Inferring Bottlenecks from Genome-Wide Samples of Short Sequence Blocks

Lynsey Bunnefeld,*[,1] Laurent A. F. Frantz,[†,2] and Konrad Lohse*
*Institute of Evolutionary Biology, University of Edinburgh, Edinburgh EH9 3FL, United Kingdom, and [†]Animal Breeding and Genomics Centre, Wageningen University, Wageningen 6708 PB, The Netherlands

ABSTRACT The advent of the genomic era has necessitated the development of methods capable of analyzing large volumes of genomic data efficiently. Being able to reliably identify bottlenecks—extreme population size changes of short duration—not only is interesting in the context of speciation and extinction but also matters (as a null model) when inferring selection. Bottlenecks can be detected in polymorphism data via their distorting effect on the shape of the underlying genealogy. Here, we use the generating function of genealogies to derive the probability of mutational configurations in short sequence blocks under a simple bottleneck model. Given a large number of nonrecombining blocks, we can compute maximum-likelihood estimates of the time and strength of the bottleneck. Our method relies on a simple summary of the joint distribution of polymorphic sites. We extend the site frequency spectrum by counting mutations in frequency classes in short sequence blocks. Using linkage information over short distances in this way gives greater power to detect bottlenecks than the site frequency spectrum and potentially opens up a wide range of demographic histories to blockwise inference. Finally, we apply our method to genomic data from a species of pig (*Sus cebifrons*) endemic to islands in the center and west of the Philippines to estimate whether a bottleneck occurred upon island colonization and compare our scheme to Li and Durbin's pairwise sequentially Markovian coalescent (PSMC) both for the pig data and using simulations.

KEYWORDS demographic inference; population bottleneck; generating function; maximum likelihood; *Sus cebifrons*

MUCH can be learned about the demographic history of a species or population from sequence variation. Bottlenecks—short periods where the population size is drastically reduced before recovering in size—are commonly detected demographic events. Because bottlenecks are often associated with range shifts caused by Pleistocene fluctuations in climate (*e.g.*, Moura *et al.* 2014), they are of particular interest to researchers studying contemporary biogeographic patterns. Bottlenecks lead to a sudden, strong increase in the rate of coalescence that may have different effects on the shape of genealogies. A strong bottleneck may trap all lineages, leading to a star-shaped genealogy. Alternatively, some lineages may "escape" weaker bottlenecks leading to longer basal branches, each connected to a clade of lineages that coalesced during the bottleneck (Figure 1). Thus, strong bottlenecks lead to an excess of rare variants, while moderate bottlenecks produce an excess of intermediate-frequency variants. These opposing effects make it difficult to reliably diagnose bottlenecks from simple summary statistics of nucleotide diversity such as Fu's *D* (Fu and Li 1993) and Tajima's *D* (Tajima 1989). The problem is exacerbated because signals of bottlenecks can be confounded by other population processes such as geographic structure or selective sweeps (*e.g.*, Nielsen and Beaumont 2009).

Given the stochasticity of the coalescent in general and the similarity in signal between selective sweeps and bottlenecks at a single locus (*e.g.*, Galtier *et al.* 2000), integrating information across multiple loci is imperative to robust demographic inference. For example, Li and Durbin's (2011) pairwise sequentially Markovian coalescent (PSMC) method models transitions in the times to the most recent common ancestor along an individual genome and from this infers a trajectory of population size

doi: 10.1534/genetics.115.179861
Manuscript received June 24, 2015; accepted for publication September 1, 2015; published Early Online September 3, 2015.
Available freely online through the author-supported open access option.

Supporting information is available online at www.genetics.org/lookup/suppl/doi:10.1534/genetics.115.179861/-/DC1.
[1]Corresponding author: Institute of Evolutionary Biology, University of Edinburgh, Kings Bldgs., Charlotte Auerbach Rd., Edinburgh EH9 3FL, United Kingdom. E-mail: lynsey.mcinnes@ed.ac.uk
[2]Present address: The Palaeogenomics and Bio-Archaeology Research Network, Research Laboratory for Archaeology and the History of Art, University of Oxford, Oxford OX1 3QY, United Kingdom.

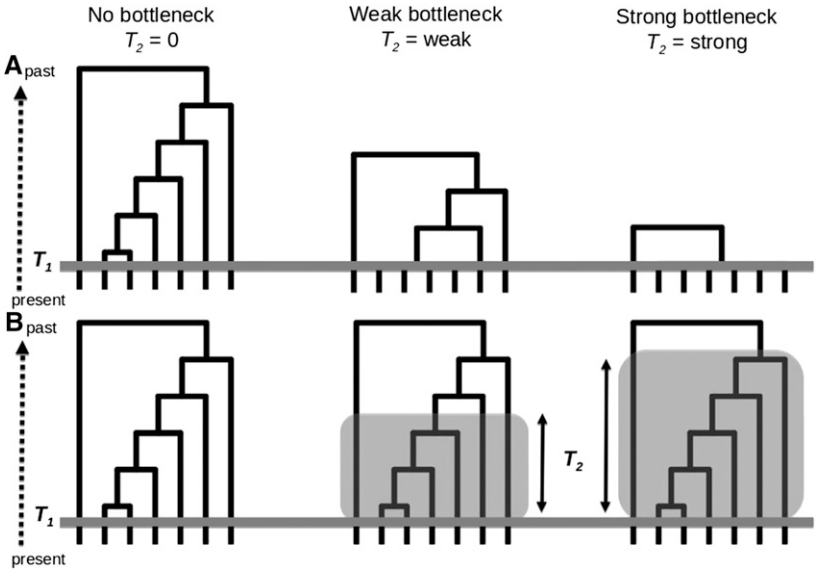

Figure 1 The effect of a bottleneck on a genealogy. (Left) Genealogy for a population of constant size. (Center) Genealogy under a weak bottleneck at time T_1 with four surviving lineages. (Right) Genealogy under a strong bottleneck at T_1 with only two surviving lineages. (A) We assume an instantaneous bottleneck that produces a sudden burst of coalescence. (B) Bottleneck strength is measured by T_2, the time necessary for the same amount of coalescence in a population of constant size.

changes. The PSMC relies on an approximation to the coalescent and, because it makes use of long-range linkage information, requires fairly well-assembled reference genomes.

At the other extreme, a number of popular methods of demographic inference are based on the site frequency spectrum (SFS)—the number of sites in a sample of n sequences with k copies of the derived allele (*e.g.*, Gutenkunst *et al.* 2009). The SFS throws away all linkage information and therefore results in a drastic loss of information. Although Bhaskar and Song (2014) showed that for the piecewise constant models of population size change, the SFS is a sufficient statistic given enough data, Terhorst and Song (2015) show that the error of SFS-based estimates converges at rate $1/\log s$, where s is the number of segregating sites. Both articles emphasize that this could be remedied by incorporating linkage information.

An alternative set of methods bases inference on many short loci (blocks of sequence) without requiring the long-range linkage information necessary for the PSMC. Considering linked sites within sequence "blocks" exploits the demographic information contained in the distribution of genealogical branches while still avoiding the need to model the ancestral recombination graph. This class of methods assumes that intrablock recombination can be ignored (Yang 2002; Hey and Nielsen 2004). Using short-read sequencing technologies, data sets containing large numbers of short blocks can easily be obtained for any organism, for example in the form of low-coverage fragmented genome assemblies, restriction site associated DNA (RAD), or transcriptome data (*e.g.*, Davey *et al.* 2011; McCormack *et al.* 2013; Hearn *et al.* 2014).

Using outgroup information to polarize mutations and assuming an infinite-sites mutation model, polymorphic sites in a nonrecombining alignment can be summarized as counts of mutations on each possible genealogical branch without loss of information. Each possible combination of mutation counts is a unique mutational configuration. Lohse *et al.* (2011) showed how the generating function (GF) of genealogies can be used to derive the probability of mutational configurations for a large range of demographic models. The probability of a mutational configuration is obtained by taking successive derivatives of the GF with respect to all relevant "dummy" variables, each corresponding to a different branch of the genealogy. This gives an efficient, maximum-likelihood scheme for estimating model parameters and comparing models from arbitrarily many sequence blocks. Although such likelihood calculations have been used to fit models of divergence and admixture from triplet samples (Hearn *et al.* 2014; Lohse and Frantz 2014), they fail for large numbers of samples ($n > 4$) because both derivation of the GF and the sheer number of possible mutational configurations become unmanageable (Lohse *et al.* 2015). Our main motivation for the present study was to develop a simple and general summary of blockwise data that (i) removes the need for phase information and (ii) captures short-range linkage information, allowing more powerful inferences than the SFS.

We first briefly describe the GF of genealogies under a bottleneck model and outline how it can be automatically generated and solved (using *Mathematica*). We then consider a new summary of blockwise data that is an extension of the SFS and summarizes polymorphic sites within blocks as counts of singletons, doubletons, etc. (*i.e.*, we lump branches with the same number of descendants, in effect generating an SFS for each block). We compare the power of this new scheme to both the likelihood-based analysis for full mutational configurations (for small samples) and the genome-wide SFS and use simulations to investigate the sensitivity of the new method to intrablock recombination and population structure. Finally, we apply our method to an example data set of Visayan warty pig (*Sus cebifrons*) genomes from the Phillippines and compare our inferences with PSMC analyses on the same data.

Materials and Methods

Outline of the model

We consider a sample of n lineages from a diploid, panmictic population with a current effective size N_e. We follow Galtier

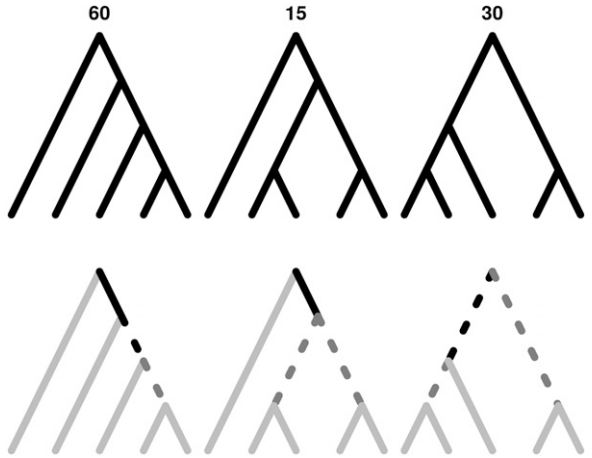

Figure 2 Combinatorial strategies to speed up likelihood calculations. For $n = 5$, there are three possible unlabeled topologies or tree shapes (left, center, and right). Because samples from the same population are exchangeable, the generating function can be written as a sum over unlabeled tree shapes (each defining an equivalence class of identically distributed genealogies). The size of each class (given above each shape) depends on the number of ways the sample labels can be permuted. The blockwise SFS introduces a further simplification (bottom) by lumping all branches with the same number of descendants (shaded line, singleton; shaded dashed line, doubleton; solid dashed line, tripleton; solid line, quadrupleton).

et al. (2000) in assuming a simple model of an instantaneous bottleneck that can be characterized by two parameters: a bottleneck start time, T_1, and a strength parameter, T_2 (Figure 1). We assume that the bottleneck is instantaneous, so that no mutations occur during it and its only effect is a sudden burst in coalescence that is measured by an imaginary time T_2. Both T_1 and T_2 are scaled in $2N_e$ generations.

Note that this two-parameter model is simpler than bottleneck models that consider step changes in N_e in real time (which requires at least three parameters, *e.g.*, Marth *et al.* 2004). Our motivation for assuming an instantaneous bottleneck is threefold. First, bottleneck duration and intensity are often confounded in practice; *i.e.*, it is generally hard to distinguish weak and long from short and strong bottlenecks (Marth *et al.* 2004). Second, an instantaneous bottleneck captures drastic events, *e.g.*, the colonization of new areas by a small founder population. Finally, instantaneous bottleneck histories extend to more general models of coalescence in which multiple mergers occur as a continuous process rather than an instantaneous event (*e.g.*, Barton *et al.* 2010; Coop and Ralph 2012).

We can describe the history of a sample of n lineages as a Markov process: going backward in time, pairs of lineages coalesce until T_1 when there is a sudden burst of coalescence due to the bottleneck. Within the bottleneck, coalescence proceeds normally, but since we ban mutations, there is no growth in genealogical branches during T_2 (Figure 1). T_2 can be thought of as the time necessary for the same amount of coalescence had the population size not changed.

We apply the general recursion for the GF of genealogical branches developed by Lohse *et al.* (2011, equation 4) to the bottleneck model. We denote the vector of all possible branches $\underline{t}$ and label branches by the individuals they are connected to. The GF of the distribution of branch lengths $P[\underline{t}]$ is defined as $\psi[\underline{\omega}] = E[e^{-\underline{\omega}\cdot\underline{t}}]$, where $\underline{\omega}$ are dummy variables corresponding to $\underline{t}$. For the bottleneck model, we need only to track coalescence events and transitions between three phases: before (1), during (2), and after the bottleneck (3). Analogous to derivations of the GF for histories involving discrete splits or admixture between populations (Lohse *et al.* 2011), we first write down a GF for a model in which transitions between phases (T_1 and T_2) are exponentially distributed with rates Λ_1 and Λ_2 (Equation 1). The GF recursion for each phase is identical apart from the Λ terms specifying the rates at which lineages enter the next phase and the fact that mutations (and hence ω terms) are banned during the bottleneck (ψ_2). We denote the GF of interest (where event times are discrete) as $P[\underline{\omega}]$ and note that $\psi[\underline{\omega}] = \int \Lambda_2\Lambda_1 P[\underline{\omega}] e^{\underline{\Lambda}\cdot\underline{T}} dT$. $P[\underline{\omega}]$ can be obtained by multiplying $\psi[\underline{\omega}]$ by $(\Lambda_2\Lambda_1)^{-1}$ and inverting once for each Λ parameter. Looking backward in time, Ω denotes the configuration of the sample before the first coalescence event i and Ω_i the configuration after it,

$$\begin{aligned}
\psi_1[\Omega] &= \frac{\Lambda_1\psi_2[\Omega] + \sum_i \psi_1[\Omega_i]}{\lambda_n + \Lambda_1 + \sum_{|S|=1}\omega_S} && T_0 < T < T_1 \\
\psi_2[\Omega] &= \frac{\Lambda_2\psi_3[\Omega] + \sum_i \psi_2[\Omega_i]}{\lambda_n + \Lambda_2} && T_1 < T < T_2 \qquad (1) \\
\psi_3[\Omega] &= \frac{\sum_i \psi_3[\Omega_i]}{\lambda_n + \sum_{|S|=1}\omega_S} && T_2 < T,
\end{aligned}$$

where $\lambda_n = n(n-1)/2$ and $\sum \omega_S$ is the sum of the ω_S that increase during that interval. For the first event, these correspond to the terminal branches; *i.e.*, $|S| = 1$. We have automated Equation 1 in *Mathematica*.

Calculating likelihoods from the GF

The general method for computing the probability of observing a particular mutational configuration, $\underline{k}$, which is defined as counts of mutations on $\underline{t}$ and can be interpreted as the likelihood, has been described in detail previously (Lohse *et al.* 2011, equation 1) and involves taking higher-order derivatives of the GF of genealogical branches. We assume a mutation rate per branch of $\theta/2$ and tabulate exact probabilities only up to a maximum number of mutations k_m per branch. Probabilities for configurations involving $> k_m$ mutations per branch are combined to avoid having to distinguish very unlikely configurations.

Although our automation allows us to generate the GF for the bottleneck model for rather large samples of individuals (n), inverting (with respect to Λ) and solving the GF are slow for $> n = 3$. We employ a set of combinatorial strategies (detailed in Lohse *et al.* 2015) to speed up these steps (and see Figure 2). In particular, we make use of the fact that lineages are exchangeable. This means that the GF (and the likelihood) can be written as a sum over unlabeled tree shapes

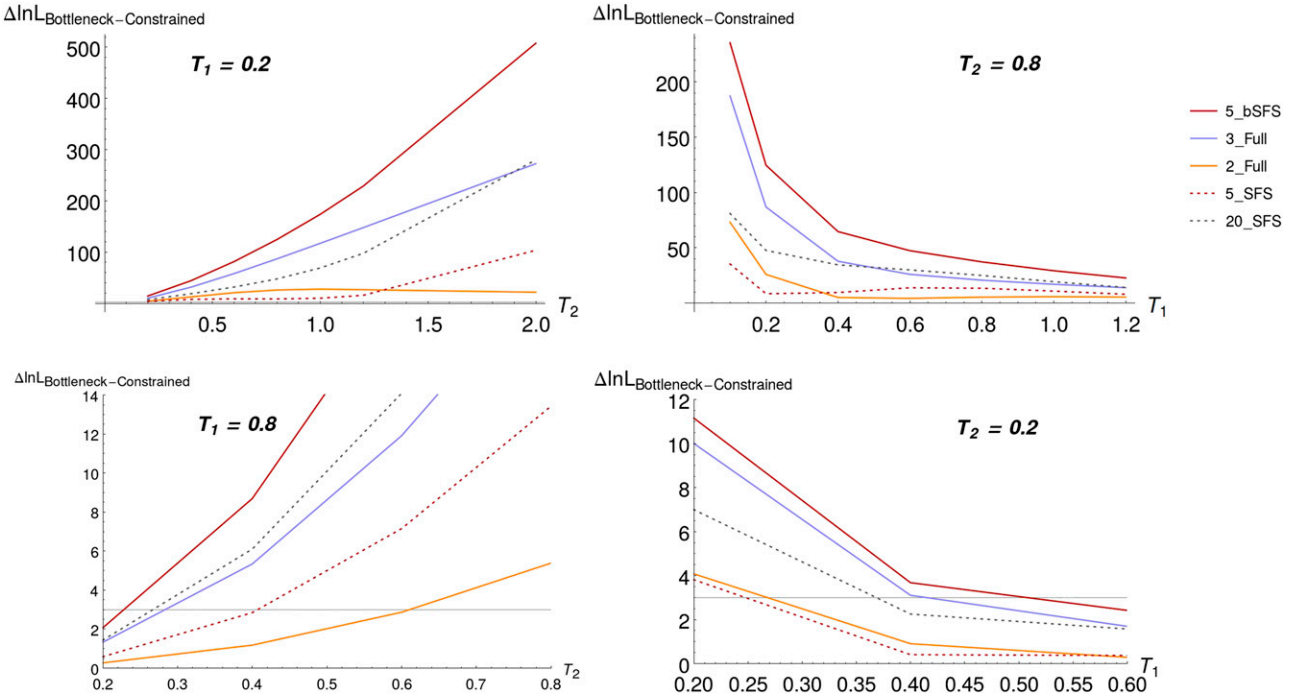

Figure 3 The expected difference in support ($E[\Delta lnL]$) between the bottleneck model ("Bottleneck") and a null model of constant population size ("Constrained") as a function of bottleneck strength (T_2, left) and start time (T_1, right). Plots are based on analytic results for the likelihood and assuming 2000 unlinked sequence blocks with one pairwise difference per block on average. The horizontal lines in the bottom panels indicate significance (at $\alpha = 0.05$) in a likelihood-ratio test with 2 d.f.

(*sensu* Felsenstein 1978, 2003) that define equivalence classes of identically distributed genealogies. For each tree shape, we can condition on a single arbitrary labeling of individuals by setting terms in the GF that are incompatible with it to zero. The full GF can be written as a sum of the GFs for the set of equivalence class representatives, each weighted by the size of its class (the number of ways the sample labels can be permuted on the tree shape, Figure 2). Second, we use *Mathematica* to successively solve the GF for increasingly large sample configurations, starting with terms for pairs of lineages, inserting this solution into the GF for $n = 3$, and so on.

Blockwise SFS simplification

The number of possible mutational configurations quickly becomes unmanageable as the number of genealogical branches increases: distinguishing mutations on each of the $2(n-1)$ branches and allowing for a maximum of k_m mutations per branch, there are $(k_m+2)^{2(n-1)}$ mutational configurations for each equivalence class (+2 stems from the configurations involving 0 and $> k_m$ mutations per branch).

We introduce a simple summary of blockwise data—which we term the blockwise SFS (bSFS)—to address this issue. Instead of distinguishing mutations on all genealogical branches, we combine branches (and their corresponding ω variables) with the same number of descendants so that the vector of branch lengths, $\underline{t}$, now distinguishes only singleton, doubleton branches, etc. (*i.e.*, $\underline{t_i} = \{t_1, t_2, \ldots, t_{n-1}\}$). In data terms, we reduce the mutational configuration of a block $\underline{k}$ to $\underline{k_i} = \{k_1, k_2, \ldots, k_{n-1}\}$ corresponding to counts of singletons, doubletons, etc. in each block. Note also that the genome-wide expected SFS ($E[\xi_i]$) is given by $E[k_i]/\sum_{n-1} E[k_i]$.

The bSFS simplification has two advantages: first, it reduces the number of branches and hence mutational configurations substantially. Because we distinguish only $n-1$ types of mutations, the number of mutational configurations goes down from $(k_m+2)^{2(n-1)}$ to $(k_m+2)^{(n-1)}$. For example, with $k_m = 3$ and $n = 5$, there are 390,625 configurations in the full scheme, but only 625 defined via the bSFS. Second, the bSFS simplification does not require phased data because we no longer distinguish in which sampled lineage a mutation occurred. However, despite these simplifications, the GF and hence the probabilities of $\underline{k_i}$ can be computed analytically only for rather limited sample sizes ($n < 6$).

Many current methods for detecting population size changes are based on the SFS (*e.g.*, Polanski and Kimmel 2003; Marth *et al.* 2004; Gutenkunst *et al.* 2009; Chen 2012). Because the bSFS considers the joint distribution of linked polymorphisms, it is, all else being equal, more powerful than the SFS. However, given that analytic likelihood calculations under the bSFS are feasible only for small n, it is important to compare their power to that of the SFS for larger n. In the *Appendix* we obtain the SFS ($E[\xi_i]$) under the bottleneck model for any n, a result that we anticipate will be useful for researchers with access exclusively to single-nucleotide polymorphism (SNP) data.

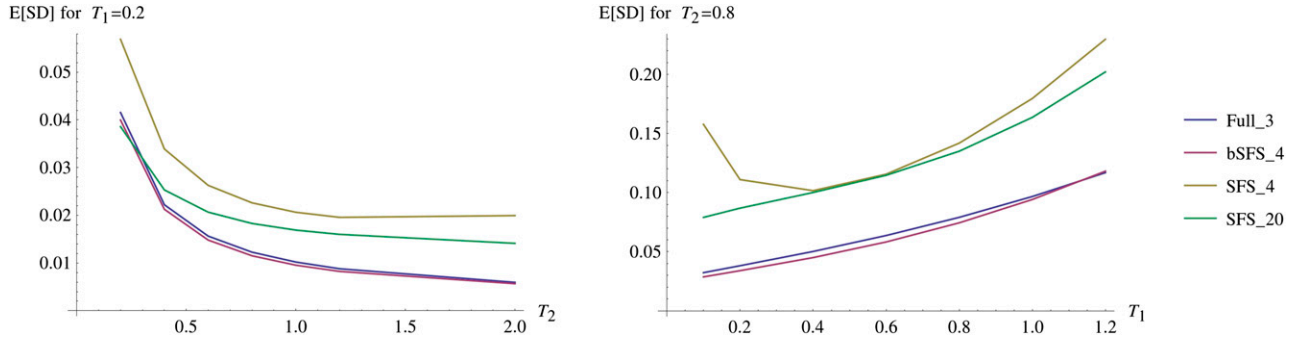

Figure 4 Expected standard deviation in T_1 (= 0.2, left) and T_2 (= 0.8, right) as a function of T_2 and T_1, respectively. As in Figure 3, we assume 2000 unlinked blocks.

Power analyses

We measured power by calculating the expected difference in support ($E[\Delta lnL]$) between the true model and a null model of constant population size for three schemes: (i) the full information for $n = 2$ or 3 ($k_m = 3$), (ii) the bSFS scheme for $n = 5$ (unfolded, $k_m = 3$, but 6 for the singleton category; and folded, $k_m = 6$), and (iii) the SFS for $n = 5,\ 10,\ 20$. The full scheme for $n = 2$ could be considered equivalent to the PSMC in the case where a continuous genome is not available.

Throughout, our choice of parameters was motivated by the pig data example, including genome size ($\approx$ 1 Gb), N_e ($\approx$ 10,000), and mutation rate ($\approx 2.5 \times 10^{-8}$ per base and generation). We explore a broad parameter space, split into four sets: we alter either T_1 or T_2 from 0.2 to 1.2, while fixing the other parameter to 0.2 or 0.8. This covers a broad range of bottleneck strengths and ages. For example, for pigs, it would include bottlenecks between 20,000 and 120,000 years ago. We additionally investigate the power to detect more recent ($T_1 = 0.1$, $T_2 = 0.8$) and very strong ($T_1 = 0.2$, $T_2 = 2$) bottlenecks that have a high probability of "trapping" all lineages.

Given the assumption of a constant μ per site, the scaled mutation rate, $\theta = 4N_e\mu$, can be thought of as the length of blocks. Since we generally have no independent knowledge of θ in practice, we conditioned on $E[S]$, the expected number of segregating sites per block, and adjusted θ to correspond to $E[S] = 1$ per block for a pairwise sample. Because the expected total length of the genealogy decreases with increasing T_2 and decreasing T_1, younger and stronger bottlenecks correspond to larger θ, *i.e.*, longer blocks (Supporting Information, Table S1;). Note that our assumption of no intralocus recombination is unaffected by conditioning on $E[S]$, because both S and the number of recombination events in a block depend on the total length of the genealogy.

We quantified the accuracy to estimate a particular parameter, using Fisher information (Edwards 1972; Lohse and Frantz 2014). This measures the sharpness of the lnL curve near the maximum and shows a linear relationship with the number of loci. Assuming parameter estimates are away from the boundaries, the inverse of Fisher information gives a lower bound on the expected variance of parameter estimates (Rao 1945). We used this to calculate the expected standard deviation $E[SD]$ for T_1 and T_2 across parameter space (and checked these using simulations).

Robustness to population structure and recombination

Previous studies have shown that simple summary statistics (Tajima's D, Fu and Li's D, etc.) cannot distinguish between demographic size changes and population structure (Nielsen and Beaumont 2009; Städler *et al.* 2009; Chikhi *et al.* 2010). In particular, intermediate migration rates and local sampling, *i.e.*, when all samples come from a single deme that is part of a much larger metapopulation, have very similar effects on these statistics as demographic size changes. This can be understood considering the separation of timescales in the structured coalescent (Wakeley 1998). Looking backward in time, during the initial "scattering" phase, lineages in the same deme either coalesce or migrate to unsampled demes. During the "collecting" phase, coalescence is much slower because lineages first need to migrate back into an occupied deme before they can coalesce. Thus, a genealogy of a sample from a single deme may look like a bottlenecked genealogy (rapid initial coalescence "during the bottleneck," followed by slow coalescence of the remaining lineages that "survived the bottleneck"). We explored the sensitivity of our method to population structure, using simulated data.

We used *ms* (Hudson 2002) to simulate data under an island model with symmetric migration at rate $M = 4N_em$. We simulated samples of four individuals in a metapopulation of 10 equally sized demes (d). Sampling was either local (all samples from the same deme) or scattered (all from a different deme). We chose a single metapopulation $\theta = 4N_ed\mu = 1$ (corresponding to a per deme θ of 0.1).

We also used simulations to assess whether ignoring intralocus recombination biases parameter values. We simulated data for $n = 4$, assuming a bottleneck at time $T_1 = 0.2$ with strength $T_2 = 1$ and a range of recombination rates: 0, 0.025, 0.125, 0.25, 1.25, 2.5, 5, 12.5, or 25 $\times 10^{-8}$ crossovers per generation per base pair. Assuming $\mu = 2.5 \times 10^{-8}$, this corresponds to $r/\mu = 0$, 0.01, 0.05, 0.1, 0.5, 1, 2, 5, 10. These ratios fall either side of the recombination estimates for mammals of 1 cM/Mb; *i.e.*, $r/\mu = 0.4$ (Tortereau *et al.* 2012).

For each parameter combination, we simulated 1,000,000 loci to obtain the expected frequencies of mutational

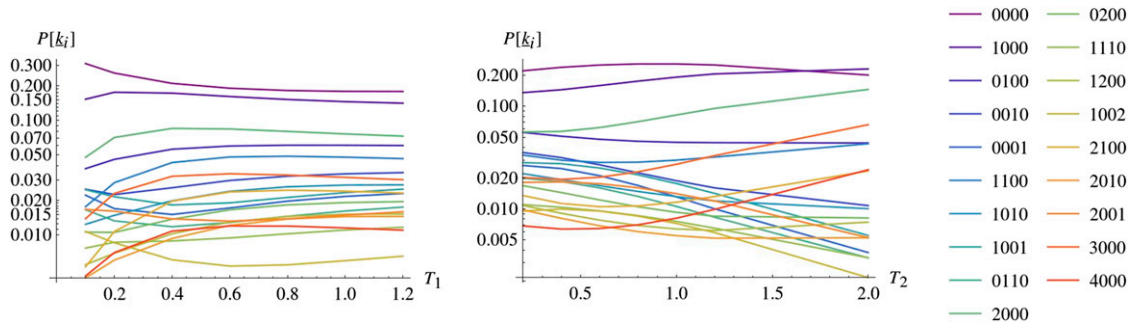

Figure 5 The probability of mutational configurations defined by the bSFS for $n = 5$ as a function (left) of T_1 (T_2 = 0.8) and (right) of T_2 (T_1 = 0.2). The key gives configurations as counts of singletons, doubletons, tripletons, and quadrupletons, respectively; *e.g.*, 1010 denotes blocks containing one singleton and one tripleton. Only configurations with probabilities >0.01 (for at least one point in parameter space) are shown.

configurations for the bSFS scheme. We maximized the *lnL* under both a null model of constant population size and a bottleneck history and calculated $E[\Delta lnL]$ between the two scenarios.

Application to island pigs

We analyzed genomic data from three Visayan warty pigs, *S. cebifrons*, a species endemic to the Philippines. Each individual genome was sequenced to 10× coverage and aligned to the *S. scrofa* reference genome (Ssc10.2) (Groenen *et al.* 2012; Frantz *et al.* 2013, 2014; Bosse *et al.* 2015). We divided the reference genome of *S. scrofa* into nonoverlapping 1000-bp blocks. To ensure enough coverage to call all heterozygous sites in each block and to remove possible copy number variants (Paudel *et al.* 2013), we filtered out blocks with an average coverage <7× or higher than twice the genome-wide average, using the pileup format in SAMtools v0.1.12 (Li *et al.* 2009). Clusters of two or more SNPs in a 10-bp window were filtered out as well as SNPs within 3 bp of an indel. We removed blocks for which <90% of sites were covered and excluded sites with a coverage <4× (Gronau *et al.* 2011). Finally, we selected only blocks that passed the above filtering criteria in all samples.

This procedure left 1,103,026 aligned 1000-bp blocks (~50% of the genome, with an average of 2.75 mutations per block). Given the mean divergence (1.8%) of the only usable outgroup, the African common warthog (*Phacochoerus africanus*), we used the folded bSFS for five samples (combining singletons with quadrupletons and doubletons with tripletons) to avoid biases due to mispolarization. To make use of all six alleles (two per individual), we averaged the counts of mutational configurations across all sample combinations, *i.e.*, after omitting one allele from each individual in turn. We used all blocks to obtain point estimates and computed confidence intervals of parameters and ΔlnL between models, using a simple correction: linkage between blocks, measured by the pairwise correlation coefficient in the number of segregating sites between blocks, dropped below 0.05 at a distance of 404 blocks (Figure S1) so we rescaled ΔlnL by a factor of 1/404. Finally, we fitted the bottleneck model to the folded genome-wide SFS for all six samples (using the same correction to rescale ΔlnL).

We conducted a PSMC analysis (Li and Durbin 2011) on the same three individuals, using the parameters $N = 25$, $T_{max} = 20$, $r = 15$, and 64 time intervals, and performed 100 bootstrap replicates for each sample. To calibrate parameter estimates, we assumed a generation time of 5 years and a mutation rate of 2.5×10^{-8} per base and generation (see Groenen *et al.* 2012; Frantz *et al.* 2013). Finally, we compared our method to the PSMC using simulations. We used Heng Li's modification of msHOT (https://github.com/lh3/foreign/tree/master/msHOT-lite) and ms to PSMC parser (https://github.com/lh3/psmc/blob/master/utils/ms2psmcfa.pl) to simulate 100 diploid sequences of 10 Mb under a variety of bottleneck parameters with an r/μ ratio of 0.4, $\mu = 2.5 \times 10^{-8}$, and $N_e = 10,000$.

Data availability

The pig data are available at the European Nucleotide Archive (http://www.ebi.ac.uk/ena/, accession no. PRJEB9326). A Mathematica notebook to fit the bottleneck model to data using the bSFS is available upon request.

Results

Power analyses

Our power analyses highlight several features of the blockwise inference method (Figure 3). First, inferences based on blockwise data have high power to correctly detect a bottleneck history (measured by $E[\Delta lnL]$) in absolute terms. We plot $E[\Delta lnL]$ for data sets consisting of 2000 unlinked blocks, but because $E[\Delta lnL]$ scales linearly with the number of blocks, the *y*-axes in Figure 3 can be rescaled for any number of blocks. Unsurprisingly, power decreases with bottleneck age and increases with bottleneck strength (Figure 3). Given data from 2000 unlinked blocks, even very weak bottlenecks would be detectable. For example, for a relatively recent and weak bottleneck ($T_1 = 0.2$ and $T_2 = 0.2$), $\Delta lnL = 11.2$ for the bSFS scheme for $n = 5$ and $\Delta lnL = 10.0$ for the full scheme for $n = 3$, which both indicate a significantly ($P < 0.0001$, likelihood-ratio test with 2 d.f.) better fit than a null model of constant size.

Second, blockwise likelihood computations are substantially more powerful than the SFS and the relative

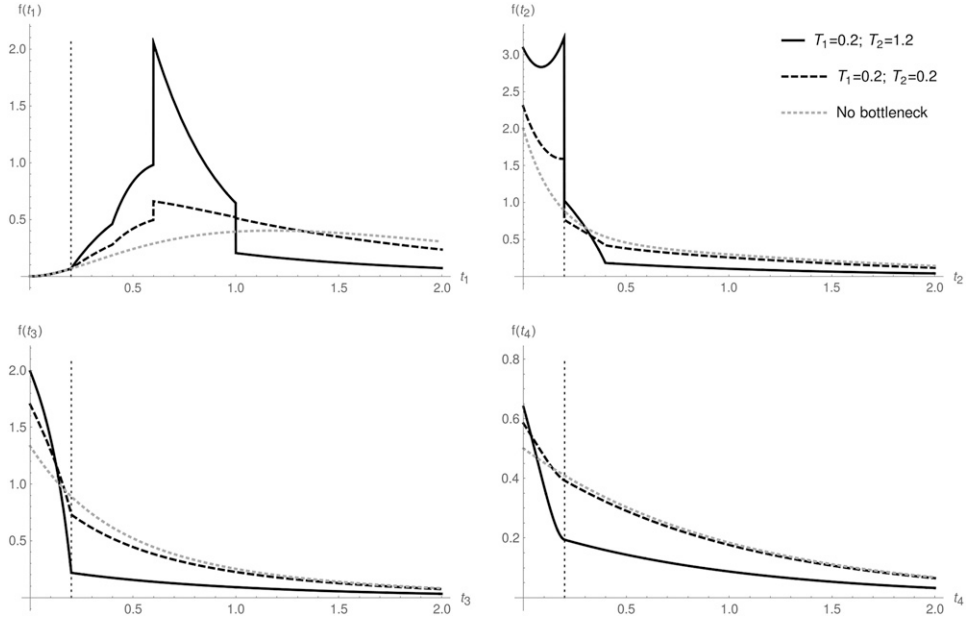

Figure 6 The length distributions of each branch type for two recent bottlenecks of different strengths and a constant size model under the blockwise-SFS scheme for five lineages. The dotted line indicates the bottleneck start time (T_1 = 0.2).

improvement is particularly large for young bottlenecks. We have highlighted the area of parameter space where the bSFS scheme for $n = 5$ is the only scheme that can still detect a bottleneck (Figure 3, bottom). As T_2 increases, the SFS scheme for $n = 20$ overtakes the full scheme for $n = 3$ in terms of power. However, the bSFS scheme for $n = 5$ remains substantially more powerful throughout.

Interestingly, the power to detect bottlenecks with the SFS does not decrease monotonically with respect to the start time, T_1. While very recent bottlenecks ($T_1 = 0.1$) are easiest to detect, older bottlenecks ($T_1 > 0.6$) are easier to detect than intermediate bottlenecks at $T_1 = 0.2$ (Figure 3, top right). In contrast, for all blockwise schemes, power decreases monotonically with start time.

To compare the different inference schemes in terms of their accuracy of parameter estimates, we calculated the expected standard deviation (E[SD]) of each parameter. The analytical results for E[SD] match the empirical SD across replicate simulated data sets and scale with the number of loci (n_l) as $\sqrt{1/n_l}$ (Figure 4, Figure S3, and Figure S4). Across schemes, E[SD] increases as bottlenecks get older and weaker. E[SD] is always less for blockwise schemes, apart from weak, recent bottlenecks (Figure 4, left) where E[SD] is large and similar for both the blockwise schemes and the SFS with 20 individuals (the SFS for 4 individuals is even worse).

To get a sense of the extra information captured by the bSFS, we plotted the probability of mutational configurations (for $n = 5$) (Figure 5; *cf.* Figure S5 for the SFS). This highlights two features of the bSFS. First, even for short blocks ($E[S] = 1$) there are many more mutational configurations than site frequency classes. For example, for a bottleneck of strength $T_2 = 0.8$ at time $T_1 = 0.2$ there are 19 configurations with probability > 0.01. Second, the probabilities of mutational configurations depend on the bottleneck parameters in a nonlinear way.

The marginal distributions of the total length of each type of n-ton branch also illustrate the complex effects bottlenecks have on genealogies (Figure 6). These can be found from the GF by differentiating with respect to the corresponding ω_i and evaluating at $\omega_i \to 0$. All distributions have a discontinuity at T_1 (Figure 6, dotted line) and at multiples of T_1 corresponding to varying numbers of coalescences occurring within the bottleneck. None of this complexity is captured by the genome-wide SFS that depends only on the mean of each distribution.

Robustness to population structure and recombination

With local sampling, population structure is difficult to distinguish from a bottleneck that always has greater support than a model of constant size (Table 1, top row). This is true for a wide range of migration rates, *i.e.*, $M = 0.05$–0.5. At the limit of low M, the chance of migration out of the focal deme is minuscule and so the individual deme functions as an isolated, constant size population. At the limit of high M, the whole metapopulation functions as a single panmictic population. Where there is support, we find that estimates for T_1 are very close to zero (see Table S2 for parameter estimates) as expected if the bottleneck mimics the scattering phase of the structured coalescent. Conversely, with scattered sampling, we never observe support for a bottleneck, irrespective of M; *i.e.*, ΔlnL is close to zero and θ estimates become very large as M decreases (not shown).

To investigate how well data generated under the island model (local sampling) actually fit a bottleneck history, we computed the expected frequency of mutational configurations under the (mis)inferred bottleneck parameters. Using ΔlnL as a measure of model fit, we compared the bottleneck model and a null model of constant population size. Although a bottleneck history fits data simulated under population structure (with intermediate M) better than a null model, the fit is poor in absolute terms. In particular, we find that

Table 1 ΔlnL between a bottleneck history and a model of constant population size for data simulated under the island model at a range of migration rates (M)

M	0.01	0.02	0.05	0.1	0.2	0.5	1	10
Island (local sampling)	0.752	1.88	4.58	6.75	7.15	4.73	2.40	0.080
Bottleneck	0.918	2.35	5.87	7.44	7.56	4.87	2.47	0.084

The top row ("Island") gives ΔlnL for the simulated data. The bottom row ("Bottleneck") gives $E[\Delta lnL]$ when assuming the bottleneck parameters estimated for the simulated data. In both cases, 2000 unlinked blocks are assumed.

more singleton mutations are seen in the data than predicted under a bottleneck history (Figure S6). Thus, it may in principle be possible to distinguish the two models.

We find that with realistic levels of recombination, parameter estimates under the bSFS are essentially unbiased. For $r/\mu > 0.5$, T_1 is increasingly overestimated, while T_2, θ, and ΔlnL between the bottleneck and the null model are underestimated. SD of parameter estimates using 2000 blocks is low for $r/\mu < 1$ (Figure S7). These results are encouraging and suggest that inferences based on the bSFS are robust to realistic levels of recombination.

Application to island pigs and comparison with the PSMC

A bottleneck history fitted the *S. cebifrons* data significantly better than a model of constant N_e, using the folded bSFS for $n = 5$ ($\Delta lnL = 110$ after correcting for linkage). We inferred a strong bottleneck ($T_2 = 1.72$) ~141 KYA and an ancestral N_e of $\approx 22,400$. While both T parameter estimates have narrow confidence intervals even after correction for linkage between blocks, there is more information about the time (T_1) than the strength (T_2) of the bottleneck (Table 2). Parameter estimates from the genome-wide SFS (for $n = 6$) agree, but have much wider confidence intervals (Table 2).

Our estimate of T_1 also broadly coincides with the period of the lowest population size in the PSMC analyses of the three *S. cebifrons* genomes (Figure 8A). Applying the PSMC to data simulated under the bottleneck history inferred for the pigs (Table 2) gives a shorter period of reduced population size than the empirical plot (Figure 8A *vs.* Figure 8B). The PSMC plot generated from the simulated data also does not include the apparent difference in population size seen in the empirical plot; *i.e.*, N_e after the bottleneck (looking backward in time) is approximately twofold greater than before. Applying the PSMC to data simulated under an instantaneous bottleneck produces a trajectory of gradual decline and rise of N_e (Figure 8B). This is true for a range of bottleneck times and strengths (Figure S8). Reassuringly, the period of smallest N_e coincides with the time of the bottleneck. An unexpected feature of all PSMC plots from simulated data is that they show a marked increase of N_e prior (looking backward in time) to the bottleneck start time (Figure 8B and Figure S8).

Discussion

Researchers have long been interested in using genetic data to infer population bottlenecks associated with landmark events in the history of species and populations, for example, the colonization of new regions or range contractions (reviewed in Gattepaille *et al.* 2013). Indeed, the demographic history of model organisms such as humans and *Drosophila* has been characterized as a series of expansions and contractions that coincide with the spread of populations across the globe (*e.g.*, Haddrill *et al.* 2005; Voight *et al.* 2005; Sjödin *et al.* 2012). However, the genetic resources for model organisms far exceed those available for most species. Thus there is a need for inference methods that do not require high-quality reference genomes, phase information, and/or large samples of individuals. Maximum-likelihood estimation based on the blockwise SFS makes optimal use of the information contained in short blocks of sequence from just a handful of individuals—data that can readily be obtained for any organism with current short-read sequencing technology. Furthermore, and in contrast to the PSMC framework (Li and Durbin 2011) that qualitatively documents changes in population size, our method explicitly assesses the statistical support for a bottleneck model over a null model (see Table S3 for the pros and cons of alternative demographic inference methods).

Table 2 Maximum-likelihood estimates for bottleneck parameters for *Sus cebifrons* (95% C.I. in parentheses)

Scheme	n	T_1	T_2
bSFS	5	0.63 (0.54–0.73)	1.72 (1.43–2.03)
SFS	6	0.46 (0.14–0.84)	1.54 (1.24–∞)

Power of the blockwise SFS

We find that our blockwise method has higher power to detect bottlenecks (Figure 3 and Figure 4) than the SFS. Perhaps surprisingly, the bSFS scheme for 5 lineages almost always contained more information about past bottlenecks than the genome-wide SFS for 20 lineages, a more typical sample size for SFS-based analyses (see Bhaskar and Song 2014, for specific bounds on sample sizes under various piecewise population size change models). Although our strategy of exploiting the symmetries of the coalescent by partitioning the GF of branch lengths into a sum over unlabeled tree shapes makes it possible to compute the probability of full mutational configurations for nontrivial sample sizes, in practice, the number of mutational configurations still explodes catastrophically for $n > 5$ (Lohse *et al.* 2015). By combining branches with the same number of descendants, the bSFS scheme introduces a further simplification that substantially decreases the dimensionality of mutational configurations and is a tractable compromise between information and complexity. It still exploits the information contained in the joint distribution of closely linked SNPs (Figure 5) and, since topologies are defined by the presence or absence of mutation categories, also retains topology information (Figure 2). On the other hand, the bSFS simplification avoids unnecessary computational complexity and removes the need for phase information. Because our likelihood calculations for the bSFS assume no intralocus recombination, they are restricted to

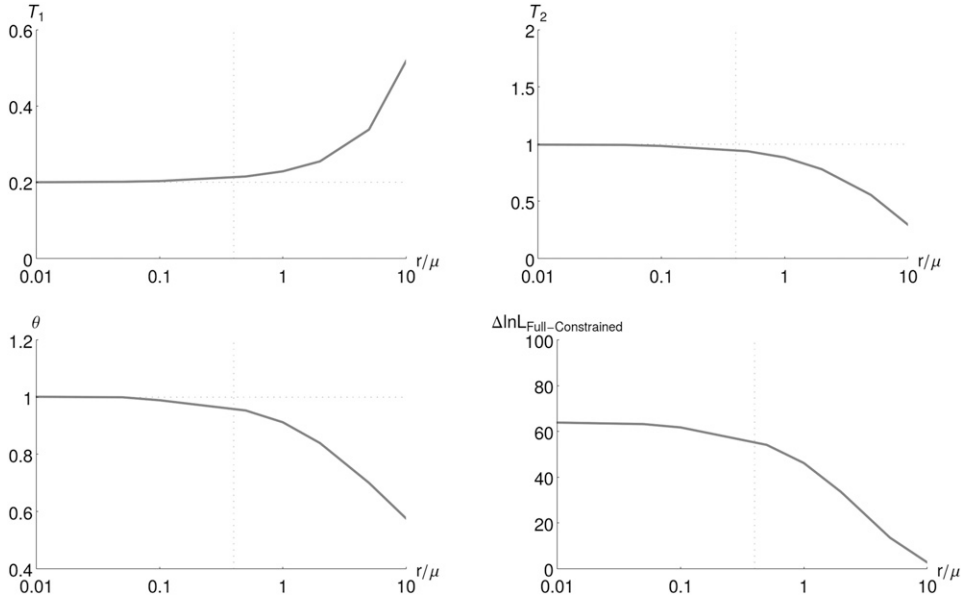

Figure 7 The effect of recombination (r/μ) on T_1 and T_2 (top) and on θ and ΔlnL (assuming 2000 unlinked blocks) between a bottleneck model and a null model (bottom). The horizontal dotted lines indicate the true parameters used for simulations: $T_1 = 0.2$, $T_2 = 1$, and $\theta = 1$. The vertical dotted line at 0.4 is a reasonable r/μ for mammals (assuming $r = 1 \times 10^{-8}$ and $\mu = 2.5 \times 10^{-8}$). Figure S7 shows the standard deviation of estimates obtained from 100 replicate data sets, each consisting of 2000 blocks.

relatively short blocks, which rarely contain sufficient mutational information to fully resolve all branches anyway.

Runs of homozygosity contain valuable information about demography (*e.g.*, Harris and Nielsen 2013) and some of the power of the blockwise schemes undoubtedly stems from the fact that some of this information is included in the frequency of monomorphic blocks that are almost always the most probable configuration (Figure 5). For example, for very recent, strong bottlenecks ($T_1 = 0.01$ and $T_2 = 1.2$), >50% of blocks are monomorphic for $n = 5$ whereas under the null model of constant size only 20% of blocks are monomorphic. This is despite the fact that we have fixed the block length to correspond to one pairwise difference per block on average in both cases.

Although we have focused on bottleneck histories because they are biologically interesting and analytically tractable, we emphasize that the bSFS simplification can be used to summarize data and compute likelihoods for any sampling scheme and demographic model. Just like the SFS, the bSFS also extends to multiple populations (Lohse *et al.* 2015). For larger samples and/or more complex models, for which analytic likelihood calculations become intractable, the bSFS could form the basis for alternative inference methods. In particular, it should be possible to approximate the bSFS using simulations and devise a composite-likelihood inference method analogous to that of Excoffier *et al.* (2013).

When can we expect to be able to detect bottlenecks?

Our power analyses indicate that young and strong bottlenecks are easiest to identify. This makes intuitive sense given that these have the strongest effect on genealogies (greatest compression of tree length). Looking backward in time, the older the bottleneck, the more likely it is that lineages have already coalesced by the time the bottleneck "starts." Even very large samples are expected to coalesce to two lineages within $2N_e$ generations (Tajima 1983; Nordborg 1998), suggesting that there is little information to infer bottlenecks approaching this age or older (as we find, Figure 3) and that larger samples of individuals add only information about recent bottlenecks (see also Terhorst and Song 2015).

A potential issue of our power analyses is that we have adjusted the length of sequence blocks (*i.e.*, θ) to a fixed average number of segregating sites per block irrespective of the severity of the bottleneck. This ignores the fact that genomes have a finite length and so overestimates the power to infer very strong and/or recent bottlenecks. However, even after correcting for this by normalizing ΔlnL by a factor, $1/\theta$, the power to detect bottlenecks still remains highest for recent and strong bottlenecks (Figure S2). In the limit of infinitely young and strong bottlenecks and assuming a finite genome, it would of course be impossible to sample enough independent blocks for inference. However, we can still detect surprisingly recent bottlenecks: *e.g.*, assuming $T_1 = 0.02$ (2000 years ago for pigs), $T_2 = 1$, and a typical pig N_e and μ, blocks would need to be ~2.5 kb long to contain one pairwise difference on average, and only nine such blocks would be sufficient to reject a null model of no population size change ($E[\Delta lnL] = 0.34$ for a single locus, nine loci would give a significant likelihood-ratio test with 2 d.f., and $\alpha = 0.05$).

Our finding that local sampling of a structured population mimics a bottlenecked population mirrors other recent analyses that conclude that population structure and demography can be hard to distinguish (*e.g.*, Nielsen and Beaumont 2009). However, we identified an excess of singletons in samples from structured compared to bottlenecked populations that could, in principle, be exploited to distinguish the two. Conversely, and in agreement with others (Wakeley 1999; Städler *et al.* 2009; Chikhi *et al.* 2010; Leblois *et al.* 2014; Mazet *et al.* 2015), we find that scattered sampling (one individual per deme) minimizes the problem of confounding size change and structure.

Finally, we show that for realistic recombination rates, blockwise inferences are robust to intralocus recombination (Figure 7). If the ratio of recombination to mutation rate is

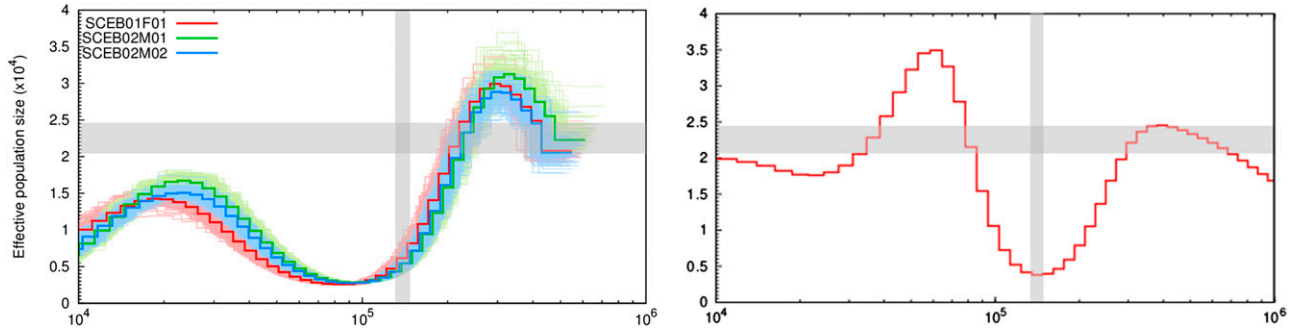

Figure 8 Comparison of our method with the PSMC. (Left) Shown are PSMC results for the three *Sus cebifrons* individuals (SCEB01 from Panay and SCEB02 from Negros). Estimates from the full data and 100 bootstrap replicates are depicted by darker and lighter lines, respectively. The gray bars indicate our estimates of T_1 and N_e (plus 95% C.I.) from the folded bSFS for $n = 5$. (Right) Using the inferred pig parameters (Table 2), we simulated 100 diploid sequences of 10 Mb with $r/\mu = 0.4$.

high, T_1 is overestimated, T_2 underestimated, and there is less power to detect a bottleneck overall. Because recombination within blocks generates a mosaic of correlated genealogies, this is expected, given the decrease in variance in branch lengths across loci (Wall 2003).

Application to island pigs and comparison with the PSMC

The instantaneous bottleneck (but constant N_e) we consider is, in a sense, the inverse of the scenario assumed by the PSMC and other piecewise models that reconstructs demography as a trajectory of changing N_e. Given these opposing assumptions, it is reassuring how well the bottleneck time we infer for *S. cebifrons* ($T_1 = 140$ KY) agrees with the PSMC analysis of the same data set (Figure 8A). In both cases, the inferred reduction in population size occurred long after the species divergence between *S. cebifrons* and *S. barbatus* ~300 KYA (Frantz *et al.* 2013) and so was not associated with the initial colonization of the Philippines by the ancestor of *S. cebifrons*. This is not to say that we can rule out any colonization bottleneck, because most of the signal of older bottlenecks would have been erased by the more recent demographic event we infer. It is likely that the inferred bottleneck coincides with the colonization of the central/western islands of Panay, Cebu, and Negros by *S. cebifrons*. This would support the findings of Lucchini *et al.* (2005) that this species is genetically and morphologically distinct from its sister species *S. philippensis* that is found all over the eastern and northern Philippines. Unfortunately, genomic data do not exist for *S. philippensis* to validate this interpretation. Finally, our result is not consistent with a bottleneck due to anthropogenic activities (*i.e.*, it is too old). This does not mean that this critically endangered species has been unaffected by human activities but rather suggests that its genetic diversity was already reduced prior to any human contact and supports current efforts for careful management of zoo populations (Bosse *et al.* 2015).

We found that the PSMC is unable to reconstruct abrupt changes in N_e (Figure S8): for example, instantaneous bottlenecks 2000 or 20,000 years ago are reconstructed as a gradual dip in N_e over a period of 5000 or 50,000 years, respectively. The PSMC also artificially infers an increase in N_e prior to the bottleneck. This increase in N_e was also apparent when applying the PSMC to data simulated under a range of bottleneck scenarios including the history we inferred for the pigs (Figure 8B). We therefore recommend caution when interpreting PSMC plots.

Conclusion

We have developed a method for inferring demographic size change from short sequence blocks for small numbers of samples. We show that the new method has higher power than inferences based on the genome-wide SFS for much larger samples. The central idea of the new framework is to summarize mutational configurations as blockwise site frequency spectra. This both results in little loss of information and enables the analysis of nontrivial samples of individuals and unphased data. We stress that the bSFS scheme could be used for inference under any model of population history. For example, one could use blockwise data to fit scenarios of population expansion, serial bottlenecks, or more general models of multiple merger coalescence (Coop and Ralph 2012). We also envisage extensions of this framework to explicitly compare the demographic histories of multiple, codistributed species.

Acknowledgments

We thank Nick Barton, Mike Hickerson, two reviewers, and associate editor Noah Rosenberg for comments on earlier versions of this manuscript and Martien Groenen and Hendrik-Jan Megens for access to the genome sequences of *Sus cebifrons* ahead of publication. This work was supported by funding from the United Kingdom (UK) Natural Environment Research Council (NERC) to Graham Stone and K.L. (NE/J010499/1) and a UK NERC fellowship to K.L. (NE/I020288/1).

Literature Cited

Barton, N. H., J. Kelleher, and A. M. Etheridge, 2010 A new model for extinction and recolonisation in two dimensions: quantifying phylogeography. Evolution 64(9): 2701–2715.

Bhaskar, A., and Y. S. Song, 2014 Descartes rule of signs and the identifiability of population demographic models from genomic variation data. Ann. Stat. 42(6): 2469–2493.

Bosse, M., H. J. Megens, O. Madsen, R. P. Crooijmans, O. A. Ryder *et al.*, 2015 Using genome-wide measures of coancestry to maintain diversity and fitness in endangered and domestic pig populations. Genome Res. (in press).

Chen, H., 2012 The joint allele frequency spectrum of multiple populations: a coalescent theory approach. Theor. Popul. Biol. 81(2): 179–195.

Chikhi, L., V. C. Sousa, P. Luisi, B. Goossens, and M. A. Beaumont, 2010 The confounding effects of population structure, genetic diversity and the sampling scheme on the detection and quantification of population size changes. Genetics 186: 983–995.

Coop, G., and P. L. Ralph, 2012 Patterns of neutral diversity under general models of selective sweeps. Genetics 192: 205–224.

Davey, J. W., P. A. Hohenlohe, P. D. Etter, J. Q. Boone, J. M. Catchen *et al.*, 2011 Genome-wide genetic marker discovery and genotyping using next-generation sequencing. Nat. Rev. Genet. 12(7): 499–510.

Edwards, A. W. F., 1972 *Likelihood*. Cambridge University Press, Cambridge, UK.

Excoffier, L., I. Dupanloup, E. Huerta-Sánchez, V. C. Sousa, and M. Foll, 2013 Robust demographic inference from genomic and SNP data. PLoS Genet. 9(10): e1003905.

Felsenstein, J., 1978 The number of evolutionary trees. Mol. Phylogenet. Evol. 27(1): 27–33.

Felsenstein, J., 2003 *Inferring Phylogenies*. Sinauer Associates, Sunderland, MA.

Frantz, L., J. Schraiber, O. Madsen, H. J. Megens, M. Bosse *et al.*, 2013 Genome sequencing reveals fine scale diversification and reticulation history during speciation in Sus. Genome Biol. 14(9): R107.

Frantz, L. A. F., O. Madsen, H. J. Megens, M. A. M. Groenen, and K. Lohse, 2014 Testing models of speciation from genome sequences: divergence and asymmetric admixture in island South-East Asian Sus species during the plio-pleistocene climatic fluctuations. Mol. Ecol. 23(22): 5566–5574.

Fu, Y. X., and W. H. Li, 1993 Statistical tests of neutrality of mutations. Genetics 133: 693–709.

Galtier, N., F. Depaulis, and N. H. Barton, 2000 Detecting bottlenecks and selective sweeps from DNA sequence polymorphism. Genetics 155: 981–987.

Gattepaille, L. M., M. Jakobsson, and M. G. Blum, 2013 Inferring population size changes with sequence and SNP data: lessons from human bottlenecks. Heredity 110(5): 409–419.

Griffiths, R., and S. Tavaré, 1998 The age of a mutation in the general coalescent tree. Stoch. Models 14(1-2): 273–295.

Groenen, M. A. M., A. L. Archibald, and H. Uenishi, C. K. Tuggle, Y. Takeuchi*et al.*, 2012 Analyses of pig genomes provide insight into porcine demography and evolution. Nature 491(7424): 393–398.

Gronau, I., M. Hubisz, B. Gulko, C. Danko, and A. Siepel, 2011 Bayesian inference of ancient human demography from individual genome sequences. Nat. Genet. 43(10): 1031–1035.

Gutenkunst, R. N., R. D. Hernandez, S. H. Williamson, and C. D. Bustamante, 2009 Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. PLoS Genet. 5(10): e1000695.

Haddrill, P. R., K. R. Thornton, B. Charlesworth, and P. Andolfatto, 2005 Multilocus patterns of nucleotide variability and the demographic and selection history of *Drosophila melanogaster* populations. Genome Res. 15(6): 790–799.

Harris, K., and R. Nielsen, 2013 Inferring demographic history from a spectrum of shared haplotype lengths. PLoS Genet. 9(6): e1003521.

Hearn, J., G. N. Stone, L. Bunnefeld, J. A. Nicholls, N. H. Barton *et al.*, 2014 Likelihood-based inference of population history from low-coverage de novo genome assemblies. Mol. Ecol. 23(1): 198–211.

Hey, J., and R. Nielsen, 2004 Multilocus methods for estimating population sizes, migration rates and divergence time, with applications to the divergence of *Drosophila pseudoobscura* and *D. persimilis*. Genetics 167: 747–760.

Hudson, R. R., 2002 Generating samples under a Wright-Fisher neutral model of genetic variation. Bioinformatics 18: 337–338.

Leblois, R., P. Pudlo, J. Néron, F. Bertaux, C. Reddy Beeravolu *et al.*, 2014 Maximum-likelihood inference of population size contractions from microsatellite data. Mol. Biol. Evol. 31(10): 2805–2823.

Li, H., and R. Durbin, 2011 Inference of human population history from individual whole-genome sequences. Nature 475(7357): 493–496.

Li, H., B. Handsaker, A. Wysoker, and T. Fennell; 1000 Genome Project Data Processing Subgroup *et al.*, 2009 The sequence alignment/map format and samtools. Bioinformatics 25(16): 2078–2079.

Lohse, K., and L. A. F. Frantz, 2014 Neanderthal admixture in Eurasia confirmed by maximum-likelihood analysis of three genomes. Genetics 196: 1241–1251.

Lohse, K., R. J. Harrison, and N. H. Barton, 2011 A general method for calculating likelihoods under the coalescent process. Genetics 58: 977–987.

Lohse, K., M. Chmelik, S. H. Martin, and N. Barton, 2015 Strategies for calculating blockwise likelihoods under the coalescent. bioRxiv**:** 016469.

Lucchini, V., E. Meijaard, C. H. Diong, C. P. Groves, and E. Randi, 2005 New phylogenetic perspectives among species of southeast asian wild pig (Sus sp.) based on mtDNA sequences and morphometric data. J. Zool. 266(1): 25–35.

Marth, G. T., E. Czabarka, J. Murvai, and S. T. Sherry, 2004 The allele frequency spectrum in genome-wide human variation data reveals signals of differential demographic history in three large world populations. Genetics 166: 351–372.

Mazet, O., W. Rodríguez, and L. Chikhi, 2015 Demographic inference using genetic data from a single individual: separating population size variation from population structure. bioRxiv**:** 011866.

McCormack, J. E., S. M. Hird, A. J. Zellmer, B. C. Carstens, and R. T. Brumfield, 2013 Applications of next-generation sequencing to phylogeography and phylogenetics. Mol. Phylogenet. Evol. 66: 526–538.

Moura, A. E., C. Janse van Rensburg, M. Pilot, A. Tehrani, and P. B. Best *et al.*, 2014 Killer whale nuclear genome and mtDNA reveal widespread population bottleneck during the last glacial maximum. Mol. Biol. Evol. 31(5): 1121–1131.

Nielsen, R., and M. A. Beaumont, 2009 Statistical inferences in phylogeography. Mol. Ecol. 18(6): 1034–1047.

Nordborg, M., 1998 On the probability of Neanderthal ancestry. Am. J. Hum. Genet. 63(4): 1237–1240.

Paudel, Y., O. Madsen, H. J. Megens, L. Frantz, M. Bosse *et al.*, 2013 Evolutionary dynamics of copy number variation in pig genomes in the context of adaptation and domestication. BMC Genomics 14(1): 449.

Polanski, A., and M. Kimmel, 2003 New explicit expressions for relative frequencies of single-nucleotide polymorphisms with application to statistical inference on population growth. Genetics 165: 427–436.

Rao, C. R., 1945 Information and the accuracy attainable in the estimation of statistical parameters. Bull. Calcutta Math. Soc. 37: 81–89.

Sjödin, P. E., A. Sjöstrand, M. Jakobsson, and M. G. Blum, 2012 Resequencing data provide no evidence for a human bottleneck in Africa during the penultimate glacial period. Mol. Biol. Evol. 29(7): 1851–1860.

Städler, T., B. Haubold, C. Merino, W. Stephan, and P. Pfaffelhuber, 2009 The impact of sampling schemes on the site frequency spectrum in nonequilibrium subdivided populations. Genetics 182: 205–216.

Tajima, F., 1983 Evolutionary relationships of DNA sequences in finite populations. Genetics 105: 437–460.

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

Terhorst, J., and Y. S. Song, 2015 Fundamental limits on the accuracy of demographic inference based on the sample frequency spectrum. Proc. Natl. Acad. Sci. USA 112(25): 7677–7682.

Tortereau, F., B. Servin, L. Frantz, H. J. Megens, D. Milan *et al.*, 2012 A high density recombination map of the pig reveals a correlation between sex-specific recombination and gc content. BMC Genomics 13(1): 586.

Voight, B. F., A. M. Adams, L. A. Frisse, Y. Qian, R. R. Hudson *et al.*, 2005 Interrogating multiple aspects of variation in a full resequencing data set to infer human population size changes. Proc. Natl. Acad. Sci. USA 102(51): 18508–18513.

Wakeley, J., 1998 Segregating sites in Wright's island model. Theor. Popul. Biol. 53(2): 166–174.

Wakeley, J., 1999 Nonequilibrium migration in human history. Genetics 153: 1863–1871.

Wall, J. D., 2003 Estimating ancestral population sizes and divergence times. Genetics 163: 395–404.

Yang, Z., 2002 Likelihood and Bayes estimation of ancestral population sizes in hominoids using data from multiple loci. Genetics 162: 1811–1823.

Communicating editor: N. A. Rosenberg

Appendix: Calculating the Site Frequency Spectrum

We denote the time during which there are p lineages present, t_p. Griffiths and Tavaré (1998, equation 1.3) have shown that for any model in which lineages are statistically exchangeable the expected site frequency $E[\xi_i]$ can be computed from $E[t_p]$ by considering the probability that branches that exist during t_p have i descendants:

$$E[\xi_i] = \frac{\theta}{i}\binom{n-1}{i}^{-1}\sum_{p=2}^{n-i+1}\binom{p}{2}\binom{n-p}{i-1}E[t_p], \quad 1 \le i \le n-1. \tag{A1}$$

We denote the vector of coalescence intervals t_p as $\underline{t_p}$.

We can write the GF of $\underline{t_p}$ as a sum over all possible ways that a sample can enter and exit the bottleneck and distinguish only the branches (or rather the corresponding ω variables) associated with each interval t_p:

$$\begin{aligned}
\psi_1[\omega_p] &= \frac{\Lambda_1}{\Lambda_1 + \lambda_{p_s} + \omega_{p_s}}\left(\prod_{r=p_s+1}^{n}\frac{\lambda_r}{\Lambda_1 + \lambda_r + \omega_r}\right) \qquad \text{if } p_s = 1, \psi_1[\omega_p] = 1 \\
\psi_2[\omega_p] &= \frac{\Lambda_2}{\Lambda_2 + \lambda_{p_e}}\left(\prod_{j=p_e+1}^{p_s}\frac{\lambda_j}{\Lambda_2 + \lambda_j}\right) \qquad \text{if } p_e = 1, \psi_2[\omega_p] = 1 \\
\psi_3[\omega_p] &= \left(\prod_{m=1}^{p_e}\frac{\lambda_m}{\lambda_m + \omega_m}\right) \qquad \text{if } p_e = 1, \psi_3[\omega_p] = 1 \\
\psi[\omega_p] &= \sum_{p_s=1}^{n}\sum_{p_e=1}^{p_s}\psi_1[\omega_p]\psi_2[\omega_p]\psi_3[\omega_p].
\end{aligned} \tag{A2}$$

Here, p_s and p_e are the numbers of ancestral lineages that enter and exit the bottleneck at T_1, respectively; n is the number of sampled lineages; and ψ_1, ψ_2, and ψ_3 give the GF before, during, and after the bottleneck, as in Equation 1. Multiplying by $(\Lambda_1\Lambda_2)^{-1}$ and inverting with respect to Λ_1 and Λ_2 gives the GF under the bottleneck model, $P[\omega_p]$, where event times are discrete. $E[t_p]$ is obtained from this as

$$E[t_p] = -\left(\frac{\partial\psi[\omega_p]}{\partial\omega_p}\right)_{\underline{\omega_p}=0}. \tag{A3}$$

The SFS can be obtained by inserting $E[t_p]$ into Equation A1 above.

GENETICS

Supporting Information
www.genetics.org/lookup/suppl/doi:10.1534/genetics.115.179861/-/DC1

Inferring Bottlenecks from Genome-Wide Samples of Short Sequence Blocks

Lynsey Bunnefeld, Laurent A. F. Frantz, and Konrad Lohse

DOI: 10.1534/genetics.115.179861

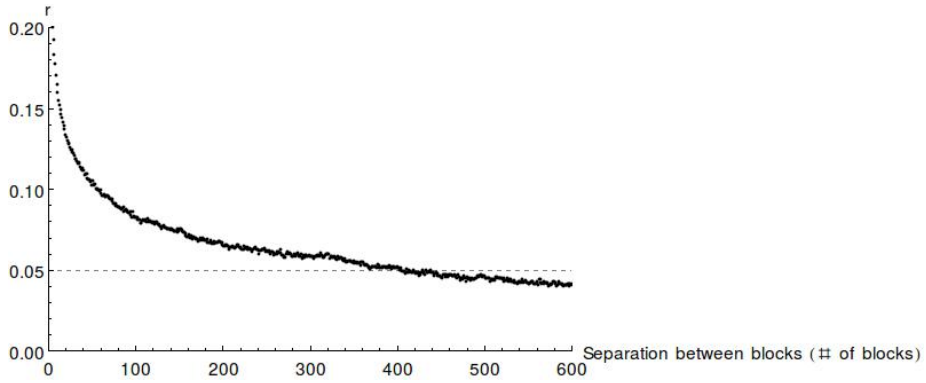

Figure S 1: Extent of correlation along the *Sus cebifrons* genome ($n = 5$). Blocks were aligned along chromosomes and Pearson's correlation coefficient (r) of total number of segregating sites per block was calculated for blocks at increasing distances from each other. The horizontal dashed line indicates where the correlation drops to 0.05.

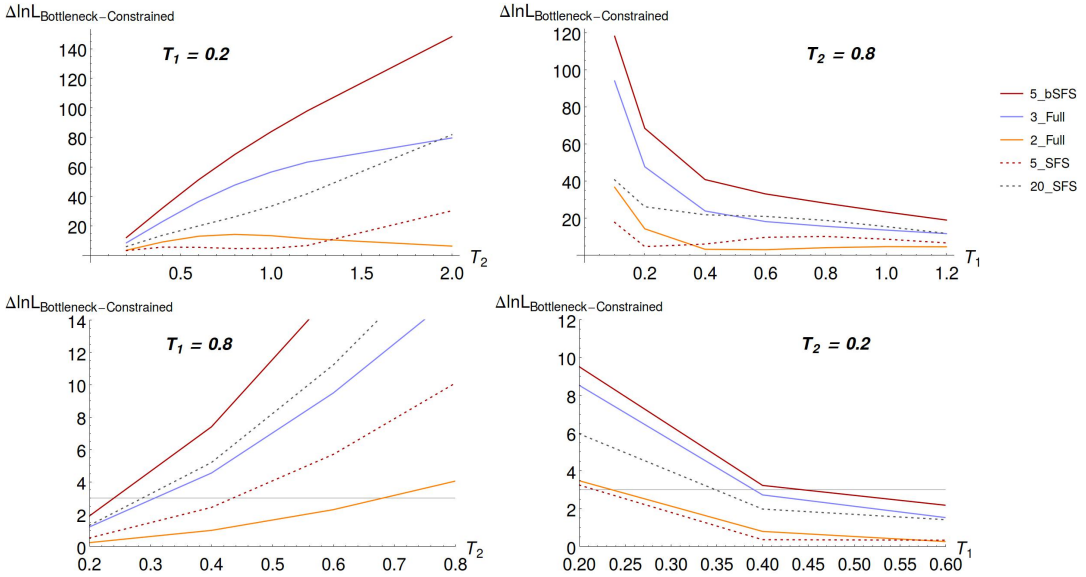

Figure S 2: The effect of conditioning on the same number of segregating sites per block irrespective of bottleneck parameters. Plotted is the expected difference in support ($E[\Delta lnL]$) between the bottleneck model and a null model of constant population size (as in Figure 3) as a function of bottleneck strength (T_2) (lefthand plots) and bottleneck start time (T_1) (righthand plots). Plots are based on analytic results for the likelihood and assuming 2,000 unlinked sequence blocks. Each $E[\Delta lnL]$ has been divided by the θ used to get the expected probabilities of mutational configurations. The horizontal line shows where $E[\Delta lnL]$) becomes insignificant (at α = 0.05) according to a likelihood ratio test with 2 degrees of freedom.

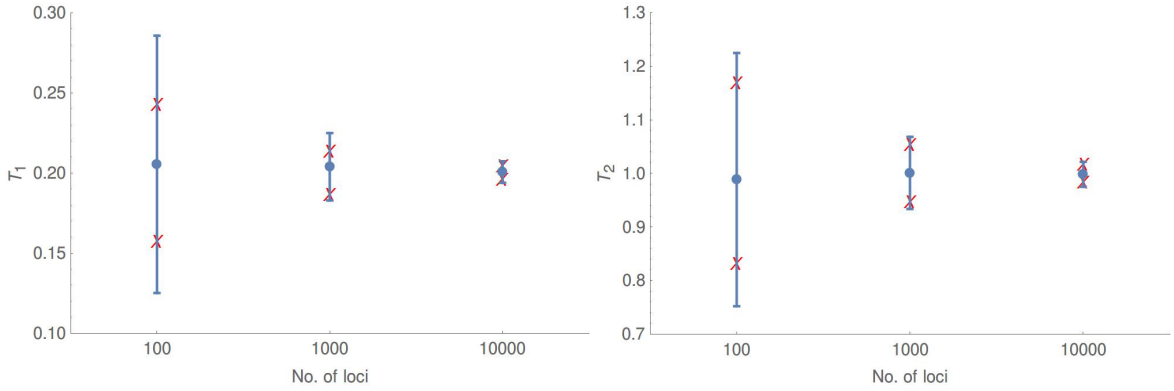

Figure S 3: Mean and standard deviation of 100 replicate model fits for 100, 1000 or 10000 simulated loci using the bSFS scheme and $n = 4$. The true parameters were $T_1 = 0.2$ and $T_2 = 1$. The red crosses correspond to the $E[SD]$ for the same number of loci.

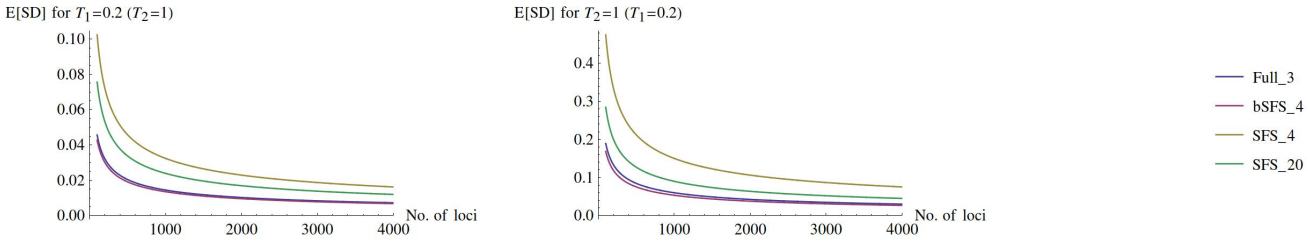

Figure S 4: $E[SD]$ as a function of number of loci for bottleneck parameters T_1 = 0.2 and T_2 = 1.

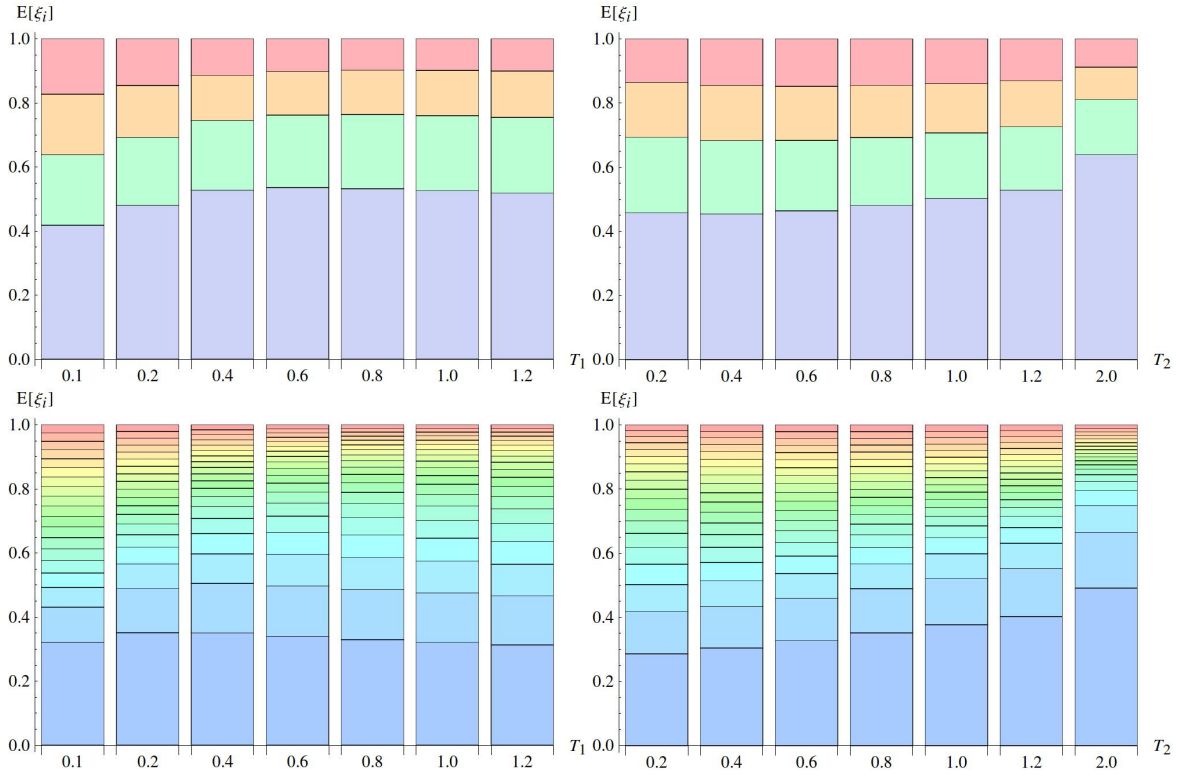

Figure S 5: The expected site frequency spectrum for n=5 and 20 under different bottleneck scenarios: top panel: n=5; bottom panel: n=20. Each bar represents a single stacked SFS for a particular bottleneck scenario with singletons at the bottom, doubletons above, and so on. Left-hand plots T_2 always 0.8; right-hand plots T_1 always 0.2.

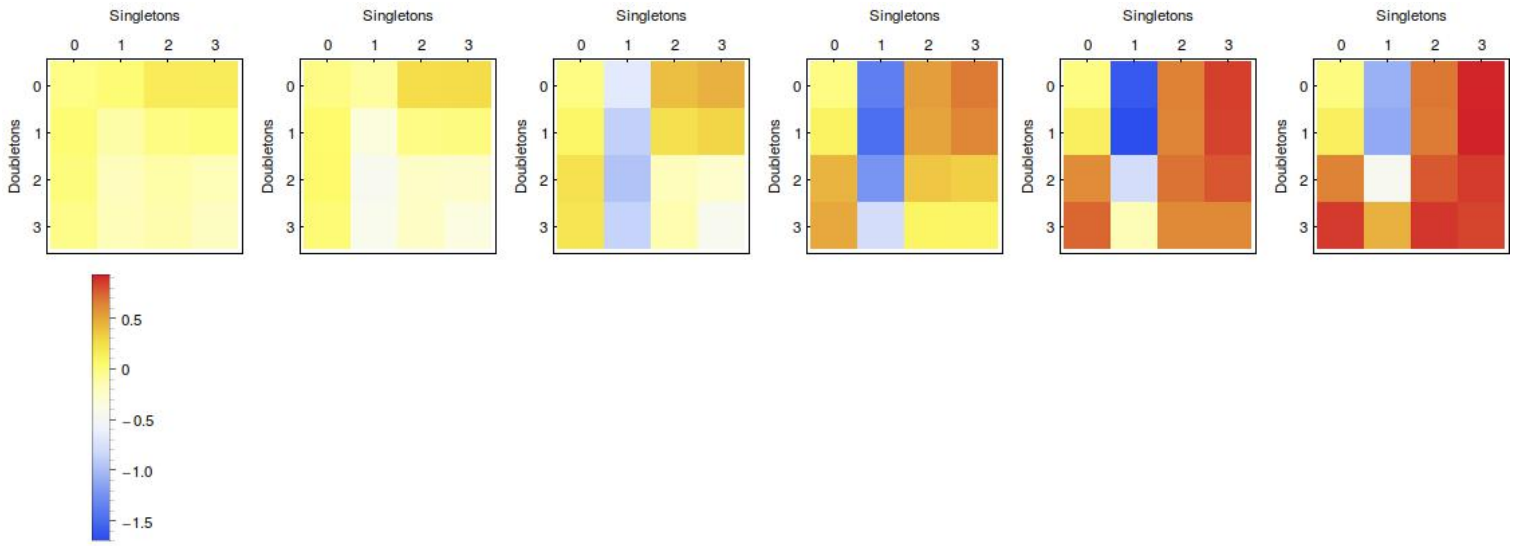

Figure S 6: Comparing mutational configurations expected from structured and bottlenecked data. The difference between the expected joint probabilities of numbers of singletons and doubletons given the inferred bottleneck parameter estimates for each migration rate and the joint probabilities of singletons and doubletons obtained from the structured data (normalised by the expected probabilities from the inferred parameters) for six migration rates (from left to right): 4 $N_0 m$ = 1, 0.5, 0.2, 0.1, 0.05 and 0.02 for n = 4. Where there is strong support for a bottleneck (Table 1), there is also a notable deficit of blocks with a single singleton mutation in the expected joint probabilities given the inferred parameters. This suggests that the bottleneck model does not completely characterise the structured data (does not account for a surplus of single singleton mutations) and that it may still be possible to distinguish population structure from size change.

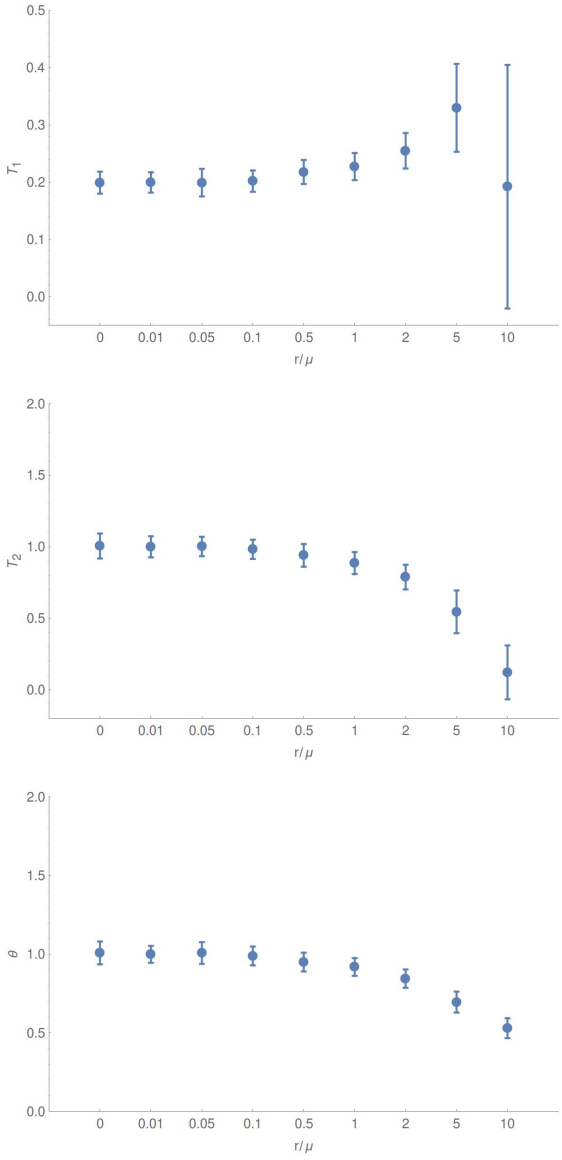

Figure S 7: Mean and standard deviation of 100 replicate model fits for 2000 simulated loci using the bSFS scheme, $n = 4$ and increasing r/μ. The true parameters were $T_1 = 0.2$, $T_2 = 1$ and $\theta = 1$.

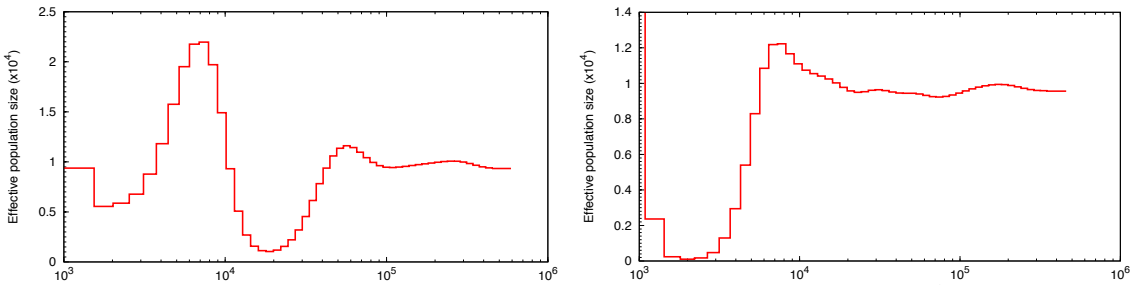

Figure S 8: PSMC plots from 100 simulated diploid sequences of 10Mb with $r/\mu = 0.4$, N_e = 10000, T_2 = 1 and T_1 = 0.2 (left) and T_1 = 0.02 (right).

Table S 1: Parameters used in the power analyses (figure 3). Given are T_1 and T_2 (both in units of $2N_e$ generations), θ adjusted to give one polymorphic site per block in a pair-wise comparison, $E[t_{total}]$ for a pair-wise sample, block size (bp) assuming pig-relevant parameters (see text) and the number of blocks of this size in a 1 Gb genome (before linkage correction).

T_1	T_2	θ	$E[t_{total}]$	Block size	No. of blocks
0.2	0.2	1.17	1.70	1174.3	851,589
0.2	0.4	1.37	1.46	1369.7	730,081
0.2	0.6	1.59	1.26	1585.1	630,874
0.2	0.8	1.82	1.10	1821.0	549,149
0.2	1.0	2.07	0.96	2072.7	482,463
0.2	1.2	2.34	0.86	2337.2	427,866
0.2	2.0	3.42	0.58	3423.8	292,072
0.1	0.8	1.99	1.00	1993.1	501,732
0.2	0.8	1.82	1.10	1821.0	549,149
0.4	0.8	1.59	1.26	1585.1	630,874
0.6	0.8	1.43	1.40	1433.1	697,785
0.8	0.8	1.33	1.51	1328.8	752,568
1.0	0.8	1.25	1.59	1254.1	797,419
1.2	0.8	1.20	1.67	1198.8	834,141

Table S 2: Parameter estimates for bottleneck models fitted to samples simulated with population structure and sampled using local sampling. We used *ms* to simulate an island model with symmetric migration at rate $M = 4N_e m$. We simulated local samples of four individuals from the same deme, part of a metapopulation of ten equally-sized demes (d). We chose a single metapopulation $\theta = 4N_e d\mu = 1$ (corresponding to a per deme θ of 0.1) and simulated 1,000,000 loci to obtain expected mutational configurations. Using the bSFS scheme, we maximised the lnL under both a null model of constant population size and a bottleneck history and calculated $E[\Delta lnL]$ between the two scenarios (shown here for 2000 unlinked blocks).

$4N_0m$	θ	T_1	T_2	ΔlnL
10	1.09	0.02	0.08	0.080
1	1.79	0.02	0.06	2.40
0.5	2.48	0.01	0.94	4.73
0.2	3.65	0.01	1.47	7.15
0.1	3.90	0.01	1.84	6.75
0.05	2.90	0.01	2.07	4.58
0.02	1.33	0.03	2.06	1.88
0.01	0.68	0.06	1.79	0.752

Table S3. Pro and cons of popular alternative methods

Method	Reference	Sample size requirements	Uses linkage information	Pros	Cons
Site frequency spectrum and summary statistics based on it (to detect departures from expectations of DNA polymorphism under neutrality)	Reviewed in e.g. Ramos-Onsins & Rozas 2002	Power generally increases with sample size up to around 20 individuals (depending on sample diversity)	No	Simple, fast, problems well-established in the literature	Different demographic or selective histories can lead to the same values of summary statistics. Ascertainment bias of SNPs if using SNP chip approaches.
Haplotype-based summary statistics (to detect departures from expectations of DNA polymorphism under neutrality)	Reviewed in e.g. Ramirez-Soriano et al. 2008	Power generally increases with sample size up to around 20 individuals (depending on sample diversity)	Yes	Simple, fast, problems well-established in the literature	Highly sensitive to how recombination is modelled.
∂a∂i	Gutenkunst et al. 2009	3-5 diploid individuals at a minimum, power depends on the specific history (Robinson et al. 2014)	No	Tests support for specific model in a composite likelihood framework	Does not make use of linkage information. Computationally intensive to obtain confidence intervals on parameter estimates.
Pairwise sequentially Markovian coalescent (PSMC)	Li & Durbin 2011	One diploid individual	Yes	Non-parametric, works with a single diploid individual	Non-parametric; restrictive data requirements
Skyline plots	Pybus et al. 2000 and extensions	Error decreases with more samples and more loci (Heled et al. 2008)	Yes	Non-parametric	Non-parametric; Very slow for multi-locus datasets
Approximate Bayesian Computation	Pritchard et al. 1999	Varies depending on summary statistics used	Can do	Provides a means to test support for more complex models than feasible with full likelihood approaches	Difficult to implement, inference is only as good as the data, but It is possible to be falsely confident of results. Very slow for large datasets and/or complex models.
Identity by descent approaches	Harris & Nielsen 2013, among others	At least one diploid individual	Yes	Good for founder events and recent bottlenecks	Restrictive data requirements; does not capture older demographic events well.

Additional references

Heled J, Drummond AJ (2008) Bayesian inference of population size history from multiple loci. *BMC Evolutionary Biology*, 8:289.

Pritchard J, Seielstad M, Perez-Lezaun A, Feldman M (1999) Population growth of human Y chromosomes: a study of Y chromosome microsatellites. *Molecular Biology & Evolution*, 16:1791–1798.

Pybus OG, Rambaut A, Harvey PH (2000) An integrated framework for the inference of viral population history from reconstructed genealogies. *Genetics*, 155:1429-1437.

Ramirez-Soriano A, Ramos-Onsins SE, Rozas J, Calafell F, Navarro A (2008) Statistical power analysis of neutrality tests under demographic expansions, contractions and bottlenecks with recombination. *Genetics*, 179:555-567.

Ramos-Onsins, SE & Rozas J (2002) Statistical properties of new neutrality tests against population growth. *Molecular Biology & Evolution*, 19:2092-2100.

Robinson JD, Coffman AJ, Hickerson MJ, Gutenkunst RN (2014) Sampling strategies for frequency-spectrum-based population genomic inference. *BMC Evolutionary Biology,* 14:254.