ANALYSIS OF DNA SEQUENCE VARIATION WITHIN MARINE SPECIES USING BETA-COALESCENTS

M. BIRKNER, J. BLATH, M. STEINRÜCKEN

ABSTRACT. We apply recently developed inference methods based on general coalescent processes to DNA sequence data obtained from various marine species. Several of these species are believed to exhibit shallow gene genealogies, potentially due to extreme reproductive behaviour, e.g. via Hedgecock's "reproduction sweepstakes". Besides the data analysis, in particular the inference of mutation rates and the estimation of the (real) time to the most recent common ancestor, we briefly address the question whether the genealogies might be adequately described by so-called Beta coalescents (as opposed to Kingman coalescents), allowing multiple mergers of genealogies.

The choice of the underlying coalescent model for the genealogy has drastic implications for the estimation of the above quantities, in particular the real-time embedding of the genealogy.

1. INTRODUCTION

Within the last decade, considerable attention has been turned to the explanation of the fact that intra-species DNA sequence variation yields shallow gene genealogies resp. low ratio between effective population size N_e and adult census size N in several marine species (see, e.g., [A04], [H94], [H05], [TWG02], [WT03]), as well as to the theory of mathematical models which may describe such genealogies in terms of so-called Λ-coalescents (see, e.g., [P99], [S99], [DK99], [MS01], [EW06], [BB08]).

Indeed, Hedgecock [H94] proposes a mechanism to describe substantial variation of reproductive success in marine species in terms of so-called "reproduction sweepstakes", to be won by highly successful individuals within each generation, i.e, a few or even a single individual may replace a large fraction of the entire population. This requires great individual fecundity and high mortality early in life as well as variable oceanographic conditions conducive to spawning, fertilization, larval survival and successful recruitment. As a result, Hedgecock claims that the population variance may be orders of magnitudes higher than what standard binomial or Poisson models predict, leading to a small N_e/N ratio.

However, if one tries to turn such a mechanism into a mathematical population model, it turns out that in order to make sense for large populations, sweepstakes whose size are a positive fraction of the present population cannot be too frequent, otherwise this would lead to vanishing genetic variability (as already pointed out by Árnason in [A04]).

To circumvent such trivialities, we first discuss two rigorous mathematical population models in the classical "Cannings' framework" ([C74], [C75]) that incorporate extreme reproductive events (due to Eldon and Wakeley [EW06] and Schweinsberg [S03]), which

can be considered as simple models of Hedgecock's sweepstakes, but still lead to non-vanishing variability. A classsification result due to Möhle and Sagitov ([MS01]) then yields the required timescales for large population size, in a way that sweepstakes neither dominate (leading to vanishing genetic variability) nor become negligible.

It is important to note that the required time-scaling is mostly non-classical (in particular non-linear in the model census population size), hence will affect the mutation rates and make the concept of the so-called (Kingman-) coalescent effective population size discussed in [SKK+05] void, since the existence of the latter depends on a linear change in time-scale.

The resulting limiting ancestral processes embedded in our population models (with extreme reproduction due to sweepstake-like behaviour) coincide with special cases of the so-called Λ-coalescents, i.e. exchangeable coalescents, which allow multiple collisions of lineages (extending the merely binary collisions in the Kingman-coalescent setup), but not simultaneous multiple collisions. Such processes were introduced and studied by Pitman [P99] and Sagitov [S99] and include the classical Kingman-coalescent as a special case. However, the class of Λ-coalescents is vast, in particular allowing for any type of probability distributions on the set of (random) sweepstakes sizes. An important pair of questions in each concrete scenario therefore is: *What is the "right" distribution on sweepstakes sizes, what is the right timescale?* In [EW06], Eldon and Wakeley discuss a simple model, in which sweepstake sizes are always the same fixed positive fraction of the population size. They then fit their model, using a maximum-likelihood method based on the number of segregating sites and total number of mutations, to mitochondrial data from Pacific Oysters (*Crassostrea gigas*), taken from [BBB94], with the result that the maximum-likelihood for the fixed sweepstakes size is 8% of the living population.

To our knowledge, this was the first time that a Λ-coalescent based model has been calibrated to real data. However, there are a few issues that should be discussed. From the modeling perspective, there is no reason why there should be a fixed sweepstakes size. Still, it is certainly infeasible to infer from from the full (non-parametric) class of Λ-coalescents. Parametric subclasses, which describe "realistic" mechanisms, would therefore be of interest (we use the class of "Beta-coalescents", cf. e.g. [BB08], [BBC+05]). Another important point is the adequacy of the equilibrium population assumption underlying the coalescent model used in [EW06]. As pointed out in [BBB94], the pacific oyster data are taken from a population which had only recently been introduced from Japan to Canada. The shallow genealogy might therefore also be explained by the presence of a relatively recent population bottleneck.

In the present article, we analyse several datasets obtained from Atlantic Cod (*Gadus morhua*), taken from geographically separated locations, under the Beta-coalescent model, taking the full information provided by the infinitely-many sites model into account (as opposed to [EW06] where the authors use summary statistics based only on the number of segregating sites and the total number of mutations).

As a result we report that in many cases, a neutral panmictic Kingman-based scenario can be rejected. Further, our maximum-likelihood estimators for the parameters of the Beta-coalescents and mutation rates are presented, as well as a real-time embedding of the genealogy and in particular the expected time to the most recent common ancestor given the data. We finally discuss the question whether there is some evidence for a sweepstake-based scenario in these datasets and propose and calibrate a potential candidate for the distribution of sweepstake sizes.

2. Methods

2.1. **Exceptional genealogies and exchangeable coalescents.** Since the early 80ies, models based on the Kingman coalescent have been successfully used to describe the genealogy of many biological populations. One of their distinguishing features is, together with exchangeability, that only binary collisions are allowed. That is, at most two ancestral lineages may coalesce at a time (exchangeability meaning that all pairs of lineages are treated equal).

However, it turns out that many species seem to exhibit "exceptional genealogies", for example "shallow genealogies", (cf. e.g., [A04], [H94], [H05], [TWG02], [WT03]), which may be appropriately described by more general exchangeable coalescents, the so-called Λ-coalescents, which allow *multiple collisions* of ancestral lineages. Indeed, under a Lambda-coalescent, given a sample of size n, each k-tuple of ancestral lineages (where $2 \leq k \leq n$) is merging to form a single lineage at rate $\lambda_{n,k}$, where

$$\lambda_{n,k} = \int_0^1 x^k (1-x)^{n-k} \frac{1}{x^2} \Lambda(dx), \tag{2.1}$$

for some finite measure Λ on the unit interval $[0,1]$, see [P99] or [S99]. Note that the family of Lambda-coalescents is rather large, and in particular cannot be parametrised by finitely-many real variables. Important examples include $\Lambda = \delta_0$ (Kingman's coalescent) and $\Lambda = \delta_1$ (star-shaped genealogies). Later, we will mainly be concerned with so-called *Beta-coalescents*, where Λ has a Beta$(2 - \alpha, \alpha)$-density for some $\alpha \in (1, 2]$ (where $\alpha = 2$ corresponds to the classical Kingman coalescent). Note that an atom in $(0, 1]$, say $\Lambda = c\delta_y, c > 0$, has a simple interpretation for the corresponding population model, namely, that on a macroscopic time-scale, at rate $c > 0$ a fraction of $y \cdot 100\%$ of the living population is replaced by the offspring of single ancestor. Such macroscopic reproduction events will be called "extreme reproductive behaviour", and leads to multiple collisions in the genealogy. See [BB09] for further details and references.

2.2. **Population models with highly skewed offspring distributions leading to exceptional genealogies.** We follow a classical framework of population genetics and regard neutral, non-overlapping discrete-generation population models with exchangeable offspring distribution, that is, *Cannings models*, see [C74], [C75]. Consider a (haploid) population of fixed size N, let $t \in \mathbb{N}$ denote the t-th generation. Let $\nu^{(t)} := \left(\nu_1^{(t)}, \ldots, v_N^{(t)}\right)$ denote the vector of offspring replacing the N individuals in the t-th generation, where $\nu_k^{(t)}$ is the number of children of individual k. We assume the $\nu^{(t)}$ to be independent, exchangeable random variables. Note that exchangeability and constant population size force that the expected number of offspring of each individual i is one, i.e. $\mathbb{E}[\nu_i] = 1$. Let

$$c_N := \frac{\mathbb{E}[\nu_i(\nu_i - 1)]}{N - 1}. \tag{2.2}$$

Note that the numerator equals the variance of the offspring distribution, i.e. $\sigma_N^2 = \mathbb{E}[\nu_i(\nu_i - 1)]$ and, of course, is independent of i due to exchangeability. A famous result by Kingman [K82] shows that if $\sigma_N^2 \to \sigma^2 \in (0, \infty)$ as $N \to \infty$ (and with suitable moment bounds), then the genealogy of a sample of size n taken from the limiting population (time-changed by $1/c_N$) will be given by Kingman's n-coalescent.

In [MS01], extending Kingman's original result, Möhle and Sagitov provide the necessary and sufficient criteria so that a limiting Cannings population model has a non-trivial genealogy given by a Λ-coalescent. In particular, if we consider the population model on the time scale $1/c_N$, then the limiting genealogy will be governed by a Λ-coalescent iff

- $c_N \to 0$, as $N \to \infty$,
- for all $y \in (0,1)$ with $\Lambda(\{y\}) = 0$, we have

$$\frac{N}{c_N} \mathbb{P}\{\nu_1 > Ny\} \longrightarrow \int_{(y,1]} \frac{1}{x^2} \Lambda(dx), \quad \text{as} \ \ N \to \infty, \tag{2.3}$$

- and finally, for $i \neq j$, $\mathbb{E}[\nu_1(\nu_1 - 1)\,\nu_2(\nu_2 - 1)]/(N^2 c_N) \to 0$ as $N \to \infty$.

Note that the second condition forces the offspring distribution *to be "highly" skewed*, i.e. there must be reproduction events where the number of offspring is of the same order as the total population size, which implies, in particular, that the variance σ_N^2 blows up as N gets large. Note that this condition also implies that these extreme reproductive events need to happen on the time-scale $1/c_N$, hence cannot be too frequent (an effect that has been pointed out, informally, in the introduction and also already in [A04]).

2.3. **Examples for models with extreme reproductive behaviour.**
Several classes of Cannings models have been considered recently in the literature in this context. For example, in [EW06], Eldon and Wakeley discuss a model of size N where, in each generation, exactly one uniformly chosen individual reproduces and becomes parent of a fixed (non-random) positive fraction of the population, leading to a measure with an atom in (0,1].

If one is interested in a model with variable (random) sweepstakes sizes, one might consider Schweinsberg's ([S03]) model, which leads naturally to Beta-coalescent based genealogies and is related to α-stable branching processes for some $\alpha \in (1,2]$. Again, consider a haploid population of size N. Reproduction is assumed to happen in two steps. First, each individual spawns (independently) $\tilde{\nu}_k^{(t)}$ offspring, according to a probability distribution with a power law tail-behaviour, i.e.

$$\mathbb{P}\{\tilde{\nu}_k^{(t)} \geq l\} \sim C l^{-\alpha}, \quad \alpha > 0, C > 0,$$

so that the amount of potential offspring has infinite variance. We denote the resulting vector of (potential) offspring by $\tilde{\nu}^{(t)} = \left(\tilde{\nu}_1^{(t)}, \ldots, \tilde{\nu}_N^{(t)}\right)$. Note that the components do not (yet) sum up to N, but to some $\tilde{N}$ typically much larger than N. Hence, in a second step, N individuals are chosen uniformly at random from the $\tilde{N}$ potential offspring particles, so that we obtain the new offspring vector $\nu^{(t)} = (\nu_1^{(t)}, \ldots, \nu_N^{(t)})$ where $\nu_i^{(t)}$ denotes the number of individuals drawn from the $\tilde{\nu}_i^{(t)}$ potential children of parent k.

This model has some resemblance of so-called "type-III survivorship": high fertility leads to excessive amount of offspring, corresponding to the first reproduction step, whereas high mortality early in life is modeled in the second step. The model is also mathematically appealing, revealing the close connection between α-stable branching processes and Beta$(2 - \alpha, \alpha)$-coalescents, which appear as limiting ancestral processes in this setup, see e.g. [BBC+05].

2.4. **Inference methods.**
Choosing a suitable limit of a Cannings population model (e.g. one of the models above) determines an explicit probabilistic mechanism for the ancestral process, i.e. the underlying class of Λ-coalescents. Still, one needs to calibrate the corresponding parameters to the observed DNA sample in questions, thus inferring "evolutionary parameters" like the mutation rate. We pursue a maximum-likelihood approach based on the full sample information (assuming the infinite-sites model).

A general probabilistic mechanism of obtaining DNA samples from a coalescent tree – which is our model for the gene genealogy of the ancestral limit process of our population models under consideration – is described in detail in [BBS10]. Here, we provide a quick

overview. To obtain a sample of size n, first run an n-Λ-coalescent to obtain a *rooted* coalescent tree. On this rooted tree with n leaves (numbered from 1 to n), place mutations along the branches at rate r to obtain a gene genealogy (here, r is the scaled mutation rate). Then, label these mutations randomly: Given there are s mutations in total, attach uniformly at random the labels from $1, \ldots, s$ to these mutations. An observed genetic *type* $\mathbf{x}$ is then given by the sequence of labels of mutations following its path backwards from a leaf to the root. Finally, we enumerate the different types randomly, from $1, \ldots, d$. We then let $[\mathbf{t}, \mathbf{n}] = \big[(\mathbf{x_1}, \ldots, \mathbf{x_d}), (n_1, \ldots, n_d)\big]$ denote the pair consisting of the observed unordered d-tuple of types $\mathbf{t}$ and their respective multiplicities $\mathbf{n}$. Note that $[\mathbf{t}, \mathbf{n}]$ equivalently describes a tree, commonly referred to as a *genetree*. We will denote the distribution of such a coalescent tree, depending on Λ and mutation rate r, by $\mathbb{P}_{\Lambda,r}$.

We may then compute the likelihood of observed data $[\mathbf{t}, \mathbf{n}]$ under the parameter' $\theta = (r, \Lambda)$, i.e. $\mathbb{P}_{\theta}([\mathbf{t}, \mathbf{n}]) := \mathbb{P}_{\Lambda,r}([\mathbf{t}, \mathbf{n}])$ recursively, conditioning on the last event in the coalescent history, see [BBS10, Section 1.3]. Such a recursion may, for small sample sizes with few mutations, be solved numerically. However, for more complex samples, Monte-Carlo methods, for example Importance Sampling, need to be employed. Such methods are being discussed in detail in [BBS10] and implemented in the program `MetaGeneTree`.[1]

3. Analysis of DNA sequence data

3.1. **Description of underlying datasets.** The Pacific Oysters dataset presented in [BBB94] was obtained as the result of a restriction-enzyme digest of mitochondrial DNA taken from 159 Pacific oysters (*Crassostrea gigas*) from British Columbia. This digest can (in an ad hoc fashion) be interpreted as sequence information resulting in 49 segregating sites or positions, where an enzyme either cuts or leaves the DNA-molecule intact, depending on the allele present. This pattern was then manually edited to resolve violations of the infinitely many sites model, resulting in the exclusion of four samples and five sites, see [S09], Chapter 5.6.2 for some details. Using Gusfiled's [G91] algorithm such sequence data can then be transformed into a genetree. This dataset has also been analysed in [EW06], where the authors already pointed out the underlying genealogy might not be adequately modelled by Kingman's coalescent.

The second set of DNA sequence data was discussed in [A04]. There, Árnason combined several datasets, published in other works, from a 250 bp region of the mitochondrial cytochrome b gene of the Atlantic cod (*Gadus morhua*). In [A04], he provided a discussion of the whole combined dataset which, unfortunately turned out to be too large to be treated by our methods. For this reason we analysed the smaller component datasets described in [AP96, APP98, APKS00, CM91, CSHW95, PC93, SA03] separately. As Árnason pointed out these samples stem from various geographic locations throughout the Atlantic. In our analysis, we choose the most abundant type to represent the ancestral type, and we also consider summing out over all possible ancestral types. Again, some of the samples violated the assumptions of the infinitely many sites model. To cope with this, we considered the combined dataset of [A04] and introduced a consistent pattern of parallel mutations to resolve all violations. This procedure lead to a dataset that corresponds to the phylogenetic maximum parsimony network from Figure 2 of [A04, p. 1875]. We then analysed the respective subsamples specified in the different publications. For a more detailed description of the datasets used for the analysis see [S09, Chapter 5.6.3, Case 2)].

---

[1]Version 0.1.1, available from `http://metagenetree.sourceforge.net`

3.2. **Rejection of the "Kingman hypothesis"?** In several datasets, in particular the one discussed in [BBB94], standard tests reject the "Kingman hypothesis", indicating that the genealogies underlying the observed datasets might not be adequately described by a Kingman-coalescent. In particular, we consider Tajima's D and Fu & Li's D in each case. Recall that for a sample of n sequences, Tajima's D (see [T89]) is based on the normalized difference between the mean number of pairwise differences Δ_n and the weighted number of segregating sites S_n. In a neutral, Kingman-coalescent based scenario, D should be approximately 0. Small values of D indicate shallow genealogies, large values indicate long internal branches. Another standard test statistic is Fu & Li's D (see [FL93]). Again, the test statistic is based on a standardized variable which is the difference between the number of mutations on external branches η_e, and the weighted number of internal branches, η_i. Values for approximate "confindence intervals" (CIs) for each D can be found in Table 2 of [FL93].

Ref.	n	d	Tajima's D	Δ_n	S_n	CI	Fu & Li's D	η_i	η_e	CI
[BBB94]	155	40	-2.65 *	0.94	44	[-1.76, 2.10]	-5.8 *	14	30	[-2.38,1.63]

TABLE 1. Statistical tests reject Kingman hypothesis for Pacific Oyster data

Ref.	n	d	Tajima's D	Δ_n	S_n	CI	Fu & Li's D	η_i	η_e	CI
[AP96]	100	11	-0.66	1.44	10	[-1.78, 2.07]	-1.50	6	4	[-2.36, 1.61]
[APP98]	109	11	-0.89	1.23	10	[-1.78, 2.07]	-1.54	6	4	[-2.36, 1.61]
[APKS00]	78	12	-1.45	1.04	11	[-1.79, 2.06]	-1.83	6	5	[-2.38, 1.59]
[CM91]	55	12	-1.87 *	0.84	11	[-1.8, 2.05]	-2.24	5	6	[-2.45, 1.57]
[CSHW95]	236	16	-2.11 *	0.39	15	[-1.75, 2.11]	-2.24	9	6	[-2.25, 1.65]
[PC93]	103	10	-2.12 *	0.46	12	[-1.78, 2.07]	-3.07 *	5	7	[-2.36, 1.61]
[SA03]	74	15	-1.28	1.59	14	[-1.79, 2.06]	-2.88 *	6	8	[-2.38, 1.59]

TABLE 2. Rejection of Kingman hypothesis for some Atlantic Cod datasets

Table 1 and Table 2 show the observed values for each D and the corresponding 95% confidence intervals for the Pacific Oyster and Atlantic Cod datasets, respectively. Since both are always negative for all datasets, there is a consistent, sometimes rather weak, sometimes significant (marked by an asterisque) indication of a "shallow" genealogy.

3.3. **Likelihood analysis.** Figures 2 and 3(a) in the appendix contain the likelihood surfaces for our pair of parameters (r, α), that is, the coalescent time mutation rate and the parameter of the underlying Beta-coalescent. Both the rooted and unrooted tree cases are presented. The results were obtained with the tool `MetaGeneTree` using the methods introduced in [BBS10] and briefly recapitulated in Section 2.4. The surfaces were calculated on a discrete grid and the maximum of the surface reported. Table 3 shows the maximum likelihood for the Pacific Oyster dataset. The surface was obtained on a discrete grid with spacing (0.2,0.05) and using importance sampling with a sufficient number of independent runs to get an estimated relative error around 0.02.

The column called "rooted" in Table 4 shows the maximum likelihood estimates for the Atlantic cod datasets. The grid-spacing was (0.1,0.05) and all datasets except [CSHW95] could be analysed using the exact recursive formula. For the latter dataset we employed importance sampling, where we used sufficiently many independent runs to get an estimated relative error of approximately 0.01.

6

In the column titled "unrooted" we present for some datasets maxima of likelihood surfaces obtained by summing the likelihoods of all different samples obtained by choosing a different type to be the ancestral one (thus the likelihood for an unrooted genetree), following the methods introduced in [GT95, Section 2.1].

Ref.	Location	$(\hat{r}, \hat{\alpha})$
[BBB94]	British Columbia	(1.2, 1.25)

TABLE 3. Estimate for r and α for the [BBB94] Pacific Oyster dataset, with the most abundant type considered ancestral.

The maximum likelihood estimates for the tree-shape parameter α for both datasets range from 1.25 to 1.65. Recalling that $\alpha = 2$ corresponds to Kingman's coalescent, these results indicate consistently that the data is better explained by a genealogy allowing for multiple merger, rather then a Kingman-based genealogy. We will briefly discuss possible explanations for this evidence of shallow genealogies in the next section. Further, the maximum likelihood estimates for the unrooted datasets do not differ substantially from the estimates in the rooted case, suggesting that the choice of the ancestral type does not have a too much impact on the estimate.

4. DISCUSSION

4.1. **Possible biological causes for shallow genealogies.** The presence of "Hedge-cock's reproduction sweepstakes" is only one possible cause for violations of the Kingman framework produced by shallow genealogies. As discussed by Árnason in [A04, pp. 1882], several effects might, a priori, also be relevant. Among them are *selection*, either acting directly on the observed part of the genome or in the "background", the presence of *frequenct selective sweeps* (indeed, thinking e.g. of [DS04, DS05], which can in principle explain multiple mergers in genealogies) and geographical subdivision (resulting e.g. from glaciation events). Finally, a potential for very high fitness variances and sweepstakes reproduction exists in cod and may account for the pattern of variability observed in the data. It is this last point which is identified by Árnason as the most plausible cause.

As discussed above, population models with Beta-coalescent based genealogies are compatible with sweepstakes-size reproduction. If one believes in such a mechanism, this has

		$(\hat{r}, \hat{\alpha})$	
Ref.	Location	rooted	unrooted
[AP96]	Norway	(0.7, 1.65)	(0.7, 1.65)
[APP98]	Baltic/ trans. area	(0.6, 1.55)	(0.5, 1.45)
[APKS00]	Greenland subsample	(0.7, 1.5)	(0.7, 1.5)
[CM91]	Norway/ Newfoundland	(0.8, 1.4)	(0.7, 1.35)
[CSHW95]	Newfoundland	(0.6, 1.5)	–
[PC93]	Newfoundland	(0.6, 1.4)	(0.6, 1.4)
[SA03]	Faroe Islands	(0.7, 1.3)	–

TABLE 4. Maximum likelihood estimates for the Atlantic cod datasets, for the "rooted" genetrees (most abundant type ancestral) and for the "unrooted" genetrees (summing over all possibilities of choosing an ancestral types). Missing values are due to limited computing resources and will be incorporated in a future version of this preprint.

|	coal. time		real time (in years)	
Ref.	est. mean	CI	est. mean	CI
[AP96]	1.59	[0.70, 3.07]	115535	[50917, 222756]
[APP98]	1.82	[0.82, 3.47]	113127	[51012, 215491]
[APKS00]	1.60	[0.68, 3.11]	115994	[49019, 225654]
[CM91]	1.38	[0.54, 2.76]	114785	[45166, 228959]
[CSHW95]	1.52	[0.55, 3.11]	94683	[34481, 193115]
[PC93]	1.86	[0.75, 3.66]	115510	[46400, 227582]
[SA03]	1.72	[0.77, 3.25]	124638	[55560, 235854]

TABLE 5. Estimates for T_{MRCA} given the different datasets assuming the Beta-coalescent as the true underlying model. The respective estimated mean is given together with the corresponding credibility interval (CI), both in coalescent time units and embedded in real time.

significant implications for the estimation of parameters and the real-time embedding of various quantities, see Section 4.2.

[A04, p. 1883] writes "Studies of temporal variation are called for to test it and better resolve the differences between historical and contemporary factors influencing variance in offspring number and effective population sizes." We heartily agree.

4.2. **Real time mutation rates.** Assuming that the Beta-coalescent and the estimated parameter values present a reasonable approximation to the real population mechanims under consideration, we calibrate our models on the Atlantic cod data, using the method described in [GT94, Section 6] adapted to Λ-coalescents. The maximum likelihood estimates from Table 4 can be used to estimate the time to the most recent common ancestor (T_{MRCA}) in coalescent-time units conditioned on the observed data. For comparison, we also estimated the time to the most recent common ancestor assuming that the Kingman coalescent is the appropriate model for the genealogy and using the corresponding estimate for the mutation rate at $\alpha = 2$. In both cases, we estimated the value of the cumulative distribution function on a discrete grid with spacing 0.1 ranging from 0.3 to 4.0 using 10^8 independent runs of the Markov chain. The two columns in the middle of Table 5 and Table 6 show the approximation of the expected value based on the empirical distribution function, as well as the corresponding 95%-credibility interval assuming the Beta-coalescent respectively Kingman prior for the genealogy. For this we interpolated the distribution function using the function `Interpolation` in Mathematica 6.0 [W07] with an `InterpolationOrder` of 3, and reported the respective 0.025- and 0.975-quantiles.

To embed these values into real time we use Árnason's estimate ([A04, p.1873]) for the mutation/substitution rate of the mitochondrial DNA for the Atlantic cod $\hat{\mu} = 3.86 \times 10^{-8}$/site/year. Since we consider a stretch of 250 bp in the analysis of the Atlantic cod data, the substitution rate for this stretch is given by $\hat{\mu} = 9.65 \times 10^{-6}$/year. For the real time embedding of the coalescent time note that $\hat{\mu} \approx$ #mut./year and $\hat{r} \approx$ #mut./coal.-time-unit. Thus the coalescent time can be transformed into real time by the relation

$$\text{coal.-time-unit} \approx \frac{\text{year}}{\hat{r}}\hat{\mu}.$$

The last two columns of Table 5 and Table 6 show the real time embeddings of the time to the most recent common ancestor and the corresponding credibility intervals for the different samples, in the Beta respectively Kingman case. Some cumulative distribution

	coal. time		real time (in years)	
Ref.	est. mean	CI	est. mean	CI
[AP96]	1.19	[0.58, 2.22]	147960	[72640, 276315]
[APP98]	1.29	[0.63, 2.41]	147124	[71545, 274445]
[APKS00]	1.05	[0.51, 1.96]	151671	[74233, 284256]
[CM91]	0.90	[0.45, 1.67]	158024	[78402, 293404]
[CSHW95]	–	–	–	–
[PC93]	1.12	[0.53, 2.12]	162794	[77366, 307582]
[SA03]	0.95	[0.49, 1.69]	186658	[97403, 333593]

TABLE 6. Estimates for T_{MRCA} given the different datasets assuming Kingman's coalescent. The respective estimated mean is given together with the corresponding credibility interval (CI), both in coalescent time units and embedded in real time. Missing values are due to limited computing resources and will be incorporated in a future version of this preprint.
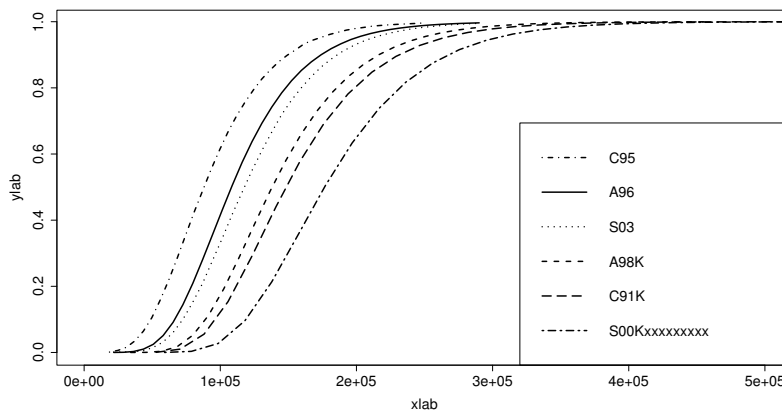
FIGURE 1. The largest, smallest and an average cumulative distribution function in the Beta- and the Kingman-case, where ($\alpha = 2$) indicates the latter.

functions are shown in Figure 1. We are not aware of any other estimates of T_{MRCA} for the presented Atlantic cod datasets in the literature, but it would be interesting to compare our results with results obtained by other methods.

More importantly, our results show that the choice of the prior distribution for the genealogy has a severe impact on the estimation of T_{MRCA}. The estimates under the Kingman coalescent are approximately 50% higher on average. This shows that, when estimating evolutionary parameters, it is crucial to verify the validity of Kingman's coalescent as an appropriate model for the underlying neutral genealogy.

4.3. **Further issues.** It is a natural question to ask which Λ-coalescents are of biological relevance. Indeed, if one believes in the presence of Hedgecocks' reproduction sweepstakes, one needs to infer a distribution of the corresponding reproduction sizes. This distribution typically depends in non-trivial ways on the offspring distribution of the species in question, but also on many other factors which cannot be derived from simple assumptions from the outset (e.g. the geography of the habitat). A starting point can be the model of Schweinsberg discussed above.

However, in other scenarios, other Lambda-coalescents might become relevant. Examples include Durrett and Schweinsberg's models of recurrent selective sweeps mentioned above. The presence of recurrent severe bottlenecks might even yield Ξ-coalescent based genealogies, admitting *simultaneous* multiple merger coalescents [BBM+09].

Finally, it is still a largely open question to assess the statistical properties of the estimator employed here, which is based on methods described in [BBS10]. Some considerations in this direction can be found in [S09].

References

[A04]   Árnason, E.: Mitochondrial Cytochrome b DNA Variation in the High-Fecundity Atlantic Cod: Trans-Atlantic Clines and Shallow Gene Genealogy, *Genetics* **166**, 1871–1885 (2004).

[AP96]   Árnason, E.; Palsson, S.: Mitochondrial cytochrome *b* DNA sequence variation of Atlantic cod, Gadus morhua, from Norway, *Mol. Ecol.* **5**, 715–724 (1996).

[APP98]   Árnason, E.; Palsson, S.; Petersen, P. H.: Mitochondria1 cytochrome *b* DNA sequence variation of Atlantic cod, Gadus murhua, from the Baltic and the White Seas, *Hereditas* **129**, 37–43, (1998).

[APKS00]   Árnason, E.; Petersen, P. H.; Kristinsson, K.; Sigurgíslason, H.: Mitochondrial cytochrome *b* DNA sequence variation of Atlantic cod from Iceland and Greenland, *J. Fish Biol.* **56**, 409–430, (2000).

[BBC+05]   Birkner, M.; Blath, J.; Capaldo, M.; Etheridge, A.; Möhle, M.; Schweinsberg, J.; Wakolbinger, A.: Alpha-stable branching and Beta-coalescents. *Electron. J. Probab.* **10**, 303–325, (2005).

[BB08]   Birkner, M.; Blath, J.: Computing likelihoods for coalescents with multiple collisions in the infinitely-many-sites model. *Journal of Mathematical Biology* **57**, no.3 , 435–465, (2008).

[BB09]   Birkner, M.; Blath, J.: Measure-valued diffusions, general coalescents and population genetic inference. *Trends in Stochastic Analysis*, LMS **353**, Cambridge University Press, 329–363 (2009).

[BBS10]   Birkner, M.; Blath, J., Steinrücken, M.: Importance Sampling for Lambda Coalescents in the infinitely many sites model, 36 pages, submitted, (2010).

[BBM+09]   Birkner, M.; Blath, J., Möhle, M.; Steinrücken, M.; Tams, J.: A modified lookdown construction for the Xi-Fleming-Viot process with mutation and populations with recurrent bottlenecks, *ALEA Lat. Am. J. Probab. Math. Stat.* **6**, 25–61 (2009).

[BBB94]   Boom, J. D. G.; Boulding, E. G.; Beckenbach, A. T.: Mitochondrial DNA variation in introduced populations of Pacific oyster, Crassostrea gigas, in British Columbia. *Can. J. Fish. Aquat. Sci.* 51:1608–1614, (1994).

[C74]   Cannings, C.: The latent roots of certain Markov chains arising in genetics: a new approach, I. Haploid models. *Adv. Appl. Prob.* **6**, 260–290, (1974).

[C75]   Cannings, C.: The latent roots of certain Markov chains arising in genetics: a new approach, II. Further haploid models. *Adv. Appl. Prob.* **7**, 264–282, (1975).

[CM91]   Carr, S. M.; Marshall, H.D.: Detection of intraspecific DNA sequence variation in the mitochondrial cytochrome *b* gene of Atlantic cod (Gadus morhua) by the polymerase chain reaction, *Can. J. Fish. Aquat. Sci.* **48**, 48–52, (1991).

[CSHW95]   Carr, S. M.; Snellen, A. J.; Howse, K. A.; Wroblewski, J. S.: Mitochondrial DNA sequence variation and genetic stock structure of Atlantic cod (Gadus morhua) from bay and offshore locations on the Newfoundland continental shelf, *Mol. Ecol.* **4**, 79–88, (1995).

[DK99]   DONNELLY, P.; KURTZ, T.: Particle representations for measure-valued population models. Ann. Probab. 27, no. 1, 166–205, (1999)

[DS04]   DURRETT, R.; SCHWEINSBERG, J.: Approximating selective sweeps, *Theor. Popul. Biol.* **66**, 129–138, (2004).

[DS05]   DURRETT, R.; SCHWEINSBERG, J.: A coalescent model for the effect of advantageous mutations on the genealogy of a population. *Stoch. Proc. Appl.* **115**, 1628–1657 (2005)

[EW06]   ELDON, B.; WAKELEY, J.: Coalescent processes when the distribution of offspring number among individuals is highly skewed. *Genetics* **172:** 2621–2633, (2006).

[G91]   GUSFIELD, D.: Efficient algorithms for inferring evolutionary trees. *Networks*, **21** (1), 19–28, (1991).

[GT94]   GRIFFITHS, R.C. AND TAVARÉ, S.: Ancestral Inference in population genetics. *Statistical Science* **9**, 307–319, (1994).

[GT95]   GRIFFITHS, R.C. AND TAVARÉ, S.: Unrooted Genealogical Tree Probabilities in the Infinitely-Many-Sites Model. *Mathematical Biosciences* **127**, 77–98, (1995).

[FL93]   FU, X. Y.; LI, W.-H.: Statistical tests of neurality of mutations. *Genetics* **133**, 693–709, (1993).

[H94]   HEDGECOCK, D.: Does variance in reproductive success limit effective population size of marine organisms?, pp. 123–134 in *Genetics and Evolution of Aquatic Organisms*, edited by A. R. Beaumont, Chapman & Hall, London, (1994)

[H05]   HEDRICK, P. W.: Large Variance in reproductive success and the N_e/N ratio. *Evolution* **59**, 1596–1599, (2005)

[K82]   KINGMAN, J. F. C.: The coalescent. *Stoch. Proc. Appl.* **13**, 235–248, (1982).

[MS01]   MÖHLE, M; SAGITOV, S.: A classification of coalescent processes for haploid exchangeable population models. *Ann. Probab.* **29**, 1547–1562, (2001).

[PC93]   PEPIN, P.; CARR, S. M.: Morphological, meristic, and genetic analysis of stock structure in juvenile Atlantic cod (Gadus morhua) from the Newfoundland shelf. *Can. J. Fish. Aquat. Sci.* **50**, 1924–1933, (1993).

[P99]   PITMAN, J.: Coalescents with multiple collisions. *Ann. Probab.* **27** (4), 1870-1902, (1999).

[P01]   POGSON, G. H.: Nucleotide polymorphism and natural selection at the pantophysin (*Pan* I) locus in the Atlantic cod, *Gadus Morhua* (L.). *Genetics* **157**, 317–330, (2001).

[S99]   SAGITOV, S.: The general coalescent with asynchronous mergers of ancestral lines. *J. Appl. Probab.* **36** (4) 1116–1125, (1999).

[S03]   SCHWEINSBERG, J.: Coalescent processes obtained from supercritical Galton-Watson processes. *Stoch. Proc. Appl.* **106**, 107–139, (2003).

[SA03]   SIGURGÍSLASON, H.; ÁRNASON, E.: Extent of mitochondrial DNA sequence variation in Atlantic cod from the Faroe Islands: a resolution of gene genealogy. *Heredity* **91**, 557–564, (2003).

[SKK+05] SJÖDIN, P. I.; KAY, I.; KRONE, S.; LASCOUX, M., NORDBORG, M.: On the meaning and existence of an effective population size. *Genetics* **105**, 437–460, (2005).

[S09]   STEINRÜCKEN, M.: *Multiple Merger Coalescents and Population Genetic Inference.* Dissertation, Technische Universität Berlin, (2009).

[T89]   TAJIMA, F.: Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* **123**, 585–595, (1989)

[TWG02]   TURNER, T. F.; WARES, P.; GOLD, J. R.: Genetic effective size is three orders of magnitude smaller than adult census size in an abundant, estuarine-dependent marine fish *(Sciaenops ocellatus)*. *Genetics* **162**, 1329–1339, (2002).

[W07]   WOLFRAM RESEARCH, INC.: Mathematica, Version 6.0, Champaign, IL (2007).

[WS09]   WAKELEY, J.; SARGSYAN, O.: Extensions of the coalescent effective population size. *Genetics* **181**, 341-345, (2009).

[WT03]   WAKELEY, J.; TAKAHASHI, T.: Gene genealogies when the sample size exceeds the effective size of the population. *Mol. Biol. Evol.* **20**, 208–213, (2003).
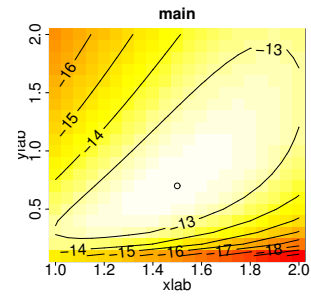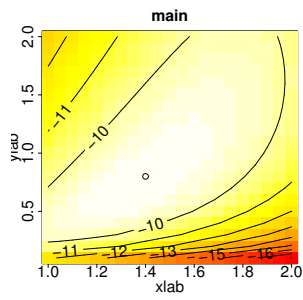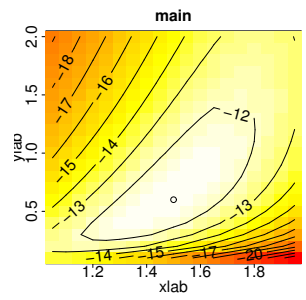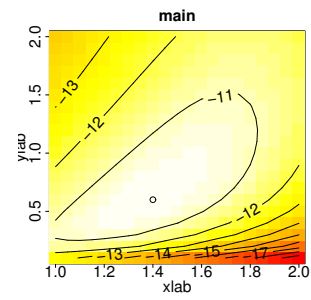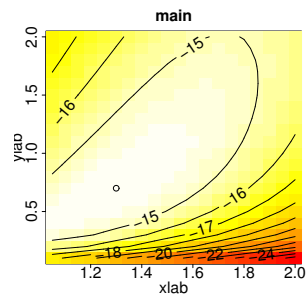
5. APPENDIX
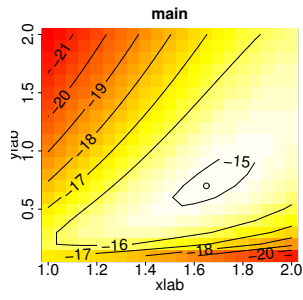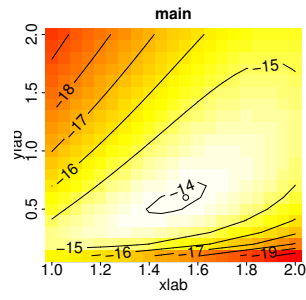
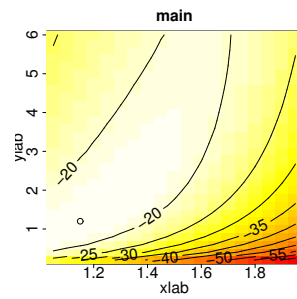(a) [AP96] (b) [APP98] (c) [APKS00]

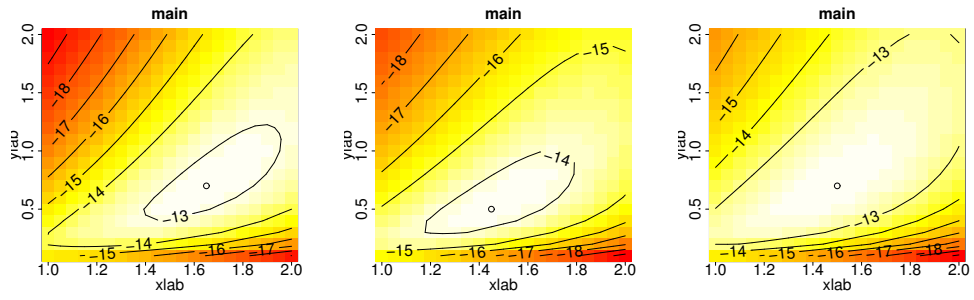(d) [CM91] (e) [CSHW95] (f) [PC93]
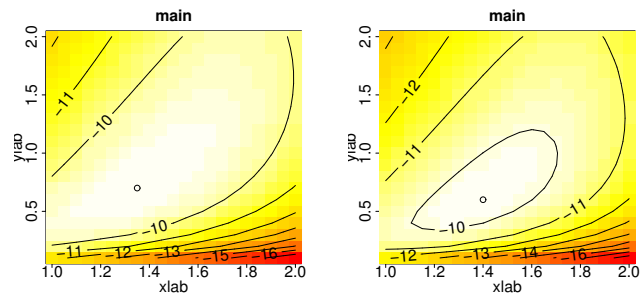
(g) [SA03] (h) ([BB08])

FIGURE 2. Log-Likelihood-surfaces for cod and oyster datasets.

(a) [AP96]   (b) [APP98]   (c) [APKS00]

(d) [CM91]   (e) [PC93]

FIGURE 3. Log-Likelihood-surfaces for some unrooted cod datasets.