Berit Lindum Waltoft^{1,2,3} / Asger Hobolth⁴

Non-parametric estimation of population size changes from the site frequency spectrum

¹ Bioinformatics Research Centre, Aarhus University, C.F. Møllers allé 8, 8000 Aarhus C, Denmark, Phone: +45 87165763, E-mail: berit@econ.au.dk

² National Centre for Register-based Research, Department of Economics and Business, Aarhus University, Fuglesangs allé 26, 8210 Aarhus V, Denmark, E-mail: berit@econ.au.dk

³ The Lundbeck Foundation Initiative for Integrative Psychiatric Research, iPSYCH, Aarhus, Denmark, E-mail: berit@econ.au.dk

⁴ Bioinformatics Research Centre, Aarhus University, 8000 Aarhus C, Denmark, E-mail: asger@birc.au.dk

Abstract:

Changes in population size is a useful quantity for understanding the evolutionary history of a species. Genetic variation within a species can be summarized by the site frequency spectrum (SFS). For a sample of size n , the SFS is a vector of length $n - 1$ where entry i is the number of sites where the mutant base appears i times and the ancestral base appears $n - i$ times. We present a new method, CubSFS, for estimating the changes in population size of a panmictic population from an observed SFS. First, we provide a straightforward proof for the expression of the expected site frequency spectrum depending only on the population size. Our derivation is based on an eigenvalue decomposition of the instantaneous coalescent rate matrix. Second, we solve the inverse problem of determining the changes in population size from an observed SFS. Our solution is based on a cubic spline for the population size. The cubic spline is determined by minimizing the weighted average of two terms, namely (i) the goodness of fit to the observed SFS, and (ii) a penalty term based on the smoothness of the changes. The weight is determined by cross-validation. The new method is validated on simulated demographic histories and applied on unfolded and folded SFS from 26 different human populations from the 1000 Genomes Project.

Keywords: Coalescent theory, population size, regularization, site frequency spectrum

DOI: 10.1515/sagmb-2017-0061

1 Introduction

The site frequency spectrum (SFS) is one of the most popular summary statistics of genetic variation within a species. We denote the observed SFS by $\xi = (\xi_1, \xi_2, \dots, \xi_{n-1})$ such that ξ_i is the number of segregating sites with i derived alleles among the n sampled sequences. In this paper we describe a novel non-parametric method CubSFS for estimating the changes in population size based on the site frequency spectrum (SFS). Polanski, Bobrowski, and Kimmel (2003) and Polanski and Kimmel (2003) provide a method for calculating the expected SFS given the changes in population size. We aim to solve the inverse problem: Given an observed SFS, we want to estimate the changes in population size. Myers, Fefferman, and Patterson (2008) showed that the solution to this problem is in general not unique [but see Bhaskar and Song (2014)], and we therefore need to regularize or constrain the changes in population size.

We assume a Wright-Fisher model of reproduction and the infinite sites mutation model [e.g. Wakeley (2009)]. The population size $N(r)$ is the number of diploid individuals in the population at generation r . We use scaled time t where one time unit corresponds to $2N(0)$ generations. In this time $\lambda(t) = N(0)/[N(2N(0)t)]$ is the coalescent rate at time t . The integrated intensity is defined as

$$\Lambda(t) = \int_0^t \lambda(u) du \quad \text{for } t \geq 0.$$

The main aim is to estimate the intensity $\lambda(t)$. This aim is achieved by finding the integrated intensity $\Lambda = \{\Lambda(t) : t \geq 0\}$ that fits the data and at the same time is regular. It is straightforward to transform the intensity to changes in population size by using $N(r) = N(0)/\lambda(r/(2N(0)))$.

Berit Lindum Waltoft, Asger Hobolth are the corresponding authors.

©2018 Walter de Gruyter GmbH, Berlin/Boston.

The CubSFS method uses a roughness penalty approach (Green & Silvermann, 1994) to solve the inverse problem (see Figure 1). The main idea is to estimate the changes in population size by weighing two opposing forces: The similarity between the expected and the observed SFS versus a slowly varying curve.

Figure 1: The elements of the score function. The effect of the smoothing parameter α on the elements of the score function [see equation (1)] for the CEU population ($m = 14$ and $t_m = 1.875$). To the left, is A: The expected and the observed SFS (black dots), and C: The residuals which is the square root of the relative distance to the observed SFS. For both plots the tail is grouped to include at most 10% of the sites. To the right, is B: The coalescent rate, i.e. the intensity, and D: The derived coalescent rate (intensity), $\lambda'(t)$, with time in coalescent units. Notice that the colour legend is not linear.

In particular, our solution is the function that minimizes the score function

$$S(\Lambda) = (1 - \alpha) \sum_{i=1}^{n-1} \frac{(E[\xi_i] - \xi_i)^2}{E[\xi_i]} + \alpha \int_0^\infty (\Lambda''(t))^2 dt. \quad (1)$$

The first term in the score function is the sum of squares of the residuals $R_i = (E[\xi_i] - \xi_i) / \sqrt{E[\xi_i]}$ measuring the distance between the observed and the expected SFS. Here $E[\xi_i]$ is the expected number of sites with i derived alleles determined from Λ using the theory from Polanski, Bobrowski, and Kimmel (2003) and Polanski and Kimmel (2003). The second term $\int_0^\infty (\Lambda''(t))^2 dt$ is a roughness penalty, and the smoothing parameter $0 < \alpha < 1$ determines the amount of regularization.

If α is large, then the most important term in the score function is the roughness penalty, and the minimizer $\hat{\Lambda}$ will display very little curvature. On the other hand, if α is small then the main contribution to the score function is the residual sum of squares, and the curve estimate $\hat{\Lambda}$ must resemble the data even if it requires a rather variable curve. The smoothing parameter is estimated using cross-validation.

In Figure 1 we illustrate the method. For a range of values of α we show the expected site frequency spectrum (top left), the corresponding residuals R_i (bottom left), the coalescent rate (top right), and the roughness penalty (bottom right) for the estimated value of Λ .

The SFS is a popular summary statistics for genetic variation, and in the last few years multiple methods have been developed to infer population histories from the SFS. These methods include estimates of the changes in population for a single population (Reppell, Boehnke & Zöllner, 2014; Bhaskar, Wang & Song, 2015; Eldon et al., 2015; Liu & Fu, 2015; Gao & Keinan, 2016), and estimates of ancestral population sizes, split times and migration rates using the joint SFS for multiple populations (Gutenkunst et al. 2009; 2010; Lukic & Hey 2011;

2012; Excoffier et al., 2013). The SFS approach most often assumes independence between the segregating sites. Methods that take recombination or linkage into account have recently been developed. Palacios and co-authors (Palacios & Minin, 2013; Palacios, Wakaley & Ramachandran, 2015) and Lan et al. (2015) developed a method based on gene genealogies. They provided a non-parametric Bayesian estimate of the population size using a Gaussian process as a prior on the changes in coalescent rate. However, the gene genealogies are not directly available and the error of inferring these have large effects on the estimated population size [Palacios, Wakaley, and Ramachandran (2015)]. Hidden Markov models (HMMs) are a popular framework for integrating out the unknown gene genealogies. The HMM framework for estimating changes in population size is applied in the PSMC method (Li & Durbin, 2011), the MSMC method (Schiffels & Durbin, 2014), The diCal method (Sheehan, Harris & Song, 2013), and the SMC++ method (Terhorst, Kamm & Song, 2017).

Terhorst, Kamm, and Song (2017) also regularize the changes in population size using a roughness penalty approach. Terhorst, Kamm, and Song (2017) use the penalty score $\int_0^\infty (\lambda''(t))^2 dt$. We suggest the penalty $\int_0^\infty (\Lambda''(t))^2 dt = \int_0^\infty (\lambda'(t))^2 dt$ since if the weight on the penalty is high, then the score function is minimized if $\lambda'(t) \approx 0$ for $t \geq 0$, implying an approximately constant population size. The smoothing parameter is a central parameter for our method, and is estimated using a cross-validation procedure. This parameter is chosen from experience in the SMC++ method. Finally, in the CubSFS method we pay detailed attention to the number and distribution of time intervals. In particular we use Akaike's Information Criteria to choose the number of time intervals.

The importance of regularizing is highlighted in the recent study by Lapierre, Lambert, and Achaz (2017). Lapierre, Lambert, and Achaz (2017) find that simple one-parameter models outperform the model-flexible method developed by Liu and Fu (2015) on the SFS of the Yoruba population. The roughness penalty approach that we advocate has the ability to find the right balance between a simple and a highly flexible model for changes in population size.

An alternative method is based on the Approximate Bayesian Computation (ABC) approach (Boitard et al., 2016). Here, the idea is to simulate changes in population size, calculate the corresponding appropriate summary statistics, and compare with the observed data summary statistics. If the distance between the summary statistics for the simulated data and the observed data is below a certain threshold, then the population size history is accepted. Finally, the posterior distribution of the population histories is determined based on the accepted samples.

We show that CubSFS is an attractive alternative to the methods mentioned above. In particular we can handle hundreds of samples through the SFS. Further, we avoid simulation-based methodology and prior parametric assumptions by detailed analytical considerations. Our method is validated through simulations and applied to data from 26 different populations from the 1000 Genomes Project (The 1000 Genomes Project Consortium, 2015).

2 Methods

2.1 Theory

Recall that $\xi = (\xi_1, \xi_2, \dots, \xi_{n-1})$ is the SFS for n sequences. Furthermore, $N(\cdot)$ is the population size, $\lambda(t) = N(0)/N(2N(0)t)$ is the coalescent rate, and $\Lambda(t) = \int_0^t \lambda(u) du$ is the integrated intensity. The score function is given by

$$S(\Lambda) = (1 - \alpha) \sum_{i=1}^{n-1} \frac{(E[\xi_i] - \xi_i)^2}{E[\xi_i]} + \alpha \int_0^\infty (\Lambda''(t))^2 dt,$$

where $0 < \alpha < 1$ is the smoothness parameter, and $E[\xi_i]$ is the expected number of sites with i derived alleles based on the integrated intensity function Λ .

The value $E[\xi_i]$ is evaluated using the probability $p_i(\Lambda)$ of observing a site with i derived alleles as

$$E[\xi_i] = S_n p_i(\Lambda),$$

where S_n is the number of segregating sites,

$$\begin{aligned}
 p_i(\Lambda) &= \frac{E[\xi_i(\Lambda)]}{\sum_{l=1}^{n-1} E[\xi_l(\Lambda)]} \\
 &= \frac{\sum_{j=2}^n \sum_{k=2}^j j(j-1) \binom{n-k}{l-1} \frac{(n-i-1)!(i-1)!}{(n-1)!} A_j^k e_j(\Lambda)}{\sum_{j=2}^n e_j(\Lambda) \sum_{k=2}^j \frac{i(j-1)}{k-1} A_j^k},
 \end{aligned} \quad (2)$$

and A_j^k and $e_j(\Lambda)$ are defined by Polanski, Bobrowski, and Kimmel (2003) as

$$A_j^k = \frac{\prod_{l=k, l \neq j}^n \binom{l}{2}}{\prod_{l=k, l \neq j}^n \left[\binom{l}{2} - \binom{j}{2} \right]}$$

for $k \leq j \leq n$, $A_n^n = 1$, and

$$e_j(\Lambda) = \int_0^\infty e^{-\binom{j}{2}\Lambda(t)} dt.$$

Equation (2) can be evaluated using constants defined by Polanski and Kimmel (2003). A straightforward proof of equation (2), different from that provided by Polanski, Bobrowski, and Kimmel (2003), is given in the Supplementary Information.

We estimate the integrated intensity by a cubic spline (Green & Silvermann, 1994) defined on $m + 1$ time points back in time (see Supplementary Figure S2). The time points $0 = t_0 \leq t_1 \leq \dots \leq t_m$ are proposed to be evenly distributed on a logarithmic scale (Li & Durbin, 2011), however, the placement should incorporate any prior knowledge of the model. Define the value of the cubic spline in each time point as $\Lambda(t_i) = a_i$, $i = 1, \dots, m$. The cubic spline has the following properties (see also the Supplementary Information):

- i. $\lambda(0) = \Lambda'(0) = 1$.
- ii. $\Lambda(0) = a_0 = 0$.
- iii. After t_m the coalescent rate is constant: $\lambda(t) = b_m$ for $t > t_m$ or, equivalently, $\Lambda(t) = a_m + b_m(t - t_m)$ for $t > t_m$.
- iv. The intensity is non-negative: $\lambda(t) > 0$ for $t > 0$ or, equivalently, $\Lambda(s) < \Lambda(t)$ for $0 < s < t$.

2.2 Implementation

Given α , $t_0, t_1, \dots, t_m$ and the values $\Lambda(t_i) = a_i$ for $i = 0, \dots, m$ the cubic spline of Λ is fully specified (see the Supplementary Information). The score function (1) is minimized with respect to the values of $0 = a_0 \leq a_1 \leq \dots \leq a_m$ given α and $t_0, t_1, \dots, t_m$. The search in the parameter space $(a_1, \dots, a_m)$, determining an increasing cubic spline that minimizes the score function $S(\Lambda)$, is performed using the augmented Lagrangian algorithm implemented in the nloptr package provided in R (Birgin & Martínez, 2008) using COBYLA (Constrained Optimization BY Linear Approximations) as the local solver (Powell 1994; 1998). The method is called successively until the distance between two successive estimates is less than 10^{-3} for all m parameters.

The implementation of the CubSFS method also applies to the folded SFS (see the Supplementary Information).

2.3 Estimation of the smoothing parameter and confidence intervals

The smoothness parameter α is determined using cross-validation (see Supplementary Figure S4). Briefly, we divide the segregating sites into K random groups. For each group we treat the SFS from that specific group as the validation data and the SFS from the remaining groups as the training data. For group k we calculate the expected SFS based on the training data $E[\xi^{(k)}]$, and compare to the observed SFS from the validation data, $\xi^{(k)}$

($k = 1, \dots, K$). We treat each group as a validation group exactly once and calculate the cross-validation mean goodness of fit (GOF). We finally estimate the smoothing parameter by minimizing the mean GOF

$$\hat{\alpha} = \operatorname{argmax}_{\alpha} \frac{1}{K} \sum_{k=1}^K \sum_{i=1}^{n-1} \frac{(\tilde{\zeta}_i^{(k)} - E[\tilde{\zeta}_i^{(k)}])^2}{E[\tilde{\zeta}_i^{(k)}]}.$$

Point-wise confidence intervals can be determined by means of bootstrapping from a multinomial distribution defined from the observed SFS and the number of monomorphic sites. For each bootstrapped SFS we find $\hat{\Lambda}_i$, and the lower and upper limits of the confidence interval at time t is then determined as the 2.5% and the 97.5% quantile, respectively, of $\{\hat{\Lambda}_i(t)\}_{i=1, \dots, B}$, where B is the number of bootstrap samples. The median is given by the 50% quantile of the bootstrap samples. We choose $B = 200$.

2.4 Akaike's Information Criteria

The number of free parameters in the model is $2m$ (the free parameters are $(t_1, t_2, \dots, t_m)$ and $(a_1, a_2, \dots, a_m)$). The final choice of parameters is determined using Akaike's Information Criteria (AIC):

$$AIC = 2 \times 2m - 2 \log \hat{L}.$$

The likelihood of the observed SFS, $\hat{L}$, is determined by the multinomial distribution defined by the probability of observing a site with i derived alleles $p_i(\Lambda)$ [see equation (2)]. This probability is either determined from the estimated integrated intensity, or from the point-wise median of the bootstrap samples (see the Supplementary Information). Note that the AIC is able to choose between fits generated using the same number of free parameters, but different distributions of the time points.

2.5 Validation

Four different models are used to validate the CubSFS method: Exponential growth, two epochs, a bottleneck, and a zigzag model (Schiffels & Durbin, 2014). Recall that we consider 200 sequences of length $5 \cdot 10^8$ and the number of segregating sites is $6 \cdot 10^6$.

In order to apply the CubSFS method we must choose the number and placement of the points $t_1, \dots, t_m$. We place the points using the same procedure as Li and Durbin (2011; see Supplementary Information eq. (33)). We use a number of points $m + 1$ equal to 4, 7, 10 or 15, and the last time point t_m is placed at 0.5, 1 or 1.5 times the expected time to the most recent common ancestor. For the zigzag model, further results are produced for 15, 20 and 30 time points placed according to knowledge of the model.

2.6 The 1000 Genomes Project

All 26 populations in the 1000 Genomes Project phase 3 are downloaded. The vcf files are parsed into HDF5 files, from which the minor allele count is calculated for each site for each population. From this the population specific folded SFS is generated. No further quality control or filtering is performed besides that done by the 1000 Genomes Project.

The SFS for 9 different populations used by Liu and Fu (2015) are kindly provided by the authors. These SFSs are generated from the 1000 Genomes Project phase 1 data.

The CubSFS method is called on all 35 obtained (folded or unfolded) SFSs to infer the changes in population size back in time. For all inferences we set the time to most recent common ancestor to be 600 thousand years ago. For the 26 populations of phase 3 data we used either 15 or 20 time points where the last time point is placed 900 thousand years ago or 1.2 million years ago – a total of 4 parameter settings. For the 9 populations of phase 1 data generated by Liu and Fu (2015), 15, 20 or 30 time points are used with the last time point placed 1, 1.5 or 2 times 600 thousand years ago – a total of 9 parameter settings. In units of years back in time, the time points are placed according to equation (33) in the Supplementary Information. The time points are transformed from units of years back in time to coalescent units prior to running the CubSFS method by means of a present population size of 10,000 and a generation time of 24 years. For each inference a total of 200 bootstrap samples are used. For each of the 35 folded or unfolded SFS, the parameter setting with the smallest AIC value based on the bootstrap median is chosen.

The final results are transformed from coalescent units back to calendar years using a genome size of $G = 3,234.83$ Mbp, 24 years per generation, and a mutation probability of $\mu = 1.2 \cdot 10^{-8}$ per generation per bp (see the Supplementary Information).

3 Results

3.1 The 1000 Genomes Project

The results of inferring the changes in population size from the folded SFS based on the 1000 Genomes Project phase 3 data are shown in Figure 2. Here we see a similar pattern of changes in population size as in the PSMC analysis given by The 1000 Genomes Project Consortium (2015). All populations coalesce approximately 400–600 thousand years ago. From that time and forward the non-African populations experience a decrease in population size until 10–40 thousand years ago, where an increase in population size raises the population size to that matching the Africans. However, the Asian populations experience a slightly later increase. For the European and west Asian populations the large bottleneck seems to have a small increase in population size at approximately 80–110 thousand years ago. The African populations experience a small decline in population size 80–110 thousand years ago, and a larger decline in population size 10–50 thousand years ago. The two periods of decline in population size for the Africans are both smaller than the overall decline in population size for the non-African populations.

Figure 2: The changes in population size for 26 populations based on the 1000 Genomes Project phase 3 data. The point-wise median from 200 bootstrap samples are shown for each of the 26 populations. Out of 4 different parameter settings the inferred population size is chosen based on the AIC of the bootstrap point-wise median. Each population is given a separate colour as indicated on the world map in the top right corner.

Figure 3 displays the inferred changes in population size for 9 populations based on the 1000 Genomes Project phase 1 data (additional results are shown in Supplementary Figures S16–S24). The results are similar to those given by Liu and Fu (2015). The results are very similar within the continents of Asia (CHB, CHS, and JPT), Europe (CEU, FIN, GBR, and TSI) and Africa (LWK and YRI). The Asian and European populations show an overall decline in population size ranging from 10 to 500 thousand years ago, with a possible slightly increase in populations size approximately 20–40 thousand years ago. The African populations all show one decline in population size coherent with the decline in population size of the other 7 populations happening furthest back in time. The African decline in population size is indicated slightly earlier than the other 7 populations.

Figure 3: The changes in population size for 9 populations based on the 1000 Genomes Project phase 1 data. The red line is the point-wise median of the bootstrap. The orange lines are the point-wise confidence intervals. Out of 9 different parameter settings the inferred population size is chosen based on the AIC of the bootstrap point-wise median. The location of each population is given in the map of Figure 2.

The timing of the slight increase in population size for the non-Africans is different between the analyses of the phase 1 and the phase 3 data from the 1000 Genomes Project. Similarly, the timing of the period between the two declines in population size for the Africans are not completely coherent between the two analyses. However, overall there is a large coherence between the results obtained from the phase 1 and phase 3 data from the 1000 Genomes Project.

3.2 Simulation studies

To verify our method we considered four different scenarios: Exponential growth, two epochs, a bottleneck, and a zigzag model (Schiffels & Durbin, 2014). We simulated 200 sequences of length $5 \cdot 10^8$ and with $6 \cdot 10^6$ segregating sites.

The population size inferred from CubSFS is shown in Figure 4 (additional results are provided in Supplementary Figures S6–S10). The method generally provides sensible results for all models. The exponential growth model is inferred particularly well. For instantaneous changes as the two epochs, the bottleneck, and the zigzag models the smoothing seems to detect the trends in general, but the instantaneous changes in population size are more difficult to detect. In general we expect our model to perform particularly well when the changes in population size are smooth. Our model is less suited for situations where the population size has substantial changes in a few generations.

Figure 4: The result from simulated demographic histories. The black line is the true model and the red line is the point-wise median of the bootstrap estimates based on the CubSFS method. The orange lines are the point-wise confidence intervals. The dotted grey vertical lines are the expected time to most recent common ancestor for each of the four models. The rug points are the time points used by the CubSFS method. For three of the models A: exponential growth, B: two epochs, and C: bottleneck the times points are placed linearly on a logarithmic scale in coalescent units prior to estimation. For model D: zigzag the time points are placed according to knowledge of the model. The final choice of parameter setting is determined by the AIC based on the point-wise median of the results from the bootstrap samples.

4 Discussion

The CubSFS method estimates the changes in coalescent rate dependent only on a set of pre-defined times placed backwards in time on the coalescent scale. A basic assumption of our roughness penalty approach is that the population size is slowly varying. This assumption is valid for the exponential growth model, and we believe this is the main reason why this model is inferred particularly well.

The ability to estimate the population size depends on the placement of the time points. Placing the last time point before the expected time to the most recent common ancestor may induce additional fluctuations in the recent past. The time points have to capture the whole period of which the SFS provides information. The CubSFS method provides plausible estimates for well placed time points, even when reaching beyond the time to the most recent common ancestor. In this case the CubSFS method recognises the lack of information after the time to most recent common ancestor by inferring a constant population size.

The score function (equation (1)) is composed of two elements, and both of them can possibly be extended or modified to improve performance. The first term, measuring the similarities between the observed statistics and the expected statistics from the model, could include other summary statistics such as LD patterns or runs of homozygosity. Hence, dependencies between sites and quality of the observed data can be included in the estimation procedure similar to the PopSizeABC (Boitard et al., 2016). The main advantage of our approach is that we avoid simulations by detailed analytical considerations.

We have assumed a Kingmans coalescent in a panmictic population. Under this assumptions the reciprocal coalescent rate is the effective population size. However, a demographic history with a single panmictic population is a very simple evolutionary scenario. Mazet et al. (2016) discusses the interpretation of the reciprocal coalescent rate in the context of two different scenarios: Either changes in population size for one panmictic population, or migration between multiple populations with constant population sizes.

Future studies would be able to build on the joint SFS of multiple populations, using the covariance between population specific SFSs as a measure of similarity, enabling inference of the different changes in size of the populations. However, adding admixture and migration to this set-up, as well as analysing the analytical expression of the covariance within changing population sizes is a topic for further research.

Likewise, the second term of the score function, i.e. the regularisation term, can be adjusted. Here we use the L^2 norm of the smoothing spline. However, other functions can be used such as the absolute value of the smoothing spline (the L^1 norm). The choice of regularisation will affect the smoothness of the estimated spline.

Green and Silvermann (1994) discuss the fact that using the L^2 norm for regularisation is similar to assigning a Gaussian prior to the space of all smooth functions in a Bayesian setting [see Green and Silvermann (1994), Section 3.8]. Hence, our method is similar to those developed by Palacios et. al. (Palacios & Minin, 2013; Palacios, Wakaley & Ramachandran, 2015), without the need to infer the gene genealogies. Actually, if the distribution of gene genealogies is known, then Gattepaille, Gunther, and Jakobsson (2016) have presented an exact solution to compute the corresponding changes in population size. In other words, they show how to invert the result of Polanski and Kimmel (2003).

The last part of the score function is the amount of smoothing, and here we use a cross-validation technique on the segregating sites to estimate the smoothing parameter. Other methods can be used, such as the leave-one-out theory used in kernel estimation. However such techniques may depend on the sample size n , of which the cross-validation technique is independent.

5 Software availability

The CubSFS method is implemented in R and is available at <https://github.com/blwaltoft/CubSFS>.

Acknowledgement

We are grateful to two anonymous reviewers who provided very careful and constructive questions and suggestions on the first version of this paper. The helpful comments resulted in a much improved paper. We thank X. Liu and Y.-X. Fu for providing the SFS of the 9 populations. We also thank J. Terhorst and Y. Song for providing key insight into the SMC++ method. We further thank Bjarni Vilhjalmsón, for providing the folded SFS of the 1000 Genomes Project phase 3 data.

Funding

BLW is funded by The Lundbeck Foundation Initiative for Integrative Psychiatric Research, iPSYCH, Denmark. Grant number R155–2014–1724.

References

- Bhaskar, A. and Y. S. Song (2014): “Descartes’ rule of signs and the identifiability of populations demographic models from genomic variation data,” *Ann. Stat.*, 42, 2469–2493.
- Bhaskar, A., Y. S. R. Wang and Y. S. Song (2015): “Efficient inference of population size histories and locus-specific mutation rate from large-sample genomic variation data,” *Genome Res.*, 25, 268–279.
- Birgin, E. G. and J. M. Martínez (2008): “Improving ultimate convergence of an augmented Lagrangian method,” *Optim. Method. Softw.*, 23, 177–195.
- Boitard, S., W. Rodriguez, F. Jay, S. Mona and F. Austerlitz (2016): “Inferring population size history from large samples of Genome-wide molecular data – an approximate Bayesian computation approach,” *PLoS Genet.*, 12, e1005877.
- Eldon, B., M. Birkner, J. Blath and F. Freund (2015): “Can the site-frequency spectrum distinguish exponential population growth from multiple-merger coalescents?” *Genetics*, 199, 841–856.
- 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, e1003905.

- Gao, F. and A. Keinan (2016): "Inference of super-exponential human population growth via efficient computation of the site frequency spectrum for generalized models," *Genetics*, 202, 235–245.
- Gattepaille, L., T. Gunther and M. Jakobsson (2016): "Inferring past effective population size from distributions of coalescent times," *Genetics*, 204, 1191–1206.
- Green, P. J. and B. W. Silvermann (1994): *Nonparametric regression and generalized linear models*, Chapman & Hall/CRC, London.
- 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, e1000695.
- Gutenkunst, R. N., R. D. Hernandez, S. H. Williamson and C. D. Bustamante (2010): "Diffusion approximations for demographic inference: $\delta a \delta i$," *Nature Precedings*. <http://precedings.nature.com/documents/4594/version/1>.
- Lan, S., J. Palacios, M. Karcher, V. N. Minin and B. Shahbaba (2015): "An efficient Bayesian inference framework for coalescent-based nonparametric phylodynamics," *Bioinformatics*, 31, 3282–3289.
- Lapierre, M., A. Lambert and G. Achaz (2017): "Accuracy of demographic inferences from the site frequency spectrum: the case of the Yoruba population," *Genetics*, 206, 439–449.
- Li, H. and R. Durbin (2011): "Inference of human population history from individual whole-genome sequences," *Nature*, 475, 493–496.
- Liu, X. and Y. Fu (2015): "Exploring population size changes using SNP frequency spectra," *Nature Genet.*, 47, 555–559.
- Lukic, S. and J. Hey (2011): "Non-equilibrium allele frequency spectra via spectral methods," *Theor. Popul. Biol.*, 79, 203–219.
- Lukic, S. and J. Hey (2012): "Demographic inference using spectral methods on SNP data, with an analysis of the human out-of-Africa expansion," *Genetics*, 192, 619–639.
- Mazet, O., W. Rodrigues, S. Gruseq, S. Boitard and L. Chikhi (2016): "On the importance of being structured: instantaneous coalescence rates and human evolution – lessons for ancestral population size inference?" *Heredity*, 116, 362–371.
- Myers, S., C. Fefferman and N. Patterson (2008): "Can one learn history from the allelic spectrum," *Theor. Popul. Biol.*, 73, 342–348.
- Palacios, J. and V. N. Minin (2013): "Gaussian process-based Bayesian nonparametric inference of population size trajectories from Gene Genealogies," *Biometrics*, 69, 8–18.
- Palacios, J. A., J. Wakeley and S. Ramachandran (2015): "Bayesian nonparametric inference of population size changes from sequential Genealogies," *Genetics*, 201, 281–304.
- 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.
- Polanski, A., A. Bobrowski and M. Kimmel (2003): "A note on distributions of times to coalescence, under time-dependent population size," *Theor. Popul. Biol.*, 63, 33–40.
- Powell, M. J. D. (1994): *Advances in Optimization and Numerical Analysis*, chapter A Direct Search Optimization Method That Models the Objective and Constraint Functions by Linear Interpolation, 51–67. Springer Netherlands, Dordrecht.
- Powell, M. J. D. (1998): "Direct search algorithms for optimization calculations," *Acta Numerica*, 7, 287–336.
- Reppell, M., M. Boehnke and S. Zöllner (2014): "The impact of accelerating faster than exponential population growth on genetic variation," *Genetics*, 196, 819–828.
- Schiffels, S. and R. Durbin (2014): "Inferring human population size and separation history from multiple genome sequences," *Nature Genet.*, 46, 919–925.
- Sheehan, S., K. Harris and Y. S. Song (2013): "Estimating variable effective population size from multiple genomes: a sequentially Markov conditional sampling distribution approach," *Genetics*, 194, 647–662.
- Terhorst, J., J. A. Kamm and Y. S. Song (2017): "Robust and scalable inference of population history from hundreds of unphased whole genomes," *Nature Genet.*, 49, 303–309.
- The 1000 Genomes Project Consortium (2015): "A global reference for human genetic variation," *Nature*, 526, 68–74.
- Wakeley, J. (2009): *Coalescent theory: an introduction*, Roberts and Company Publishers, Greenwood Village, Colorado 80111, USA.

Supplemental Material: The online version of this article offers supplementary material (<https://doi.org/10.1515/sagmb-2017-0061>).