Approximate Likelihood Inference of Complex Population Histories and Recombination from Multiple Genomes

Champak R. Beeravolu[1,*], Michael J. Hickerson[1,2,3], Laurent A.F. Frantz[4] and Konrad Lohse[5]

[1] Biology Department, The City College of New York, New York, New York 10031, USA
[2] The Graduate Center, The City University of New York, New York, New York 10016, USA
[3] Division of Invertebrate Zoology, American Museum of Natural History, New York, New York 10024, USA
[4] Paleogenomics and Bio-Archaeology Research Network, Research Laboratory for Archeology and History of Art, University of Oxford, Oxford OX1 3QY, UK
[5] Institute of Evolutionary Biology, University of Edinburgh, King's Buildings, Edinburgh EH9 3FL, UK

* champak.br@gmail.com

Abstract

The wealth of information contained in genome-scale datasets has substantially encouraged the development of methods inferring population histories with unprecedented resolution. Methods based on the Site Frequency Spectrum (SFS) are computationally efficient but discard information about linkage disequilibrium, while methods making use of linkage and recombination are computationally more intensive and rely on approximations such as the Sequentially Markov Coalescent. Overcoming these limitations, we introduce a novel Composite Likelihood (CL) framework which allows for the joint inference of arbitrarily complex population histories and the average genome-wide recombination rate from multiple genomes. We build upon an existing analytic approach that partitions the genome into blocks of equal (and arbitrary) size and summarizes the polymorphism and linkage information as blockwise counts of SFS types (bSFS). This statistic is a richer summary than the SFS because it retains information on the variation in genealogies contained in short-range linkage blocks across the genome. Our method, ABLE (Approximate Blockwise Likelihood Estimation), approximates the CL of arbitrary population histories via Monte Carlo simulations form the coalescent with recombination and overcomes limitations arising from analytical likelihood calculations. ABLE is first assessed by comparing it to expected analytic results for small samples and no intra-block recombination. The power of this approach is further illustrated by using whole genome data from the two species of orangutan and comparing our inferences under a series of models involving divergence and various forms of continuous or pulsed admixture with previous analyses based on the SFS and the SMC. Finally, we explore the effects of sampling (different block lengths and number of individuals) and find that accurate inference of demography and recombination can be achieved with reasonable computational effort. Our approach is also notably adapted to unphased data and fragmented assemblies making it particularly suitable for model as well as non-model organisms.

Author Summary

In this paper, we make use of the distribution of blockwise SFS (bSFS) patterns for the inference of arbitrary population histories from mutliple genome sequences. The latter can be whole genomes or of a fragmented nature such as RADSeq data. Our method notably allows for the simultaneous inference of demographic history and the genome-wide historical recombination rate. Additionally, we do not require phased genomes as the bSFS approach does not distinguish the sampled lineage in which a mutation occurred. As with the Site Frequency Spectrum (SFS), we can also ignore outgroups by folding the bSFS. Our Approximate Blockwise Likelihood Estimation (ABLE) approach implemented in C/C++ and taking advantage of parallel computing power is tailored for studying the population histories of model as well as non-model species.

Introduction

The current surge in genomic data from a broad array of taxa provides huge potential for describing complex demographic and evolutionary histories across the tree of life. However, using this information in an efficient way to fit even very simple models of population history remains challenging. One such challenge is a complete description of the population history of genomic samples with both the ancestral process of coalescence and recombination, i.e. the ancestral recombination graph (ARG). However, as the correlation structure generated by the ARG is complex and the number of recombination and coalescent events in any stretch of genome generally exceeds the information (number of mutations) available to reconstruct these events, it is currently impossible to integrate demographic inference over all possible realizations of the ARG that are compatible with the data [1].

Current coalescent-based methods dealing with genomic data simplify this inherent problem and are broadly differentiated based on how they deal with linkage and account for recombination [2]. Methods based on the site frequency spectrum (SFS) [3–5] tend to ignore linkage information altogether. As the SFS is only a function of the expected length of genealogical branches [6, 7], likelihood computations are greatly simplified but come at the expense of inferential power. For example, Lapierre *et. al.* (*pers. comm.*) have recently shown that the SFS cannot distinguish between even simple one-parameter models of population growth in Human African populations and analytic studies show that many demographic histories are not identifiable using the SFS [8].

Some methods incorporate linkage by exploiting the haplotype structure while simplifying the coalescent with recombination. The Sequential Markov Coalescent (SMC, [9]) approximates recombination by modelling the sequential transitions between local genealogies along the genome as a Markov process [10, 11]. These methods are computationally intensive, mostly limited to relatively simple models [12] or small samples [10, 13, 14] and require good genome assemblies which are presently available only for a handful of species. While the simultaneous inference of recombination (using the SMC) and demography has been made possible via Hidden Markov Models (HMMs, [13, 14]), they remain limited to pairs of whole genomes and a restricted set of population histories.

Ignoring recombination altogether, *multi-locus* methods make use of relatively short blocks of sequence [15–19], thereby allowing the analysis of RAD data or fragmented genome assemblies that can now be generated for any species [20–22]. However, this approach is limited in several fundamental ways. First, the assumption of complete linkage over short distances (within blocks) restricts such analyses to very short blocks and introduces potential bias. Thus information contained in longer range linkage disequilibrium (LD), which for example is expected to result from historical

admixture or drastic changes in population size, is lost. Second, given the combinatoric explosion of the number of possible genealogies with sample size, analytic multi-locus methods that can efficiently deal with large data sets are necessarily restricted to small samples of individuals [19]. Further, similar methods which can incorporate recombination within loci and deal with larger samples of individuals are still too computationally intensive to scale to whole genome data [23].

Yet other inference methods focus solely on recombination and account for little or no demographic history [24, 25]. For example, Composite Likelihood (CL) estimates of recombination based on pairs of SNPs generally assume a single panmictic population of constant size [26, 27]. Extensions of these approaches, infer the fine scale variation in historic recombination along the genome [24, 28], though simplifying assumptions about the population demography can be misleading [29]. Recently, some progress has been made for incorporating known single population demographic events, such as bottlenecks and population growth [30]. However, it remains unclear how such approaches can be extended to complex histories such as subdivided populations undergoing gene flow or admixture.

To overcome these limitations, we introduce in this paper an approximate CL framework that allows for the joint inference of both arbitrary demographic histories and the average genome-wide recombination rate from multiple genomes. We build upon an existing analytic approach [17, 19] that partitions the genome into blocks of equal (and arbitrary) size and summarizes the genome-wide pattern of linked polymorphism as a frequency distribution of blockwise site frequency spectra. We refer to this data summary as the distribution of blockwise SFS (bSFS) configurations. The distribution of bSFS configurations is a straightforward extension of the SFS, yet it is a richer summary of sequence variation than the SFS because it retains information about the variation in genealogies which is contained in short-range linkage information. We implement a software framework named Approximate Blockwise Likelihood Estimation (ABLE) which approximates the bSFS using Monte Carlo simulations from the coalescent with recombination. Our method allows for arbitrary models of population history and overcomes the limitations of analytic likelihood calculations [19, 31] (based on the same blockwise data summary) by enabling the analysis of larger samples of genomes and including recombination within blocks as a free parameter.

We describe how the distribution of bSFS patterns can be approximated for samples from single and multiple populations both with and without recombination. We first assess the accuracy of this approximation by comparing it to analytic results for small samples and no intra-block recombination under simple demographic models including a size change in a single population, isolation with continuous migration (IM) between two populations and divergence and instantaneous admixture between three populations. We then demonstrate how our method performs on real data by re-analysing whole genome data from the two species of orangutan and compare our inferences with previously analyses (based on the SFS and the SMC [32, 33]) under a complex history involving a population split followed by a pulse of bidirectional admixture. Finally, we investigate how the sampling scheme (i.e. block length and number of genomes) affects demographic inference.

Results

The blockwise SFS (bSFS)

Consider a random sample of sequence blocks of fixed length. In practice, such sequence blocks (coloured segments in Fig. 1) may be obtained by partitioning an available reference genome [31, 34] or from reduced representation sequencing strategies, such as restriction site associated DNA (RADSeq, [20]).

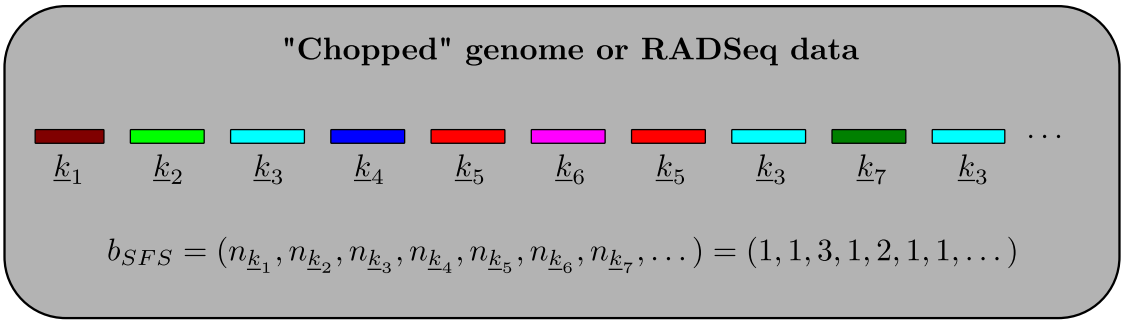

Fig 1. The blockwise SFS (bSFS). The bSFS is computed by partitioning sequences into short blocks, identifying mutation configurations (see Fig. 2) and noting their respective counts.

Given a sample of b genomes, the polymorphic sites in each sequence block can be summarized by a vector $\underline{k}$ of length $b-1$ (Fig. 2). For a single panmictic population, $\underline{k}$ is the SFS of the block and summarizes polymorphic sites within it as counts of singletons, doubletons *etc.* Following [31], the bSFS is essentially a frequency spectrum of site frequency spectra types across blocks (i.e. a histogram of histograms) and works as a straightforward extension of the SFS that accounts for the linked polymorphisms within sequence blocks of fixed length (Fig. 1).

The bSFS readily extends to samples from multiple populations where the entries of $\underline{k}$ are counts of mutation types defined by the joint SFS [7]. One advantage of the bSFS is that we only require *unphased* data as mutations are not distinguished based on unique branches but branch classes (singletons, doubletons, *etc.*, see Fig. 2). In the absence of outgroup information and/or to avoid biases due to errors when polarizing with distant outgroups, the bSFS may be *folded*. The analytical treatment of Lohse *et. al.* [19, 31] assumes non-recombining blocks and uses a recursion for the generation function of genealogies to derive the probability of bSFS configurations for small samples and simple demographic histories involving one or two populations. This allows for a direct comparison with the approximate composite likelihood developed here.

Approximating the bSFS

The bSFS can be approximated given an arbitrary population history while accommodating for intra-block recombination (see Methods). In summary, we use Monte Carlo sampling of blockwise data ARGs. The probability of bSFS configurations conditional on a particular simulated blockwise genealogy/ARG can be computed analytically. Thus, each simulation replicate contributes to the approximate likelihood of all configurations compatible with it. With the approximate CL at any point in parameter space we used a two step optimization procedure to hone in on the Maximum Likelihood Estimate (MLE) for a given demographic model (see Methods).

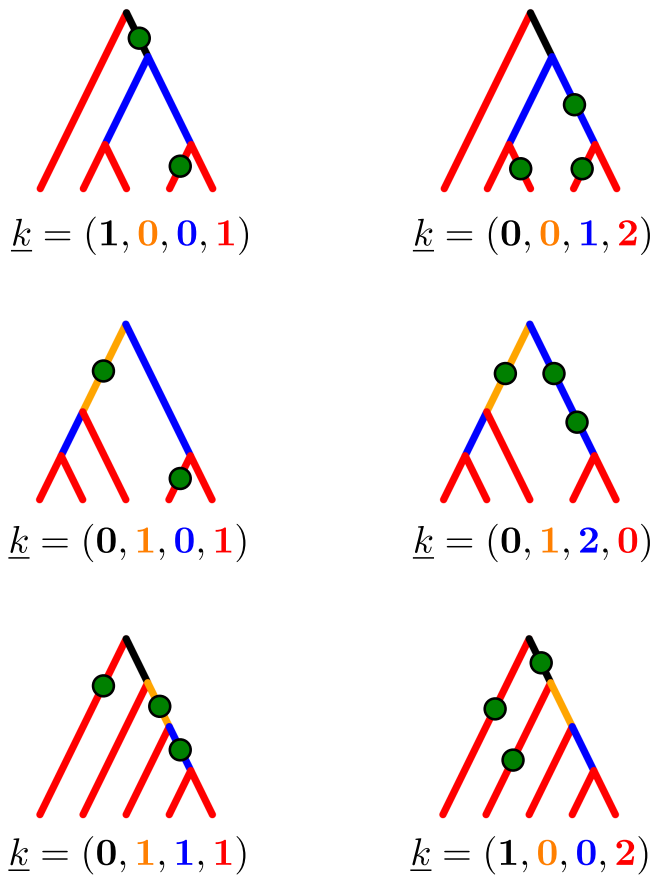

Fig 2. Example bSFS configurations for a sample from a single population. Genealogical relationships for a sample of size 5 can be generated by three types of topologies (top, middle and bottom rows). Ignoring information on the phase, the branches can be classified by the number of tips they are ancestral to; i.e. singletons (red), doubletons (blue), tripletons (orange) and quadrupletons (black). Further, mutations on these branches give rise to different bSFS configurations, i.e. vectors $\underline{k}$.

Comparison with analytic results

To study how the number of sampled ARGs summarized by the bSFS affects the convergence of the approximate CL to the analytical expectations, we considered small samples under three simple demographic models: a single population ($b = 4$, no outgroup) which doubled in effective size (N_e) at time $T = 0.2$ (Fig. 3a), a history of isolation between two populations A and B (at time $T = 1.2$) followed by continuous unidirectional migration (IM) at a rate $M = 4N_e m = 0.5$ migrants per generation from A to B ($b = 2$ per population, no outgroup, Fig. 3b) and a history of isolation between three populations ($b = 1$ per population with outgroup) with a recent instantaneous and unidirectional admixture (IUA) that transfers a fraction f of lineages from population A to B

(Fig. 3c). Parameters under the latter model were chosen to correspond roughly to the divergence and admixture history of humans and Neandertals: $f = 0.6$, $T_2 = 0.6$, $T_1 = 0.15$, $T_{gf} = 0.125$ [34]. All times are measured in $2N_e$ generations. For the sake of simplicity, the models in Figs. 3a,b assume identical N_e for all current and ancestral populations (see also [35,36]). The analytic solution for the bSFS under these models were previously obtained using an automation for the generating function implemented in *Mathematica* [19,31,34].

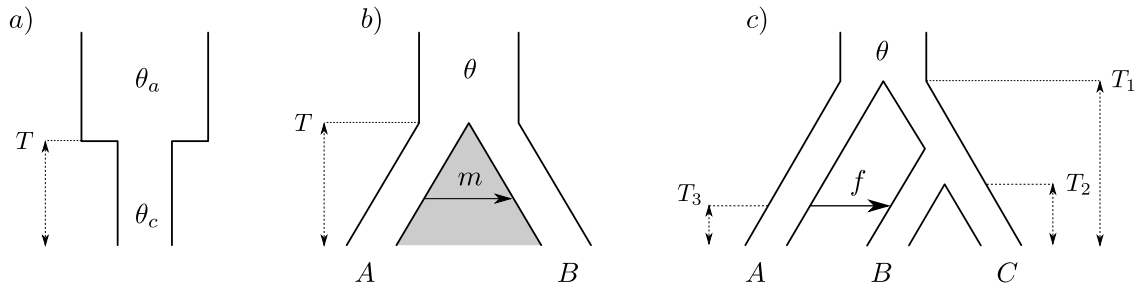

Fig 3. Three demographic models for which ABLE was compared against analytic expectations for the bSFS. (a) a single population with a sudden reduction in N_e, (b) IM: isolation between populations A and B followed by continuous unidirectional migration (from A to B) at rate M migrants per generation and (c) IUA: isolation between three populations A, B and C followed be unidirectional admixture of a fraction f from A to B. Analytical expectations for these models can be found in [19,31,34].

The Monte Carlo approximation to the distribution of bSFS configurations matches the analytic prediction extremely well (Fig. 4) even when only small samples of genealogies are used, e.g. 1000 simulated replicates. This is perhaps surprising, given that this sample size is on the same order as the number of unique bSFS configurations. For example, for a sample of $b = 2$ from the two populations IM model in Fig. 3b and counting up to $k_{max} = 4$ mutations per SFS-type and block, there are 396 unique bSFS configurations. Interestingly, the probability of bSFS configurations involving fixed differences (Fig. 4; green middle row) can be approximated accurately with fewer sampled genealogies than the probability of rare configurations that include shared polymorphism (Fig. 4; blue middle row). The latter are due to either incomplete lineage sorting or admixture. Likewise, for the IUA model, the probability of bSFS configurations involving mutations shared by A and B was harder to approximate than that of $(B, (A, C))$ configurations (blue vs. magenta in Fig. 4, bottom row).

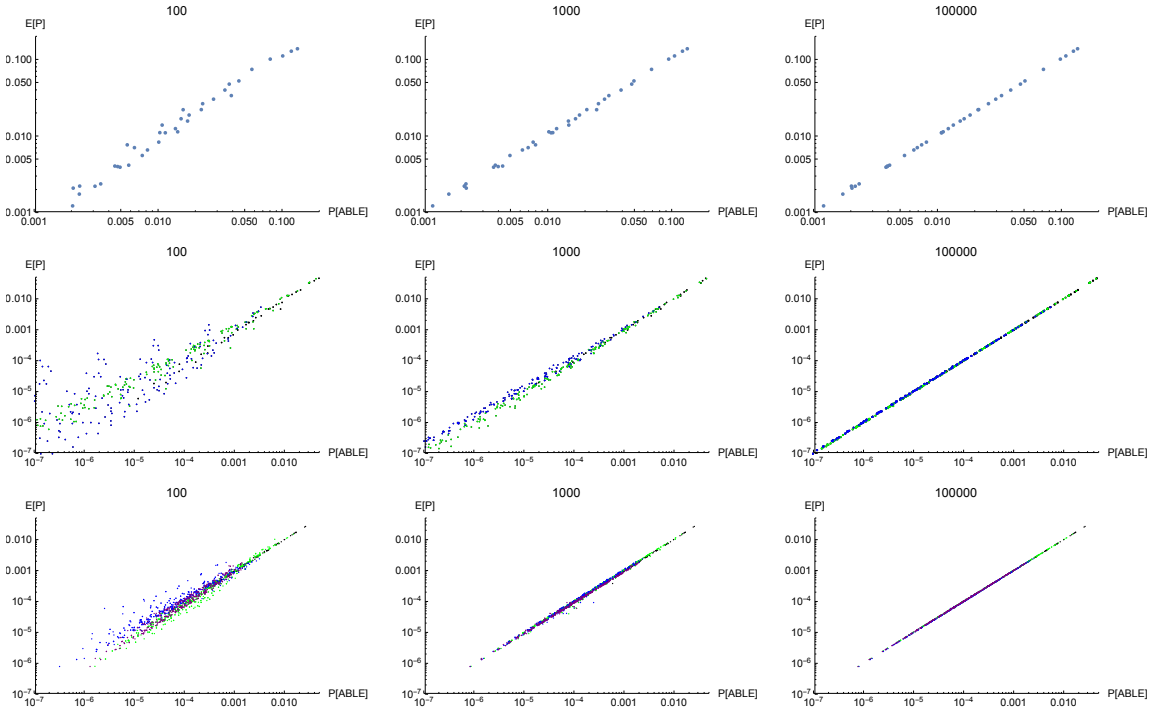

Fig 4. ABLE converges to the expected analytic probabilities with increasing numbers of simulated genealogies (100, 1000 and 10000). Results are shown for models specified in Figs. 3a,b c (top, middle and bottom row) assuming blocks of length $\theta = 0.6$, 1 and 2.4 respectively. For the IM model (Fig. 3b) (middle row) bSFS configurations with shared polymorphisms are shown in blue, those involving fixed differences in green, those with neither in black. For the IUA model (Fig. 3c, here in the bottom row) blocks with topology $(A,(B,C))$, $(C,(A,B))$ and $(B,(A,C))$ are shown in green, blue and magenta respectively. Topologically uninformative blocks are in black.

Orangutan analyses

To demonstrate the performance of our new bSFS approach using the ABLE framework on real data, we obtained whole genomes from previous analyses [32, 33] for the two species of orangutan (*Pongo pygmaeus* and *P. abelii*) which inhabit Borneo and Sumatra respectively. These sister taxa are an excellent test case given that their demographic history has been the subject of several previous analyses [13, 14, 23, 32, 33, 37]. We selected a subsample consisting of two diploid genomes per species (*i.e.* $b = 4$ per island) and partitioned the entire autosome into blocks of 2kb (on average 8.22 SNPs/block). After filtering, a total length of 163 Mb of sequence was retained in the final dataset (see Data processing for details), which consisted of 36,544 unique bSFS configurations. To investigate the effect of block size on our inference, all analyses were repeated using shorter blocks (500bp; 9,085 unique bSFS configurations) which were obtained by dividing each 2kb block.

To facilitate comparison with previous studies, we fitted a series of increasingly complex models of divergence with gene flow (Fig. 5) to these datasets and inferred demographic parameters along with the genome-wide recombination rate r under each model. All demographic models included an

instantaneous split at time T. We allowed effective population sizes N_e to differ between the two island populations and the ancestral population (M2 - M6). Additionally, we considered a model of divergence followed by exponential growth (or decline) in each population given by population specific growth rates α (M3). Asymmetric, bidirectional gene flow was modelled either as a continuous process occurring at a constant rate of $M = 4N_A m$ migrants per generation (M4 and M5), or as an instantaneous (bidirectional) admixture pulse affecting a fraction f of the admixted population (M6). We considered both an IM model with gene flow from time T to the present (M4) and a more complex history of isolation with initial migration (IIM) which assumes that migration ceases at time T_2 (M5) [38].

To convert time estimates (scaled in $4N_A$ generations) into absolute time, we followed [32] and assumed a generation time of 20 years and a mutation rate $\mu = 2 \times 10^{-8}$ bp^{-1} per generation.

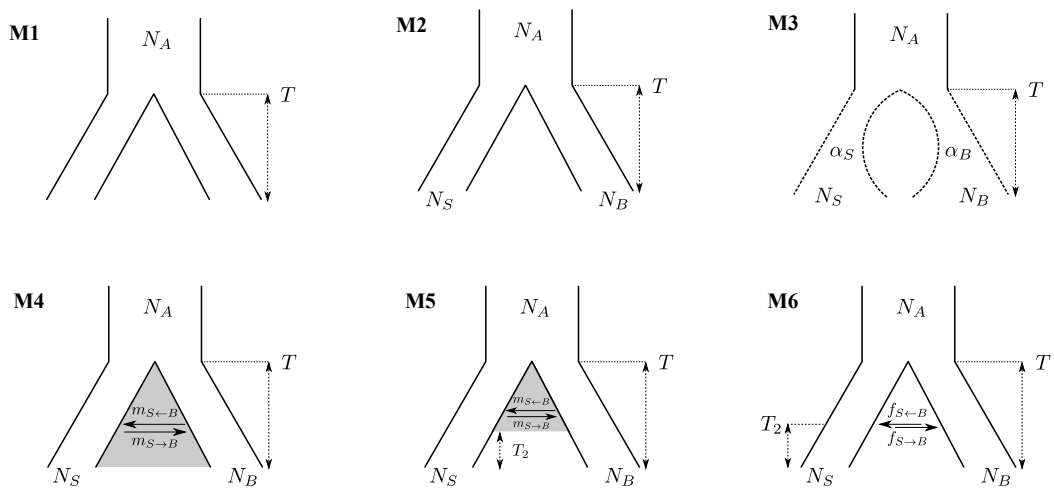

Fig 5. Models describing orangutan demography considered in this paper. All models describe a split between two populations at time T with an effective ancestral population size N_A. In M2 - M6, the current population sizes are additional free parameters. A model of divergence followed by exponential growth (or decline) in each population is considered in M3. Continuous asymmetrical gene flow since the time of split is considered up to the present or with a stopping time T_2 respectively in M4 and M5. Alternatively, an asymmetric *pulse* admixture at time T_2 is considered in M6.

As expected, model support increased with increasing complexity for nested models (i.e. M1 vs. M2 and M4 vs. M5) (Fig. 6 and Table 1). The only exception was the IM model (M4) which did not increase support compared to a strict divergence history (M2). Interestingly, we found greater support for instantaneous admixture (M6) compared to a history of Isolation and Initial Migration (IIM) up to a time T_2 (M5).

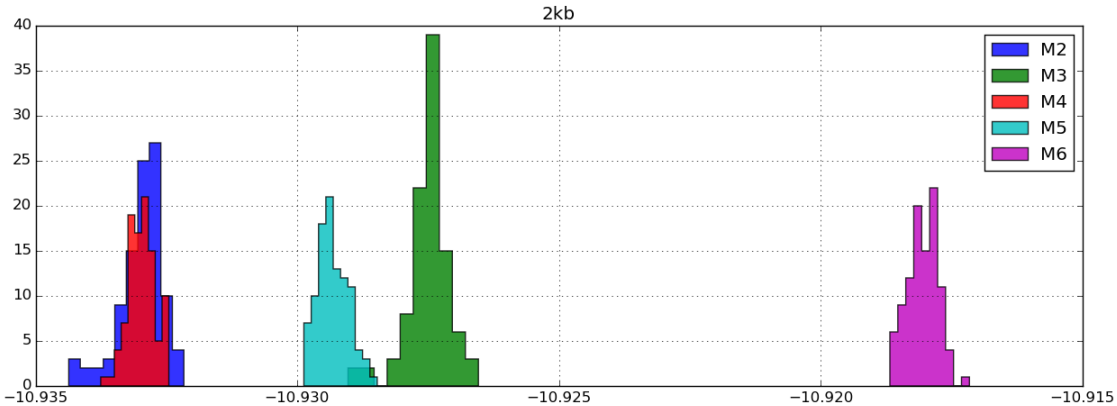

Fig 6. Relative fit of demographic models. Histograms of 100 evaluations of the Composite Likelihood (CL) using 1 million ARGs at the MLE for each model (Table 1). Note that the x-axis gives the *per block* CLs *i.e.* downscaled with respect to the number of blocks. Models further to the right fit relatively better than those to the left. Model M1 has the worst fit and is not shown (appears much further to the left)

Regardless of whether gene flow was modelled as a continuous process (M5) or a discrete admixture event (M6), our analyses reveal greater gene flow from Borneo into Sumatra than in the reverse direction. The maximum likelihood estimate (MLE) under M6 (Table 1), the best supported model, suggests a much higher admixture fraction ($f_{B\to S} \approx 0.2$) and no significant admixture in the reverse direction ($f_{S\to B} \approx 0.07$).

Likewise, independent of any particular model, the effective size of the Sumatran species was 2.5 fold greater than that of the Bornean species. This is in agreement with previous studies [32] and mirrors the relative diversity in each species as measured by Watterson's θ_W [39]. We respectively found $\theta_W = 2.19$, $\theta_W = 2.91$ and $\theta_W = 3.18$ in 2kb blocks for the Bornean, Sumatran and for both samples combined.

Table 1. Point estimates for the demographic history of orangutan species obtained from 2kb blockwise data (*cf.* Fig. 5)

Model	N_A	r	T	N_S	N_B	α_S	α_B	$4N_Am_{S\to B}$	$4N_Am_{S\leftarrow B}$	T_2	$f_{S\to B}$	$f_{S\leftarrow B}$	lnL
M1	18 200	1.58×10^{-8}	387 000										-907 477
M2	1 380	2.06×10^{-8}	294 000	22 100	8 610								-891 341
M3	2 180	2.09×10^{-8}	306 000	21 800	5 490	-0.003	-0.728						-891 308
M4	1 260	2.11×10^{-8}	320 000	22 300	8 210			0.025	0.000				-892 423
M5	1 280	1.87×10^{-8}	1 807 000	21 600	8 850			1.568	2.202	274 000			-892 225
M6	1 420	2.73×10^{-8}	816 000	22 400	8 910					295 000	0.121	0.267	-891 139

To determine the confidence in MLE under M6, we carried out a full parametric bootstrap using data simulated with linkage and determined 95 % Confidence Intervals (CI) as ±2 SD (standard deviations) across bootstrap replicates (see Methods and Fig. S6 for details). The CIs in Table 2 indicate that we have relatively greater power to infer more recent aspects of orangutan history (N_S, N_B and T_2) compared to the time of initial divergence (T) and the size of the common ancestral population (N_A). The admixture fraction from Sumatra to Borneo ($f_{S\to B}$) displayed

more uncertainty (CIs spanning an order of magnitude) than in the reverse direction which clearly excluded a zero level of admixture.

Table 2. 95% confidence intervals obtained via a parametric bootstrap. 100 datasets were simulated given the point estimates of the 2kb analysis and model M6 (*cf.* Table 1). Bootstrap replicates were generated by cutting long (0.5 Mb) contiguous sequences into 2kb blocks. The confidence intervals were calculated as 2 standard deviations on either side of the maximum likelihood estimate.

Parameter	MLE ± 2SD
N_A	1,180 - 1,670
$r \times 10^8$	2.5 - 3
T	695,000 - 936,000
N_S	21,200 - 23,600
N_B	8,400 - 9,420
T_2	284,000 - 306,000
$f_{S \rightarrow B}$	0.03 - 0.21
$f_{S \leftarrow B}$	0.2 - 0.33

Power and sensitivity analyses

We assessed how sampling affects the power of ABLE to infer population history and recombination in two ways: First, we repeated the orangutan analyses using shorter blocks (500bp). Second, we used simulations to investigate how sampling additional genomes per population affects power.

Block length

Comparing estimates based on 2kb blocks (1) to shorter 500bp blocks (Table S1) suggest that most, but not all, aspects of the inference were fairly robust to block length. As expected, shorter blocks led to a greater uncertainty in model and parameter estimates (Table S2). Importantly however, M6 was also the best fitting model for 500bp blocks and we found broad overlap in 95 % CIs of parameter estimates with the 2kb analysis.

Both the divergence time T and the genome-wide recombination rate r were poorly estimated with 500bp blocks. While the 2kb analyses resulted in fairly stable inferences for r ($\approx 2 \times 10^{-8}$ bp^{-1} per generation) that agree with recombination estimates for humans [40], the 500bp estimates were 2-4 times greater and had very wide 95 % CIs (Table S2). However, note that the 95 % CIs of T under 2kb and 500bp do overlap for periods < 1 Mya.

To test whether the method has any inherent bias to underestimate recombination with shorter blocks, we simulated block-wise data under model M6 and the r estimates obtained from the 2kb data (Table 1). Applying ABLE to these simulated datasets confirmed that the increase in r estimates for 500bp blocks (M6 (2dp), Table 3) reflects a genuine signal in the data (see Discussion).

Table 3. Power analysis of M6 (*cf.* Fig. 5) using 500bp blocks and 1/2/3 diploid genomes per population respectively.

Model	N_A	r	T	N_S	N_B	T_2	$f_{S\rightarrow B}$	$f_{S\leftarrow B}$	lnL
M6	1 250	2.50×10^{-8}	1 000 000	16 250	8 750	500 000	0.200	0.500	
M6 (1dp)	1 352	12.2×10^{-8}	1 732 345	13 343	8 404	439 427	0.389	0.295	-3.314791
M6 (2dp)	1 588	2.89×10^{-8}	1 080 289	16 360	9 055	502 330	0.283	0.426	-5.531729
M6 (3dp)	1 857	2.77×10^{-8}	1 095 628	15 655	8 598	492 806	0.177	0.446	-7.040805

The MLE for each sampling scheme is shown in gray, the true values used for the simulations in white.

Sample size

As expected, point estimates and power generally improved (Table 3 and Fig. 7) with increasing sample sizes. While some parameters, in particular r, appear non-identifiable with minimal sampling (a single diploid genome per species), all eight parameters of M6 are well estimated with just two or three diploid genomes. We observed a five fold improvement for r and an up to 2 fold improvement for demographic parameters when moving from a single diploid genome to 2 diploid genomes per population.

Perhaps surprisingly, Fig. 7 suggests that we can expect at best slight improvements in power when adding a third diploid genome per population. For example, estimates for N_A, T_2 and $f_{S\rightarrow B}$ respectively had a SD of 351, 18 457 and 0.121 for two diploid genomes – the sampling scheme used in the real data – compared to 257, 12 492 and 0.117 when three diploid genomes were sampled. As such, adding the third sample almost triples computation times (Fig. S7) and suggests that (at least in the case of orangutans) analysing a total of four diploid genomes is a good compromise between information gain and computational cost.

Discussion

We have introduced a flexible, efficient and widely applicable simulation-based approach to infer complex demographic histories while co-estimating genome-wide recombination rates under the full ARG. This overcomes the limitations of previous approaches that either ignore recombination [3, 4], use fixed estimates [23], or approximate recombination as a Markov process along the genome [10, 13, 14]. Using the bSFS as a data summary, ABLE captures linkage information at the scale of hundreds to thousands of base pairs across the genome, and allows researchers to efficiently fit realistic demographic models across the variety of genome scale data sets that are becoming available for a rapidly growing number of species.

The quick asymptotic convergence of the bSFS approximated by ABLE to the expected bSFS under various demographic scenarios (Fig. 4) in the absence of recombination is reassuring and distinguishes our method from related multi-locus approaches that integrate over possible genealogies locus by locus [23]. Moreover, the fit of the expected bSFS to data simulated under long range linkage and recombination is encouraging (Fig. S3) and warrants the use of the bSFS for purposes of demographic inference and beyond as outlined below.

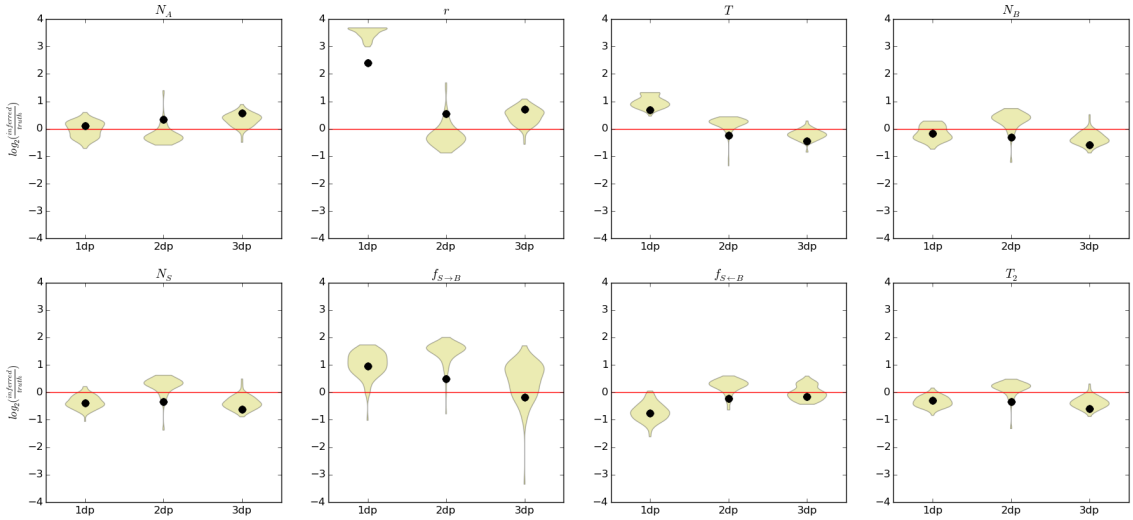

Fig 7. Parametric bootstrap results for model M6 and varying sample size. The x-axis indicates the number of diploid genomes per population. Accuracy for each parameter is given as the base-2 logarithm of the relative error (*i.e.* estimated value divided by the truth): 0 (red line) corresponds to the truth, increments of +1/-1 to two-fold over- or under-estimates. The black dots show the inferred point estimates (Table 3).

Orangutan history

The best fitting demographic model (M6) suggests that the two *Pongo* species diverged 650 - 1000 kya and experienced a burst of admixture around 240 kya. Given the Pleistocene history of periodic sea-level changes in South East Asia [41], such a scenario of secondary contact seems biogeographically more plausible than histories of continued migration. Interestingly, our estimates of the divergence time under M6 are consistent with previous estimates based on the SMC [10, 33] and also agree well with species splits estimated for other island-endemic mammals in SE Asia [41].

Overall, our results are in general agreement with previous analyses regarding the absence of recent gene flow ($<$ 250 kya) between Bornean and Sumatran orangutans (see Tab. S21-1 & S21-2 in [32]). Likewise, our inference of a larger N_e in Sumatran compared to Bornean orangutans agrees with relative measures of nucleotide diversity and previous analyses using various types of data [13, 23, 32, 37]. While we infer a contraction for the Bornean population under M3, in agreement with the simpler models used in [37], this issue lacks the required sampling resolution (see further below).

Reassuringly, the time of secondary admixture under M6 agrees with our estimated split time between the two *Pongo* species for simpler models M1-M4 (Table 1) which are similar to those considered by Locke *et. al.* [32]. Using the joint SFS ($\delta a \delta i$, [3]), Locke *et. al.* [32] estimate a species divergence time of 400 kya, which is somewhat older than our estimate (250-300 kya) under M1-M4. However, a similar difference in estimates has already been noted by the Hidden Markov Model approach of Mailund *et. al.* (see Suppelental Text S2, [14]) which models a simplified demography of speciation with gene flow and recombination using whole genome data.

Absolute model fit and the effect of selection

Like most demographic inference methods, ABLE assumes selective neutrality. Furthermore, efficient calculation or approximation of the bSFS relies on the assumption that blocks are statistically exchangeable, and so ignores heterogeneity in mutation and recombination rates.

We can visualize the absolute fit of our demographic model to the data by comparing the observed distribution of bSFS configurations to that expected under M6 (obtained using 50 million simulated blockwise ARGs). If the data were generated entirely by the inferred demographic history alone, we would expect the most common bSFS configurations to fit this expectation most closely (see Fig. S3). In contrast, figure 8 shows that, irrespective of which demographic model we assume, some aspects of the data are poorly captured. In particular, the most likely bSFS configurations under neutrality are over-represented in the data. This effect is strongest for configurations with few (or no) mutations (shown in blue), which is compatible with background selection [42] and/or positive selection affecting a fraction of blocks.

If background selection at linked sites is indeed responsible and a fraction of selective targets is weakly deleterious, we would expect increasingly biased N_e estimates going backwards in time. This could explain why we estimated a much smaller effective size for the ancestral population (Table 1 and Table S1 estimates are smaller than previous studies [13, 23, 32, 37]), while our estimates for the two current N_es agree fairly well [14]. Going further, it will be interesting to explore the possibility of jointly inferring background selection and demography using the bSFS [43].

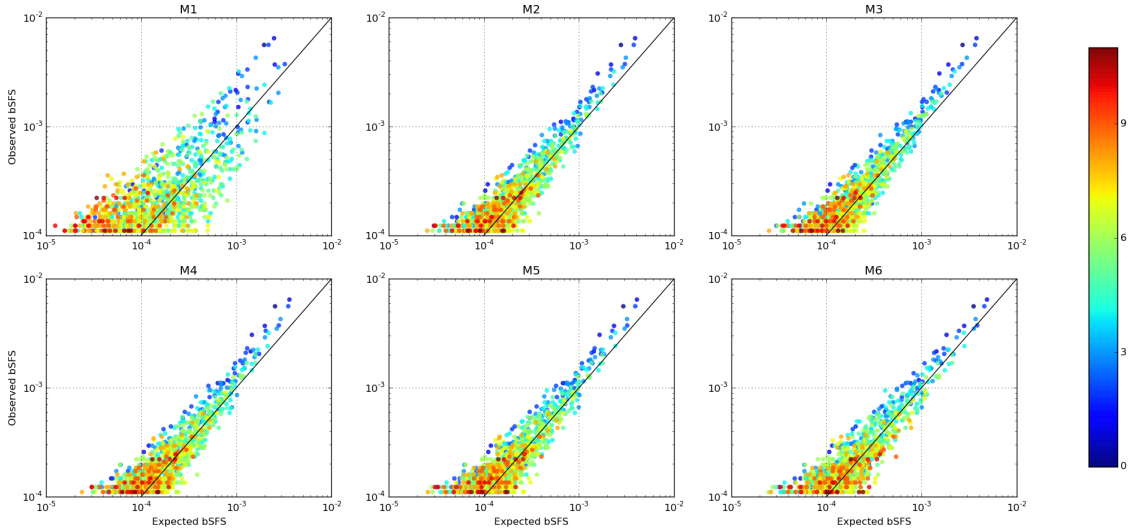

Fig 8. Absolute model fit to the observed 2kb bSFS. The expected bSFS (x-axis) was generated with ABLE using 50 million ARG's at the MLE for each model (Table 1) and plotted against the observed bSFS (y-axis) in the orangutan data. The diagonal black line indicate a perfect match between the expected and observed. bSFS configurations are colored by the total number of SNPs.

Effect of block length and sample size

An interesting property of the bSFS is that it collapses to the SFS in both the limits of minimal block length (one base) and maximal block length (all data in a single block). At both extremes, all linkage information is lost and so the information contained in the distribution of bSFS types must be optimal at an intermediate block length. A sensible starting point to finding an informative block length for a particular dataset is to ask at what length the number of unique bSFS configurations is maximised (Fig. S5). Crucially however, the most informative block length depends on μ/r and the demographic history and parameters in question. For example, a burst of recent admixture generates an excess of long blocks with excessive LD. Thus, an interesting direction for future work would be to integrate CL estimates based on the bSFS over a range of block sizes which should improve the power to infer recent admixture.

Our finding of larger r estimates for shorter blocks was surprising. Given that our method ignores heterogeneity in both r and μ, both of which increase auto-correlation across short distances, we expected to find the opposite, i.e. a decrease in r estimates for shorter blocks. Our power analyses showed that ABLE gives unbiased estimates of r even for short (500bp) blocks. Gene conversion, which inflates r at short blocks but has a diminishing effect on the bSFS for blocks that are longer than the typical conversion tract length of several hundred bases (see Table 2 in [44]) is a plausible explanation. It may be possible to use this dependence on block length to develop explicit estimators that can distinguish gene conversion and cross-over.

Even under a complex demography such as M6, our simulation-based power analyses indicate that most demographic parameters can be reasonably recovered with only a single diploid genome per population (Table 3). Increasing sample size to two diploid genomes more than halved the standard deviation in estimates for some parameters, most notably the recombination rate (Fig. 7). However, a further increase in sample size gave a negligible improvement, despite the considerable computational cost (Fig. S7 involved (the number of unique bSFS configurations increased more than three fold with three rather than two diploid genomes per population). This diminishing return with increasing sample size (in terms of sequences), given genome scale data, makes intuitive sense and has been commented upon since the pre-genomic era [45, 46]. Going backwards in time, larger samples in each species are likely to have coalesced down to a small number of lineages (see Fig. 3 in [46]) before the admixture event and could not have added much information about older demographic processes.

Limits to inference

While our choice of models was guided by previous knowledge of the demographic history of orangutans [14, 32, 33, 37], it remains to be determined what the limits of model complexity and identifiability are with our approach and to what degree the distribution of bSFS patterns overcomes the non-identifiability of the SFS [8, 47]. Unlike analytic likelihood calculations (*e.g.* [19]), there is no significant increase in computational cost with increasing model complexity when approximating the likelihood for a given point in parameter space. However, searching parameter space carries an obvious and rapidly increasing cost with greater model complexity. Like all approximate likelihood approaches, ABLE requires the user to make careful choices about the number of parameters, the number of genealogies to sample per point in parameter space, the search bounds for the MLE all of which are crucial elements of the optimization strategy [4]. In this regard, we suggest that simple pilot analyses varying some or all of the factors mentioned above (see Fig. S4 and Fig. S5) should help to inform the inference strategy.

It is also clear that, independent of the inference approach, the information in the data is finite so there must be a hard limit on how realistic a history one can hope to infer. Thus, the fact that ABLE can, in principle, be used for fittitng any demographic model puts the onus of constraining inference to scenarios that are both statistically identifiable and biologically interpretable on the user. Evaluating the relative fit of simpler nested models is an important sanity check on the limits of information in the data. For instance, comparing relative model fits between 2kb and 500 bp datasets (Fig. 6 and Fig. S1 respectively) highlights the apparent limits of the latter in discriminating between models. With regards to complex models not explored here, such as those covered by [37], these cannot be addressed with the data at hand (whole genomes from [32]) as sampling at finer spatial scales is needed to pick up the signal of substructure in both the Sumatran and Bornean populations.

The inferential approach implemented in ABLE makes use of the coalescent simulator ***ms*** [48] for sampling blockwise genealogies or ARGs. In principle ABLE can accommodate other simulators and is thus amenable to include additional processes such as linked selection [49, 50]. Another interesting avenue for further research is to apply approximate composite likelihoods based on the bSFS in sliding windows along the genome. Such an approach would not only help improve upon recombination maps for non-model organisms but could also provide a robust framework to identify *outlier regions* of the genome under positive selection and/or introgression from another species.

Methods

Approximating the bSFS

A single population

It is easiest to first consider the simpler case of non-recombining blocks and a sample of b genomes from a single panmictic population. We assume an arbitrary population history which is entirely described by a vector of parameters $\mathbf{\Theta}$. In the simplest case, $\mathbf{\Theta}$ consists of the scaled mutation rate $\theta = 4N_e\mu$, where N_e is the effective population size and μ the mutation rate per site per generation.

The branches of a given genealogy corresponding to our population sample can be partitioned into a vector $\underline{t}$ whose entries $t_i \in [t_1, t_2, \ldots, t_{b-1}]$ denote the total length of all branches with i descendants (Fig. 2). The probability of observing k_i mutations on a branch class t_i is given by a Poisson distribution with rate parameter $\theta t_i > 0$:

$$p(k_i \mid t_i) \sim \frac{(\theta t_i)^{k_i} e^{-\theta t_i}}{k_i!}. \tag{1}$$

Because mutations occur independently on different branch types, the joint probability of a seeing a specific bSFS $\underline{k}_j = \{k_{1,j}, k_{2,j}, \ldots, k_{b-1,j}\}$ in a sequence block j and for a given branch length vector $\underline{t}$ is then a product of Poisson distributions

$$p(\underline{k}_j \mid \underline{t}) = \prod_{i=1}^{b-1} p(k_{i,j} \mid t_i). \tag{2}$$

The likelihood $\mathcal{L}(\mathbf{\Theta})$ at a point in parameter space $\mathbf{\Theta} \in \mathbb{R}^+$ is calculated as

$$\mathcal{L}(\mathbf{\Theta}) \propto p(\mathcal{D} \mid \mathbf{\Theta}) = \sum_{\mathcal{G}} p(\mathcal{D} \mid \mathcal{G}, \mathbf{\Theta}) p(\mathcal{G} \mid \mathbf{\Theta}), \tag{3}$$

where $\mathcal{G}$ is the (unknown) genealogy and $\mathcal{D}$ the data [51]. Summarizing genealogies $\mathcal{G}$ by $\underline{t}$ and $\mathcal{D}$ by $\underline{k}_j$ and drawing $\mathcal{M}$ random samples of $\underline{t}$ from $p(\underline{t} \mid \mathbf{\Theta})$, the Monte Carlo approximation of Eq. 3 can be obtained

$$\hat{p}(\underline{k}_j \mid \mathbf{\Theta}) \approx \frac{1}{\mathcal{M}} \sum_{d=1}^{\mathcal{M}} p(\underline{k}_j \mid \underline{t}_d, \mathbf{\Theta}). \tag{4}$$

In theory, each block in a dataset might have a unique bSFS configuration. In practice however, for short blocks spanning that span a handful of SNPs (*e.g.* < 10), the number of observed bSFS configurations will be much smaller than the number of blocks. Assuming that blocks are equivalent and independent, that is, they have the same length, per base mutation and recombination rates and are unlinked, we can summarize the entire genome into *blockwise data* (Fig. 1) by counting the number of each unique bSFS type $n_{\underline{k}_j}$. Thus, the approximate joint composite log likelihood for a sample of n genomes is given as

$$\ln(\mathcal{L}(\mathbf{\Theta})) = \sum_{\underline{k}_j} \ln(\hat{p}(\underline{k}_j)) n_{\underline{k}_j}. \tag{5}$$

Multiple populations

The Monte Carlo approximation detailed above extends itself easily to the joint bSFS [7, 19] for multiple populations. Assuming a sample from X populations, the (unfolded) joint bSFS defines $(\prod_{x=1}^{X} b_x + 1) - 2$ site types, where b_x denotes the number of genomes sampled from population x. Some branches will be specific to a single population while others are shared between populations. Thus the vectors $\underline{t}$ and $\underline{k}$ have entries corresponding to the joint bSFS. Note that one specific configuration which we denote as k_0 refers to monomorphic blocks.

The Ancestral Recombination Graph

In the presence of recombination, the ancestry of a sequence block is described by the *ancestral recombination graph* $\mathcal{A}$ [1] which can be partitioned into a set of marginal genealogies corresponding to the non-recombining segments that make up the block [52]. Here, $\mathbf{\Theta}$ consists of the scaled mutation rate θ and the scaled recombination rate $\rho = 4N_e r$, where r is the recombination rate per site per generation. For a given $\mathcal{A}$, let S be the number of non-recombining blocks with respective (sequence) lengths $w_1, w_2, \ldots, w_S$ such that the size of the sequence block $L = \sum_{p=1}^{S} w_p$. Let $\underline{t}_p$ be the marginal branch length vector for each non-recombining segment p. The total length of the ith branch class over the graph $\mathcal{A}$ is then given by

$$t_{\{i,\mathcal{A}\}} = \frac{1}{L} \sum_{p=1}^{S} w_p t_{\{i,p\}} \tag{6}$$

Following Eq. 1, we can write the joint probability of observing a specific bSFS configuration over the entire recombining block as $p(k_{\{i,\mathcal{A}\}} \mid t_{\{i,\mathcal{A}\}}) \sim Poisson(k_{\{i,\mathcal{A}\}}; \theta t_{\{i,\mathcal{A}\}})$ (analogous to Eq. 2). Drawing $\mathcal{M}$ random samples of $\mathcal{A}$ from $p(\mathcal{A} \mid \mathbf{\Theta})$ and replacing $p(\underline{k}_j \mid \underline{t}_d, \mathbf{\Theta})$ with $p(\underline{k}_{\mathcal{A}_j} \mid \underline{t}_{\mathcal{A}_j}, \mathbf{\Theta}, \rho)$ in Eqs. 4 and 5 gives the approximate likelihood for a point in parameter space $\mathbf{\Theta}, \rho \in \mathbb{R}^{2+}$ (see also [23]). However note that $\mathbf{\Theta} \in \mathbb{R}^{2+}$ can be too restrictive a criterion for some parameters of complex demographies such as coefficients of exponential population expansion/contraction where $\alpha_S, \alpha_B \in \mathbb{R}$ (see Fig. 5).

An intriguing property of the distribution of bSFS patterns is that it converges to the SFS both in the limit of minimally short blocks (a single base) and maximally long blocks (all data are contanied in a single block). Thus, information must be maximised at some intermediate, optimal blocklength, which is a function both of the demographic history and the ratio of mutation and recombination events μ/r (Fig. S5).

Implementation

The ABLE implementation includes a seamless integration (invisible to the user) of the simulator ***ms*** [48] for sampling genealogies from $p(\mathcal{G} \mid \boldsymbol{\Theta})$ or $p(\mathcal{A} \mid \boldsymbol{\Theta})$. Crucially, for each simulated genealogy, we only record the total branch lengths of all SFS classes $t_{\{i,\mathcal{A}\}}$ in each ARG. This is a sum over marginal genealogies contributing to the ARG, each weighted by its length. From these we can tabulate the probabilities (conditional on $\mathcal{G}$) of all bSFS patterns compatible with that ARG. This task is extremely efficient compared to previous *multi-locus* methods that sample $\mathcal{G}$ separately for each locus (see [23, 53]).

Note that ABLE differs from previous, analytic calculations based on the distribution bSFS configurations in an important way. Lohse *et al.* [19] tabulate probabilities of all bSFS configurations up to a maximum number of mutations (k_{max}) in each category and lump all configurations $> k_{max}$ mutations.

$$p(k_i > k_{max} \mid t_i) = 1 - \sum_{k_i=0}^{k_{max}} p(k_i \mid t_i), \tag{7}$$

and

$$\begin{matrix} p(k_i = 0 \mid t_i) = 1 \\ p(k_i > 0 \mid t_i) = 0 \end{matrix} \; , \forall\, t_i = 0. \tag{8}$$

Bounding the table of mutational configuration in this way makes analytic computations feasible and ensures that the table of probabilities sums to unity. However, choosing k_{max} involves a trade-off between computational efficiency (low k_{max}) and information (high k_{max}). In contrast, ABLE only computes probabilities for mutational configurations that are observed in the data without setting any bounds on the space of possible configurations.

ABLE is implemented in C/C++, follows closely the command-line structure of ***ms*** [48] along with a brief configuration file with additional instructions and is freely available for download from `https://github.com/champost/ABLE`.

Data processing

Raw reads were downloaded from the NCBI Short Read Archive (SRA) for two individual genomes each from Borneo (B) and Sumatra (S): KB5405 (B, male, SRS009466), KB4204 (B, male, SRS009464), KB9258 (S, male, SRS009469), KB4361 (S, female, SRS009471). Mean Depth of Coverage was between 7.25 and 8.06 per individual. The alignment was performed using BWA-MEM, with a re-alignment step using GATK. For each sample we estimated a 95 % Depth of Coverage interval using BEDTools [54]. This was used to filter out any region that had too high coverage for SNP calling (*e.g.* copy number variable region). Genotype calls were then performed using a custom approach. In brief, a call was made when at least 4 reads of base quality (BQ) and mapping quality (MQ) > 30 were available. We also masked any non reference allele that were within 3bp of

an indel and checked for allele balance at heterozygous calls (e.g. each allele was covered by $> 20\%$ of the reads at that position). Thereafter, we binned the genome into non-overlapping blocks of fixed length $l = 2kb$ and sampled the first $0.8 \times l = 1600$ bases in each block that passing filtering in all individuals (a python script is available upon request). Blocks with fewer bases post filtering were excluded. The 500bp dataset was generated by partitioning each post-filtered 1.6kb block into four blocks of equal size. The 500bp and 2kb block datasets used in this study (in bSFS format and only genotypes) are available for download from the aforemetnioned website.

Optimization

Because ABLE approximates the likelihood function (Eq. 5) using Monte Carlo simulations – which induces some variability in the CL obtained (Fig. S4) – algorithms based on the gradient of the CL surface (*e.g.* [3, 11]) are not reliable [4]. In addition, due to the possibility of multiple local optima in the likelihood surface, we adopted a two-step, heuristic search strategy.

We initially searched parameter space between broad, user-specified non-linear bounds as part of a *global search* step. Search bounds during this step spanned several orders of magnitude for all parameters. Upper bounds of some parameters were set on the basis of simple data summaries: *e.g.* effective population sizes were bounded by Watterson's θ_W [39]. 50,000 ARGs were used to approximate the CL at each point in 10 replicate global searches. These were then used to set narrower bounds for a *local search* based on 500,000 ARGs/point which was repeated 20 times. In Table 1 and Table S1, we report the best MLEs whose likelihoods have been evaluated using 1M ARGs. For some models for which replicate local searches did not converge sufficiently, a second round of local searches was used.

ABLE employs several search algorithms implemented in the Non-Linear optimization library (NLopt version 2.4.2, [55]). Both global and local searches used the improved penalization provided by the Augmented Lagrangian algorithm [56] to navigate the non-linear delimitation of parameter space. A Controlled Random Search with rules for the evolution of a population of points given by the Local Mutation algorithm [57] was used for global searches. Local searches used the Subplex algorithm [58], a variant of the Nelder-Mead simplex with start points that were randomly chosen within the parameter bounds set by the global searches.

Finally, tolerances for terminating MLE searches were determined by probing the CL surface (*e.g.* Fig. S4). All settings used in this paper are available upon request from the authors.

Parametric bootstrap and power analysis

While the CL is a statistically consistent estimator of demographic parameters and recombination (in the limit of large data, [59]), it suffers from severe overconfidence because correlations between blocks due to their physical linkage are ignored. To obtain meaningful measures of confidence, we conducted a full parametric bootstrap under the best fitting model (M6) and parameter estimates (Table 1). We simulated 100 replicate datasets of 164 Mb each using a modified version of ***ms*** [48] and under the best model (*i.e.* M6) and MLE (Table 1 and Table S1). Blocks in each dataset were assumed to be completely linked (given our estimate of per site r) across 0.5 Mb stretches of sequence. These simulations represent an extreme case of linkage and are thus conservative. Indeed our real data contain large gaps between blocks especially due to the highly repetitive nature of the orangutan genome. As we wish to know the *local variability* of the bootstrap inferences around the MLE obtained from the orangutan data, we only carried out local searches for each bootstrap

replicate (using the boundaries and step sizes obtained in the analysis of real data, see above).

The power test exploring the effect of sample size (1 - 3 diploid genomes per population) was based on inferences using simulated data followed by a full parametric bootstrap. Given the computational effort required (see Fig. S7), we restricted our study to 500bp blocks with values for the demographic parameters chosen to represent the results inferred from the real data under M6 (Table 1 and Table S1). Parametric bootstrap datasets were generated with linkage exactly analogous to the bootstrap in the real data analysis.

Acknowledgments

We thank Stuart Baird, Lynsey Bunnefeld, Brian Charlesworth and Graham Stone for helpful discussions throughout this project. KL was supported by an Independent Research fellowship from the Natural Environment Research Council (NE/L011522/1). CBR and MJH were supported by grants from FAPESP (BIOTA, 2013/50297-0 to M.J.H. and Ana Carnaval), NASA through the Dimensions of Biodiversity Program, and National Science Foundation (DOB 1343578 and DEB-1253710 to MJH). LAFF was supported by a Junior Research fellowship from Wolfson College (University of Oxford) and a European Research Council grant (ERC-2013-StG-337574-UNDEAD). This research was supported by National Science Foundation Grants CNS-0958379, CNS-0855217, ACI-1126113 to the City University of New York High Performance Computing Center at the College of Staten Island. We also wish to thank Silicon Mechanics and their Research Cluster Grant program for the donation of the high-performance computing cluster that was used in support of this research.

Author Contributions

Conceived and designed the experiments: CRB, MJH, KL. Performed the experiments: CRB. Analyzed the data: CRB. Contributed reagents/materials/analysis tools: CRB, MJH, LAFF, KL. Wrote the paper: CRB, MJH, LAFF, KL.

References

1. Griffiths RC, Marjoram P. An ancestral recombination graph. Institute for Mathematics and its Applications. 1997;87:257.

2. Schraiber JG, Akey JM. Methods and models for unravelling human evolutionary history. Nature Reviews Genetics. 2015;.

3. Gutenkunst RN, Hernandez RD, Williamson SH, Bustamante CD. Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. PLoS Genet. 2009;5(10):e1000695.

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

5. Xue AT, Hickerson MJ. The aggregate site frequency spectrum for comparative population genomic inference. Molecular ecology. 2015;24(24):6223–6240.

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

7. Chen H. The joint allele frequency spectrum of multiple populations: a coalescent theory approach. Theoretical population biology. 2012;81(2):179–195.

8. Terhorst J, Song YS. Fundamental limits on the accuracy of demographic inference based on the sample frequency spectrum. Proceedings of the National Academy of Sciences. 2015;112(25):7677–7682.

9. McVean GA, Cardin NJ. Approximating the coalescent with recombination. Philosophical Transactions of the Royal Society of London B: Biological Sciences. 2005;360(1459):1387–1393.

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

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

12. Schiffels S, Durbin R. Inferring human population size and separation history from multiple genome sequences. Nature genetics. 2014;46(8):919–925.

13. Mailund T, Dutheil JY, Hobolth A, Lunter G, Schierup MH. Estimating divergence time and ancestral effective population size of Bornean and Sumatran orangutan subspecies using a coalescent hidden Markov model. PLoS Genet. 2011;7(3):e1001319.

14. Mailund T, Halager AE, Westergaard M, Dutheil JY, Munch K, Andersen LN, et al. A new isolation with migration model along complete genomes infers very different divergence processes among closely related great ape species. PLoS Genet. 2012;8(12):e1003125.

15. Hey J, Nielsen R. Multilocus methods for estimating population sizes, migration rates and divergence time, with applications to the divergence of Drosophila pseudoobscura and D. persimilis. Genetics. 2004;167(2):747–760.

16. Yang Z. Likelihood and Bayes estimation of ancestral population sizes in hominoids using data from multiple loci. Genetics. 2002;162(4):1811–1823.

17. Lohse K, Harrison R, Barton NH. A general method for calculating likelihoods under the coalescent process. Genetics. 2011;189(3):977–987.

18. Gronau I, Hubisz MJ, Gulko B, Danko CG, Siepel A. Bayesian inference of ancient human demography from individual genome sequences. Nature genetics. 2011;43(10):1031–1034.

19. Lohse K, Chmelik M, Martin SH, Barton NH. Strategies for calculating blockwise likelihoods under the coalescent. Genetics. 2016;accepted.

20. Davey JW, Hohenlohe PA, Etter PD, Boone JQ, Catchen JM, Blaxter ML. Genome-wide genetic marker discovery and genotyping using next-generation sequencing. Nature Reviews Genetics. 2011;12(7):499–510.

21. Hearn J, Stone GN, Bunnefeld L, Nicholls JA, Barton NH, Lohse K. Likelihood-based inference of population history from low-coverage de novo genome assemblies. Molecular ecology. 2014;23(1):198–211.

22. Harris SE, Xue AT, Alvarado-Serrano D, Boehm JT, Joseph T, Hickerson MJ, et al. Urbanization shapes the demographic history of a native rodent (the white-footed mouse, Peromyscus leucopus) in New York City. Biology letters. 2016;12(4):20150983.

23. Becquet C, Przeworski M. A new approach to estimate parameters of speciation models with application to apes. Genome research. 2007;17(10):1505–1519.

24. Auton A, McVean G. Recombination rate estimation in the presence of hotspots. Genome research. 2007;17(8):1219–1227.

25. Rasmussen MD, Hubisz MJ, Gronau I, Siepel A. Genome-wide inference of ancestral recombination graphs. PLoS Genet. 2014;10(5):e1004342.

26. Hudson RR. Two-locus sampling distributions and their application. Genetics. 2001;159(4):1805–1817.

27. Fearnhead P, Donnelly P. Approximate likelihood methods for estimating local recombination rates. Journal of the Royal Statistical Society: Series B (Statistical Methodology). 2002;64(4):657–680.

28. McVean GA, Myers SR, Hunt S, Deloukas P, Bentley DR, Donnelly P. The fine-scale structure of recombination rate variation in the human genome. Science. 2004;304(5670):581–584.

29. Johnston HR, Cutler DJ. Population demographic history can cause the appearance of recombination hotspots. The American Journal of Human Genetics. 2012;90(5):774–783.

30. Kamm JA, Spence JP, Chan J, Song YS. Two-Locus Likelihoods Under Variable Population Size and Fine-Scale Recombination Rate Estimation. Genetics. 2016;p. genetics–115.

31. Bunnefeld L, Frantz LA, Lohse K. Inferring Bottlenecks from Genome-Wide Samples of Short Sequence Blocks. Genetics. 2015;201(3):1157–1169.

32. Locke DP, Hillier LW, Warren WC, Worley KC, Nazareth LV, Muzny DM, et al. Comparative and demographic analysis of orang-utan genomes. Nature. 2011;469(7331):529–533.

33. Prado-Martinez J, Sudmant PH, Kidd JM, Li H, Kelley JL, Lorente-Galdos B, et al. Great ape genetic diversity and population history. Nature. 2013;499(7459):471–475.

34. Lohse K, Frantz LA. Neandertal admixture in Eurasia confirmed by maximum-likelihood analysis of three genomes. Genetics. 2014;196(4):1241–1251.

35. Durand EY, Patterson N, Reich D, Slatkin M. Testing for ancient admixture between closely related populations. Molecular biology and evolution. 2011;28(8):2239–2252.

36. Green RE, Krause J, Briggs AW, Maricic T, Stenzel U, Kircher M, et al. A draft sequence of the Neandertal genome. Science. 2010;328(5979):710–722.

37. Nater A, Greminger MP, Arora N, Schaik CP, Goossens B, Singleton I, et al. Reconstructing the demographic history of orang-utans using Approximate Bayesian Computation. Molecular ecology. 2015;24(2):310–327.

38. Wilkinson-Herbots HM. The distribution of the coalescence time and the number of pairwise nucleotide differences in a model of population divergence or speciation with an initial period of gene flow. Theoretical population biology. 2012;82(2):92–108.

39. Watterson G. On the number of segregating sites in genetical models without recombination. Theoretical population biology. 1975;7(2):256–276.

40. Kong A, Thorleifsson G, Gudbjartsson DF, Masson G, Sigurdsson A, Jonasdottir A, et al. Fine-scale recombination rate differences between sexes, populations and individuals. Nature. 2010;467(7319):1099–1103.

41. Frantz LA, Madsen O, Megens HJ, Groenen MA, Lohse K. Testing models of speciation from genome sequences: divergence and asymmetric admixture in Island South-East Asian Sus species during the Plio-Pleistocene climatic fluctuations. Molecular ecology. 2014;23(22):5566–5574.

42. Charlesworth B, Morgan M, Charlesworth D. The effect of deleterious mutations on neutral molecular variation. Genetics. 1993;134(4):1289–1303.

43. Ewing GB, Jensen JD. The consequences of not accounting for background selection in demographic inference. Molecular ecology. 2016;25(1):135–141.

44. Padhukasahasram B, Rannala B. Meiotic gene-conversion rate and tract length variation in the human genome. European Journal of Human Genetics. 2013;.

45. Pluzhnikov A, Donnelly P. Optimal sequencing strategies for surveying molecular genetic diversity. Genetics. 1996;144(3):1247–1262.

46. Felsenstein J. Accuracy of coalescent likelihood estimates: do we need more sites, more sequences, or more loci? Molecular biology and evolution. 2006;23(3):691–700.

47. Myers S, Fefferman C, Patterson N. Can one learn history from the allelic spectrum? Theoretical population biology. 2008;73(3):342–348.

48. Hudson RR. Generating samples under a Wright–Fisher neutral model of genetic variation. Bioinformatics. 2002;18(2):337–338.

49. Barton NH. The effect of hitch-hiking on neutral genealogies. Genetical Research. 1998;72(02):123–133.

50. Coop G, Ralph P. Patterns of neutral diversity under general models of selective sweeps. Genetics. 2012;192(1):205–224.

51. Felsenstein J. Phylogenies from molecular sequences: inference and reliability. Annual review of genetics. 1988;22(1):521–565.

52. Hudson RR. Properties of a neutral allele model with intragenic recombination. Theoretical population biology. 1983;23(2):183–201.

53. Tellier A, Pfaffelhuber P, Haubold B, Naduvilezhath L, Rose LE, Städler T, et al. Estimating parameters of speciation models based on refined summaries of the joint site-frequency spectrum. PloS one. 2011;6(5):e18155.

54. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010;26(6):841–842.

55. Johnson SG. The NLopt nonlinear-optimization package;. Available from: `http://ab-initio.mit.edu/nlopt`.

56. Birgin EG, Martínez JM. Improving ultimate convergence of an augmented Lagrangian method. Optimization Methods and Software. 2008;23(2):177–195.

57. Kaelo P, Ali M. Some variants of the controlled random search algorithm for global optimization. Journal of optimization theory and applications. 2006;130(2):253–264.

58. Rowan TH. Functional stability analysis of numerical algorithms. Ph.D. thesis, Department of Computer Sciences, University of Texas at Austin; 1990.

59. Wiuf C. Consistency of estimators of population scaled parameters using composite likelihood. Journal of mathematical biology. 2006;53(5):821–841.

Supporting Information

Tables

Table S1 Point estimates for the demographic history of orangutan species obtained from 500bp blockwise data (*cf.* Fig. 5)

Model	N_A	r	T	N_S	N_B	α_S	α_B	$4N_A m_{S\rightarrow B}$	$4N_A m_{S\leftarrow B}$	T_2	$f_{S\rightarrow B}$	$f_{S\leftarrow B}$	lnL
M1	18 000	3.34×10^{-8}	413 000										-1 653 152
M2	1 640	4.68×10^{-8}	314 000	21 900	9 090								-1 632 879
M3	1 970	4.39×10^{-8}	347 000	23 000	4 120	0.019	-1.132						-1 632 852
M4	1 920	5.81×10^{-8}	317 000	21 900	8 020			0.027	0.000				-1 633 028
M5	1 800	5.13×10^{-8}	2 800 000	21 200	8 530			4.961	6.148	289 000			-1 631 767
M6	1 480	8.67×10^{-8}	1 345 000	22 500	9 500					329 000	0.010	0.167	-1 628 548

Table S2 95% confidence intervals obtained via a parametric bootstrap. 100 datasets were simulated given the point estimates of the 500bp analysis and model M6 (Table S1). Bootstrap replicates were generated by cutting long (0.5 Mb) contiguous sequences into 500bp blocks. The confidence intervals were calculated as 2 standard deviations on either side of the maximum likelihood estimate.

Parameter	MLE ± 2SD
N_A	1,160 - 1,810
$r \times 10^8$	5.8 - 11.5
T	822,000 - 1,869,000
N_S	19,400 - 25,700
N_B	8,100 - 10,890
T_2	285,000 - 374,000
$f_{S\rightarrow B}$	0 - 0.06
$f_{S\leftarrow B}$	0.08 - 0.26

Figures

Fig. S1 Relative fit between models using 500bp blocks. Histograms of 100 evaluations of the Composite Likelihood (CL) using 1 million ARG's and at the MLE corresponding to each model (Table Table S1). Note that the x-axis represents the *per block* CLs *i.e.* downscaled with respect to the number of blocks found in each dataset. Models further to the right are a relatively better fit than those to the left. Model M1 is ignored (appears much further to the left) as it has the worst fit among all models.

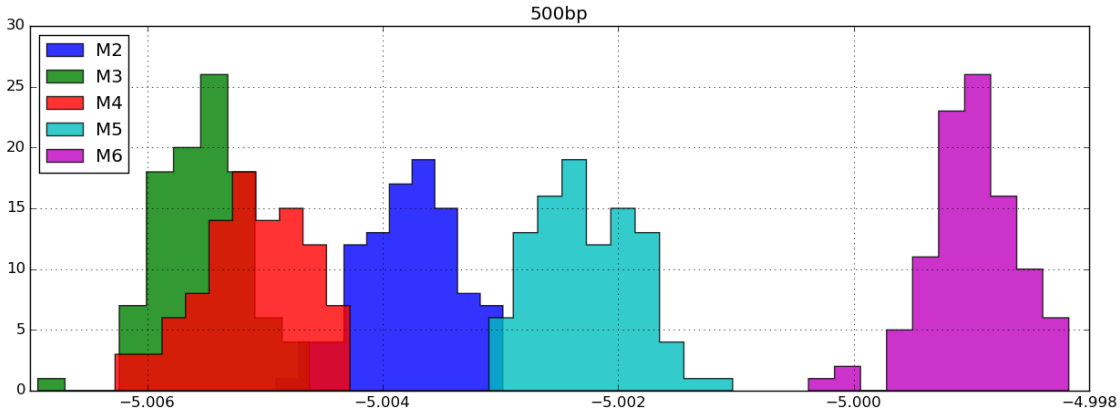

Fig. S2 Absolute model fit to the observed 500bp bSFS for the most common configurations. The expected bSFS (x-axis) was generated with ABLE using 50 million ARG's at the MLE corresponding to each model (Table S1) and plotted against the observed bSFS (y-axis) obtained from the orangutan data. Only those bSFS configurations which have been observed at least 32 times in the 500bp dataset have been shown. The colour legend indicates the number of SNPs associated with each configuration of the bSFS. Those points that fall on the black line ($x = y$) indicate a perfect match between the expected and observed.

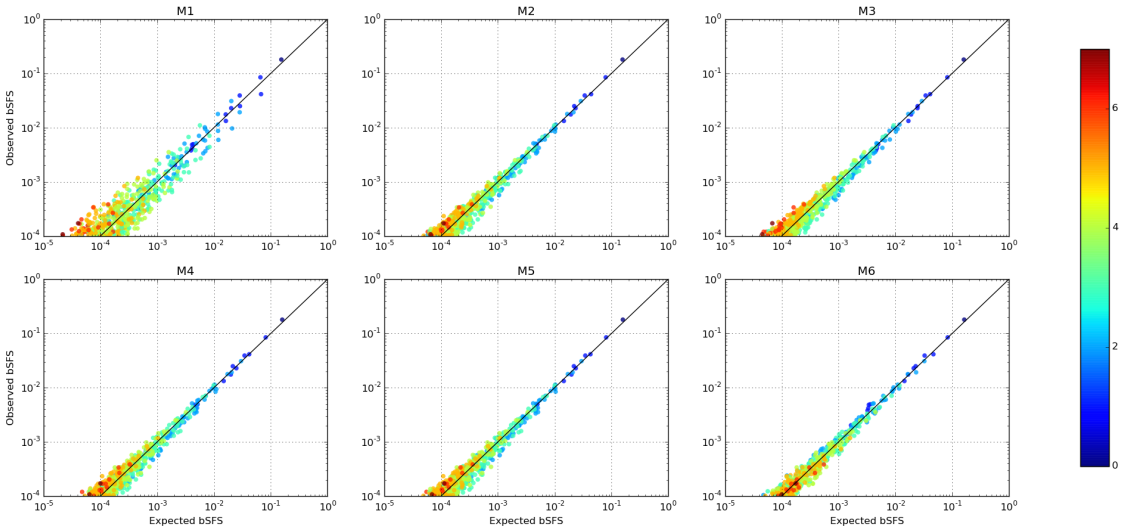

Fig. S3 Effect of long range linkage on the absolute model fit using simulations. Two datasets were simulated under M6 using long (0.5 Mb) contiguous sequences totalling 200 Mb (cut into 500bp and 2kb blocks respectively) and with the inferred MLEs from the orangutan dataset (Table S1 and Table 1 respectively). The expected bSFS (x-axis) was generated by sampling 50 million ARG's at the inferred MLEs and using the same values as for the simulated bSFS (y-axis). Only the most common bSFS configurations have been shown and the colour legend indicates the number of SNPs associated with each configuration of the bSFS. Those points that fall on the black line ($x = y$) indicate a perfect match between the expected and simulated.

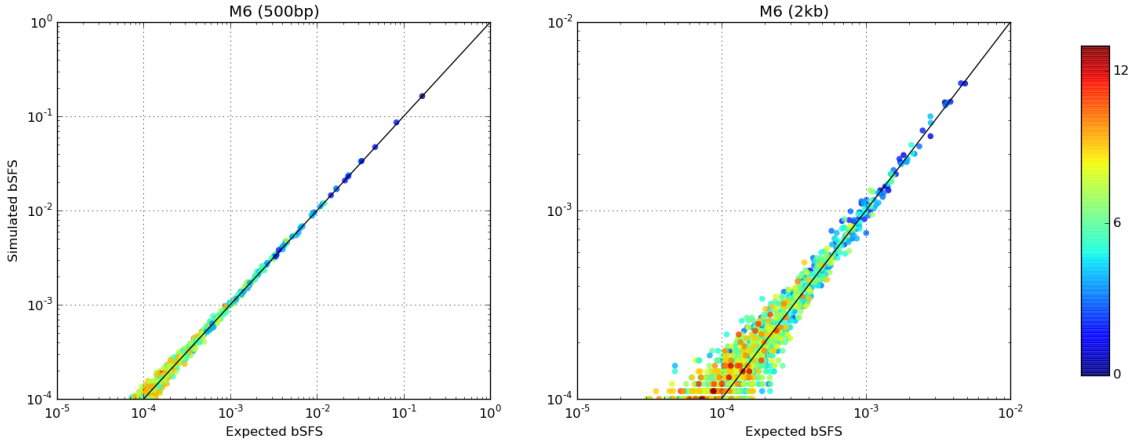

Fig. S4 Accuracy of the log likelihood (LnL) approximation for M6. The LnL was evaluated at the MLEs obtained for the 500bp (Table S1) and 2kb datasets (Table 1) for an increasing number of sampled ARGs (x-axis). Violin plots show the entire spread around the respective averages (red dots) over 100 log likelihood evaluations. Note that the y-axis represents the *per block* CLs *i.e.* downscaled with respect to the number of blocks found in each dataset.

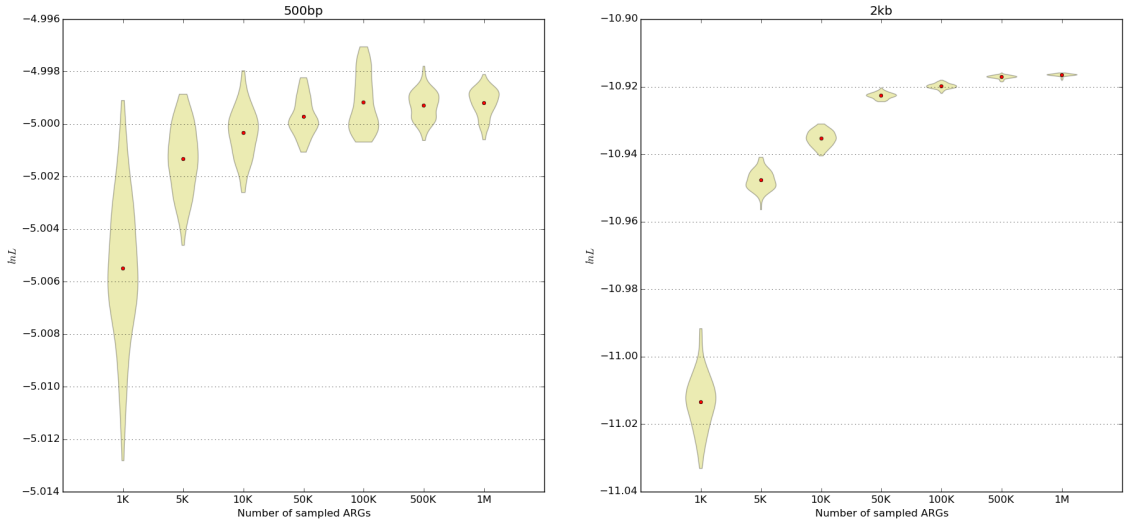

Fig. S5 The number of unique bSFS configurations for different block sizes. Data sets of a total length of 200Mb sized were simulated as independent blocks of a fixed length (x-axis) under model M6 estimated from the 2kb MLE data (see Table 1). For each block length the number of unique bSFS configurations is plotted on the y-axis. For this history, the number of configurations defined by the bSFS is maximized for blocks between 1kb - 10kb.

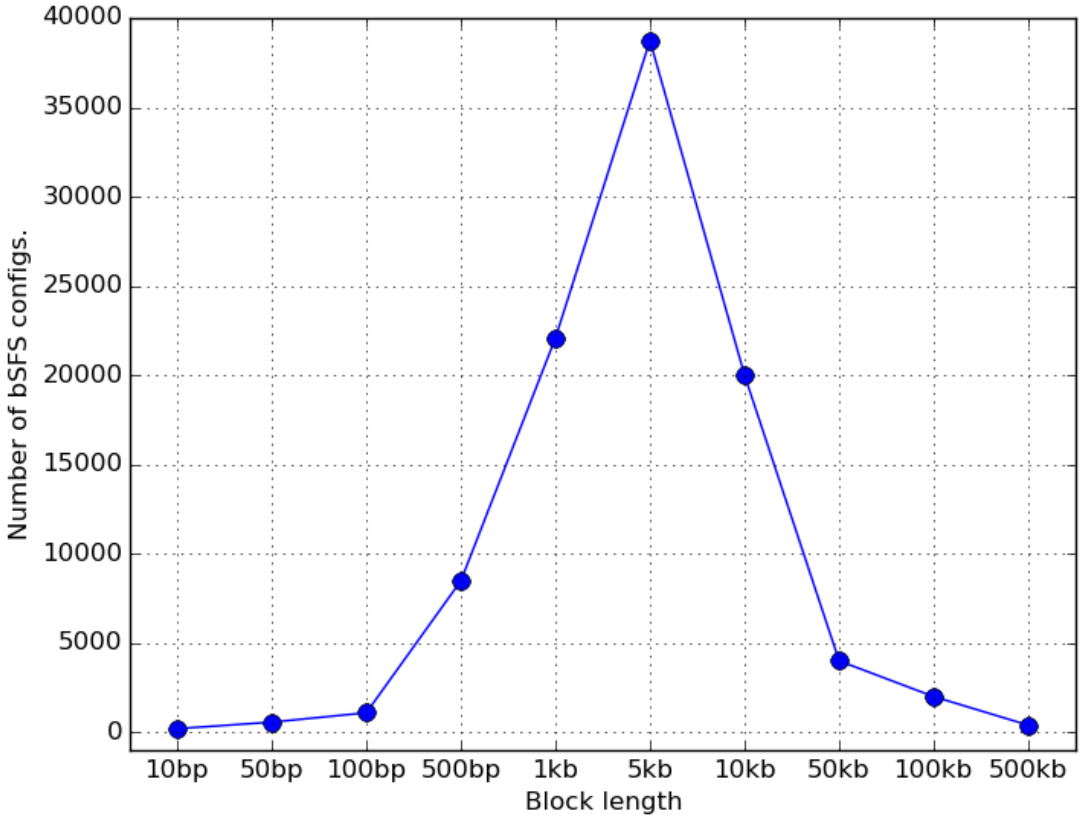

Fig. S6 Correlations between the parametric bootstrap estimates obtained for the 2kb dataset (Table 2). The scatter plots in the lower and upper half of the figure (are symmetrically identical) show the pairwise correlations among parameter estimates under M6. Density plots for each parameter are respectively given along the diagonal.

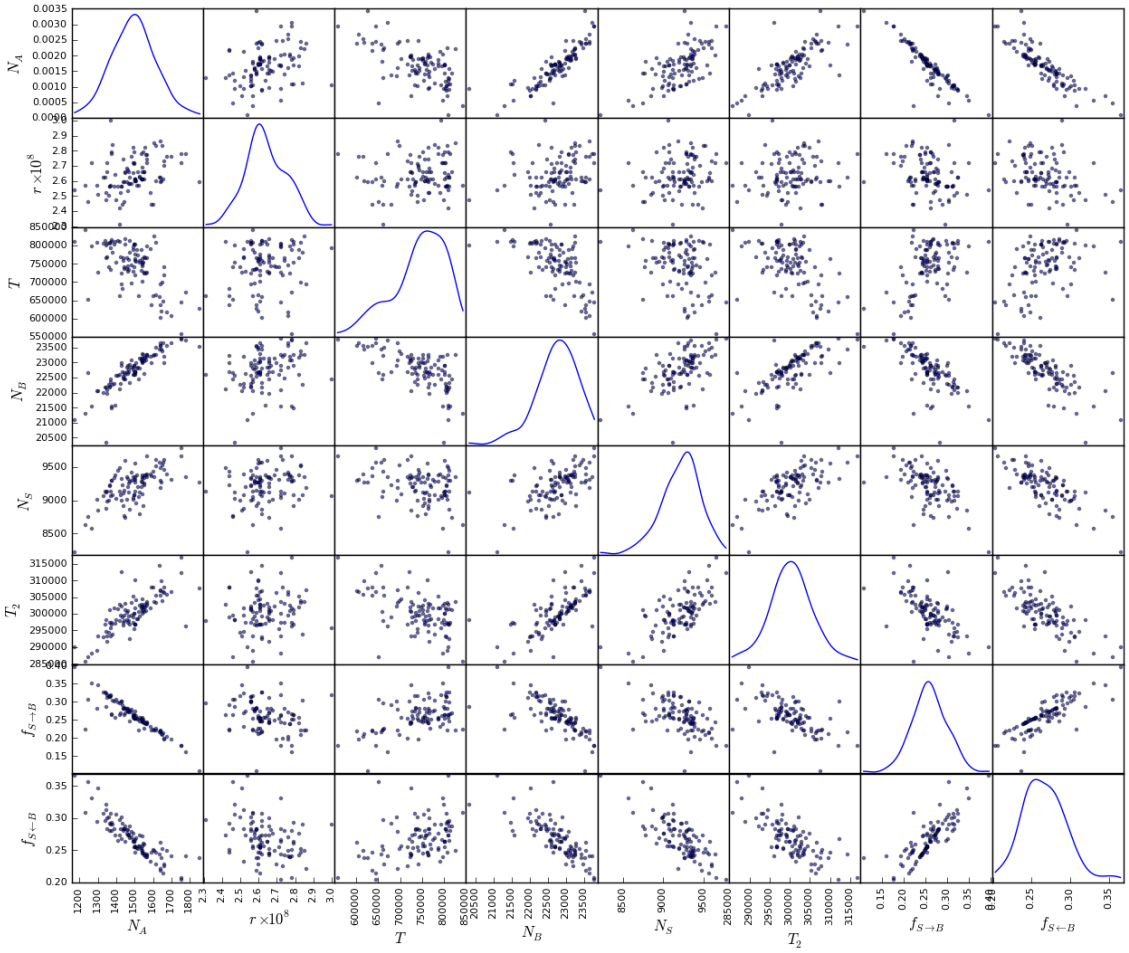
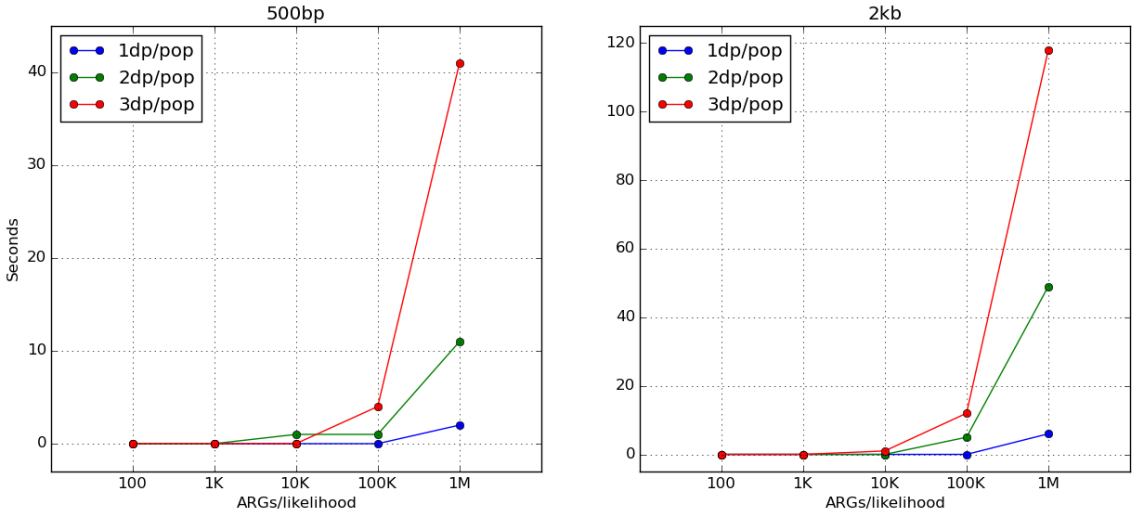
Fig. S7 Computation times with varying sample and block sizes. Simulated datasets were generated by varying the number of diploid genomes per population under model M6 (for values used, see Table 3). Two types of block sizes were considered (500bp and 2kb) by keeping the whole genome size fixed at 200Mb. A single likelihood was evaluated at the true MLE for each sampling configuration and block size by varying the number of sampled ARGs (x-axis). Times have been approximated to the nearest second.