Joint inference of demography and mutation rates from polymorphism data and pedigrees

Florence Parat[*,1], **Sándor Miklós Szilágyi**[†,‡,§], **Daniel Wegmann**[§,**,a] **and Aurélien Tellier**[*,a]

[*]Centre of Life and Food Sciences Weihenstephan, Technische Universität München, 85354 Freising, Germany, [†]Department of Informatics, Faculty of Sciences and Letters, Petru Maior University, 540088 Tîrgu Mureş, Romania, [‡]Faculty of Electrical Engineering and Informatics, Department of Control Engineering and Information Technology, Budapest University of Technology and Economics, H-1117 Budapest, Hungary, [§]Department of Biology, University of Fribourg, Fribourg, Switzerland, [**]Swiss Institute of Bioinformatics, Fribourg, Switzerland, [a]These authors contributed equally

ABSTRACT Inference of demography and mutation rates is of major interest but difficult because genetic data is only informative about the population mutation rate, the product of the effective population size times the mutation rate, and not about these quantities individually. Here we show that this limitation can be overcome by combining genetic data with pedigree information. To successfully use pedigree data, however, important aspects of real populations such as the presence of two sexes, unbalanced sex ratios and overlapping generations have to be taken into account. We present here an extension of the classic Wright-Fisher model accounting for these effects and show that the coalescent process under this model reduces to the classic Kingman coalescent with specific scaling parameters. We further derive the probability of a pedigree under that model and show how pedigree data can thus be used to infer demographic parameters. Finally, we present a computationally efficient inference approach combining pedigree information and genetic data summarized by the site frequency spectrum (SFS) that allows for the joint inference of the mutation rate, sex-specific population sizes and the fraction of overlapping generations. Using simulations we then show that these parameters can be accurately inferred from pedigrees spanning just a few generations, as are available for many species. We finally discuss future possible extensions of the model and inference framework necessary for applications to wild and domesticated species, namely the account for more complex demographies and the uncertainty in assigning pedigree individuals to specific generations.

KEYWORDS likelihood, domesticated species, sex-ratio bias

Demographic processes shape the genetic variation and genetic structure of populations and species, but also affect the efficacy of selection. There is thus considerable interest in inferring past demographic events, not least to serve as a null model for the identification of markers under selection. To this end, several methods have been proposed to estimate past demography from genetic data, and many of which are based on coalescent theory using either likelihood methods (e.g., Hey and Nielsen 2007; Hey 2011; Excoffier *et al.* 2013) or simulations embedded in Approximate Bayesian Computation (e.g., Beaumont *et al.* 2002; Wegmann *et al.* 2009). In these approaches, the coalescent model is used to link the genetic information to demographic parameters by integrating over the unobserved genetic relationships between individuals in a population in a computationally efficient way.

While the standard coalescent approach assumes no prior information on the true (parent-offspring) relationship between genetic lineages, knowledge on this would bring complementary information and hence increase estimation accuracy. Indeed, several methods have been proposed to use pedigree information to infer demographic processes using the increase of inbreeding over time under a given reproduction model (Falconer and Mackay 1996; Gutiérrez *et al.* 2008). Additionally, a method that allows the simulation of pedigrees under a given demographic and reproductive model and draws genealogies inside these pedigrees was developed (Gasbarra *et al.* 2005), and could be

doi: 10.1534/genetics.XXX.XXXXXX
Manuscript compiled: Wednesday 30th November, 2016%
[1]Centre of Life and Food Sciences Weihenstephan, Technische Universität München, 85354 Freising, Germany, parat@wzw.tum.de

used, in an ABC framework, for inference. However, there is currently no general inference framework for such data.

Here we develop a maximum-likelihood method to infer demographic parameters from both pedigree information and genetic data summarized by the site frequency spectrum (SFS) of a sample. We postulate that the pedigree of a sample contains information about the demography of the population at least partially complementary to the information contained in the genetic data and a method exploiting the full information should therefore improve the inference of demographic parameters.

An additional advantage of our framework is its ability to infer effective population sizes (N_e) and mutation rates (μ) jointly. Under the standard coalescent framework, both parameters are simply scaling the coalescent tree and hence only their product can be estimated, usually in the form $\theta = 4N_e\mu$. Wakeley and Takahashi (2003) showed that a joint estimation becomes feasible if the sample size n exceeds N_e since the rate of coalescence in the first few generation is a function of the ratio n/N_e and hence contains information about N_e regardless of μ. This was later used to infer gene specific mutation rates in humans from deep sequencing data Nelson *et al.* (2012); Schaibley *et al.* (2013). As we show here, pedigree information also contains information about N_e independent of μ, enabling the joint inference of both demography and mutation rate even in case where $n \ll N_e$.

Pedigree information is available for many populations or species, in particular for managed populations under conservation management or domesticated animals under active breeding (e.g., Clutton-Brock *et al.* 1982; Ellegren 1999; Cunningham *et al.* 2001; Mc Parland *et al.* 2007). Yet many of these species have important life history traits that are not reflected in the standard Wright-Fisher model, including overlapping generations and two sexes. Additionally, in most domesticated species, fewer males are reproducing than females but males can reproduce over a longer time period, spanning several generations. While such life history traits have an impact on the response of the population to selection and are therefore accounted for in breeding programs Hill (1974), changes in allele frequencies due to drift remain well described by scaling the models with an appropriate effective population size (N_e) Wright (1931); Engen *et al.* (2007). As a consequence, demographic inference in domesticated species using coalescent theory does usually not model overlapping generation or sex biased population sizes.

However, simple scaling does not extend to models incorporating pedigree information. To address this, we present here a Wright-Fisher-based diploid two-sex model with overlapping generation well describing pedigrees observed from domesticated breeding programs. We then show how this model results in a simple scaling of the standard coalescent in the absence of pedigree information and derive some analytical and numerical solutions to obtain estimates of the model parameters using the information contained in the SFS and pedigree data jointly. This also allows for the estimation of important life history characteristics such as the degree of overlapping generations and sex specific population sizes from such data.

The Model

To account for life history traits very common in animal populations, we extend the classic Wright-Fisher model to a diploid species with two sexes and overlapping generations. Our model, which is schematically depicted in Figure 1, assumes discrete generations consisting of N_f female and N_m male individuals. We further assume random mating in that in each generation,

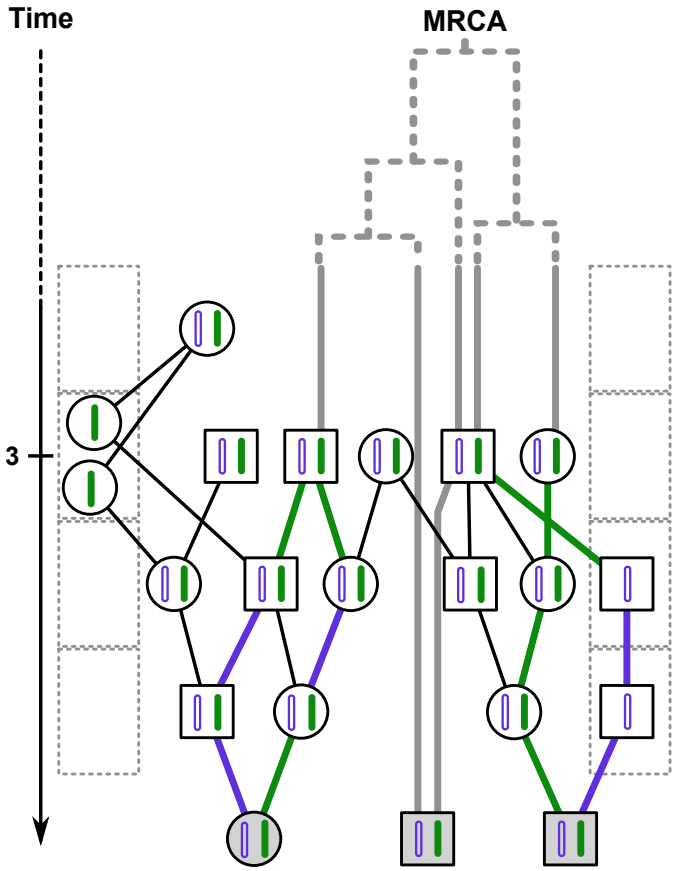

Figure 1 Graphical representation of the proposed two-sex model with overlapping generations. Shown are female (circles) and males (squares) individuals along with their parent-offspring relationships (black edges). The doted boxes on each side represent the female and male gamete storages populated with gametes of individuals from the pedigree. or the individuals of generation 0 (grayed) genetic data is available and the thick lines represent the genealogy of these individual. Dotted lines indicate the part of the genealogy before g_{Max} for which the continuous time approximation is used. The tick mark on the time scale represent the depth of this pedigree, $d = 3$.

and going backward in time, each individual picks a random female and male individual of the previous generation as mother and father, respectively.

We model overlapping generations by allowing individuals to pick their parents not from the directly preceding generation but from an earlier one with probabilities b_f and b_m for female and male parents, respectively. In this case, however, the choice of the actual distant parent is delayed and the lineage is just stored. In biological terms, these stored lineages thus represent gametes of a defined sex from previous generations, and we will refer to this compartment as "gamete storage" in the following. At the beginning of a generation, the so stored gametes then pick a parent in the current generation with probabilities $1 - b_f$ and $1 - b_m$, and otherwise remain in storage, which implies that the number of generations between parents of offspring are exponentially distributed.

For simplicity, we will considered here only the case of constant population sizes N_f, N_m and probabilities to jump a generation b_m, b_f, but we note that relaxing this assumption is straight

forward under the inference framework introduced below.

Derivation of the coalescent

Two-sex models were previously shown (Möhle 1998) to be approximated accurately by the time-changed Kingman coalescent (1982). Similarly, Blath *et al.* (2013) recently showed that models with overlapping generations also result in a simple scaling of the classic coalescent if the average time lineages spend in storage is relatively small compared to the waiting time between coalescent events. Here we derive the appropriate scaling for the model introduced above.

We begin with the rate of coalescent and note that only the lineages that are in individuals of the real populations can coalesce, whereas lineages currently stored for later use have first to reenter the real populations. To obtain the fraction of lineages that can coalesce with each other, we obtain their relative distribution at equilibrium in the following four possible compartments: the real female (R_f) and male (R_m) populations as well as the female (B_f) and male (B_m) gamete storages (Supplementary Figure S1) using the following system of difference equations:

$$\begin{cases} \Delta B_m = R_m \frac{b_m}{2} + R_f \frac{b_m}{2} - B_m(1-b_m) \\ \Delta R_m = B_m(1-b_m) + R_f \frac{1-b_m}{2} - R_m \frac{b_m}{2} - R_m \frac{b_f}{2} - R_m \frac{1-b_f}{2} \\ \Delta B_f = R_f \frac{b_m}{2} + R_m \frac{b_m}{2} - B_f(1-b_f) \\ \Delta R_f = B_f(1-b_f) + R_m \frac{1-b_f}{2} - R_f \frac{b_f}{2} - R_f \frac{b_m}{2} - R_f \frac{1-b_m}{2}. \end{cases}$$

The global rate of coalescence $\Pr(\text{Coal})$ is then given by sum of the per compartment rates weighted by the fraction of lineages residing in them. Since the coalescent rates are zero in the gamete storages and $1/2N_f$ and $1/2N_m$ in R_f and R_m, respectively, we have

$$\mathbb{P}(\text{Coal}) = \frac{1}{2N_f}\left[\frac{(1-b_f)^2}{(1-b_m)+(1-b_f)}\right]^2 + \frac{1}{2N_m}\left[\frac{(1-b_m)^2}{(1-b_m)+(1-b_f)}\right]^2.$$

In accordance with previous results for two-sex (Möhle 1998) or seed bank (Kaj *et al.* 2001) models, the obtained rate reduces to

$$\mathbb{P}(\text{Coal}) = \frac{1}{2}\frac{N_m+N_f}{4N_mN_m},$$

if $b_f = b_m = 0$, and to

$$\mathbb{P}(\text{Coal}) = \frac{(1-b)^2}{2}\frac{N_m+N_f}{4N_mN_m}.$$

if $b_f = b_m = b$.

Following Kingman's approach, the distribution of time of coalescent under our model is $T_i \sim \exp\binom{i}{2}$ with time scaled in $2N_e$ with

$$N_e = \frac{N_fN_m(2-b_m-b_f)^2}{N_f(1-b_m)^4 + N_m(1-b_f)^4}.$$

We next derive the rate of novel mutations in the presence of overlapping generations. Importantly, the number of germ-line mutations may not scale linearly with time, as, especially in females, most of these mutations occur during early development (Crow 2000). We model this effect using two mutation rates: μ for the part of the branches connecting real individuals and $\mu^* = \varepsilon\mu$ for the time spent in the gamete storage. From the compartment model introduced above, we obtain the average fraction of time t_b that linages spend in one of the gamete storages as

$$t_b = \frac{B_f+B_m}{R_f+R_m+B_f+B_m} = \frac{b_f-2b_fb_m+b_m}{(1-b_m)+(1-b_f)},$$

which results in an average effective mutation rate $\bar{\mu}$ per generation:

$$\bar{\mu} = \mu + \frac{b_f-2b_fb_m+b_m}{(1-b_m)+(1-b_f)}(1-\varepsilon)\mu.$$

Inference

We introduce here a Maximum Likelihood (MLE) method to infer jointly the demographic $\theta_d = \{N_f, N_m, b_f, b_m\}$ and mutational $\theta_m = \{\mu, \varepsilon\}$ parameters of the model introduced above. This estimation is based on genetic data summarized by the site frequency spectrum (SFS) and available pedigree information in terms of child-parent relationships (filiation) that form one or several connected networks spanning two or more generations ($\mathcal{P}$). The relevant likelihood function can be decomposed as

$$\begin{aligned} L(\mathcal{M}) &= \Pr(\text{SFS}|\mathcal{P},\theta_d,\theta_m)\Pr(\mathcal{P}|\theta_d) \\ &= \sum_G \left[\Pr(\text{SFS}|G,\theta_m)\Pr(G|\mathcal{P},\theta_d)\right]\Pr(\mathcal{P}|\theta_d), \end{aligned} \quad (1)$$

where the sum runs over the unknown genealogies G representing the genetic relationships between all sampled individuals up to the the most recent common ancestor (MRCA). While the pedigree and the genealogies share similar features, they should not be confused.

In the following sections, we will first derive each term of the likelihood function individually, and then give a detailed description of an inference framework under this model.

The Pedigree

Let $\mathcal{P}_g$ be the way in which the individuals of generation $g-1$ in the pedigree are assigned to their parents in generation g. Note that generations as well as the choice of the mother and the father are independent, and hence that

$$\Pr(\mathcal{P}|\theta_d) = \prod_{g\geq 1}\Pr(\mathcal{P}_g|\theta_d) = \prod_{g\geq 1}\Pr(\mathcal{P}_{f,g}|\theta_d)\Pr(\mathcal{P}_{m,g}|\theta_d), \quad (2)$$

where $\mathcal{P}_{f,g}$ and $\mathcal{P}_{m,g}$ represent the assignment of individuals to their mothers and fathers, respectively.

The pedigree spans between the generation of the most recent individual $g=0$ and the generation of the last known parent that we call g_{Max}. To derive the probability of the pedigree, we consider all individuals at generation 0 in the pedigree as numbered (i.e. identifiable). These individuals then choose their parents from the previous generation, but are constrained in their choices by the pedigree. The so chosen parents become identifiable themselves and in turn choose the parents from the previous generation according to the pedigree information of that generation. This process continues until the top of the pedigree is reached.

Here we will derive $\Pr(\mathcal{P}_{f,g}|\theta_d)$ for this process for the individuals of generation $g-1$, of which exactly $B_{f,g-1}$ will enter the gamete storage with probability b_f as their mother is from a distant generation. Among the $\bar{B}_{f,g-1}$ that choose a mother

from generation g, the first individual of each of the M_g groups of siblings chooses a distinct mother from the population, which they do in turn with probabilities 1, $\frac{N_f-1}{N_f}, \dots \frac{N_f-M_g}{N_f}$. The M_g so chosen mothers, which have become identifiable themselves, are chosen by their remaining offspring with probability $\frac{1}{N_f}$ each. The resulting probability of this process is

$$\mathbb{P}(\mathcal{P}_{f,g}|\theta_d) = \alpha_{fg} \left(\frac{1}{N_f}\right)^{\bar{B}_{f,g-1}-M_g} b_f{}^{B_{f,g-1}} (1-b_f)^{\bar{B}_{f,g-1}}, \quad (3)$$

where we used the notation

$$\alpha_{fg} = \frac{N_f!}{N_f^{M_g}(N_f-M_g)!}.$$

The same holds true analogously for $\mathbb{P}(\mathcal{P}_{m,g}|\theta_d)$ by replacing the subscript f by m and using F_g, the number of fathers in generation g, instead of M_g.

A maximum likelihood estimate of N_f, N_m, b_f and b_m is easily obtained by taking the first derivative of the logarithm eq. 2. For b_f, thies yields

$$\frac{\mathrm{d}}{\mathrm{d}b_f} \log \mathbb{P}(\mathcal{P}|\theta_d) = \frac{\sum_{g=1}^{g_{\mathrm{Max}}} B_{f,g-1}}{b_f} - \frac{\sum_{g=1}^{g_{\mathrm{Max}}} \bar{B}_{f,g-1}}{(1-b_f)},$$

which admits the maximum likelihood estimate

$$\hat{b_f} = \frac{\sum_{g=1}^{g_{\mathrm{Max}}} B_{f,g-1}}{\sum_{g=1}^{g_{\mathrm{Max}}} \bar{B}_{f,g-1} + B_{f,g-1}},$$

and analogously for b_m. For N_f, the first derivative is

$$\frac{\mathrm{d}}{\mathrm{d}N_f} \log \mathbb{P}(\mathcal{P}|\theta_d) = \sum_{g=1}^{g_{\mathrm{Max}}} \mathcal{F}(N_f) - \mathcal{F}(N_f-M_g) - \frac{\bar{B}_{f,g-1}}{N_f}, \quad (4)$$

where $\mathcal{F}$ is the digamma function defined as the logarithmic derivative of the factorial function. As the population size is a finite natural number, the maximum of this probability, if finite, can be easily found numerically using dichotomy.

Genetic data

Coalescence is the merging of two or more genetic lineages. In a diploid population, an offspring may inherit one of two possible chromosomes of each parent. There are thus 2^l ways in which l offspring lineages can be assigned to the two chromosomes of a single parent (Figure 1). Enumerating all possible genealogies constraint by even a small but fully resolved pedigree, as done for two lineages in Wakeley *et al.* (2012), is computationally already very challenging for large sample sizes, and easily becomes prohibitive if the pedigree is only partially known. We thus chose to turn to simulations to evaluate the sum in eq. 1, as is commonly done in the absence of pedigree information (e.g. Excoffier *et al.* 2013; Nielsen 2000; Nelson *et al.* 2012):

$$\sum_G \left[\Pr(\mathrm{SFS}|G,\theta_m)\Pr(G|\mathcal{P},\theta_d)\right] \approx \frac{1}{N_g}\sum_g \Pr(\mathrm{SFS}|G=g,\theta_d,\theta_m),$$

where the genealogies $g \sim \Pr(G|\mathcal{P},\theta_d,\theta_m)$ are simulated under model parameters $\mathcal{M}$ and constrained by the pedigree $\mathcal{P}$.

Simulating genealogies inside a pedigree is straight forward and only requires binary choices when following lineages backward in time through the pedigree. In case of only partial pedigree information, lineages reaching parents of which only one or none of the parents are known choose their unknown parents randomly from the whole population, or enter the gamete storage (Figure 1). Since it is required to keep track of all these choices, the simulations become rather time consuming in case of limited pedigree information. At a certain generation in the past we term g_{Max}, the pedigree does not contain any information about ancestors anymore and the genealogy is then only constraint by the parameters of the model. As g_{Max} is often reached long before the MRCA, we make use of the appropriately scaled coalescent approximation introduced above to simulate the genealogies from g_{Max} backwards to the MRCA (Figure 1).

To calculate $\Pr(\mathrm{SFS}|G=g,\theta_m)$, the probability of the genetic data summarized by the SFS given a genealogy g, we use the classic infinite site mutation model with Poisson distributed mutations at rate $\bar{\mu}$ per site. Under this model, the probability that a mutation results in a derived sample allele frequency of i is given by the summed length L_i of all branches with i leaves and the probability of the SFS is thus given by a multinomial distribution

$$\Pr(\mathrm{SFS}|G=g,\mathcal{M}) = e^{-\bar{\mu}L}\frac{\bar{\mu}^S L_1^{S_1}\dots L_{n-1}^{S_{n-1}}}{S_1!\dots S_{n-1}!}, \quad (5)$$

where S_i is the number of segregating sites being shared by i chromosomes in the sample of size n and L the total length of the genealogy g (Fu 1998). We note that the branch lengths L_i are measured in mutational generations, and hence all generations spent in the gamete storage only add ε to the total length.

The maximum likelihood estimate of $\bar{\mu}$ can be obtained analytically by differentiating the logarithm of eq. 5, which yields the estimator

$$\hat{\mu} = \frac{S}{L},$$

where L is the total length of the genealogy in mutational generations. In the absence of pedigree information, for instance for the part of the genealogy simulated under the coalescent approximation, the total length of the genealogy is only available measured in generations. In this case, the ML estimate becomes

$$\hat{\mu} = \frac{(1-\hat{b})S}{4\hat{N}_e L_c}, \quad (6)$$

where L_c is the total length of the genealogy in coalescent time (i.e., in θ generations) and $\hat{b}$ and $\hat{N}_e$ are the ML estimates of b and N_e, respectively.

Inference algorithm

An exact analytical or numerical solution for the joint maximum likelihood of all parameters is not available. We will therefore combine some of our analytical derivations with numerical evaluations to perform an MCMC-MLE inference as follows:

1. We sample vectors of demographic parameters $\theta_d^{(i)} \sim \Pr(\theta_d|\mathcal{P}), i = 1,\dots,I$ from their joint posterior distribution given the pedigree using an MCMC framework.

2. For each sampled vector of parameters $\theta_d^{(i)}$, we simulate $G = 100$ genealogies constraint by the pedigree $\mathcal{P}$.

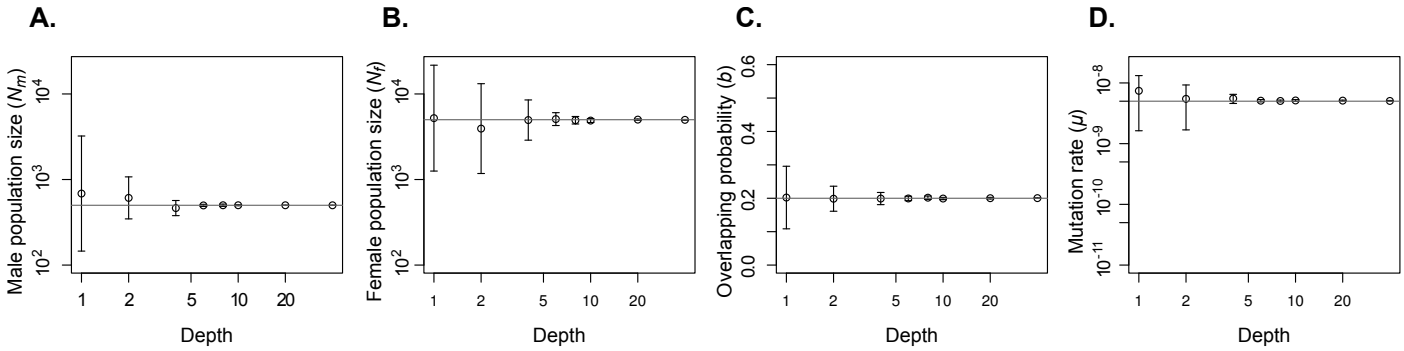

Figure 2 Power of parameter inference as a function of pedigree depth. Shown are the mean and standard deviation over 10 simulated datasets. The true parameter values used for all simulations (A) $N_m = 500$, (B) $N_f = 500$, (C) $b = 0.2$, and (D) $\mu = 5 \times 10^{-9}$ are indicated by gray horizontal lines.

3. For each $\theta_d^{(i)}$ we then compute the MLE estimate $\bar{\mu}^{(i)}$ according to eq. 6 using the sampled genealogies.

4. Finally, we compute the joint likelihood of all model parameters for each pair of $\theta_d^{(i)}$ and $\bar{\mu}^{(i)}$ according to eq. 1, again using the simulated genealogies.

Our inference scheme is thus closely related to a grid search on the model parameters where we make use of the pedigree information to conduct the simulation based likelihood evaluation at promising locations only. The proposed combination of MCMC sampling and MLE is possible because the population size is constant and the maximum likelihood of $\bar{\mu}$ does only depend on the total length of the genealogies and not on their topology. As shown in the results, our MCMC-MLE method is an efficient compromise between speed and accuracy. Further advantages and limitations are discussed in the last section of the article.

The MCMC sampling in step 1 is implemented using a standard Metropolis algorithm (Metropolis *et al.* 1953) in which a single parameter is updated per iteration using a Gaussian proposal kernel mirrored at prior limits. We use uniform priors on all parameters except the population sizes for which we use log-uniform priors and propose updates on the logarithmic scale during the MCMC to account for their prior easily spanning several orders of magnitude.

For all the runs presented here, we run the MCMC for 4.5×10^6 steps thinned out to keep only every 500th parameter combination, of which the first 200 were discarded as a burn-in (resulting in 8800 sampled parameter vectors θ_d). We use a normal distribution for the kernel of all three estimated demographic parameters (N_f,N_m and b). The MCMC parameters relative to each of these demographic parameters can be found in the supplementary Table S1. The implementation of the method in C++ is available upon request.

Simulations

To test the performances of our inference method, we used a custom R script to simulate pseudo observed datasets (PODs) consisting of a pedigree and a corresponding SFS for a sample of 50 individuals, unless specified differently. The pedigree includes all ancestors of the sampled individuals until the predefined depth d as well as the parents of all lineages in gamete storage at generation d (Figure 1). Thus, the generation of the oldest individual contained in the pedigree g_{Max} is such that $g_{\text{Max}} \geq d$. We set $b_f = b_m = b$ for all simulations and generated an SFS by simulating 2000 loci of 10 kb each with $\mu = 5 \times 10^{-9}$ and $\varepsilon = 0$.

We first used simulations to assess the benefit of having pedigree data as a function of the pedigree depth across 10 independently generated PODs with demographic parameters realistic for domesticated breeds ($N_f = 5000$, $N_m = 500$ and $b = 0.2$). As shown in Figure 2, our method is capable of accurately disentangling the effects of the mutation rate and population sizes on genetic diversity already if limited pedigree information is available. Indeed, reliable estimates are obtained for all parameters including sex-specific population sizes, the frequency of overlapping generations as well as the mutation rate if a pedigree of depth four or more is used (Figure 2).

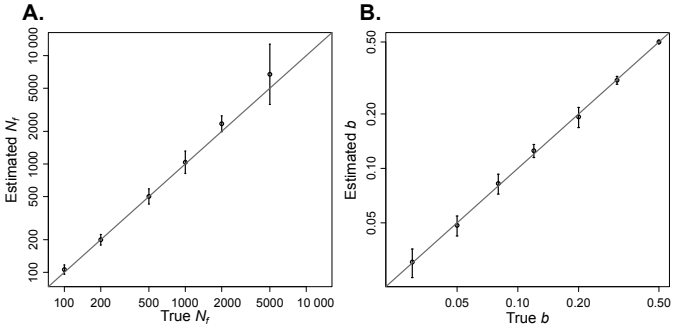

Figure 3 Estimated (A) N_e in function of simulated N_e and (B) b in function of simulated b, in log-log scale. The vertical bares represent the standard deviation over 10 datasets. The gray diagonal marks the identity line.

Interestingly, the rate of overlapping generations b is estimated well across the whole parameter range in the presence of sufficient pedigree data (Figure 3), but smaller population size seem to be consistently estimated more accurately than lager sizes. This is visible as a reduced accuracy in the inference of the female compared to male size in Figure 2, but also occurs if the population sizes of both sexes are equal (Figure 3 and 4). We explain this as follows. The information about the population size of a pedigree is mostly contained in individuals sharing parents (i.e., half or full siblings). If the population is large but the number of individuals in the pedigree relatively small, few to no siblings are observed and the power to estimate the population size decreases. Indeed, when there are no

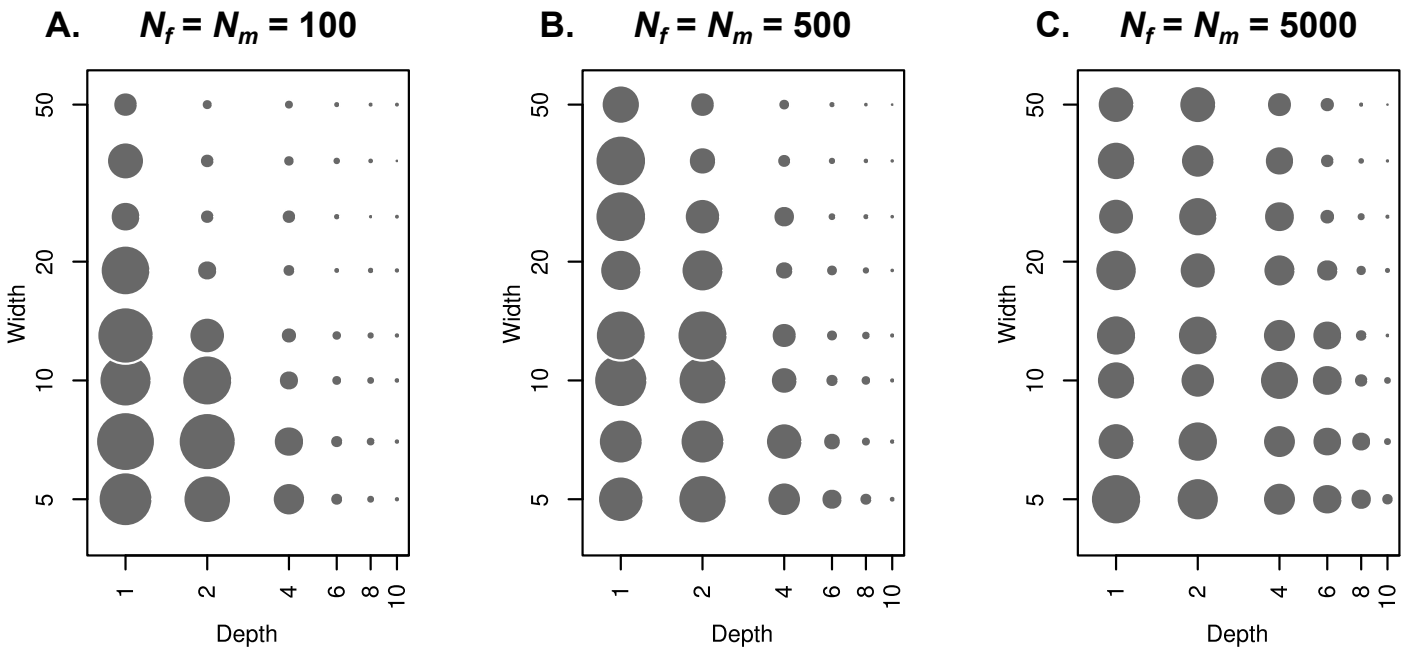

Figure 4 Power to infer population sizes as a function of pedigree width and depth. The surface of each dot represents the root mean squared errors (RMSE) over 10 simulations with population sizes (A) $N_f = N_m = 100$, (B) $N_f = N_m = 500$ and (C) $N_f = N_m = 5,000$. The RMSE is comprised between 1.681×10^{-3} for depth $= 40$ and width $= 36$ in (C) and 3.716 for depth $= 1$ and width $= 7$ in (A).

siblings in the pedigree, the likelihood of the population size increases monotonously but reaches a kind of plateau before the true value is reached (e.q. 4). This leads to an overestimation of the population size in the absence of genetic information and the inability to disentangle N_e from μ if such data is available. As an example, consider the posterior distributions shown in Supplementary Figure S2 for the case of $N_f = 5,000$ and a pedigree depth of one.

We next quantified the effect of pedigree depth and width (number of individuals at $g = 0$) on the accuracy of inferring population sizes. Maybe not surprisingly we found that much more information is contained in small but deep compared to large but shallow pedigrees (Figure 4). Indeed, increasing the width beyond just a handful of individuals seems to hardly reduce estimation accuracy except for very small population sizes, probably due to the oversampling effect described by Wakeley and Takahashi (2003).

The reason for this is that the number of individuals included in a completely resolved pedigree is growing rapidly with each generation going further back into the past (Derrida *et al.* 2000, Supplementary Figure S3), and so are the number of observed parent-offspring relationships informative about population size. Indeed, around 80% of the whole population is included in a complete pedigree of width 50 individuals at only few generations in the past, depending on the population size. At a depth of four, which we found to result in good estimates, about 7.5% or 750 individuals will be part of the complete pedigree of 50 individuals from a population of 10,000 individuals (Supplementary Figure S3).

Discussion

Here we developed a model explicitly accounting for two sexes and overlapping generations. Under this model, genealogies follow a standard coalescent provided that time is rescaled appropriately and that expected coalesce times are much larger than the expected number of generations between parents and offspring, which is generally true under realistic parameter values. This new model allowed us to infer parameters jointly from genetic data and pedigree information. Using simulations we then show that including pedigree information not only improves the estimates of demographic parameters, but also allows to disentangle the effects of demographic and mutational processes on genetic diversity and hence to estimate these processes jointly. Indeed, our simulations show that with pedigree information of 50 individuals tracing back four generations is sufficient to obtain accurate joint estimates of male and the female effective population sizes, the proportion of overlapping generations and the mutation rate. Importantly, obtaining this amount of pedigree information is realistic for many populations of interest. For example, such pedigrees are available for several human populations (e.g., Hussin *et al.* 2015), for many domesticated animals breed of cattle and horses (e.g., Cunningham *et al.* 2001; Mc Parland *et al.* 2007) and for some wild animals (e.g., Clutton-Brock *et al.* 1982; Ellegren 1999).

However, we note that the amount of pedigree information required for accurate inference does depend on the population size with more data being required for larger populations. This stems from the fact that most of the information about the population size contained in a pedigree depends on the number of individuals sharing common ancestors. As a random sample is expected to contain less such individuals in a large population than in a smaller one, it will contain less information. Since the number of common ancestors increases more rapidly with the depth than the width of a pedigree, deep pedigrees of a few individuals contain much more information than shallow pedigrees of many individuals. As we discussed, the number of distinct ancestors in previous generations rapidly decreases

with depth and reaches about 80% of the population within only few generations (Derrida *et al.* 2000, , Supplementary Figure S3). However, these results consider a complete pedigree and are expected to be mitigated in presence of missing information.

Having only little pedigree data available will make it difficult to disentangle the effect of mutation and drift. A particular characteristic of such a situation is that the posterior distribution of the population sizes given the pedigree data alone will be very flat and often extend to very large population sizes. In such cases, the samples generated with our MCMC will likely not be distributed densely enough around the joint MLE to warrant accurate inference. In the extreme case of no pedigree information, the joint likelihood surface of the mutation rate and population sizes will form a ridge and the estimate produced by our stochastic inference method will single out a random combination not necessarily reflective of the true parameters. However, as we have show, already limited pedigree information of a few individuals over a few generations is sufficient to result in accurate inference.

While our theoretical and simulation results are very promising, we note that its application to real data may present some challenges. Firstly, the concept of generation, while convenient, is an artificial construct to discretize time that has little biological meaning for many long lived species. As a consequence, attributing the individuals of a pedigree to specific generations can be difficult. However, it is possible to extend our inference framework to also integrate over the attribution of individuals to generations as

$$L(\mathcal{M}) = \int \sum_{G} [\mathbb{P}(\mathrm{SFS}|G,\mathcal{M})\mathbb{P}(G|\mathcal{P},\mathcal{M})]\mathbb{P}(\mathcal{P}|\mathcal{M})\mathbb{P}(\mathcal{P}^*|\mathcal{P})d\mathcal{P},$$

where we denote by $\mathcal{P}^*$ the pedigree data without generation information (hence only relationships). Here, $\Pr(\mathcal{P}^*|\mathcal{P}) = 1$ if the pedigree $\mathcal{P}$ is compatible with $\mathcal{P}^*$, that is, if all parents are from an older generation than all of their offspring and the most recently born individual is from generation 0, and $\Pr(\mathcal{P}^*|\mathcal{P}) = 0$ otherwise. Unfortunately, none of the parameters' MLE is trivial to derive because finding the maximum of this likelihood function implies finding the optimal set of pedigrees $\mathcal{P}$. However, an MCMC method sampling such pedigrees can be envisioned to infer parameters under such an extended model.

Secondly, demographic events such as population size changes, migration between populations or complex mating systems, e.g., monogamy or harem models, may be needed to describe real populations. The introduction of such demographic events in the discrete generation pedigree model is fairly easy. For example, population size changes can be directly implemented in eq. 3 by using generation specific values of N_f and N_m. The way demographic events shape coalescent processes is well described for many cases and they apply to our model if appropriately scaled. Complex mating systems or reproductive skew are well described for generation by generation models (Gasbarra *et al.* 2005) and some non-standard coalescent models are known to arise in these cases (Eldon and Wakeley 2006). In less extreme cases, specific mating systems can be well approximated by strongly skewed sex ratio (Nunney 1993) which our model already incorporates in its current form. It is important to note that inference under complex demographic scenarios is unlikely to work well with the MCMC-MLE approach introduced here as the pedigree may not have information for all demographic parameters, resulting in a bad proposal for the grid search. However, it is straight-forward to embed our model in an MCMC framework sampling from the joint posterior distribution.

In conclusion, we presented here a new model and some theoretical results on how to combine pedigree and genetic information for the inference of demographic and mutational process and showed that these processes can be disentangled if sufficient pedigree information is available. This is widely unexplored territory as most methods use individual or genealogy based models. But the availability of both pedigree and genetic data for many species, in particular domesticated animals, motivates the development of methods that combines such data. While an application to real data may pose additional challenges, our work is a first step towards such a method and extensions of our approach to more complex demographies and other features of real populations are readily possible. If done properly, the application of these to real data has the potential to give us deep insight into the mutational process in natural populations.

Acknowledgements

We thank Dr. Christoph Leuenberger for helpful discussions on the early version of the model. FP and AT acknowledge support from the German Federal Ministry of Education and Research (BMBF) within the AgroClustEr 'Synbreed–Synergistic plant and animal breeding' (grant no. 0315528I). The work of S. M. Szilágyi was supported by the János Bolyai Fellowship Program of the Hungarian Academy of Sciences.

Literature Cited

Beaumont, M. A., W. Zhang, and D. J. Balding, 2002 Approximate Bayesian Computation in Population Genetics. Genetics **162**: 2025–2035.

Blath, J., A. G. Casanova, N. Kurt, and D. Spanò, 2013 The ancestral process of long-range seed bank models. Journal of Applied Probability **50**: 741–759.

Clutton-Brock, T. H., F. E. Guinness, and S. D. Albon, 1982 *Red Deer: Behavior and Ecology of Two Sexes*. University of Chicago Press.

Crow, J. F., 2000 The origins, patterns and implications of human spontaneous mutation. Nature Reviews Genetics **1**: 40–47.

Cunningham, E. P., J. J. Dooley, R. K. Splan, and D. G. Bradley, 2001 Microsatellite diversity, pedigree relatedness and the contributions of founder lineages to thoroughbred horses. Animal Genetics **32**: 360–364.

Derrida, B., S. C. Manrubia, and D. H. Zanette, 2000 On the Genealogy of a Population of Biparental Individuals. Journal of Theoretical Biology **203**: 303–315.

Eldon, B. and J. Wakeley, 2006 Coalescent Processes When the Distribution of Offspring Number Among Individuals Is Highly Skewed. Genetics **172**: 2621–2633.

Ellegren, H., 1999 Inbreeding and Relatedness in Scandinavian Grey Wolves Canis Lupus. Hereditas **130**: 239–244.

Engen, S., T. H. Ringsby, B.-E. Sæther, R. Lande, H. Jensen, M. Lillegård, and H. Ellegren, 2007 Effective Size of Fluctuating Populations with Two Sexes and Overlapping Generations. Evolution **61**: 1873–1885.

Excoffier, L., I. Dupanloup, E. Huerta-Sánchez, V. C. Sousa, and M. Foll, 2013 Robust Demographic Inference from Genomic and SNP Data. PLoS genetics **9**: e1003905.

Falconer, D. S. and T. F. C. Mackay, 1996 *Introduction to Quantitative Genetics*. Longman, Google-Books-ID: 7ASZNAEACAAJ.

Fu, Y.-X., 1998 Probability of a Segregating Pattern in a Sample of DNA Sequences. Theoretical Population Biology **54**: 1–10.

Gasbarra, D., M. J. Sillanpää, and E. Arjas, 2005 Backward simulation of ancestors of sampled individuals. Theoretical Population Biology **67**: 75–83.

Gutiérrez, J. P., I. Cervantes, A. Molina, M. Valera, and F. Goyache, 2008 Individual increase in inbreeding allows estimating effective sizes from pedigrees. Genetics, selection, evolution: GSE **40**: 359–378.

Hey, J., 2011 Isolation with Migration Models for More Than Two Populations. Molecular Biology and Evolution **27**: 905–920.

Hey, J. and R. Nielsen, 2007 Integration within the Felsenstein equation for improved Markov chain Monte Carlo methods in population genetics. Proceedings of the National Academy of Sciences **104**: 2785–2790.

Hill, W. G., 1974 Prediction and Evaluation of Response to Selection with Overlapping Generations. Animal Production **18**: 117–139.

Hussin, J. G., A. Hodgkinson, Y. Idaghdour, J.-C. Grenier, J.-P. Goulet, E. Gbeha, E. Hip-Ki, and P. Awadalla, 2015 Recombination affects accumulation of damaging and disease-associated mutations in human populations. Nature Genetics **47**: 400–404.

Kaj, I., S. M. Krone, and M. Lascoux, 2001 Coalescent Theory for Seed Bank Models. Journal of Applied Probability **38**: 285–300.

Kingman, J., 1982 The coalescent. Stochastic Processes and their Applications **13**: 235–248.

Mc Parland, S., J. F. Kearney, M. Rath, and D. P. Berry, 2007 Inbreeding trends and pedigree analysis of Irish dairy and beef cattle populations. Journal of Animal Science **85**: 322–331.

Metropolis, N., A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller, 1953 Equation of State Calculations by Fast Computing Machines. The Journal of Chemical Physics **21**: 1087–1092.

Möhle, M., 1998 A convergence theorem for Markov chains arising in population genetics and the coalescent with selfing. Advances in Applied Probability **30**: 493–512.

Nelson, M. M. R., D. Wegmann, M. G. M. Ehm, D. Kessner, P. S. Jean, C. Verzilli, J. Shen, Z. Tang, S.-A. Bacanu, D. Fraser, L. Warren, J. Aponte, M. Zawistowski, X. Liu, H. Zhang, Y. Zhang, J. Li, Y. Li, L. Li, P. Woollard, S. Topp, M. D. Hall, K. Nangle, J. Wang, G. Abecasis, L. R. Cardon, S. Zöllner, J. C. Whittaker, S. L. Chissoe, J. Novembre, and V. Mooser, 2012 An abundance of rare functional variants in 202 drug target genes sequenced in 14002 people. Science **337**: 100–104.

Nielsen, R., 2000 Estimation of population parameters and recombination rates from single nucleotide polymorphisms. Genetics **154**: 931–42.

Nunney, L., 1993 The Influence of Mating System and Overlapping Generations on Effective Population Size. Evolution **47**: 1329–1341.

Schaibley, V. M., M. Zawistowski, D. Wegmann, M. G. Ehm, M. R. Nelson, P. L. St Jean, G. R. Abecasis, J. Novembre, S. Zöllner, and J. Z. Li, 2013 The influence of genomic context on mutation patterns in the human genome inferred from rare variants. Genome research **23**: 1974–84.

Wakeley, J., L. King, B. S. Low, and S. Ramachandran, 2012 Gene Genealogies Within a Fixed Pedigree, and the Robustness of Kingman's Coalescent. Genetics **190**: 1433–1445.

Wakeley, J. and T. Takahashi, 2003 Gene genealogies when the sample size exceeds the effective size of the population. Molecular biology and evolution **20**: 208–213.

Wegmann, D., C. Leuenberger, and L. Excoffier, 2009 Efficient approximate Bayesian computation coupled with Markov chain Monte Carlo without likelihood. Genetics **182**: 1207–18.

Wright, S., 1931 Evolution in Mendelian Populations. Genetics **16**: 97–159.