MOLECULAR ECOLOGY
RESOURCES
Molecular Ecology Resources (2017) **17**, 631–641 doi: 10.1111/1755-0998.12606

PSMC (pairwise sequentially Markovian coalescent) analysis of RAD (restriction site associated DNA) sequencing data

SHENGLIN LIU and MICHAEL M. HANSEN
Section for Genetics, Ecology and Evolution, Department of Bioscience, University of Aarhus, Ny Munkegade 114-116, DK-8000 Aarhus C, Denmark

Abstract

The pairwise sequentially Markovian coalescent (PSMC) method uses the genome sequence of a single individual to estimate demographic history covering a time span of thousands of generations. Although originally designed for whole-genome data, we here use simulations to investigate its applicability to reference genome-aligned restriction site associated DNA (RAD) data. We find that RAD data can potentially be used for PSMC analysis, but at present with limitations. The key factor is the proportion (*p*) of the genome that the RAD data covers. In our simulations, a proportion of 10% can still retain a substantial amount of coalescent information, whereas for 1% estimation becomes unreliable. The performance depends strongly on mutation rate (*μ*) and recombination rate (*r*) and is proportional to *μp/r*. When the value of this term is low, increasing the amount of data and number of iterations helps restoring the power of the estimation. We subsequently analyse one whole-genome-sequenced and 17 RAD-sequenced three-spined sticklebacks (*Gasterosteus aculeatus*) from a lake in Greenland. The whole-genome sequence suggests a relatively recent expansion and decline within ca. 4000–40 000 generations ago, possibly reflecting postglacial expansion and founding of the lake population. RAD data, where chromosomes from 10 individuals are combined, identify a similar pattern. Our study provides guidance about the use of PSMC analysis and suggests measures that can improve its utility for RAD data. Finally, the study shows that RAD loci in general contain coalescent information that can be used for developing more targeted methods.**

Keywords: demographic history, pairwise sequentially Markovian coalescent (PSMC), RAD sequencing, three-spined stickleback, whole-genome sequencing

Received 26 October 2015; revision received 21 September 2016; accepted 4 October 2016

Introduction

The accelerating development and accessibility of next-generation sequencing (NGS) has led to increasing use of reduced representation (e.g. RAD sequencing) and whole-genome sequencing in population studies, allowing for targeting both classical and novel research questions (Davey *et al.* 2011; Ellegren 2014). Recently, a number of new methods have become available that make use of whole-genome sequencing data for tracking demographic history through time (Li & Durbin 2011; Sheehan *et al.* 2013; Schiffels & Durbin 2014; Liu & Fu 2015). Among these, pairwise sequentially Markovian coalescent (PSMC) (Li & Durbin 2011) is an analytical method to infer the demographic history of a population using the genome of a single individual. It models the coalescence time across the genome with a hidden Markov model, and based on its distribution, it deduces the demographic history. It is an approximated version of genome-wide coalescent with recombination (McVean & Cardin 2005). While exploiting the most of the demographic information from the genome, it also shows extremely high efficiency, and it is model flexible and does not rely on a predefined evolutionary model for a population (Liu & Fu 2015). In humans, it provides reliable results on a timescale extending from a few millions to a few tens of thousands years ago (Li & Durbin 2011), whereas for organisms with shorter generation time, the method is expected to provide results on a correspondingly more recent timescale. It therefore represents a highly useful method for inferring demographic history of populations and has been widely used for a range of species with available reference genomes, ranging from insects to vertebrates (Moura *et al.* 2014; Wallberg *et al.* 2014; Nadachowska-Brzyska *et al.* 2015; Palkopoulou *et al.* 2015; Xue *et al.* 2015; Liu *et al.* 2016). PSMC was originally designed for whole-genome sequence data

Correspondence: Shenglin Liu, Fax: +45 87150201;
E-mail: liu.shenglin@bios.au.dk

and has so far been exclusively used for this type of data. However, the question arises whether PSMC can also be used for methods that provide a reduced representation of the genome, such as restriction site associated DNA (RAD) sequencing (Baird *et al*. 2008) and related methods (Elshire *et al*. 2011; Peterson *et al*. 2012; Poland *et al*. 2012). Obviously, such RAD data will have to be aligned to a reference genome.

Given a certain restriction enzyme, a RAD data set is a collection of short sequences (at present typically ca. 50–100 bp), flanking all the restriction sites in the genome. Assuming the restriction sites are randomly distributed, a RAD data set can be viewed as a subset of the genome obtained by randomly sampling many short sequences. By choosing different restriction enzymes, one can adjust the sampling density, with longer recognition sites of the enzyme generally resulting in lower sampling density. Compared to whole-genome sequencing, RAD sequencing is considerably cheaper, and it is easy to obtain a large quantity of high-quality SNPs for many individuals. These features underlie the increasing popularity of RAD sequencing in population genomics.

One major drawback of RAD data sets consists in strongly reduced linkage information compared to whole-genome sequencing data. This is due to the random sampling of sequences along the genome, which increases distances between detected SNPs and causes them to show reduced linkage relationships. This is particularly obvious when the sampling density is low, such as when employing a restriction enzyme with long (e.g. 8 bp) recognition sequence, which again imposes limitations on the methods that can be applied to analyse RAD data. In particular, it reduces the power of coalescence-based methods as they depend on multiple linked SNPs to provide coalescent information, and the number of linked SNPs has to be sufficiently high to yield good resolution. This is probably the reason why many coalescence-based methods such as PSMC have not been used to analyse RAD data. Instead, most methods so far applied for analysing demographic history using RAD data are frequency based (Gutenkunst *et al*. 2009; Excoffier *et al*. 2013), as exemplified by several recent studies (Moura *et al*. 2014; Lanier *et al*. 2015; Ferchaud & Hansen 2016). Nevertheless, reduced linkage in RAD data does not necessarily mean that there is no linkage (or coalescent) information left at all. The genome of an individual generally can be seen as consisting of many recombination units. Recombination units show different lengths depending on the coalescence time of each. If the time is ancient, the unit will be short due to many historical recombination events. On the other hand, if the time is recent, the unit will be long and RAD data may still preserve sufficient information to retrieve this part of the demographic history.

In this study, we assessed the applicability of PSMC for analysing RAD data through simulation. Possible approaches for improvements were investigated by examining the influences of different parameters on the performance of the analysis, for instance, sampling density of the genome, recombination rate, mutation rate, population size, number of iterations and number of chromosomes. From a technical perspective, there is a continuing development for most sequencing platforms towards longer reads, so we also assessed the influence of longer reads on the performance of the method. Furthermore, we also evaluated with simulations whether PSMC can be used to estimate divergence time between populations with RAD data. Finally, we analysed RAD data and a whole-sequenced genome from a three-spined stickleback (*Gasterosteus aculeatus*) population and compared PSMC results from the two types of data. Both simulated and empirical results showed that RAD data contain coalescent information that can be used for reconstructing demographic history using PSMC, although presently with limitations. The critical factor is the proportion of genome covered by RAD data. However, adjustment of certain parameters can improve the performance of the analysis.

Materials and methods

Simulation of RAD data and PSMC analysis

We generated RAD data through coalescent simulation as implemented in the software ms by Hudson (2002). For each simulation, we first defined a demographic history and accordingly generated a genome of a single diploid individual. By default, each individual had 10 pairs of chromosomes, and each chromosome was 30 Mb long. The mutation rate (denoted as μ) was by default set to 1e−8 per nucleotide per generation, and the recombination rate (denoted as r) to 2e−9 per generation between neighbouring nucleotide sites, as in the simulations by Li & Durbin (2011). A RAD data set was subsequently obtained by randomly sampling RAD loci from the genome with a specified sampling density. The sampling density (denoted as d) was measured as the number of the RAD loci divided by the size of the genome. For example, with $d = 0.001$, it means that on average there is one RAD locus in every 1000 bp of the genome. The read length of each RAD locus (denoted as L) was by default set to 100 bp, approximating the general length of RAD loci in real data. The sampling of RAD loci from the genome was implemented in R v3.1.1 (R Core Team 2014) (code provided in DRYAD).

The RAD data were then transformed into a '.psmcfa' file as the input for PSMC. The default bin size was set to 100 bp with each RAD locus

representing a bin. The bins were ordered according to their positions in the genome and were directly concatenated by omitting the missing sequences between neighbouring RAD tags. PSMC was run with the same parameters as used by Li & Durbin (2011) for autosomal data except that we used the results from the 25th iteration by default. The time was measured in number of generations. The transformation of data and the visualization of the results in graphs were conducted in R (code provided in DRYAD).

Assessing the applicability and testing the influence of the parameters

We assessed the applicability of PSMC for RAD data using various sampling densities. For this, we adopted three demographic history scenarios previously used by Li & Durbin (2011), that is Sim-1, Sim-2 and Sim-3. Briefly, Sim-1 is based on a generalized demographic history of non-African human populations, with an out-of-Africa population decline and a recent population expansion. Sim-2 reflects a history of expansion and a decrease in population size, whereas Sim-3 reflects an acute bottleneck followed by recent expansion. For each scenario, three data sets were created: (i) whole genome, (ii) RAD with $d = 0.001$ and (iii) RAD with $d = 0.0001$, among which the whole-genome data served as control. Each data set from each demographic scenario was replicated 30 times to account for the stochasticity of the estimation.

We subsequently tested whether increased read length could improve the performance of the analysis. Therefore, a RAD data set with $d = 0.0001$ and $L = 1000$ bp was simulated under the demographic history of Sim-1. The bin size remained as 100 bp; hence, each RAD locus represented 10 bins. The simulation was repeated 10 times. Similar tests were conducted for a series of other parameters, such as mutation rate, recombination rate, population size (denoted as N), number of iterations (denoted as m) and number of chromosomes (denoted as k).

Additionally, PSMC can be used for detecting the divergence time of two populations (Li & Durbin 2011). This is achieved by implementing the analysis for F1 hybrids between the two populations, either real F1 hybrids or synthetic hybrids obtained by merging one chromosome from two individuals from each population; the latter procedure requires phasing. The PSMC plots of the hybrid should show a blow-up of the population size at the time point when the two populations diverged from each other. We tested the applicability of this method for the RAD data as well. The detailed procedures are provided in Supporting Information, Note 1.

Empirical data from three-spined stickleback

We analysed three-spined sticklebacks (*Gasterosteus aculeatus*) from a freshwater lake (Qarajat; ca. 0.4 km^2), situated in the Nuuk Fjord region of western Greenland (coordinates 63.991°, -51.446°). All individuals exhibited a low number of lateral plates (<10) as is often observed among freshwater sticklebacks (Colosimo *et al.* 2005). This species has a genome size of ca. 0.46 Gb, and a well-assembled genome sequence is available (Jones *et al.* 2012). Seventeen individuals were RAD sequenced (90-bp paired-end sequencing) at an average depth of 25× using the six-base cutter *Eco*RI. RAD sequencing was outsourced to Beijing Genomics Institute (BGI, Hong Kong, China), and the procedures were the same as those described in Pujolar *et al.* (2013), including subsequent trimming of read lengths to 75 bp and quality filtering. Reads were aligned to the three-spined stickleback genome (Jones *et al.* 2012) using Bowtie v1.1.1 (Langmead *et al.* 2009) and thus allowing for two mismatches (v-option, $v = 2$). RAD loci and SNPs for each individual were identified using the pstacks pipeline in Stacks v1.13 (Catchen *et al.* 2013) with a minimum depth of 5, and the output was transformed to '.psmcfa' files using R (code provided in DRYAD). Only the 20 autosomal chromosomes were used; that is, the sex chromosome (group XIX) and the smaller scaffolds were excluded. Each locus was set as a bin (i.e. bin size = 75 bp), and the bins were ordered according to their positions in the genome. The missing sequences between neighbouring RAD tags were omitted. The parameters for running PSMC were identical to those used in the simulations.

We furthermore sequenced the whole genome of one individual at a depth of 23×. Briefly, 100-bp paired-end sequencing with an average insert size of 374 bp was conducted using the TruSeq PCR Free DNA Sample Prep kit and TruSeq v4 Chemistry on the HiSeq 2500 platform (Illumina, San Diego, CA, USA). This work was outsourced to AROS Applied Biotechnology (Aarhus, Denmark). The reads were first assessed for quality with FastQC v0.11.3 (http://www.bioinformatics.bbsrc.ac.uk/projects/fastqc/). The last five base pairs of each read were trimmed off with FastX v0.0.13 (http://hannonlab.cshl.edu/fastx_toolkit/index.html) because of high 'N' content. The reads were then mapped to the three-spined stickleback genome using the 'aln' and 'sampe' subfunctions of BWA-0.7.12 (Li & Durbin 2009) by allowing a maximum of two mismatches. The SAM format alignment file was filtered for a minimum mapping quality of 20 and was sorted and converted into a BAM file which was further transformed into a '.psmcfa' file following the instructions in the GitHub page of PSMC (https://github.com/lh3/psmc). Again, only the 20 autosomal chromosomes were used. The bin size was set to

75 bp to correspond to the RAD data, and the parameters for running PSMC were the same as above, except that 100 bootstraps were performed.

Results

Assessing the applicability and testing the influence of the parameters

The simulation results of RAD data with various sampling densities are shown in Fig. 1. Compared to the results from whole-genome data (Fig. 1a,d,g), those from RAD of $d = 0.001$ (Fig. 1b,e,h) still could detect all the major demographic events in the history, although the magnitude of the immediate change in population size and the timing of each event were somewhat biased. This suggests that despite the sampling effect, there is still a considerable amount of coalescent information left. However, when sampling density was reduced to 0.0001 (Fig. 1c,f,i), almost no demographic events could be detected. Overall, this implies that the power of PSMC in reconstructing the demographic history decays with the reduction of the sampling density of RAD data, and the performance seems consistent across different patterns of demographic history, for example Sim-1, Sim-2 and Sim-3.

However, the performance can be improved by increasing the read length of each RAD locus. For example, the results from RAD with $d = 0.0001$ and $L = 1000$ (Fig. 2) resembled those from RAD with $d = 0.001$ and $L = 100$ (Fig. 1b). This could indicate that for a given species (i.e. μ and r are fixed), the proportion of the genome covered by the RAD data (denoted as p) is the limiting factor for the applicability of PSMC. The value of p is numerically equal to d^*L. Both Figs 1b and 2 have $p = 0.1 = 10\%$.

Increasing the mutation rate or lowering the recombination rate also provided similar improvement (Fig. 3a–f). When the mutation rate was increased by 10 times, the RAD data with $d = 0.001$ (Fig. 3b) yielded results equivalent to those from whole-genome data with default mutation rate (Fig. 1a), and RAD with $d = 0.0001$ (Fig. 3c)

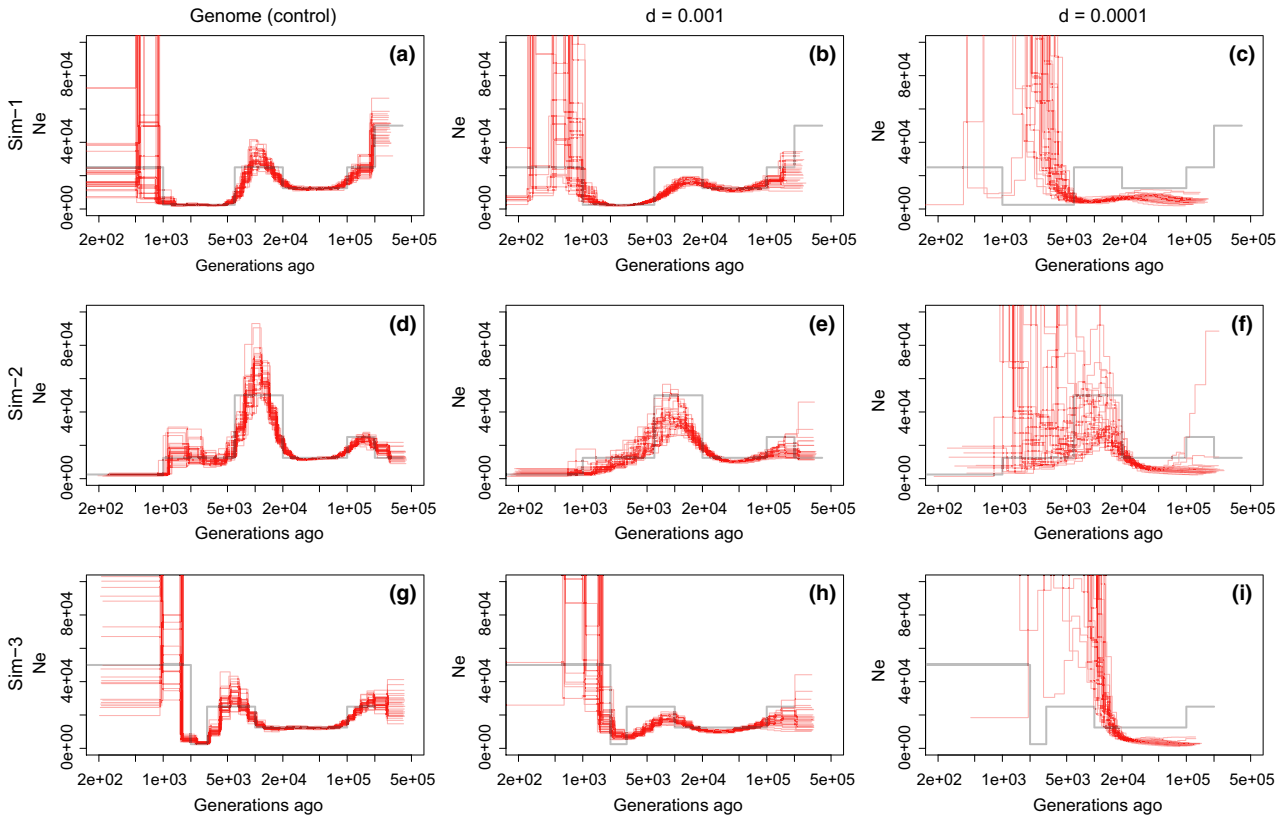

Fig. 1 PSMC analyses for data simulated under different demographic scenarios and various sampling densities. Each row of plots was generated under the same demographic scenario with the name of the scenario indicated on the left. Each column of plots was produced under the same sampling density with the value of the density specified on the top. The first column was produced with whole-genome data. The grey lines represent the demographic histories from which the data were simulated, whereas the red lines denote estimations obtained by running PSMC using the simulated data. Each simulation was repeated 30 times.

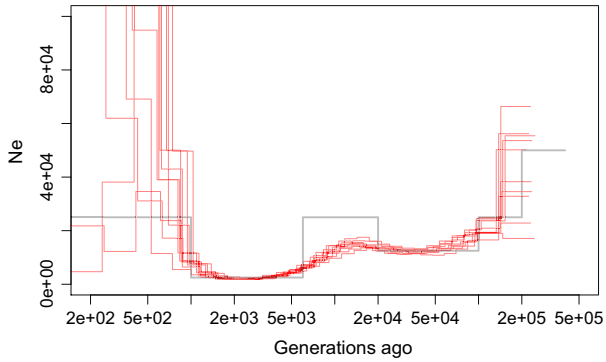

Fig. 2 PSMC analyses of simulated RAD data with long-sequencing reads. The grey lines represent the demographic histories from which the data were simulated, whereas the red lines denote estimations obtained by running PSMC on the simulated data. Each simulation was repeated 10 times. Demographic scenario: Sim-1; *L*: 1000 bp; *d*: 0.0001.

equivalent to the one with $d = 0.001$ (Fig. 1b). Lowering the recombination rate by 10 times resulted in the same pattern, although the variance of the estimation became higher (Fig. 3d–f). Increasing the population size (Fig. 3g–i) nevertheless did not improve the power of inferring the true demographic history, but it decreased variance of the estimation.

We further found that the performance of PSMC for RAD data could be improved by increasing the number of iterations and the number of chromosomes (Fig. 4). However, these two treatments must be carried out at the same time. Our simulations showed that increasing the number of iterations alone could improve the accuracy of the inference, but it compromised the precision of the estimation (Fig. S1, Supporting information). Here, by high accuracy, we refer to high convergence of the inference to the true demographic history, and low precision refers to high variance of the estimation. In contrast, increasing the number of chromosomes alone could improve precision, but it had no influence on the accuracy (Fig. S2, Supporting information). In Li & Durbin (2011), the low precision caused by prolonged cycles of iterations was referred to as overfitting. However, our study indicates that this low precision can be compensated by increasing the number of chromosomes. For example, for RAD with $d = 0.001$, 250 iterations with 100 chromosomes (Fig 4b) yielded results comparable to

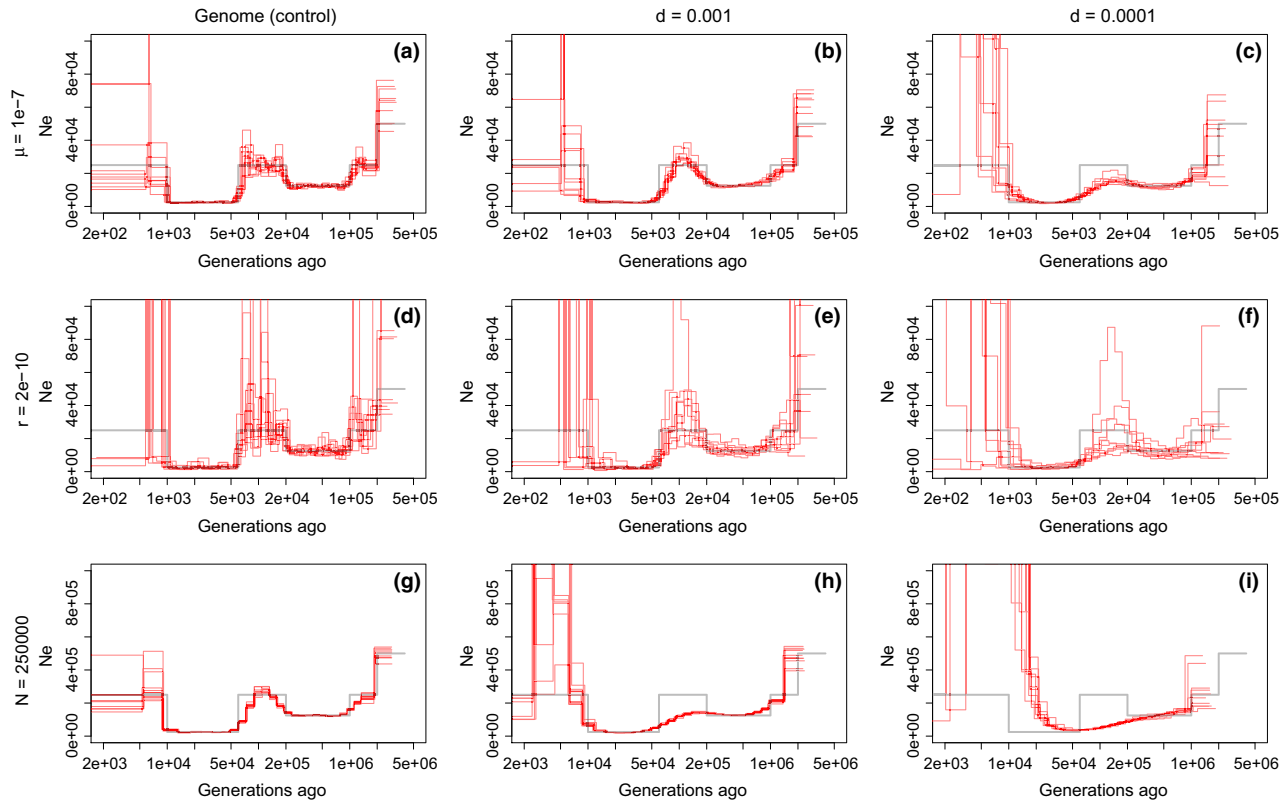

Fig. 3 Influence of high mutation rate (a–c) or low recombination rate (d–f) or high population size (g–i) on the performance of PSMC analysis of simulated RAD data. The grey lines represent the demographic histories from which the data were simulated, whereas the red lines denote estimations obtained by running PSMC on the simulated data. Each simulation was repeated 10 times. Demographic scenario: Sim-1; *N* (default): 25 000. For a–c and g–i, the bin size was set as 10 bp due to the increased genetic diversity.

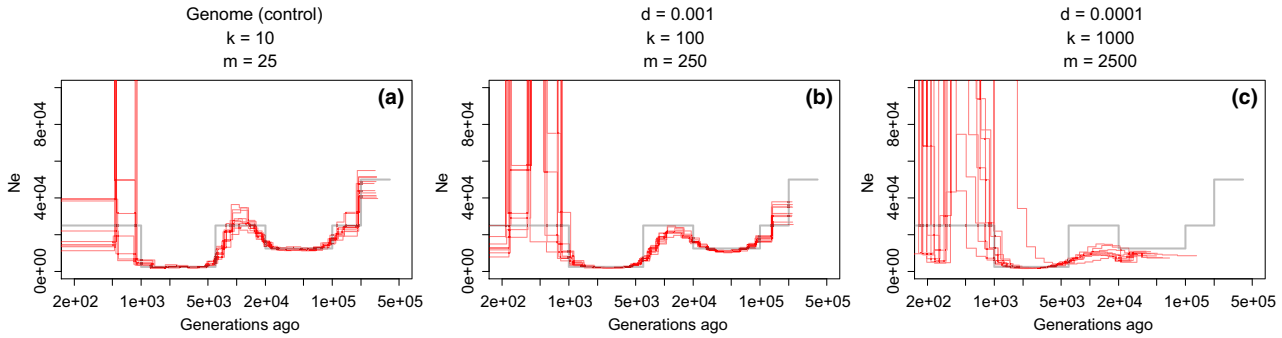

Fig. 4 Improving the performance of PSMC analysis of simulated RAD data by increasing the number of iterations and the number of chromosomes. The grey lines represent the demographic histories from which the data were simulated, whereas the red lines denote estimations obtained by running PSMC on the simulated data. Each simulation was repeated 10 times. Demographic scenario: Sim-1.

those obtained from whole genome with 10 chromosomes and 25 iterations (Fig. 4a). Similar improvement was observed for RAD with $d = 0.0001$, $m = 1000$, $k = 2500$ (Fig. 4c).

When PSMC is used to estimate the divergence time between two populations, a similar positive relationship as above between the power of inference and the sampling density was observed (Fig. S3, Supporting information). The blow-ups indicating the time of divergence between the populations were observed in both whole-genome and RAD data. However, the time of the blow-up shifted further back in time with the decrease in sampling density, that is divergence time was systematically overestimated using RAD data. The absence of phase information compromised the estimation even more (Fig. S3d–f, Supporting information), although phasing with software improved the results (Fig. S3g–i, Supporting information). The bias of the divergence time estimate was decreased by incorporating more chromosomes and increasing the number of iterations (Fig. S4, Supporting information). In particular, the estimate from RAD with $d = 0.001$, $m = 100$, $k = 250$ (Fig. S4b, Supporting information) became comparable to that from the whole genome. However, the estimate from RAD with $d = 0.0001$, $m = 1000$, $k = 2500$ (Fig. S4c, Supporting information) was still strongly biased towards high estimates.

Empirical data from three-spined stickleback

A high-quality genome sequence was obtained for a three-spined stickleback from Lake Qarajat, Greenland, with 97% of all reads were paired in mapping. The coverage distribution and the insert size distribution showed regular pattern and fit with expectation (Fig. S5, Supporting information). The proportion of missing nucleotides was as low as 6.3%. PSMC analysis based on this genome sequence and assuming a mutation rate of 1e−8 per generation per nucleotide (Fig. 5a) suggested a drastic population expansion starting ca. 4×10^4 generations ago, which was followed by a decline from ca. 1.2×10^4 generations ago. For the last ca. 4×10^3 generations towards the present, the effective population size was low but stable.

RAD sequencing for the other 17 individuals using the restriction enzyme *Eco*RI resulted in an average of 114 819 RAD loci per individual, equivalent to a sampling density of 0.00025, hence $p \approx 2\%$. The distribution of the distance between consecutive RAD loci fits well with expectations from random sampling (Fig. S6, Supporting information). PSMC analysis to some extent retrieved the same population history as the whole-genome sequence (Fig. 5b), that is an expansion followed by a decline. However, the effective population size estimates and the timing of the demographic events were heavily biased, and when it came to the more recent part of the history, the variance increased massively across individuals making it difficult to draw reliable conclusions.

We tried to improve the analysis for the RAD data by increasing both the number of iterations and the number of chromosomes. The number of chromosomes was increased by pooling the chromosomes of 10 individuals to form the so-called pooled individual, considering that each chromosome pair independently reflects the same evolutionary history no matter whether it comes from the same individual or different individuals from the same population. Ten pooled individuals were created by sampling different subsets of individuals among the 17 RAD-sequenced sticklebacks. The results of the PSMC analyses involving 2500 iterations are shown in Fig. 5c. These produced estimates of demographic history with much higher accuracy, showing high resemblance to the results from the whole-genome data. The recent expansion and decline were clearly indicated, although the timing of the events was still somewhat biased. The most

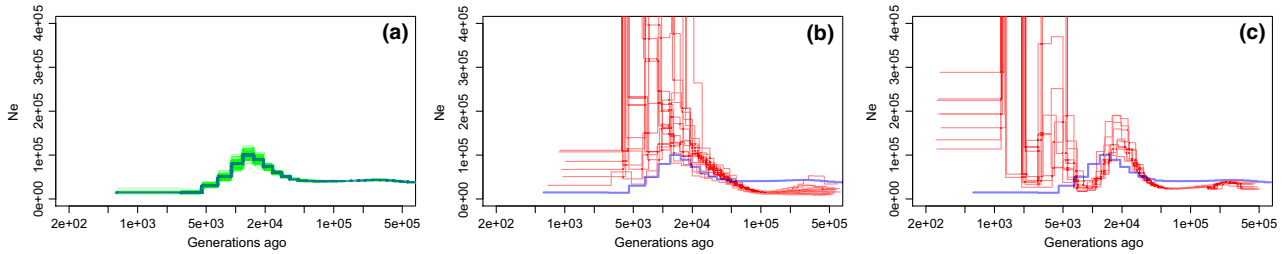

Fig. 5 Inference of the demographic history of three-spined sticklebacks from Lake Qarajat using PSMC. (a) Estimation based on the whole-genome sequence from a single individual. The blue line is the estimated history, whereas the green lines represent the 100 bootstrap runs. (b) Inference based on RAD data. Each red line represents estimates from one individual. The superimposed blue line is the estimate obtained from whole-genome sequencing. (c) Inference based on ten pooled individuals and 2500 iterations. Each red line represents the result from a pooled individual created by combining RAD data from 10 randomly chosen individuals. The superimposed blue line is the estimate obtained from whole-genome sequencing.

recent part of the history from ca. 5000 generations ago could not be resolved due to the increase in the variance.

Discussion

Both simulations and empirical results showed that a dilution of the genomic information associated with RAD sequencing still left coalescent information that could be used for analysing demographic history using PSMC. The proportion of the genome covered by RAD data was the limiting factor for the applicability. Both mutation rate and recombination rate could influence the threshold of this limit. The estimation could be improved by increasing both the number of iterations and the number of chromosomes at the same time. In particular, it helped improve the PSMC results from the RAD data of the stickleback population included in this study. Nevertheless, the dilution of genomic information in RAD sequencing incurs loss of information which particularly reduces the time span over which demographic history can be estimated. In the following, we elucidate the mechanisms behind the observations above and discuss the general prospects for using PSMC for RAD sequencing data.

Explaining the performance of PSMC for RAD data

To better understand our results, we here introduce two measures: the number of SNPs per recombination unit (SPR) and the number of recombination units (NR). SPR reflects the amount of coalescent information contained in each recombination unit. It determines whether the data can find the correct demographic history, namely the accuracy. NR can be viewed as the sample size of the recombination units and determines the variance of the estimate, namely the precision. RAD sequencing does not necessarily decrease NR, especially when the average length of the recombination units is large. However, it makes SPR decline drastically. For example, in Fig. 1, the estimation deviated increasingly from the real history with the decrease in sampling density of RAD. Nevertheless, the variance of estimates seemed to remain at the same level. For RAD data of a given species, SPR is correlated with p.

To create a visual representation of the recombination units and to see how well the RAD loci cover the recombination units, we plotted the posterior probability distribution of TMRCA (time to most recent common ancestor) across the genome (Figs S7 and S8, Supporting Information). A group of neighbouring nucleotides with similar TMRCA can be approximately regarded as belonging to the same recombination unit. The larger the TMRCA is, the smaller the recombination unit tends to be, and the harder it becomes for RAD loci to cover the unit. With the decrease in the sampling density, there are fewer RAD loci in each unit, reflecting the decrease in SPR.

Figure 3 can also be explained by SPR and NR. Higher mutation rate can increase SPR because of increased genetic diversity, but NR stays the same because the number of recombination events across the history does not depend on mutation rate. This has the effect that estimation based on whole-genome sequences converges even better towards the real history, whereas the variance of the estimation remains the same, as seen in a comparison of Fig. 1a vs. 3a. The RAD data in Fig. 3b have similar SPR as the whole-genome data in Fig. 1a and therefore display the same accuracy. Similarly, the RAD data in Figs 1b and 3c have the same accuracy. Lowering the recombination rate, on the other hand, cannot change genetic diversity, but it can increase the average length of the recombination units. It consequently yields higher SPR as well, whereas NR becomes lower. That is why Fig. 3a–c and d–f are similar in terms of accuracy, and yet, the latter shows lower precision. As for Fig. 3g–i, where the population size was higher, the genetic diversity increases, but it does not increase SPR. This is because a large population size also brings more

recombination events into the history and NR therefore increases, shortening the average length of the recombination units. As a result, the precision increases, while the accuracy stays unchanged. From these interpretations, we deduce that decreasing p has the same effect as increasing recombination rate or decreasing mutation rate. This is demonstrated in Fig. S9 (Supporting information), where we analysed simulated whole-genome data with various recombination and mutation rates.

SNPs per recombination unit determines the accuracy of PSMC estimation, and it can be quantified with the following term, μ^*p/r. The genetic diversity is positively correlated with $\boldsymbol{\Theta} = 4\mu N$. NR is positively correlated with $\rho = 4rN$. As SPR is the amount of genetic diversity per recombination unit, it is approximately equal to the ratio between $\boldsymbol{\Theta}$ and ρ, which can be reduced to μ/r. Because RAD has a diluting effect on SPR, we incorporate p into the term, which gives μ^*p/r. This can measure SPR in both whole-genome and RAD data, and can be used to predict the performance of PSMC on the data and provide guidance for the proper conduct of the analysis. The higher μ^*p/r is, the better the performance will be. Conversely, if the value is low, the accuracy of the estimation will decrease. According to our simulations (Fig. S10, Supporting Information), when μ^*p/r decreases until 0.5, PSMC can still provide relatively good estimates. But once the value drops below 0.5, the estimation becomes unreliable, and some extra measures need to be implemented to restore the performance.

Mutation rate and recombination rate estimates are hard to obtain in reality. However, the ratio between them can be obtained through certain software (Li & Durbin 2011; Schiffels & Durbin 2014). PSMC itself can estimate this ratio, although not accurately as it tends to provide overestimates (Li & Durbin 2011) (Fig. S10, Supporting Information). However, from the results in Fig. S10 (Supporting information), the estimation still seems informative as long as estimates are higher than 1. In the case of the stickleback genome sequence, PSMC estimated the ratio to be ca. 3, indicating a good performance of PSMC for the whole-genome data of this species.

When μ^*p/r is low, the accuracy can be improved by increasing the number of iterations and the number of chromosomes. For example, in our simulations, for RAD data with $d = 0.001$ we can still obtain a good estimation of the demographic history by increasing m from 25 to 250 and k from 10 to 100 (Fig. 4b). For RAD data with $d = 0.0001$, m will need to increase to 2500 or perhaps even more, while k has to be increased as well (Fig. 4c). In practice, the amount of data can be increased by pooling individuals, which is equivalent to increasing the number of chromosomes within an individual. As long as the individuals are randomly chosen from a population, their chromosomes should independently reflect the evolutionary history of the population. This solution is also practically feasible because sample sizes in most studies employing RAD sequencing are relatively high (>10). Yet, the drawback of such measures involves an increase in the computational time by magnitudes. Perhaps, modification in the expectation maximization algorithm of this software could make the process faster.

Empirical results from three-spined sticklebacks

Similar to the simulated data, the whole-genome data and RAD data from the sticklebacks converged towards the same population history when pooled individuals were created and the number of iterations was increased. Hence, the recent expansion ca. 4×10^4 generations ago and the decline starting ca. 1.2×10^4 generations ago were picked up by both data sets.

When interpreting the results, it should be noted that they are scaled by mutation rate. For simplicity, we assumed a rate of 10^{-8} per generation per nucleotide, and a higher rate would shift the demographic events more towards the present and result in lower effective population size estimates. In humans, genome-wide mutation rates of 1.25×10^{-8} and 2.5×10^{-8} have been assumed in various studies (Nachman & Crowell 2000; Altshuler *et al.* 2010). In a recent study using PSMC analysis of whole-genome sequences of three-spined sticklebacks to reconstruct demographic history, a range of mutation rates from 1.4×10^{-8} to 6.6×10^{-8} were considered, leading to qualitatively highly different interpretations of results (Liu *et al.* 2016). Also, generation time can be difficult to estimate and may differ across time and geographical regions. In three-spined stickleback, DeFaveri & Merila (2013) estimated generation length to be between 2 and 4 years in Fennoscandia, which also implies that it could be shorter in more southern regions with higher temperature leading to faster growth and earlier age at reproduction. Finally, when analysing historical demographic history it is often the case that the most recent estimates pertain to the sampled population, but as estimates move back in time they are likely to encompass the whole species or perhaps even several related species prior to their divergence.

With these caveats in mind, we suggest that the most recent population expansion inferred from the PSMC plots reflects population expansion of sticklebacks towards the end of the last (Weichselian) glaciation (ca. 115 000–11 700 years bp). The subsequent decline would then reflect colonization of the Lake Qarajat by marine sticklebacks, following which the lake population would harbour a lower effective population size than the ancestral marine and presumably widespread population.

This interpretation is supported by findings of subfossil remains of sticklebacks in lakes in the Nuuk Fjord region dating back to ca. 9000 years bp (Bennike 1997).

Using PSMC analysis for RAD sequencing data

Our results show that PSMC analysis has potentials for analysis of RAD data, and they provide suggestions for how to optimize its performance. First, it is helpful to have knowledge about mutation and recombination rates for the study species. If this information is unavailable, it is recommended to sequence the whole genome of at least one individual of the species and conduct PSMC analysis, which will provide an estimate of the ratio between mutation rate and recombination rate. Knowing this ratio can assist in determining the range of p that we should aim for in the RAD data to optimize the performance of PSMC, because we know that its performance is determined by μ^*p/r. According to this expectation and the read lengths that can be obtained from the sequencing platform, we can then choose which restriction enzyme to use. Simulation tools such as simRAD (Lepais & Weir 2014) are highly useful for this purpose. If, after all, the value for μ^*p/r is still below the range expected to lead to good performance of PSMC, one can consider improving the power of inference by increasing the number of iterations and creating pooled individuals. The single factor that can most realistically lead to higher p is read length. Fortunately, there is a continuing development towards increasing the length of reads for virtually all-sequencing platforms, and for instance (at the time of writing), read lengths of 300 bp can now be obtained using the Illumina MiSeq platform.

Using RAD data rather than whole genome sequencing obviously involves some loss of resolution, particularly by decreasing the timescale over which demographic history can be inferred. However, if addressing the general scientific question in a project requires RAD sequencing of several individuals, then our results show that PSMC analysis may still be a possibility, which can provide information over a timescale of relevance to phylogeography, although cautions must be taken when interpreting the results. Finally, our results may provide inspiration for testing the usefulness of RAD data for other coalescent-based statistical methods and developing similar tools for RAD data.

Acknowledgements

We thank Annie Brandstrup for Technical Assistance, Michael Glad for maintaining computers, Rasmus Nygaard and Rasmus Hedeholm for assistance with sampling sticklebacks in Greenland, Joshua Schraiber, three anonymous referees and the Subject Editor for constructive comments on previous versions of the manuscript, and the Danish Council for Independent Research| Natural Science for funding (Grant no. 1323-00158A to MMH).

References

Altshuler D, Durbin RM, Abecasis GR *et al*. (2010) A map of human genome variation from population-scale sequencing. *Nature*, **467**, 1061–1073.

Baird NA, Etter PD, Atwood TS *et al*. (2008) Rapid SNP discovery and genetic mapping using sequenced RAD markers. *PLoS ONE*, **3**, e3376.

Bennike O (1997) Quaternary vertebrates from Greenland: a review. *Quaternary Science Reviews*, **16**, 899–909.

Catchen J, Hohenlohe PA, Bassham S, Amores A, Cresko WA (2013) Stacks: an analysis tool set for population genomics. *Molecular Ecology*, **22**, 3124–3140.

Colosimo PF, Hosemann KE, Balabhadra S *et al*. (2005) Widespread parallel evolution in sticklebacks by repeated fixation of ectodysplasin alleles. *Science*, **307**, 1928–1933.

Davey JW, Hohenlohe PA, Etter PD *et al*. (2011) Genome-wide genetic marker discovery and genotyping using next-generation sequencing. *Nature Reviews Genetics*, **12**, 499–510.

DeFaveri J, Merila J (2013) Variation in age and size in Fennoscandian three-spined sticklebacks (*Gasterosteus aculeatus*). *PLoS ONE*, **8**, e80866.

Ellegren H (2014) Genome sequencing and population genomics in non-model organisms. *Trends in Ecology & Evolution*, **29**, 51–63.

Elshire RJ, Glaubitz JC, Sun Q *et al*. (2011) A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PLoS ONE*, **6**, e19379.

Excoffier L, Dupanloup I, Huerta-Sanchez E, Sousa VC, Foll M (2013) Robust demographic inference from genomic and SNP data. *Plos Genetics*, **9**, e1003905.

Ferchaud AL, Hansen MM (2016) The impact of selection, gene flow and demographic history on heterogeneous genomic divergence: three-spine sticklebacks in divergent environments. *Molecular Ecology*, **25**, 238–259.

Gutenkunst RN, Hernandez RD, Williamson SH, Bustamante CD (2009) Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. *Plos Genetics*, **5**, e1000695.

Hudson RR (2002) Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics*, **18**, 337–338.

Jones FC, Grabherr MG, Chan YF *et al*. (2012) The genomic basis of adaptive evolution in threespine sticklebacks. *Nature*, **484**, 55–61.

Langmead B, Trapnell C, Pop M, Salzberg SL (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology*, **10**, R25.

Lanier HC, Massatti R, He QX, Olson LE, Knowles LL (2015) Colonization from divergent ancestors: glaciation signatures on contemporary patterns of genomic variation in Collared Pikas (*Ochotona collaris*). *Molecular Ecology*, **24**, 3688–3705.

Lepais O, Weir JT (2014) SimRAD: an R package for simulation-based prediction of the number of loci expected in RADseq and similar genotyping by sequencing approaches. *Molecular Ecology Resources*, **14**, 1314–1321.

Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, **25**, 1754–1760.

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

Liu XM, Fu YX (2015) Exploring population size changes using SNP frequency spectra. *Nature Genetics*, **47**, 555–559.

Liu S, Hansen MM, Jacobsen MW (2016) Region-wide and ecotype-specific differences in demographic histories of threespine stickleback populations, estimated from whole genome sequences. *Molecular Ecology*, **25**, 5187–5202.

McVean GAT, Cardin NJ (2005) Approximating the coalescent with recombination. *Philosophical Transactions of the Royal Society B-Biological Sciences*, **360**, 1387–1393.

Moura AE, Janse van Rensburg C, Pilot M *et al.* (2014) Killer whale nuclear genome and mtDNA reveal widespread population bottleneck during the last glacial maximum. *Molecular Biology and Evolution*, **31**, 1121–1131.

Nachman MW, Crowell SL (2000) Estimate of the mutation rate per nucleotide in humans. *Genetics*, **156**, 297–304.

Nadachowska-Brzyska K, Li C, Smeds L, Zhang GJ, Ellegren H (2015) Temporal dynamics of avian populations during Pleistocene revealed by whole-genome sequences. *Current Biology*, **25**, 1375–1380.

Palkopoulou E, Mallick S, Skoglund P *et al.* (2015) Complete genomes reveal signatures of demographic and genetic declines in the woolly mammoth. *Current Biology*, **25**, 1395–1400.

Peterson BK, Weber JN, Kay EH, Fisher HS, Hoekstra HE (2012) Double Digest RADseq: an inexpensive method for de novo SNP discovery and genotyping in model and non-model species. *PLoS ONE*, **7**, e37135.

Poland JA, Brown PJ, Sorrells ME, Jannink JL (2012) Development of high-density genetic maps for barley and wheat using a novel two-enzyme genotyping-by-sequencing approach. *PLoS ONE*, **7**, e32253.

Pujolar JM, Jacobsen MW, Frydenberg J *et al.* (2013) A resource of genome-wide single-nucleotide polymorphisms generated by RAD tag sequencing in the critically endangered European eel. *Molecular Ecology Resources*, **13**, 706–716.

R Core Team (2014) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna. URL: http://www.R-project.org/.

Schiffels S, Durbin R (2014) Inferring human population size and separation history from multiple genome sequences. *Nature Genetics*, **46**, 919–925.

Sheehan S, Harris K, Song YS (2013) Estimating variable effective population sizes from multiple genomes: a sequentially Markov conditional sampling distribution approach. *Genetics*, **194**, 647–662.

Wallberg A, Han F, Wellhagen G *et al.* (2014) A worldwide survey of genome sequence variation provides insight into the evolutionary history of the honeybee *Apis mellifera*. *Nature Genetics*, **46**, 1081–1088.

Xue YL, Prado-Martinez J, Sudmant PH *et al.* (2015) Mountain gorilla genomes reveal the impact of long-term population decline and inbreeding. *Science*, **348**, 242–245.

S.L and M.M.H. conceived and designed the study. S.L. conducted simulations, bioinformatics and statistical analyses. S.L. and M.M.H. wrote the manuscript and both of them approved the final version.

Data accessibility

The RAD sequencing reads from empirical data are deposited in the NCBI Sequence Read Archive under project number PRJNA343338. The whole-genome sequencing reads from empirical data are deposited in the EBI Sequence Read Archive with Accession no. ERR1599907. The scripts mentioned in the text and the PSMC input files for empirical data are deposited in DRYAD (doi: 10.5061/dryad.0618v).

Supporting Information

Additional Supporting Information may be found in the online version of this article:

Fig. S1 The influence of number of iterations on the accuracy and the precision of PSMC inference based on the simulated data. The grey lines represent the demographic histories from which the data were simulated, whereas the red lines denote estimations obtained by running PSMC for the simulated data. Each simulation was repeated 10 times. Demographic scenario: Sim-1; whole genome data.

Fig. S2 The influence of numbers of chromosomes on the accuracy and the precision of PSMC inference based on the simulated data. The grey lines represent the demographic histories from which the data were simulated, whereas the red lines denote estimations obtained by running PSMC for the simulated data. Each simulation was repeated 10 times. Demographic scenario: Sim-1; whole genome data.

Fig. S3 The simulation results for divergence time estimation with RAD data under various sampling density. The grey lines specify the expected plot under the demographic history Sim-split. The vertical part of these lines points to the time when the two populations split. The red lines are the estimations using PSMC. a-c. Results of simulations with known phase, based on whole genome sequences, RAD with $d = 0.001$ and RAD with $d = 0.0001$, respectively. d-f. Results of simulations with unknown phase, based on whole genome sequences, RAD with $d = 0.001$ and RAD with $d = 0.0001$, respectively. g-i. Results of simulations where phasing has been conducted, based on whole genome sequences, RAD with $d = 0.001$ and RAD with $d = 0.0001$, respectively. Each simulation was repeated 10 times.

Fig. S4 Improving the divergence time estimation by increasing the number of iterations and the number of chromosomes. The grey lines specify the expected plot under the demographic history Sim-split. The vertical part of these lines points to the time when the two populations split. The red lines are the estimations using PSMC. Each simulation was repeated 10 times.

Fig. S5 Coverage distribution (a) and the insert size distribution (b) of the threespine stickleback whole genome sequence from Lake Qarajat in Nuuk Fjord, Greenland after mapping the sequencing reads to the reference genome (Jones *et al.* 2012).

Fig. S6 Distribution of the logarithm (base = 10) of distance between consecutive RAD loci. The plot on the left is the observed distribution from the RAD data obtained from the 17 threespine sticklebacks from Lake Qarajat, Greenland. The plot on the right is the expected distribution assuming random sampling of RAD loci from the genome with the same sampling density as the RAD data obtained from sticklebacks.

Fig. S7 Posterior probability distribution of TMRCA (time to most recent common ancestor) across the genome inferred by PSMC. The genome comes from simulations under the scenario of Sim-1. Only the first 300 Kb of Chromosome 1 is shown here. The bars above the plot indicates the location of the RAD loci under sampling density of $d = 0.001$ (lower) and $d = 0.0001$ (upper).

Fig. S8 Posterior probability distribution of TMRCA (time to most recent common ancestor) across a part of the genome inferred by PSMC. The genome comes from the whole genome

sequencing result of a threespine stickleback from Lake Qarajat, Greenland. Only the first 300 Kb of "groupI" is shown here. The bars above the plot indicates the location of the RAD loci obtained from the 17 sticklebacks from the same lake.

Fig. S9 The performance of PSMC under low mutation rate or high recombination rate. a-b. Decreasing mutation rates to 1e-9 and 1e-10, respectively. c-d. Increasing recombination rates to 2e-8 and 2e-7, respectively. The grey lines represent the simulated population histories, whereas the red lines denote the estimations using PSMC. Each simulation was repeated 10 times. Demographic scenario: Sim-1; whole genome data.

Fig. S10 The performance of PSMC under various μ/r. The grey lines represent the simulated population histories, whereas the red lines denote the estimations using PSMC. Above each plot, "μ/r" specifies the μ/r applied in the simulation, and "Estimated" shows the μ/r estimated by PSMC. Each simulation was repeated 10 times. Demographic scenario: Sim-1; whole genome data (hence $p = 1$).