alpha, \alpha)$ coalescent

At the time of writing there are no off-the-shelf inference packages capable of estimating $\alpha$ or a population size profile from whole-genome data under the $\Xi$-Beta$(2-\alpha, \alpha)$ coalescent. However, synthetic data from the model can be simulated using `msprime` (Kelleher et al., 2016). Hence, we fit our model using approximate Bayesian computation (ABC), in which model fitting is accomplished by comparing summary statistics of simulated and observed data under various parameters.

We used the logit transform of the normalized site frequency spectrum (SFS) as our summary statistic. The `msprime` package is not well-optimized for simulating multiple chromosomes, so we used chromosome 4 as our observed data. To simulate observations, we set the chromosome length to 3.5 megabases, and used respective per-site per-generation mutation and recombination probabilities of $10^{-7}$ and $10^{-8}$ respectively.

A proposed parameter combination was accepted whenever the simulated statistic was within a specified tolerance of the observed statistic. To avoid tuning the tolerance and other hyperparameters, and to focus computational effort on regions of $\Theta_B$ of good model fit automatically, we used the adaptive ABC-MCMC method of (Vihola and Franks, 2020) with a target acceptance rate of 10%, which the authors recommend.

### Durrett-Schweinsberg coalescent

To our knowledge, there are no off-the-shelf inference packages for the Durrett-Schweinsberg model, and also no packages for simulating it. Hence we implemented a basic, single locus simulator in `C++`, based on the exact rejection sampling mechanism which is used in both the `msprime` and `Beta-Xi-Sim` simulation packages (see the Appendix in (Koskela, 2018)). Since the Durrett-Schweinsberg coalescent is a single locus model, we computed an observed site frequency spectra separately for 25kb lengths of genome separated by 500kb gaps. This was done across all 19 non-inversion chromosomes, and the mean of the resulting ensemble was taken to be the observed SFS. Simulated values were calculated as the mean of 10,000 independent, single-locus replicates. This number was found to be high enough in trial runs to avoid zero entries in the averaged SFS, and hence infinite values in the logit transform.

Then we used the same ABC-MCMC pipeline outlined above for the $\Xi$-Beta$(2-\alpha, \alpha)$ coalescent to infer an approximate posterior distribution of values for the compound parameter $c$ of the Durrett-Schweinsberg model.

## Computations

The computations in this paper were run on the Odyssey cluster supported by the FAS Division of Science, Research Computing Group at Harvard University. Some computations were run on the Mimir bioinformatics server and the Assa bioinformatics computer at the University of Iceland.

# Acknowledgments

We thank John Wakeley and W. Stewart Grant for comments on the manuscript and Kristján Kristinsson and the Icelandic Marine Research Institute for help in sampling. **Funding:** The work was supported by an Icelandic Research Fund Grant of Excellence no. 185151-051 to EÁ, KH, Alison Etheridge, Wolfgang Stephan, and BE. BE also acknowledges financial support by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) Project number 273887127 through DFG grant STE 325/17 to Wolfgang Stephan through DFG Priority Program (SPP) 1819: Rapid evolutionary adaptation, a DFG SPP1819 start-up module grant to JK, Maite Wilke Berenguer, and BE, and JK acknowledges financial support from Engineering and Physical Sciences Research Council (EPSRC) grants EP/R044732/1 and EP/V049208/1. **Author contributions:** Design of the experiment: all authors. Molecular work, data acquisition, bioinformatic work, estimation of the site-frequency spectra: EÁ and KH. Data analysis: all authors. Writing manuscript: all authors. **Competing interests:** The authors declare that they have no competing interests.

## Data and materials availability

All data needed to evaluate the conclusions of the paper are presented in the paper and/or the supplementary materials. The `bam` files of the whole genome sequencing of each individual aligned to the Gadmor3 reference genome (NCBI accession ID: GCF_902167405.1) are available from the NCBI SRA Sequence Read Archive under accession number BioProject ID: PRJNA663624 at time of publication.

For the purpose of open access, the authors have applied a Creative Commons Attribution (CC BY) licence to any Author Accepted Manuscript version arising from this submission.

## Code availability

Simulations of background selection were done with `SLiM 3` (Haller and Messer, 2019) available at `https://messerlab.org/slim/`. Estimates of population size histories for the Kingman coalescent were produced using the `stairwayplot` (Liu and Fu, 2015; Liu, 2020) and `smc++` (Terhorst et al., 2016) available via Github at `https://github.com/xiaoming-liu/stairway-plot-v2` and `https://github.com/popgenmethods/smcpp` respectively. Based on the estimated population size histories site frequency spectra under the Kingman and the $\Xi$-Beta$(2-\alpha,\alpha)$ coalescents were simulated using `msprime`, available via GitHub at `https://github.com/tskit-dev/msprime`, with documentation at `https://tskit.dev/msprime/docs/stable/`. Our msprime simulations also make use of the tskit library, available via GitHub at `https://github.com/tskit-dev/tskit`, with documentation at `https://tskit.dev/tskit/docs/`. To our knowledge, no prior implementation of the Durrett-Schweinsberg coalescent is available. Hence, we wrote a simulator, which is available via GitHub at `https://github.com/JereKoskela/ds-tree`. This repository also contains documentation of the Durrett-Schweinsberg implementation, as well as Python and shell scripts for the i) ABC pipelines we used to conduct model fitting for both the $\Xi$-Beta$(2-\alpha,\alpha)$ and Durrett-Schweinsberg coalescents, and ii) the simulation pipelines for sampling site frequency spectra under the best-fit Kingman, $\Xi$-Beta$(2-\alpha,\alpha)$ , and Durrett-Schweinsberg coalescents. C++

code and python scripts implementing the sampling schemes described in `https://github.com/eldonb/coalescents`. C code using recursions (Birkner et al., 2013a) for computing the exact expected branch length spectrum for Examples 2.3 and 2.4 of the Durrett-Schweinsberg model (Durrett and Schweinsberg, 2005) is available at `https://github.com/eldonb/Durrett_Schweinsberg_Expected_SFS`.

# References


N. Alachiotis, A. Stamatakis, and P. Pavlidis. OmegaPlus: a scalable tool for rapid detection of selective sweeps in whole-genome datasets. *Bioinformatics*, 28(17):2274–2275, July 2012. doi: 10.1093/bioinformatics/bts419. URL `https://doi.org/10.1093/bioinformatics/bts419`.

Nikolaos Alachiotis and Pavlos Pavlidis. RAiSD detects positive selection based on multiple signatures of a selective sweep and SNP vectors. *Communications Biology*, 1(1), June 2018. doi: 10.1038/s42003-018-0085-8. URL `https://doi.org/10.1038/s42003-018-0085-8`.

Einar Árnason. Mitochondrial cytochrome *b* DNA variation in the high fecundity Atlantic cod: Trans-Alantic clines and shallow gene-genealogy. *Genetics*, 166:1871–1885, 2004. doi: 10.1534/genetics.166.4.1871. URL `https://doi.org/10.1534/genetics.166.4.1871`.

Einar Árnason and Katrín Halldórsdóttir. Nucleotide variation and balancing selection at the *Ckma* gene in Atlantic cod: analysis with multiple merger coalescent models. *PeerJ*, 3:e786, 2015. doi: 10.7717/peerj.786. URL `http://dx.doi.org/10.7717/peerj.786`.

Einar Árnason and Katrín Halldórsdóttir. Codweb: Whole-genome sequencing uncovers extensive reticulations fueling adaptation among Atlantic, Arctic, and Pacific gadids. *Science Advances*, 5(3): eaat8788, March 2019. doi: 10.1126/sciadv.aat8788. URL `https://doi.org/10.1126/sciadv.aat8788`.

Geraldine A. Auwera, Mauricio O. Carneiro, Christopher Hartl, Ryan Poplin, Guillermo del Angel, Ami Levy-Moonshine, Tadeusz Jordan, Khalid Shakir, David Roazen, Joel Thibault, Eric Banks, Kiran V. Garimella, David Altshuler, Stacey Gabriel, and Mark A. DePristo. From FastQ data to high-confidence variant calls: The genome analysis toolkit best practices pipeline. *Current Protocols in Bioinformatics*, 43(1), October 2013. doi: 10.1002/0471250953.bi1110s43. URL `https://doi.org/10.1002/0471250953.bi1110s43`.

N. H. Barton. The effect of hitch-hiking on neutral genealogies. *Genetical Research*, 72(2):123–133, October 1998. doi: 10.1017/s0016672398003462. URL `https://doi.org/10.1017/s0016672398003462`.

Franz Baumdicker, Gertjan Bisschop, Daniel Goldstein, Graham Gower, Aaron P Ragsdale, Georgia Tsambos, Sha Zhu, Bjarki Eldon, E Castedo Ellerman, Jared G Galloway, Ariella L Gladstein, Gregor Gorjanc, Bing Guo, Ben Jeffery, Warren W Kretzschmar, Konrad Lohse, Michael Matschiner, Dominic Nelson, Nathaniel S Pope, Consuelo D Quinto-Cortés, Murillo F Rodrigues, Kumar Saunack, Thibaut Sellinger, Kevin Thornton, Hugo van Kemenade, Anthony W Wohns, Yan Wong, Simon Gravel, Andrew D Kern, Jere Koskela, Peter L Ralph, and Jerome Kelleher. Efficient ancestry and mutation simulation with msprime 1.0. *Genetics*, December 2021. doi: 10.1093/genetics/iyab229. URL `https://doi.org/10.1093/genetics/iyab229`.

Paul R. Berg, Bastiaan Star, Christophe Pampoulie, Marte Sodeland, Julia M. I. Barth, Halvor Knutsen, Kjetill S. Jakobsen, and Sissel Jentoft. Three chromosomal rearrangements promote genomic divergence between migratory and stationary ecotypes of Atlantic cod. *Scientific Reports*, 6:23246, mar 2016. doi: 10.1038/srep23246. URL `http://dx.doi.org/10.1038/srep23246`.

M Birkner, J Blath, and M Steinrücken. Analysis of DNA sequence variation within marine species using Beta-coalescents. *Theoretical Population Biology*, 87:15–24, 2013a. doi: 10.1016/j.tpb.2013.01.007. URL `https://doi.org/10.1016/j.tpb.2013.01.007`.

Matthias Birkner, Jochen Blath, and Bjarki Eldon. An ancestral recombination graph for diploid populations with skewed offspring distribution. *Genetics*, 193(1):255–290, January 2013b. doi: 10.1534/genetics.112.144329. URL `https://doi.org/10.1534/genetics.112.144329`.

Matthias Birkner, Huili Liu, and Anja Sturm. Coalescent results for diploid exchangeable population models. *Electronic Journal of Probability*, 23(0), 2018. doi: 10.1214/18-ejp175. URL `https://doi.org/10.1214/18-ejp175`.

Jochen Blath, Mathias Christensen Cronjäger, Bjarki Eldon, and Matthias Hammer. The site-frequency spectrum associated with $\xi$-coalescents. *Theoretical Population Biology*, 110:36–50, 2016. doi: 10.1016/j.tpb.2016.04.002. URL `https://doi.org/10.1016/j.tpb.2016.04.002`.

Brian Charlesworth. Why we are not dead one hundred times over. *Evolution*, 67(11):3354–3361, July 2013. doi: 10.1111/evo.12195. URL `https://doi.org/10.1111/evo.12195`.

Deborah Charlesworth. Balancing selection and its effects on sequences in nearby genome regions. *PLoS Genetics*, 2(4):e64, April 2006. doi: 10.1371/journal.pgen.0020064. URL `https://doi.org/10.1371/journal.pgen.0020064`.

P. Cingolani, A. Platts, M. Coon, T. Nguyen, L. Wang, S.J. Land, X. Lu, and D.M. Ruden. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain $w^{1118}$; *iso-2*; *iso-3*. *Fly*, 6(2):80–92, 2012.

Graham Coop and Peter Ralph. Patterns of neutral diversity under general models of selective sweeps. *Genetics*, 192(1):205–224, September 2012. doi: 10.1534/genetics.112.141861. URL https://doi.org/10.1534/genetics.112.141861.

D. H. Cushing. The regularity of the spawning season of some fishes. *ICES Journal of Marine Science*, 33(1):81–92, November 1969. doi: 10.1093/icesjms/33.1.81. URL https://doi.org/10.1093/icesjms/33.1.81.

Ivana Cvijović, Benjamin H Good, and Michael M Desai. The effect of strong purifying selection on genetic diversity. *Genetics*, 209(4):1235–1278, May 2018. doi: 10.1534/genetics.118.301058. URL https://doi.org/10.1534/genetics.118.301058.

Robert R Dickson, Jens Meincke, Svend-Aage Malmberg, and Arthur J Lee. The "great salinity anomaly" in the Northern North Atlantic 1968–1982. *Progress in Oceanography*, 20(2):103–151, January 1988. doi: 10.1016/0079-6611(88)90049-3. URL https://doi.org/10.1016/0079-6611(88)90049-3.

P Donnelly and T G Kurtz. Particle representations for measure-valued population models. *Annals of Probability*, 27:166–205, 1999. doi: 10.1214/aop/1022677258. URL http://doi.org/10.1214/aop/1022677258.

Rick Durrett and Jason Schweinsberg. A coalescent model for the effect of advantageous mutations on the genealogy of a population. *Stochastic Processes and their Applications*, 115(10):1628–1657, October 2005. doi: 10.1016/j.spa.2005.04.009. URL https://doi.org/10.1016/j.spa.2005.04.009.

B. Eldon, M. Birkner, J. Blath, and F. Freund. Can the site-frequency spectrum distinguish exponential population growth from multiple-merger coalescents? *Genetics*, 199(3):841–856, jan 2015. doi: 10.1534/genetics.114.173807. URL http://dx.doi.org/10.1534/genetics.114.173807.

Bjarki Eldon. Evolutionary genomics of high fecundity. *Annual Review of Genetics*, 54(1):213–236, November 2020. doi: 10.1146/annurev-genet-021920-095932. URL https://doi.org/10.1146/annurev-genet-021920-095932.

Bjarki Eldon and John Wakeley. Coalescent processes when the distribution of offspring number among individuals is highly skewed. *Genetics*, 172(4):2621–2633, February 2006. doi: 10.1534/genetics.105.052175. URL https://doi.org/10.1534/genetics.105.052175.

Akira Endo, Sam Abbott, Adam J. Kucharski, and Sebastian Funk and. Estimating the overdispersion in COVID-19 transmission using outbreak sizes outside China. *Wellcome Open Research*, 5:67, July

2020. doi: 10.12688/wellcomeopenres.15842.3. URL https://doi.org/10.12688/wellcomeopenres.15842.3.

Rui Faria, Kerstin Johannesson, Roger K. Butlin, and Anja M. Westram. Evolving inversions. *Trends in Ecology & Evolution*, 34(3):239–248, March 2019. doi: 10.1016/j.tree.2018.12.005. URL https://doi.org/10.1016/j.tree.2018.12.005.

Justin C Fay and Chung-I Wu. Hitchhiking under positive Darwinian selection. *Genetics*, 155(3): 1405–1413, July 2000. doi: 10.1093/genetics/155.3.1405. URL https://doi.org/10.1093/genetics/155.3.1405.

Joseph Felsenstein. On the biological significance of the cost of gene substitution. *The American Naturalist*, 105(941):1–11, January 1971. doi: 10.1086/282698. URL https://doi.org/10.1086/282698.

R. A. Fisher. *The Genetical Theory of Natural Selection*. Clarendon Press, Oxford, 1930.

Fabian Freund. Cannings models, population size changes and multiple-merger coalescents. *Journal of Mathematical Biology*, 80(5):1497–1521, February 2020. doi: 10.1007/s00285-020-01470-5. URL https://doi.org/10.1007/s00285-020-01470-5.

Fabian Freund, Elise Kerdoncuff, Sebastian Matuszewski, Marguerite Lapierre, Marcel Hildebrandt, Jeffrey D. Jensen, Luca Ferretti, Amaury Lambert, Timothy B. Sackton, and Guillaume Achaz. Interpreting the pervasive observation of U-shaped site frequency spectra. *BioRxiv*, April 2022. doi: 10.1101/2022.04.12.488084. URL https://doi.org/10.1101/2022.04.12.488084.

Y X Fu and W H Li. Statistical tests of neutrality of mutations. *Genetics*, 133(3):693–709, March 1993. doi: 10.1093/genetics/133.3.693. URL https://doi.org/10.1093/genetics/133.3.693.

Nicolas Galtier, Frantz Depaulis, and Nicholas H Barton. Detecting bottlenecks and selective sweeps from DNA sequence polymorphism. *Genetics*, 155(2):981–987, June 2000. doi: 10.1093/genetics/155.2.981. URL https://doi.org/10.1093/genetics/155.2.981.

John M. Gaspar. NGmerge: Merging paired-end reads via novel empirically-derived models of sequencing errors. *BMC Bioinformatics*, 19(1), December 2018. doi: 10.1186/s12859-018-2579-2. URL https://doi.org/10.1186/s12859-018-2579-2.

Peter R. Grant and B. Rosemary Grant. *40 Years of Evolution: Darwin's Finches on Daphne Major Island*. Princeton University Press, Princeton, NJ, 2014.

John Burton Sanderson Haldane. The cost of natural selection. *Journal of Genetics*, 55:511–524, 1957.

Katrín Halldórsdóttir and Einar Árnason. Whole-genome sequencing uncovers cryptic and hybrid species among Atlantic and Pacific cod-fish. *bioRxiv*, 2015. doi: 10.1101/034926. URL `https://dx.doi.org/10.1101/034926`.

Benjamin C Haller and Philipp W Messer. SLiM 3: Forward genetic simulations beyond the wright–fisher model. *Molecular Biology and Evolution*, 36(3):632–637, January 2019. doi: 10.1093/molbev/msy228. URL `https://doi.org/10.1093/molbev/msy228`.

W. He, S. Zhao, X. Liu, S. Dong, J. Lv, D. Liu, J. Wang, and Z. Meng. ReSeqTools: an integrated toolkit for large-scale next-generation sequencing based resequencing analysis. *Genetics and Molecular Research*, 12(4):6275–6283, 2013. doi: 10.4238/2013.december.4.15. URL `https://doi.org/10.4238/2013.december.4.15`.

Dennis Hedgecock. Does variance in reproductive success limit effective population size of marine organisms? In A Beaumont, editor, *Genetics and Evolution of Aquatic Organisms*, pages 122–134. Chapman & Hall, London, 1994.

Dennis Hedgecock and Alexander I Pudovkin. Sweepstakes reproductive success in highly fecund marine fish and shellfish: A review and commentary. *Bulletin of Marine Science*, 87(4):971–1002, October 2011. doi: 10.5343/bms.2010.1051. URL `https://doi.org/10.5343/bms.2010.1051`.

G. M. Hewitt. Genetic consequences of climatic oscillations in the Quaternary. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, 359(1442):183–195, February 2004. doi: 10.1098/rstb.2003.1388. URL `https://doi.org/10.1098/rstb.2003.1388`.

Jeffrey A. Hutchings. Collapse and recovery of marine fishes. *Nature*, 406(6798):882–885, August 2000. doi: 10.1038/35022565. URL `https://doi.org/10.1038/35022565`.

K K Irwin, S Laurent, S Matuszewski, S Vuilleumier, L Ormond, H Shim, C Bank, and J D Jensen. On the importance of skewed offspring distributions and background selection in virus population genetics. *Heredity*, 117(6):393–399, September 2016. doi: 10.1038/hdy.2016.58. URL `https://doi.org/10.1038/hdy.2016.58`.

Paul Jay, Mathieu Chouteau, Annabel Whibley, Héloïse Bastide, Hugues Parrinello, Violaine Llaurens, and Mathieu Joron. Mutation load at a mimicry supergene sheds new light on the evolution of inversion polymorphisms. *Nature Genetics*, 53(3):288–293, January 2021. doi: 10.1038/s41588-020-00771-1. URL `https://doi.org/10.1038/s41588-020-00771-1`.

Mamoru Kato, Daniel A. Vasco, Ryuichi Sugino, Daichi Narushima, and Alexander Krasnitz. Sweepstake evolution revealed by population-genetic analysis of copy-number alterations in single genomes of breast cancer. *Royal Society Open Science*, 4(9):171060, September 2017. doi: 10.1098/rsos.171060. URL https://doi.org/10.1098/rsos.171060.

Jerome Kelleher, Alison M Etheridge, and Gilean McVean. Efficient coalescent simulation and genealogical analysis for large sample sizes. *PLoS Computational Biology*, 12(5):1–22, 05 2016. doi: 10.1371/journal.pcbi.1004842. URL http://doi.org/10.1371/journal.pcbi.1004842.

Yuseob Kim. Allele frequency distribution under recurrent selective sweeps. *Genetics*, 172(3): 1967–1978, March 2006. doi: 10.1534/genetics.105.048447. URL https://doi.org/10.1534/genetics.105.048447.

Yuseob Kim and Wolfgang Stephan. Selective sweeps in the presence of interference among partially linked loci. *Genetics*, 164(1):389–398, 2003.

Motoo Kimura. Evolutionary rate at the molecular level. *Nature*, 217(5129):624–626, February 1968. doi: 10.1038/217624a0. URL https://doi.org/10.1038/217624a0.

Jack Lester King. Continuously distributed factors affecting fitness. *Genetics*, 55(3):483–492, March 1967. doi: 10.1093/genetics/55.3.483. URL https://doi.org/10.1093/genetics/55.3.483.

J.F.C. Kingman. The coalescent. *Stochastic Processes and their Applications*, 13(3):235–248, September 1982. doi: 10.1016/0304-4149(82)90011-4. URL https://doi.org/10.1016/0304-4149(82)90011-4.

Tina Graceline Kirubakaran, Harald Grove, Matthew P. Kent, Simen R. Sandve, Matthew Baranski, Torfinn Nome, Maria Cristina De Rosa, Benedetta Righino, Torild Johansen, Håkon Otterå, Anna Sonesson, Sigbjørn Lien, and Øivind Andersen. Two adjacent inversions maintain genomic differentiation between migratory and stationary ecotypes of Atlantic cod. *Molecular Ecology*, 2016: 2130–2143, feb 2016. doi: 10.1111/mec.13592. URL http://dx.doi.org/10.1111/mec.13592.

Thorfinn Sand Korneliussen, Anders Albrechtsen, and Rasmus Nielsen. ANGSD: Analysis of next generation sequencing data. *BMC Bioinformatics*, 15(1), nov 2014. doi: 10.1186/s12859-014-0356-4. URL http://dx.doi.org/10.1186/s12859-014-0356-4.

Jere Koskela. Multi-locus data distinguishes between population growth and multiple merger coalescents. *Statistical Applications in Genetics and Molecular Biology*, 17(3), June 2018. doi: 10.1515/sagmb-2017-0011. URL https://doi.org/10.1515/sagmb-2017-0011.

Jere Koskela and Maite Wilke Berenguer. Robust model selection between population growth and multiple merger coalescents. *Mathematical Biosciences*, 311:1–12, May 2019. doi: 10.1016/j.mbs.2019.03.004. URL https://doi.org/10.1016/j.mbs.2019.03.004.

Michael Lawrence, Wolfgang Huber, Hervé Pagès, Patrick Aboyoun, Marc Carlson, Robert Gentleman, Martin T. Morgan, and Vincent J. Carey. Software for computing and annotating genomic ranges. *PLoS Computational Biology*, 9(8):e1003118, August 2013. doi: 10.1371/journal.pcbi.1003118. URL https://doi.org/10.1371/journal.pcbi.1003118.

Vincent Lefort, Richard Desper, and Olivier Gascuel. FastME 2.0: Comprehensive, accurate, and fast distance-based phylogeny inference program: Table 1. *Molecular Biology and Evolution*, 32(10): 2798–2800, June 2015. doi: 10.1093/molbev/msv150. URL https://doi.org/10.1093/molbev/msv150.

H. Li and R. Durbin. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, 25(14):1754–1760, may 2009. doi: 10.1093/bioinformatics/btp324. URL http://dx.doi.org/10.1093/bioinformatics/btp324.

Yong Fuga Li, James C. Costello, Alisha K. Holloway, and Matthew W. Hahn. "reverse ecology" and the power of population genomics. *Evolution*, 62(12):2984–2994, December 2008. doi: 10.1111/j.1558-5646.2008.00486.x. URL https://doi.org/10.1111/j.1558-5646.2008.00486.x.

Xiaoming Liu and Yun-Xin Fu. Exploring population size changes using SNP frequency spectra. *Nature Genetics*, 47(5):555–559, April 2015. doi: 10.1038/ng.3254. URL https://doi.org/10.1038/ng.3254.

Xiaoming Liu and Yun-Xin Fu. Stairway plot 2: Demographic history inference with folded SNP frequency spectra. *Genome Biology*, 21(1), November 2020. doi: 10.1186/s13059-020-02196-9. URL https://doi.org/10.1186/s13059-020-02196-9.

Sebastian Matuszewski, Marcel E. Hildebrandt, Guillaume Achaz, and Jeffrey D. Jensen. Coalescent processes with skewed offspring distributions and nonequilibrium demography. *Genetics*, 208(1): 323–338, November 2017. doi: 10.1534/genetics.117.300499. URL https://doi.org/10.1534/genetics.117.300499.

John H. McDonald and Martin Kreitman. Adaptive protein evolution at the *Adh* locus in *Drosophila*. *Nature*, 351(6328):652–654, June 1991. doi: 10.1038/351652a0. URL https://doi.org/10.1038/351652a0.

Gilean A.T McVean and Niall J Cardin. Approximating the coalescent with recombination. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 360(1459):1387–1393, July 2005. doi: 10.1098/rstb.2005.1673. URL `https://doi.org/10.1098/rstb.2005.1673`.

Jonas Meisner and Anders Albrechtsen. Inferring population structure and admixture proportions in low depth NGS data. *Genetics*, 2018. ISSN 0016-6731. doi: 10.1534/genetics.118.301336. URL `http://www.genetics.org/content/early/2018/08/21/genetics.118.301336`.

F. Menardo, S. Gagneux, and F. Freund. Multiple merger genealogies in outbreaks of Mycobacterium tuberculosis. *BioRxive*, December 2019. doi: 10.1101/2019.12.21.885723. URL `https://doi.org/10.1101/2019.12.21.885723`.

M Möhle and S Sagitov. Coalescent patterns in diploid exchangeable population models. *Journal of Mathematical Biology*, 47:337–352, 2003. doi: 10.1007/s00285-003-0218-6. URL `http://doi.org/10.1007/s00285-003-0218-6`.

R. A. Neher and O. Hallatschek. Genealogies of rapidly adapting populations. *Proceedings of the National Academy of Sciences*, 110(2):437–442, December 2013. doi: 10.1073/pnas.1213113110. URL `https://doi.org/10.1073/pnas.1213113110`.

Richard A. Neher. Genetic draft, selective interference, and population genetics of rapid adaptation. *Annual Review of Ecology, Evolution, and Systematics*, 44(1):195–215, November 2013. doi: 10.1146/annurev-ecolsys-110512-135920. URL `https://doi.org/10.1146/annurev-ecolsys-110512-135920`.

Masatoshi Nei, Takeo Maruyama, and Ranajit Chakraborty. The bottleneck effect and genetic variability in populations. *Evolution*, 29(1):1–10, March 1975. doi: 10.1111/j.1558-5646.1975.tb00807.x. URL `https://doi.org/10.1111/j.1558-5646.1975.tb00807.x`.

Rasmus Nielsen. Molecular signatures of natural selection. *Annual Review of Genetics*, 39(1):197–218, December 2005. doi: 10.1146/annurev.genet.39.073003.112420. URL `https://doi.org/10.1146/annurev.genet.39.073003.112420`.

Hiro-Sato Niwa, Kazuya Nashida, and Takashi Yanagimoto. Reproductive skew in Japanese sardine inferred from DNA sequences. *ICES Journal of Marine Science*, 2016:fsw070, may 2016. doi: 10.1093/icesjms/fsw070. URL `http://dx.doi.org/10.1093/icesjms/fsw070`.

Leonard Nunney. The influence of variation in female fecundity on effective population size. *Biological Journal of the Linnean Society*, 59(4):411–425, December 1996. doi: 10.1111/j.1095-8312.1996.tb01474.x. URL `https://doi.org/10.1111/j.1095-8312.1996.tb01474.x`.

David Pimentel, Lori Lach, Rodolfo Zuniga, and Doug Morrison. Environmental and economic costs of nonindigenous species in the United States. *BioScience*, 50(1):53, 2000. doi: 10.1641/0006-3568(2000)050[0053:eaecon]2.3.co;2. URL `http://doi.org/10.1641/0006-3568(2000)050[0053:eaecon]2.3.co;2`.

J Pitman. Coalescents with multiple collisions. *Annals of Probability*, 27:1870–1902, 1999. doi: 10.1214/aop/1022874819. URL `https://doi.org/10.1214/aop/1022874819`.

Fanny Pouyet, Simon Aeschbacher, Alexandre Thiéry, and Laurent Excoffier. Background selection and biased gene conversion affect more than 95% of the human genome and bias demographic inferences. *eLife*, 7:e36317, aug 2018. ISSN 2050-084X. doi: 10.7554/eLife.36317. URL `https://doi.org/10.7554/eLife.36317`.

Molly Przeworski. The signature of positive selection at randomly chosen loci. *Genetics*, 160(3): 1179–1189, 2002. ISSN 0016-6731. URL `https://www.genetics.org/content/160/3/1179`.

Aaron R. Quinlan and Ira M. Hall. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, 26(6):841–842, January 2010. doi: 10.1093/bioinformatics/btq033. URL `https://doi.org/10.1093/bioinformatics/btq033`.

D. M. Rand and L. M. Kann. Excess amino acid polymorphism in mitochondrial DNA: contrasts among genes from Drosophila, mice, and humans. *Molecular Biology and Evolution*, 13(6):735–748, July 1996. doi: 10.1093/oxfordjournals.molbev.a025634. URL `https://doi.org/10.1093/oxfordjournals.molbev.a025634`.

David Reznick. Hard and soft selection revisited: How evolution by natural selection works in the real world. *Journal of Heredity*, 107(1):3–14, January 2016. doi: 10.1093/jhered/esv076. URL `https://doi.org/10.1093/jhered/esv076`.

Andrew M. Sackman, Rebecca B. Harris, and Jeffrey D. Jensen. Inferring demography and selection in organisms characterized by skewed offspring distributions. *Genetics*, 211(3):1019–1028, January 2019. doi: 10.1534/genetics.118.301684. URL `https://doi.org/10.1534/genetics.118.301684`.

S Sagitov. The general coalescent with asynchronous mergers of ancestral lines. *Journal of Applied Probability*, 36:1116–1125, 1999. doi: 10.1239/jap/1032374759. URL `https://doi.org/10.1239/jap/1032374759`.

Ori Sargsyan and John Wakeley. A coalescent process with simultaneous multiple mergers for approximating the gene genealogies of many marine organisms. *Theoretical Population Biology*, 74

(1):104–114, August 2008. doi: 10.1016/j.tpb.2008.04.009. URL https://doi.org/10.1016/j.tpb.2008.04.009.

J Schweinsberg. Coalescents with simultaneous multiple collisions. *Electronic Journal of Probability*, 5:1–50, 2000. doi: 10.1214/EJP.v5-68. URL http://doi.org/10.1214/EJP.v5-68.

Jason Schweinsberg. Coalescent processes obtained from supercritical Galton–Watson processes. *Stochastic Processes and their Applications*, 106(1):107–139, July 2003. doi: 10.1016/s0304-4149(03)00028-0. URL https://doi.org/10.1016/s0304-4149(03)00028-0.

Jason Schweinsberg. Rigorous results for a population model with selection II: genealogy of the population. *Electronic Journal of Probability*, 22(0), 2017. doi: 10.1214/17-ejp58. URL https://doi.org/10.1214/17-ejp58.

John A Sved, T Edward Reed, and Walter F Bodmer. The number of balanced polymorphisms that can be maintained in a natural population. *Genetics*, 55(3):469–481, March 1967. doi: 10.1093/genetics/55.3.469. URL https://doi.org/10.1093/genetics/55.3.469.

F Tajima. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics*, 123(3):585–595, November 1989. doi: 10.1093/genetics/123.3.585. URL https://doi.org/10.1093/genetics/123.3.585.

Jonathan Terhorst, John A Kamm, and Yun S Song. Robust and scalable inference of population history from hundreds of unphased whole genomes. *Nature Genetics*, 49(2):303–309, dec 2016. doi: 10.1038/ng.3748. URL https://doi.org/10.1038/ng.3748.

Andrea Timm and John Yin. Kinetics of virus production from single cells. *Virology*, 424(1):11–17, mar 2012. doi: 10.1016/j.virol.2011.12.005. URL https://doi.org/10.1016%2Fj.virol.2011.12.005.

David L. J. Vendrami, Lloyd S. Peck, Melody S. Clark, Bjarki Eldon, Michael Meredith, and Joseph I. Hoffman. Sweepstake reproductive success and collective dispersal produce chaotic genetic patchiness in a broadcast spawner. *Science Advances*, 7(37), September 2021. doi: 10.1126/sciadv.abj4713. URL https://doi.org/10.1126/sciadv.abj4713.

Filipe G. Vieira, Florent Lassalle, Thorfinn S. Korneliussen, and Matteo Fumagalli. Improving the estimation of genetic distances from next-generation sequencing data. *Biological Journal of the Linnean Society, London*, 117:139–149, mar 2015. doi: 10.1111/bij.12511. URL http://dx.doi.org/10.1111/bij.12511.

Matti Vihola and Jordan Franks. On the use of approximate Bayesian computation Markov chain Monte Carlo with inflated tolerance and post-correction. *Biometrika*, 107(2):381–395, February 2020. doi: 10.1093/biomet/asz078. URL `https://doi.org/10.1093/biomet/asz078`.

S. Wahlund. Zuzammensetzung von populationen und korrelationserscheinungen vom standpunkt der vererbungslehre aus betrachtet. *Hereditas*, 11:65–106, 1928.

John Wakeley. *Coalescent Theory: An Introduction*. Roberts & Company Publishers, Greenwood Village, CO, 2007. ISBN 978-0974707754.

Bruce Wallace. Hard and soft selection revisited. *Evolution*, 29(3):465–473, September 1975. doi: 10.1111/j.1558-5646.1975.tb00836.x. URL `https://doi.org/10.1111/j.1558-5646.1975.tb00836.x`.

Y. Wang, J. Lu, J. Yu, R. A. Gibbs, and F. Yu. An integrative variant analysis pipeline for accurate genotype/haplotype inference in population NGS data. *Genome Research*, 23(5):833–842, jan 2013. doi: 10.1101/gr.146084.112. URL `http://dx.doi.org/10.1101/gr.146084.112`.

George C Williams. *Sex and Evolution*. Princeton University Press, Princeton, New Jersey, 1975.

Erik S. Wright and Kalin H. Vetsigian. Stochastic exits from dormancy give rise to heavy-tailed distributions of descendants in bacterial populations. *Molecular Ecology*, 28(17):3915–3928, August 2019. doi: 10.1111/mec.15200. URL `https://doi.org/10.1111/mec.15200`.

Sewall Wright. Evolution in Mendelian populations. *Genetics*, 16:97–159, 1931.

Rolf J. F. Ypma, Hester Korthals Altes, Dick van Soolingen, Jacco Wallinga, and W. Marijn van Ballegooijen. A sign of superspreading in tuberculosis. *Epidemiology*, 24(3):395–400, May 2013. doi: 10.1097/ede.0b013e3182878e19. URL `https://doi.org/10.1097/ede.0b013e3182878e19`.

Guangchuang Yu, David K. Smith, Huachen Zhu, Yi Guan, and Tommy Tsan-Yuk Lam. Ggtree: an R package for visualization and annotation of phylogenetic trees with their covariates and other associated data. *Methods in Ecology and Evolution*, 8(1):28–36, September 2016. doi: 10.1111/2041-210x.12628. URL `https://doi.org/10.1111/2041-210x.12628`.

Kai Zeng, Yun-Xin Fu, Suhua Shi, and Chung-I Wu. Statistical tests for detecting positive selection by utilizing high-frequency variants. *Genetics*, 174(3):1431–1439, November 2006. doi: 10.1534/genetics.106.061432. URL `https://doi.org/10.1534/genetics.106.061432`.

Chi Zhang, Shan-Shan Dong, Jun-Yang Xu, Wei-Ming He, and Tie-Lin Yang. PopLDdecay: a fast and effective tool for linkage disequilibrium decay analysis based on variant call format files. *Bioinformatics*, 35(10):1786–1788, October 2018. doi: 10.1093/bioinformatics/bty875. URL https://doi.org/10.1093/bioinformatics/bty875.