MOLECULAR ECOLOGY

Molecular Ecology (2014) **23**, 2637–2652 doi: 10.1111/mec.12755

INVITED REVIEWS AND SYNTHESES

Coalescence 2.0: a multiple branching of recent theoretical developments and their applications

AURÉLIEN TELLIER* and CHRISTOPHE LEMAIRE† ‡ §

*Section of Population Genetics, Center of Life and Food Sciences Weihenstephan, Technische Universität München, 85354 Freising, Germany, †LUNAM, UMR1345 Institut de Recherche en Horticulture et Semences, Université d'Angers, SFR 4207 QUASAV, 49045 Angers, France, ‡INRA, UMR1345 Institut de Recherche en Horticulture et Semences, 49071 Beaucouzé, France, §AgroCampus-Ouest, UMR1345 Institut de Recherche en Horticulture et Semences, 49045 Angers, France

Abstract

Population genetics theory has laid the foundations for genomic analyses including the recent burst in genome scans for selection and statistical inference of past demographic events in many prokaryote, animal and plant species. Identifying SNPs under natural selection and underpinning species adaptation relies on disentangling the respective contribution of random processes (mutation, drift, migration) from that of selection on nucleotide variability. Most theory and statistical tests have been developed using the Kingman coalescent theory based on the Wright-Fisher population model. However, these theoretical models rely on biological and life history assumptions which may be violated in many prokaryote, fungal, animal or plant species. Recent theoretical developments of the so-called multiple merger coalescent models are reviewed here (Λ-coalescent, beta-coalescent, Bolthausen-Sznitman, Ξ-coalescent). We explain how these new models take into account various pervasive ecological and biological characteristics, life history traits or life cycles which were not accounted in previous theories such as (i) the skew in offspring production typical of marine species, (ii) fast adapting microparasites (virus, bacteria and fungi) exhibiting large variation in population sizes during epidemics, (iii) the peculiar life cycles of fungi and bacteria alternating sexual and asexual cycles and (iv) the high rates of extinction-recolonization in spatially structured populations. We finally discuss the relevance of multiple merger models for the detection of SNPs under selection in these species, for population genomics of very large sample size and advocate to potentially examine the conclusion of previous population genetics studies.

Keywords: genetic drift, natural selection, parasite evolution, rapid evolution

Received 23 January 2014; revision received 8 April 2014; accepted 13 April 2014

Introduction

Since the end of the 20th century, and increasingly recently, molecular data are being used to reveal the evolutionary history of populations. Population genetics and genomic approaches provide answers to key evolutionary questions such as understanding which evolutionary forces drive genome evolution, or pinpointing the molecular bases for species or population adaptation to their environment (biotic or abiotic). Population genetics is firmly grounded on the mathematical theory founded by the seminal work of Wright (1931), Fisher (1930), Malecot (1941) and Kimura (1954) amongst others, stating that genomes evolve by the action of random neutral processes (mutation, drift and migration) and natural selection (positive or negative). In this respect, the so-called Kingman coalescent (Kingman 1982) and the Wright–Fisher (Fisher 1930; Wright 1931), and Moran (1958) models (see description below) have been instrumental for connecting mathematical and stochastic theory with polymorphism data. The Kingman model allows us to interpret the observed genetic diversity using the population genealogy (the so-called

Correspondence: Aurélien Tellier, Fax: +498161712589; E-mail: tellier@wzw.tum.de

n-coalescent) of a sample of individuals from present (and past) DNA polymorphisms, opening the possibility of model-based statistical inference (Rosenberg & Nordborg 2002).

An important result of population genetics (Kingman 1982; Wakeley 2008) is that neutral random processes, such as past demographic expansion, can generate similar patterns of nucleotidic variability in the genome as those resulting from natural selection, such as positive selection (Tajima 1989). However, it is also assumed that demographic events affect the whole genome, whereas selection affects potentially only few loci. All genomic studies using statistical methods to detect natural selection, which are based on polymorphism data (Tajima's D, Fay and Wu's H, McDonald-Kreitman), or drawing inference of past demography and/or selection rely extensively on the predictions from theoretical population genetics, sometimes more specifically from the coalescent theory. This theoretical framework and developed statistical inference methods have been extensively used to study demography of species and populations (Nelson *et al.* 2012), recent speciation events (review in Sousa & Hey 2013), existence of seed banks (Tellier *et al.* 2011) and the occurrence of natural selection (Hernandez *et al.* 2011). Population genetic studies have been conducted using the Wright–Fisher and Kingman coalescent framework on a wide range of organisms, ranging from mammals to bacteria, fungi or plants which vary greatly in their generation time, population sizes, overlapping of generations, mutation rates, spatial structures, ecological characteristics and population dynamics. Although the Kingman coalescent can handle many departures to the assumptions of the Wright–Fisher model such as population subdivision and variable population size, it assumes chiefly that offspring numbers are uncorrelated between parents and offspring with a variance independent of population size. The Kingman coalescent allows, moreover, only two lineages to coalesce at a time which may be a strong constraint in the case of species experiencing high variances in reproductive success or periods of asexual dissemination as described below. Even though much theoretical work has been carried out on extending these models contributing to our understanding of the evolutionary forces shaping patterns of polymorphism at the population or species level, it is, yet, not always clear to what extent these violations affect statistical inferences from polymorphism data.

It becomes of interest to investigate recent theoretical developments as a source for improving our current genomics and model-based inference methods. Starting with the Bolthausen–Sznitman coalescent (Bolthausen & Sznitman 1998), followed by the Λ-coalescent (Donnelly & Kurtz 1999; Pitman 1999; Sagitov 1999) and Ξ-coalescent (Schweinsberg 2000; Möhle & Sagitov 2001), there is a recent burst of mathematical theory, which aims to generalize the Kingman model (Wakeley 2013). In a nutshell, these theories allow for the merging (coalescence) of multiple lineages (more than two) at a given generation, and possibly several simultaneous mergers (Fig. 1). They are thus referred thereafter as Multiple Merger Coalescent (MMC) models. A common theme in these studies is the demonstration that the Kingman coalescent is a peculiar case of the general class of MMC models. As such, these models may present a greater range of applicability for genomic analysis of

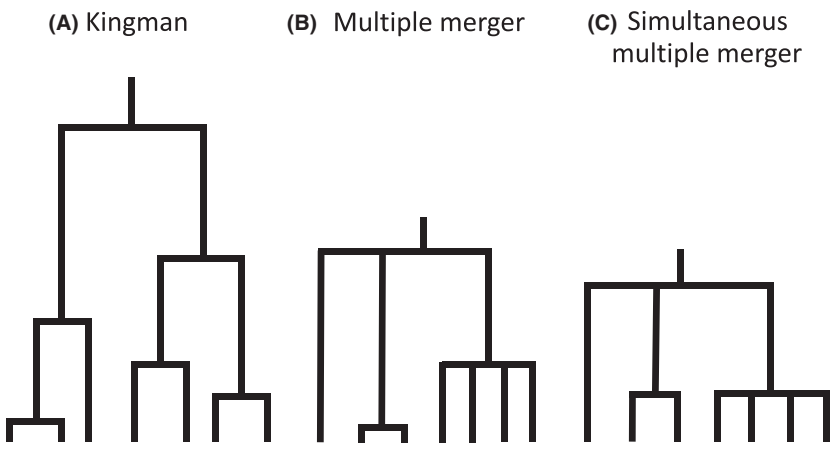

Fig. 1 Schematic genealogies for three types of coalescent models: (A) Genealogy based on the Kingman coalescent with at most only one coalescent event per generation between two lineages. (B) Genealogy with a maximum of one coalescent event per generation involving two or more lineages applicable to the Λ-coalescent, beta-coalescent, Bolthausen-Sznitman and ψ-coalescent. (C) The general genealogy with simultaneous multiple coalescent of two or more lineages per generation as found in the Ξ-coalescent. Note that the length of genealogies may vary considerably between multiple merger coalescent trees and is here not indicative.

species with peculiar life cycles, which violate the assumptions of the Wright–Fisher or Moran models and Kingman coalescent.

After summarizing the main assumptions and limitations of the current coalescent theory based on the Wright–Fisher model, we turn our attention to the MMC models. We have five major aims (i) to provide the first overview of this burgeoning literature on MMC for a nonmathematical audience, (ii) to analyse the assumptions and limitations of these models in comparison to the Kingman coalescent, (iii) to review the current applications of MMC models for data analysis, (iv) to advocate that MMC models are especially suitable for coalescent studies in a wide range of species (plant, animal, fungi, bacteria, viruses) which exhibit peculiar life cycle or life history, and (v) to highlight the need to use MMC models in such species to improve the statistical analysis of genomic data.

The Kingman coalescent: assumptions, extensions and limitations

The Kingman coalescent (Kingman 1982) is obtained as a backward in time continuous limit of the discrete Wright–Fisher (Fisher 1930; Wright 1931) or Moran (1958) models. The Wright–Fisher model assumes no overlapping generations and a panmictic population of $2N$ haploid individuals, or N diploid individuals. All individuals reproduce by producing gametes and then die at each generation t. Panmixia and constant population size are obtained assuming that at the generation $t + 1$, each offspring individual picks one ancestor at random in the parental generation t. Two properties of the Wright–Fisher model are crucial to derive the Kingman coalescent: the large population size N compared with the sampling size n from which polymorphism data are obtained, and the small and finite variance of offspring number per parental individual.

When modelling a population from one generation to the next, parental individuals produce a certain number of offspring. By definition, this number of offspring varies randomly between parents, some contributing to the next generation and others not. The variance in offspring number among parents generates genetic drift, and therefore, the genealogy of a population can be traced back from a sample size of n individuals, to a most recent common ancestor (MRCA; *e.g.* Kingman 1982; Wakeley 2008). Assuming a sample size of n individuals at present, the Kingman coalescent is obtained as a limiting genealogy when the population size N is sufficiently large, especially compared with the sample size n ($N \to \infty$ and $n \ll N$ in mathematical terms). In the Wright–Fisher model, the offspring distribution per parental individual follows a binomial distribution with mean 1, and variance $(1-1/2N)$, which is approximately equivalent to a Poisson distribution with mean 1 and variance 1. A small variance in offspring production is realistic for species with physiologically limited reproductive output such as humans or other mammals. A coalescent event is defined as the merging of k lineages at some point in time (with $k = 2$). The two key assumptions above assure that both the probability that more than two lineages coalesce at once (a so-called multiple merger event) and that of simultaneous events to occur are negligible (on the order of $O(1/N^2)$ when N is large enough). [In mathematics, $O(1/x^2)$ would capture all the terms that decrease to zero at rate $1/x^2$ or faster, as x becomes very large (goes to infinity, $x \to \infty$)]. The property of a coalescent tree, that is, its topology and the size distribution of all branches, is thus defined by the frequency of merging of lineages (the coalescent rate) and how many lineages coalesce, with time being scaled by the population size N. To obtain possible polymorphism data under a given model, mutations can be thrown on the genealogy following a Poisson process with the population mutation rate $\theta = 4N\mu$ where μ is the mutation rate per base pair per generation (description in Wakeley 2008). Note that in the Wright–Fisher model, the effective population size (Ne) is equal to the observable population size (N).

The Kingman coalescent has been instrumental in analysing polymorphism data because it can be easily modified to relax the assumptions of constant population size in time (*e.g.* Watterson 1984; Kaj & Krone 2003), single panmictic population (*e.g.* Wakeley & Aliacar 2001; Charlesworth *et al.* 2003) and non-overlapping generations (Tellier *et al.* 2011). Modifications of the Kingman coalescent are obtained using time-rescaling argument to accommodate variable rates of coalescence in time (Kaj & Krone 2003). For example, an increase in population size is accommodated by shortening coalescent times and increasing the coalescent rates, assuming that all individuals present a similar increase in their per capita offspring production and that the average offspring numbers is still much smaller than N. Analogues of the Kingman coalescent were also built to accommodate sexual reproduction and intralocus recombination (the Ancestral Recombination Graph, Hudson 1983) and natural selection (the Ancestral Selection Graph, Krone & Neuhauser 1997). We redirect interested readers to in depth reviews about the Kingman coalescent (Wakeley 2008) and the use of coalescent simulators (Hoban *et al.* 2012).

However, recent studies in various marine organisms (sardines, cods, salmon, oysters) have suggested that some individuals produce a number of surviving offspring on the order of magnitude of N, that is, the variance in offspring number becomes population size

limited. Indeed, very high fecundities but also high early mortality characterize most of the marine organisms. This 'sweepstake reproduction' effect is independent of natural selection and entirely driven by a large variance of reproductive success (Beckenbach 1994; Hedgecock 1994; Li & Hedgecock 1998; Hedgecock & Pudovkin 2011; Harrang *et al.* 2013). It has been noted as well that under the action of natural selection, during so-called selective sweeps, favoured individuals may produce broad offspring numbers (on the order of N) over a short period of time (Schweinsberg & Durrett 2005; Coop & Ralph 2012). The classical Kingman coalescent may become inadequate to study populations of many animal, plant, bacterial or fungal species, or under strong positive selection. This is the starting point for the development of MMC models.

Overview of existing multiple merger coalescent models

Our objective here is to describe the general framework of multiple merger coalescent models and not their rigorous mathematical description. We redirect the interested reader to the cited literature for rigorous mathematical definitions (see Berestycki 2009; Etheridge 2011). All the models detailed below are summarized in Table 1. MMC models are derived from the general Cannings model of population dynamics (Cannings 1974; Sagitov 1999; Möhle & Sagitov 2001), from which the Moran and Wright–Fisher models are specific cases. We describe here the recent model of Schweinsberg (2003) which has been used to derive the Λ-coalescent, but which idea captures the essence of the population model underlying the MMC models. In a population of size N, each haploid individual independently produces a random number of juvenile offsprings following a given probability distribution, allowing for large number of offsprings. At each generation, density-dependent regulation operates in the population so that exactly N juveniles, sampled at random, will survive to maturity and constitute the next generation. The mean number of juveniles produced by each parental individual is thus assumed to be >1, and there are thus always many more than N juveniles to choose from. MMC models deal thus with the case where individual offspring distributions have a large variance, that is, the number of chosen juveniles from a given individual parent can approach N with a non-negligible probability. Note, however, that after population regulation, the mean number of offspring per parental individual is assumed to be one, because population size is constant in time as in the Wright–Fisher or Moran models.

The most general model of multiple merger coalescent is the so-called Ξ-coalescent in which simultaneous multiple mergers of lineages are possible, or in other words several collisions of k lineages ($k \geq 2$; Fig. 1). This class of model was introduced by Möhle and Sagitov (2001) and described in its full generality by Schweinsberg (2000). We mention this class only briefly in this review because it is less studied for applications to polymorphism data analysis than the following Λ-coalescent models (but see *e.g.* Birkner *et al.* 2009; Taylor & Véber 2009).

A second class of model is the Λ-coalescent, which allows only one merging of multiple k lineages ($k \geq 2$) at any point in time (Fig. 1). This class of models was established independently by Donnelly and Kurtz (1999), Pitman (1999) and Sagitov (1999). It specifies the genealogy of a sample from a Cannings population model based on the so-called Λ-Fleming-Viot process (Bertoin & Gall 2003). Based on the definition of the rate of coalescence (see Appendix I), the Kingman coalescent is a special case of Λ-coalescent where only two lineages are allowed to merge at a time ($k = 2$). The Λ-coalescent can be defined by several distributions of the frequency of multiple mergers, that is, how often multiple coalescent events occur, and the size of merging events, that is, the number of coalescent lineages k. Recently, properties of Λ-coalescent genealogies have been studied such as the expected site frequency spectrum (SFS) and number of segregating sites (Birkner *et al.* 2011, 2013; Berestycki *et al.* 2014).

Table 1 Overview of the different models and their suggested generating biological processes

Coalescent model	Variance in offspring production	Multiple merging events (>2 lineages)	Simultaneous merging events	Proposed biological processes
Kingman	Small	No	No	Wright–Fisher model
Λ-coalescent	Large	Yes	No	Sweepstake, recurrent bottlenecks
ψ-coalescent	Large	Yes	No	Sweepstake
Beta-coalescent	Large	Yes	No	Sweepstake, recurrent bottlenecks
Bolthausen- Sznitman	Large	Yes	No	Rapid and recurrent positive selection
Ξ-coalescent	Large	Yes	Yes	Selective sweeps, recurrent bottlenecks, spatial extinction – recolonization

The beta-coalescent is a subclass of Λ-coalescent models which were obtained by Schweinsberg (2003). It is defined with the rate of multiple coalescent events following a beta-distribution of parameters α and 2-α (see Appendix 1). In mathematical terms, multiple merger events occur on the timescale of $O(1/N^{\alpha})$ (with $\alpha < 2$) as described in Birkner and Blath (2008). The beta-coalescent is obtained as a specific case of Λ-coalescent assuming a beta-distribution, with parameters α and 2-α, of multiple merger probabilities (Schweinsberg 2003). The properties of the beta-coalescent are well studied and various aspects of its genealogies computed such as length of coalescent trees, the expected site frequency spectrum (SFS) and number of segregating sites [Fig. 2, (Berestycki *et al.* 2007, 2008; Birkner & Blath 2008)]. These results are key for the potential application to analysis of polymorphism data. For example, in a Kingman coalescent, the ratio $R_1 = T_1/T$,

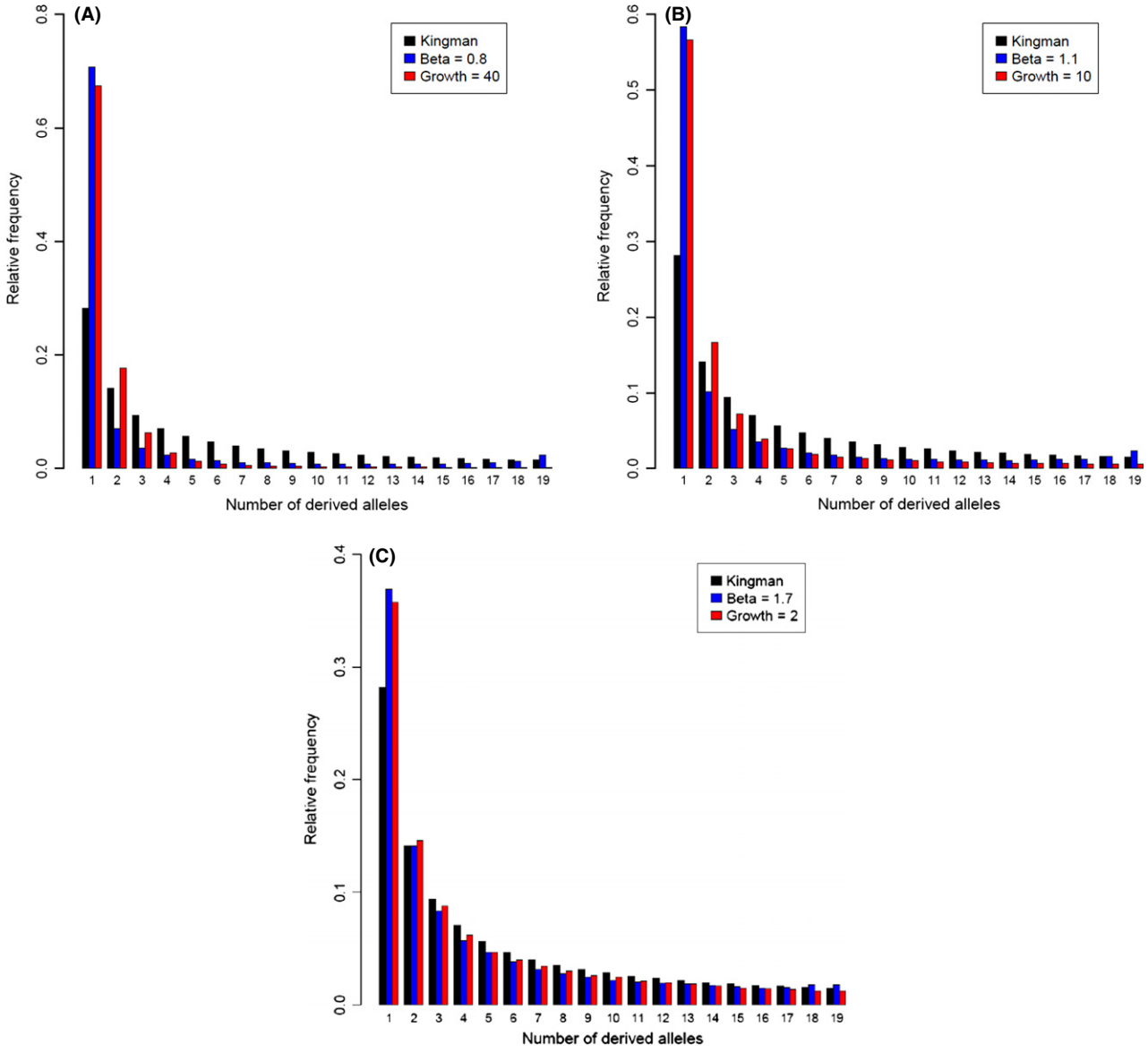

Fig. 2 Site frequency spectrum at one locus with sample size $n = 20$ for different coalescent models. Three models are compared: the neutral model of Kingman with constant population size (black bars), the beta-coalescent (blue bars) and the Kingman coalescent with stepwise expanding population (red bars). (A) SFS for beta-coalescent with parameter $\alpha = 0.8$ (blue) and population expansion with a growth factor of 40 (red). (B) SFS for beta-coalescent with parameter $\alpha = 1.1$ (blue) and population expansion with a growth factor of 10 (red). (C) SFS for beta-coalescent with parameter $\alpha = 1.7$ (blue) and population expansion with a growth factor of 2 (red). The average SFS over 10 000 replicates is shown, assuming a population mutation rate $\theta = 30$ for the Kingman models, and $r = 10$ for the beta-coalescent model. The beta-coalescent simulations are computed using the recursion (R-code) from Birkner and Blath (2008).

where T_1 is the expected external branch lengths of a coalescent tree and T is the expected total branch lengths of a tree, has been show to be equivalent to the expected proportion of singletons $R_1 \approx \xi_1/S$ where ξ_1 is the number of segregating sites at frequency one (singletons) and S is the total number of segregating sites. For the beta-coalescent, the expected proportion of singletons (R_1) tends to the value 2-α for large sample size ($n \to \infty$; Berestycki *et al.* 2007, 2008). In practice, inference of α is possible based on the SFS and the total number of segregating sites of a large enough sample n of individuals (Birkner & Blath 2008; Birkner *et al.* 2011; Steinrücken *et al.* 2013).

A peculiar case of beta-coalescent is the so-called Bolthausen–Sznitman model, although it originates from spin glasses models in physics (Bolthausen & Sznitman 1998). This model is obtained assuming a uniform distribution for the measure of intensity at which mergers of a certain size (number of coalescing lineages) happen (see Appendix I) and defined as a beta-coalescent with $\alpha = 1$. Note that this definition implies that the ratio R_1 tends in this model to one (when $n \to \infty$). Therefore, the ratio R_1 would not be a useful quantity to calculate from sequence data in this case, as it does not correlate with the model parameter. Expected statistics on the genealogies have also been derived (Basdevant & Goldschmidt 2008). This model was found to reflect the genealogy of models with population under rapid positive selection (Brunet *et al.* 2007; Brunet & Derrida 2012; Neher & Hallatschek 2013; Neher *et al.* 2013).

A final type of model, the ψ-coalescent, was derived by Eldon and Wakeley (2006, 2008, 2009) and more recently (Eldon & Degnan 2012). This model has for starting point the biological assumptions of sweepstake reproduction. The parameter ψ defines the proportion of offspring in the population, which originate from one parent at the previous generation (see Appendix I). This model is a peculiar case of a Λ-coalescent (Der *et al.* 2012), as its behaviour allows only one at a time multiple merger of k lineages ($k \geq 2$; Fig. 1). It has been used to infer the strength and occurrence of sweepstake events in semelparous fishes and marine organisms and is thus influential for applications of MMC models (Eldon 2009, 2011). Note nevertheless that the ratio R_1 also tends to one (when $n \to \infty$) in this ψ-coalescent model (Eldon 2009, 2011).

Patterns of polymorphism under multiple merger coalescent

In the following, we describe and highlight the main differing features of MMC models in terms of genetic diversity, patterns of polymorphism, linkage disequilibrium and population differentiation, compared with expectations under the classic Kingman coalescent. Furthermore, we discuss for various effects of MMC on genealogies, how the given polymorphism pattern can lead to bias and erroneous conclusions if such pattern is interpreted under the classic Wright–Fisher assumptions.

For the multiple merger coalescent genealogies to exhibit differences from those under the Kingman model, MMC events must be frequent enough. Two extreme situations are predicted. On the one hand, when the rate of multiple merging coalescent events ($k \geq 2$) is much smaller than the rate of binary coalescence ($k = 2$), the coalescent process approaches the Kingman model, and no signature of MMC would be observable. On the other hand, if the rate of multiple merging is much higher than the rate of binary coalescence, a decrease in the amount of genetic diversity may be observed (Eldon & Wakeley 2006, 2008). This is easily understandable from the biological point of view (see Arnason 2004). If the sweepstake events are very seldom and of small size, that is, most individuals in the population reproduce and leave one viable juvenile at most generation, the population's genealogy can be reconstructed by a classic Wright–Fisher model. Conversely, if at almost every generation, one or very few individuals produce all surviving juvenile offsprings, the population will be composed of very few genotypes. These genotypes are very similar from each other because they share a recent common ancestor, and there is very little time for new mutations to appear. A consequence of the sweepstake reproduction mode is then that the effective population size (*Ne*) can be significantly smaller than the total observable population size (*N*; Nunney 1995). Mathematically, this is described as follows: in a Kingman coalescent, *Ne* scales linearly with *N*, whereas for MMC models, *Ne* can be a fraction of *N* or scales on the order of log(*N*) (Huillet & Möhle 2011). In addition, as the scaling of time differs in the Kingman coalescent from that of MMC models, caution is required to interpret biologically the link between the sum of branch lengths of a genealogy and the population size. In other words, it is expected that under MMC models, the effective population size measured from polymorphism data yields significantly smaller values than the number of observable individuals *N* in the population or the species (see below for example). Namely genetic diversity scales only weakly with *N*. In contrast, the observation of small *Ne* compared with observable *N* is yet often currently interpreted under the Wright–Fisher model as a signal of population extinction or strong bias in sex ratio. We suggest here that these conclusions may be incorrect when studying species undergoing regular sweepstake reproduction with constant population size and balanced sex ratio (see Hedgecock 1994; Arnason 2004).

Effect on genetic diversity and the site frequency spectrum

Coalescent trees obtained under MMC models exhibit more star-like-shaped genealogies and skews in the resulting SFS with an excess of low (*e.g.* singletons) and high frequency variants (Fig. 2 blue bars) generating a more negative Tajima's D (Birkner *et al.* 2013) than under the Kingman model with constant population size (Fig. 2 black bars). The star-like phylogeny and skew in SFS will be more pronounced with increasing rates of multiple merging, for example, for smaller values of α (with $\alpha < 2$) in the beta-coalescent model or higher values of ψ (the proportion of offspring derived from one individual in the population) in the ψ–coalescent (Fig. 2). MMC models with constant population size generate in fact an excess in length of external branches in the genealogies, which can be compared to that arising from Kingman coalescent model with strong recent population expansion (Fig. 2 red bars, Birkner *et al.* 2013; Steinrücken *et al.* 2013). Moreover, note that the Kingman coalescent under arbitrary population size fluctuations does not generate an excess of high frequency variants, namely a u-shape SFS (Sargsyan & Wakeley 2008), contrary to MMC models. This excess of high-frequency SNPs is an important observable feature which can be used to differentiate Kingman with varying population size from MMC models. The usual interpretation that lack of genetic diversity and excess of low frequency SNPs (and negative Tajima's D) is a signal for past bottleneck or at least very strong population expansion (Fig. 2), may be often incorrect, when studying species undergoing sweepstake reproduction.

Effect on genetic drift and linkage disequilibrium

The structure of genealogies and resulting SFS (Der *et al.* 2011, 2012) deriving from genetic drift and the amount of linkage disequilibrium (Eldon & Wakeley 2008) are also affected in MMC models compared with the Kingman expectations. Assuming that the rate at which MMC events occur is high enough, that is, on the order of the coalescent rate, the strength of genetic drift decreases in MMC models compared with the Wright–Fisher model (Der *et al.* 2011, 2012). In a Wright–Fisher model with a biallelic system (two alleles *a* and *A*), at every generation each offspring chooses randomly a parent with a probability equal to that of each allele frequency. The frequencies of *a* and *A* are then given by binomial sampling at every generation. Such a case represents an upper limit to the strength and occurrence of drift (Der *et al.* 2011). In a model where both alleles *a* and *A* are neutral and their frequencies are only driven by drift, MMC models generate long period of frequency stasis (unchanged values) only interrupted by multiple merger events. Nonetheless, the fixation probability of a new allele is unchanged in MMC models compared with a Wright–Fisher model and is equal to its initial frequency (namely $1/2N$, because of the exchangeability assumption of the model; Der *et al.* 2011), whereas the time to fixation of a mutant allele entering the population at frequency $1/2N$ decreases from $2N$ in the Wright–Fisher model to $N \times \log(N)$ in MMC models.

Multiple merger coalescent models generate more pronounced star-like genealogies (see above), and consequently events affecting the genealogy, such as intralocus recombination, can affect a smaller or larger number of branches compared to that expected under the Kingman coalescent (Eldon & Wakeley 2008; Birkner *et al.* 2012). In biological terms, this means that if sweepstake reproduction events are frequent, the efficiency of recombination and meiosis in reshuffling genotypes is very limited as present genotypes descent from a very recent ancestor (Eldon & Wakeley 2008). As a result under MMC models, the amount of linkage disequilibrium (LD) can be uncorrelated to the genomic recombination rate (the Ancestral Recombination Graph for MMC model; Birkner *et al.* 2012). For example, high LD can be observed despite high genomic rates of recombination and *vice* and *versa*, depending on where multiple merging events occur in the coalescent tree (Eldon & Wakeley 2008). Moreover, genealogies for loci far apart on the same chromosome may remain correlated, and LD is a function of the rate of recombination and of the reproduction parameter (of the skew in offspring distribution; Birkner *et al.* 2012). These counter-intuitive results suggest that for species undergoing sweepstake reproduction, recombination hot spots may not be detectable by measuring recombination rates based on SNP frequencies. Conversely, the analysis of such species under the paradigm of the Kingman model would lead to over- or underestimate the rates of recombination and potentially misestimate the adaptive potential of a given species (Eldon & Wakeley 2008). Additionally, the misestimate of recombination rates would strongly affect the outcomes of analysis of hitchhiking in genomic islands of differentiation. Indeed, as the width of hitchhiked zones depends mainly of recombination and selection parameters (Barton 2000), effects of MMC on the shape of genomic islands should be taken into account.

Effect on F_{ST} and measure of population differentiation

Multiple merger coalescent models also affect the shape of genealogies of a sample from several populations. A

key factor to consider when comparing MMC and Kingman models is that the time and effective population size would scale differently because of the different coalescent rates (e.g. Donnelly & Kurtz 1999; Pitman 1999; Sagitov 1999; Eldon & Wakeley 2009). In a model with several population linked by migration, F_{ST} is a commonly computed index of allelic fixation. In the Wright–Fisher model, for a given mutation rate, F_{ST} is a function of the product $N \times m$ where N is the population size and m is the migration rate. Due to the difference in timescaling in Kingman and MMC models, Eldon and Wakeley (2009) show that a similar value of F_{ST} can be obtained for a given $N^2 \times m$ in a Moran model, and for $N^{\gamma} \times m^*$ in a MMC model, with $m^* > m$. (Note that following a time-rescaling argument, migration depends on N^2 in the Moran model). The scaling parameter of the MMC model is γ, regulating the frequency of large offspring production ($0 < \gamma < 2$) and for which high values tend to produce Kingman coalescent genealogies, and m^* is the migration rate (in a ψ-coalescent, but also valid for a beta-coalescent; Eldon & Wakeley 2009). In biological terms, this means that in spatially structured populations of a species with sweepstake reproduction, interpreting values of F_{ST} as under the Kingman coalescent would yield a clear underestimation of both N and the migration rate (m). In other words, for a given identical scaled migration and N in both MMC and Kingman models, higher rates of alleles fixation in each population and thus higher F_{ST} are expected to occur in MMC models.

In models of recent speciation where two incipient populations split from an ancestral one without post-divergence migration (Wakeley & Hey 1997), Eldon and Degnan (2012) show for a Λ-coalescent and ψ-coalescent models that the probabilities of monophyly or paraphyly of each population vary compared with the Kingman expectations. The number of incomplete lineage sorting, which is used by several methods to determine the structure of populations and times of split, is affected by MMC. In practice, this means that two species with sweepstake reproduction, which have split at time τ ago, would be estimated to have an older time of split and more incomplete lineage sorting when analysed under the classic assumptions of the Kingman setting (*e.g.* Wakeley & Hey 1997; Sousa & Hey 2013). This is because the time of allele fixation in MMC is shorter, so that more divergence (and substitutions) accumulates between the two incipient populations without gene flow than expected under the Kingman model.

Effect on natural selection

As MMC models affect the rate and strength of genetic drift, Der *et al.* (2011, 2012) demonstrate that the time of fixation and probability of fixation of alleles under positive or negative selection differ from the classic Wright–Fisher expectations. Sweepstake reproduction can either suppress or amplify the strength of selection (Der *et al.* 2011). A surprising result is, for example, that alleles under positive selection may have a probability of one to become fixed under MMC models, where it is of $2s$ in a classic Wright–Fisher model (where s is the coefficient of selection; Kimura 1954). In other words, as drift becomes weaker under MMC models, the efficacy of selection increases, because selection acts almost in a deterministic manner on allele frequencies during the phases between multiple merger events. Practically, these results point out two interesting features and potential bias in our current interpretation of population genetics results. (i) Natural selection coefficients estimated under Wright–Fisher model are overestimates of the real coefficients for population undergoing sweepstake reproduction. (ii) The efficacy of positive selection to favour alleles or that of negative selection to remove deleterious ones is greatly underestimated when ignoring the sweepstake reproduction in a population.

Current and potential applications of multiple merger coalescent

We present here a series of ecological set-ups, life history traits, species and biological systems for which genomic evolution in populations is likely better pictured by MMC rather than the currently used Kingman and Wright–Fisher models. We first describe neutral and then selective mechanisms which generate MMC genealogies. It is important to distinguish between these two types of mechanisms as neutral random processes affect the dynamics of diversity over the whole genome, whereas selective processes generate MMC at few loci or genomic regions.

Neutral mechanism: skew in offspring production per capita

Sweepstake reproduction defined as the high variance of reproductive success amongst individuals at every generation by random effect and density-dependent regulation of population has been suggested early on as a key characteristic of marine species (Beckenbach 1994; Hedgecock 1994; Fig. 3A and the model of Schweinsberg 2003). This has led to the development of empirical studies using genotyping in fish and crustaceans to confirm this effect for example in Pacific Oysters (Boom *et al.* 1994; Boudry *et al.* 2002) or atlantic cod (Li & Hedgecock 1998; Arnason *et al.* 2000; Arnason 2004). These studies have been recently reviewed in

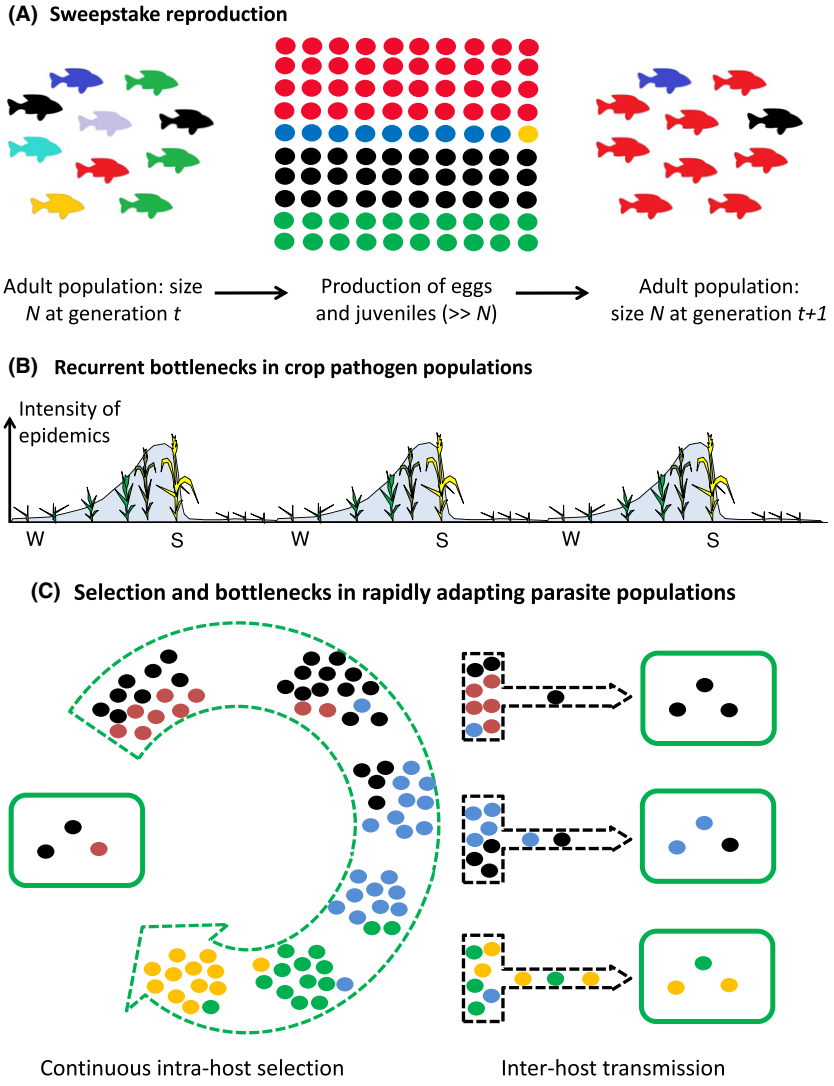

Fig. 3 Schematic representation of key life cycles and life history traits giving rise to multiple merger coalescent models. (A) Model of sweepstake reproduction common to many in marine organisms with N being the adult population size, which is also the carrying capacity of the population. Variance in offspring production is very large, and the production of eggs and juveniles largely exceeds the carrying capacity. Density-dependent regulation thus maintains the population at size N. (B) Demographic events drive random neutral evolution in crop pathogens over 3 years with seasonality (winter = W, summer = S). As the host plants grow during the year, the pathogen population grows. Harvest decreases dramatically the availability of hosts, as pathogen can only survive then on volunteer crops or wild relatives, generating regular periodic bottlenecks in pathogen density (y-axis). (C) Events in the parasite population: neutral stochastic events during interhost transmission and selective events during intrahost dynamics. Host individuals are denoted by the green solid rectangles, while parasite particles are the small filled circles. Adaptation in the parasite population is continuous in maintaining the fittest individuals within each infected host (green dotted arrow), while recurrent bottlenecks occur during interhost transmission (black dotted arrows).

Hedgecock and Pudovkin (2011). This group of organisms and genotyping data has been instrumental in the development of the statistical application of MMC models. The aim is here to infer from SNP data the rate of sweepstake events, that is, how often they occur and which percentage of the population is affected. Likelihood methods have thus been developed for the Λ-coalescent (Birkner *et al.* 2011), beta-coalescent (Birkner & Blath 2008; Steinrücken *et al.* 2013) or ψ-coalescent (Eldon & Wakeley 2006; Cenik & Wakeley 2010; Eldon 2011) to infer the parameters of interest based on the site frequency spectrum.

A population with strong density-dependent size regulation between juveniles and adults is included in Schweinsberg (2003) or in Eldon and Wakeley ψ-coalescent models. These are based on the classic sweepstake

hypothesis. However, this mechanism, and these models, may also apply generally to plant populations where the production of seeds exceeds largely the carrying capacity of a population. This is suggested for trees (Ingvarson 2010), and it may be a general feature of annual plants with fast growth colonizing disturbed habitats and producing numerous seeds (such as *Arabidopsis thaliana*). This model can also be extended to many species such as insects that experience several population boosts and bursts (Wallner 1987). The suitability of MMC models to explain patterns of diversity can be tested using genomic data in numerous plant populations (*e.g.* in *A. thaliana*) using the existing likelihood methods (Birkner & Blath 2008; Steinrücken *et al.* 2013). The applicability of MMC models for given plant and marine species has two impacts. First, this may drive us to reconsider predictions and analyses of polymorphism patterns in response to local adaptation and natural selection (*e.g.* Savolainen *et al.* 2013). Second, it would impact conservation biology measures to be adopted as the population size may not be a good indicator of future persistence if sweepstake events occur and can deplete the population from its genetic diversity (Hedgecock & Pudovkin 2011).

Neutral mechanism: parasitic life cycle

Parasites and pathogens such as viruses, fungi, oomycetes or bacteria may often present several life history traits and life cycles which promote MMC genealogies such as (i) sweepstake reproduction and (ii) alternate asexual and sexual phases.

First, pathogens, in particular plant or insects parasites producing aerial infectious spores, may present typical sweepstake reproduction. Spores production is well studied in crop pathogens, which can generate several thousands or millions of spores per lesion (Agrios 2005), as shown for two widely spread pathogen of wheat *Zymoseptoria tritici* (septoria of wheat, Simon & Cordo 1997) and *Puccinia striiformis* f. sp. *tritici* (wheat yellow rust, Hovmøller *et al.* 2011). However, when measuring effective population sizes using DNA polymorphism data, population sizes on the order of only few thousands are typically found at the continental scale: 5000 for *Z. tritici* (Stukenbrock *et al.* 2011, 2012) and 25 000 for yellow rust (Duan *et al.* 2009). Many plant pathogens may thus fulfil the criteria of sweepstake reproduction and of a large difference between observable population size and effective size (Fig. 3A), possibly due to a large majority of spores being unable to find available hosts in many geographical areas (the oasis in the desert metaphor, Brown *et al.* 2002).

Second, the life -cycle of many fungal plant pathogens (rusts, mildew, oomycetes), but also some plant species, crustaceans (Daphnia) or insects (aphids) comprising alternating phases of clonal reproduction and one or few sexual events per year, is expected to generate typical MMC genealogies. This occurs as the clonal phases generate a large amount of offsprings with a large reproductive variance between genotypes, with few clones possibly increasing in frequency due to random processes and clonal interference (Neher 2013). Moreover, the sexual phase, which occurs rarely and only depending on the environmental conditions, may represent a recurrent bottleneck for the survival of individuals to the next generation. Some rust pathogens, for example, reproduce sexually on a secondary aecidian host with a bottleneck possibly occurring at this stage (Duan *et al.* 2009). Understanding MMC models generated under such life cycles is of importance for predicting plant, animal, pests and pathogen adaptation in space and time, and the emergence of aggressive or virulent clonal lineages (Stukenbrock & McDonald 2008; Pybus & Rambaut 2009).

Demographic mechanism: bottlenecks and spatial structure with extinction/recolonization

Many parasites exhibit recurrent strong variation in population sizes. Strong recurrent bottlenecks will affect the whole genome, and the underlying genealogies are shown to converge to a Ξ-coalescent (Birkner *et al.* 2009). A first type of bottleneck occurs due to limited host availability. Crop pathogens in particular exhibit reduced population sizes at the end of the growing season when hosts are not available post-harvest (Fig. 3B; Stukenbrock & McDonald 2008). Similarly, host populations in the wild (insects or plants) are often unavailable or very drastically reduced during the winter (unfavourable) season. A second type of bottlenecks, which is common to many parasites, occurs during between host transmissions. Animal (e.g. malaria, virus) or plant (viruses or bacteria) parasites, which are transmitted by vectors (aphids, mosquitoes, etc.), experience strong bottlenecks as few viral particles or bacteria are stochastically transmitted to the next host (Fig. 3C; Moury *et al.* 2007; Pybus & Rambaut 2009). For a review and meta-analysis, see the recent study by Gutiérrez *et al.* (2012). A third type of bottleneck is produced inherently by epidemiological disease dynamics over time, where parasite population sizes show regular expansions and contractions (Pybus & Rambaut 2009). Balloux and Lehmann (2012) show that the interaction of varying population sizes due to strong recurrent bottlenecks with overlapping generations generates a neutral increase of substitution rates in the genome. This case applies particularly to many human parasites such as influenza or plague (Morelli *et al.* 2010), but

potentially to several crop and wild plant pathogens. The result is that variable mutation rates along branches of a genealogy of various parasite strains may be generated by these neutral random stochastic processes as emergent properties of MMC models, and calls for caution when inferring signatures of selection in the genome (for example using dN/dS ratio).

Most species live as a metapopulation which consists of demes, that is, local panmictic populations, connected by migration and subjected to regular extinction and recolonization events (Hanski 1998). This is a feature of many plant species (Freckleton & Watkinson 2002) and crustaceans such as *Daphnia* living in ephemeral habitats (Haag *et al.* 2005). The Kingman coalescent applied to a metapopulation suggests that the genealogy is divided into a short scattering phase and a long collecting phase (Wakeley & Aliacar 2001), while the rates of coalescence in these phases depend on the deme size and level of gene flow between demes. Using different sampling strategies, it is possible to take into account the genealogies of these phases and to study neutral and selective processes acting at the local (scattering phase) and the whole species (collecting phase) level (Städler *et al.* 2009; Tellier *et al.* 2011).

However, in a model with regular extinction and recolonization and a very large number of demes, the genealogy can present locally simultaneous multiple mergers (Ξ-coalescent; Limic & Sturm 2006; Taylor & Véber 2009; Barton *et al.* 2010). MMC genealogies are thus expected over the whole genome in each deme and would affect the study of evolutionary processes locally, that is, the inference of demography and selection in the scattering phase. Interestingly, in a model assuming a Λ-coalescent in each deme, the genealogy underlying a species-wide sampling of individuals (Städler *et al.* 2009), that is, reflecting the collecting phase, tends to converge to the classic Kingman coalescent (Heuer & Sturm 2013). It follows that the collecting phase may be well approximated by the classic Kingman coalescent, even though MMC genealogies occur locally within each deme. The occurrence of Λ- or Ξ-coalescents within structured populations following different population dynamics, especially strong extinction-recolonization, and sampling schemes has thus consequences for our understanding of polymorphism data (*e.g.* measure of F_{ST}) and for correctly estimating gene flow among populations (see above). Using an ad hoc model to represent the genealogies may be important in conservation biology for studying species in fragmented habitats and the consequences of the decrease or increase in gene flow on the genetic diversity. Additionally, MMC models may be relevant to study the population structure of crop pathogens which undergo long distance dispersal and regular extinction-recolonization such as wheat yellow rust (*P. striiformis* f. sp. *tritici*; Brown & Hovmøller 2002).

Consequence of oversampling

Multiple merger coalescent models appear under another violation of the Kingman assumptions, namely when the sample size (n) is larger or on the same order of magnitude than the effective population size (N). Wakeley and Takahashi (2003) demonstrate that when $n \gg N$, there is a short phase of numerous simultaneous multiple coalescent events, much like a Ξ-coalescent model. Interestingly, using such large sample size, it is possible to estimate the population size (N) and the mutation rate separately, because the number of singletons is then an estimator of n/N (Wakeley & Takahashi 2003). This result has been recently used in the human population to estimate the actual European effective population size to be around three millions and the ancestral size of around 7700 (Coventry *et al.* 2010; Nelson *et al.* 2012). We suggest here that this method can be used for other plant or animal species where the effective population size estimated from sequence data is small (*e.g.* if the species is endangered), and sampling of numerous individuals is feasible.

Selective mechanisms: Selective sweeps and rapid adaptation

Multiple merger coalescent models are also shown to represent genealogies at loci under positive selection (Durrett & Schweinsberg 2005; Schweinsberg & Durrett 2005). Selective sweeps generate the so-called hitch-hiking effect around the selected site (Maynard-Smith & Haigh 1974). Genome scans are used extensively to detect loci underlying species adaptation to their environment, based on the principle that these few loci show an outlier genealogy compared with the rest of the genome. Durrett and Schweinsberg (2005) show that genealogies during and shortly after selective sweeps are well approximated by certain Ξ-coalescents. This is not surprising as positive selection creates an excess of low and high frequency variants due to the high rates of coalescence observed during the selection phase (Coop & Ralph 2012). Detecting recent positive selection at a given locus is thus equivalent to estimate whether its genealogy fits better to an MMC model rather than the classic Kingman compared with other loci in the genome. Existing likelihood methods for MMC models may be used for genome-wide scans for loci under positive selection, and estimates of MMC parameters would permit inferences of the strength and time of selection. In the future, inferences of demographic parameters

under the MMC models would allow to improve the detection of outlier loci (potentially under selection) at genome-wide scale.

A final important occurrence of MMC models is observed in parasites which undergo several multiplications per host generations, especially via long period of intrahost evolution, and are subjected to strong within host selection. This intrahost evolution in facultative asexual parasites is driven by strong genetic draft, due to clonal interference and continuous positive selection as shown in Fig. 3C (Neher & Shraiman 2011; Desai *et al.* 2013; reviewed in Neher 2013). Viruses, for example, undergo strong bottlenecks during between host transmissions and exhibit huge subsequent diversification driven by intrahost selective processes (Fig. 3C; Moury *et al.* 2007; Pybus & Rambaut 2009; Neher & Hallatschek 2013). The model of selection for fast adapting parasites clearly violates the classical models based on the Kingman coalescent and is shown to follow a Bolthausen–Sznitman coalescent (Neher *et al.* 2013). The genealogy is driven by strong directional selection with continuous adaptation for a given fitness trait (Brunet *et al.* 2007; Brunet & Derrida 2012; Neher & Hallatschek 2013). The Bolthausen–Sznitman coalescent has been applied to polymorphism data from human viruses (Neher & Hallatschek 2013; Neher *et al.* 2013), but may be in principle relevant to polycyclic crop viruses, bacteria and fungi (for example the aforementioned *Z. tritici* and *P. striiformis* f. sp. *tritici*, but also other rusts or oomycetes).

Note also that the strong efficacy of selection under MMC models (see above, Der *et al.* 2011) and the strong selective pressure exerted by the use of uniform crop genotypes within fields (Stukenbrock & McDonald 2008) may explain in part the abundance of signatures of positive selection and the pervasive purifying selection found in population polymorphism data of various crop pathogens such as *Z. tritici* (Stukenbrock *et al.* 2011, 2012; Brunner *et al.* 2013) and other species (Wicker *et al.* 2013). The proportion of the genome which is affected by genetic draft due to selection (Neher & Hallatschek 2013; Neher *et al.* 2013), and thus the extent of linkage disequilibrium (Eldon & Wakeley 2008; Birkner *et al.* 2012), will depend on the frequency of recombination in a given parasite species. A future challenge remains, nevertheless, to distinguish such signatures of natural selection (positive or negative) from the genome-wide variance in topology and length under occurring MMC genealogies. The large variance in possible MMC genealogy originates from random neutral processes such as sweepstake reproduction, recurrent bottlenecks and alternate asexual and sexual phases as discussed above.

Conclusions

We highlight here several species and life history traits which promote variance in offspring production and thus depart from the assumptions of the classic WF model and Kingman coalescent. Analysing genomic data in such species may thus lead to potential source of error, misinterpretation of past demography and false detection of genes under selection. We advocate here that this burst of new MMC coalescent models represents a chance to (i) improve model-based inference in many bacterial, fungal, plant and animal species and (ii) to connect population genetics models with life history traits. However, efforts remain to be made to test the relevance of these various MMC models (Λ-coalescent, beta-coalescent, Bolthausen–Sznitman, Ξ-coalescent) by integrating realistic variable population size and population structure in order to compare with existing extensions of the Kingman model. Finally, coalescent theory does not only allow us to draw inference from polymorphism data about past demography and the action of selection in the genome; it also provides a theoretical framework to develop new statistical methods and a renewed understanding of genome data in numerous bacteria, fungi, viruses, marine organisms or plant species with peculiar life cycles.

Acknowledgements

The authors are indebted to Matthias Birkner, Bjarki Eldon, Fabian Freund for their patience in answering questions, for useful discussions and comments. We thank three anonymous reviewers for helpful comments on the manuscript. We thank M. Birkner for sharing R codes for preparing Fig. 2, and Jerome Enjalbert for inspiration for Fig. 3B. AT acknowledges support the German Federal Ministry of Education and Research (BMBF) within the AgroClustEr 'Synbreed – Synergistic plant and animal breeding' (FKZ: 0315528I).

References

Agrios GN (2005) *Plant Pathology*, 5th edn. Academic Press, San Diego.

Arnason E (2004) Mitochondrial cytochrome b DNA variation in the high-fecundity Atlantic cod: trans-Atlantic clines and shallow gene genealogy. *Genetics*, **166**, 1871–1885.

Arnason E, Petersen PH, Kristinsson K, Sigurgíslason H, Pálsson S (2000) Mitochondrial cytochrome b DNA sequence variation of Atlantic cod from Iceland and Greenland. *Journal of Fish Biology*, **56**, 409–430.

Balloux F, Lehmann L (2012) Substitution rates at neutral genes depend on population size under fluctuating demography and overlapping generations. *Evolution*, **66**, 605–611.

Barton NH (2000) Genetic hitchhiking. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, **355**, 1553–1562.

Barton NH, Kelleher J, Etheridge AM (2010) A new model for extinction and recolonization in two dimensions: quantifying phylogeography. *Evolution*, **64**, 2701–2715.

Basdevant A-L, Goldschmidt C (2008) Asymptotics of the allele frequency spectrum associated with the Bolthausen-Sznitman coalescent. *Electronic Journal of Probability*, **13**, 486–512.

Beckenbach AT (1994) Mitochondrial haplotype frequencies in oysters: neutral alternatives to selection models. In: *Non-Neutral Evolution: Theories and Molecular Data* (ed. Golding B), pp. 188–198. Chapman & Hall, New York, USA.

Berestycki N (2009) Recent progress in coalescent theory. *Ensaios Matematicos*, **16**, 1–193.

Berestycki J, Berestycki N, Schweinsberg J (2007) Beta-coalescents and continuous stable random trees. *The Annals of Probability*, **35**, 1835–1887.

Berestycki J, Berestycki N, Schweinsberg J (2008) Small-time behavior of beta coalescents. *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, **44**, 214–238.

Berestycki J, Berestycki N, Limic V (2014) Asymptotic sampling formulae for Λ-coalescents. *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*. in press.

Bertoin J, Gall J-FL (2003) Stochastic flows associated to coalescent processes. *Probability Theory and Related Fields*, **126**, 261–288.

Birkner M, Blath J (2008) Computing likelihoods for coalescents with multiple collisions in the infinitely many sites model. *Journal of Mathematical Biology*, **57**, 435–465.

Birkner M, Blath J, Moehle M, Steinruecken M, Tams J (2009) A modified lookdown construction for the Xi-Fleming-Viot process with mutation and populations with recurrent bottlenecks. *Alea-Latin American Journal of Probability and Mathematical Statistics*, **6**, 25–61.

Birkner M, Blath J, Steinrücken M (2011) Importance sampling for Lambda-coalescents in the infinitely many sites model. *Theoretical Population Biology*, **79**, 155–173.

Birkner M, Blath J, Eldon B (2012) An ancestral recombination graph for diploid populations with skewed Offspring distribution. *Genetics*, **193**, 255–290.

Birkner M, Blath J, Eldon B (2013) Statistical properties of the site-frequency spectrum associated with lambda-coalescents. *Genetics*, **195**, 1037–1053.

Bolthausen E, Sznitman AS (1998) On Ruelle's probability cascades and an abstract cavity method. *Communications in Mathematical Physics*, **197**, 247–276.

Boom JDG, Boulding EG, Beckenbach AT (1994) Mitochondrial DNA variation in introduced populations of Pacific Oyster, *Crassostrea gigas*, in British Columbia. *Canadian Journal of Fisheries and Aquatic Sciences*, **51**, 1608–1614.

Boudry P, Collet B, Cornette F, Hervouet V, Bonhomme F (2002) High variance in reproductive success of the Pacific oyster (*Crassostrea gigas*, Thunberg) revealed by microsatellite-based parentage analysis of multifactorial crosses. *Aquaculture*, **204**, 283–296.

Brown JKM, Hovmøller MS (2002) Aerial dispersal of pathogens on the global and continental scales and its impact on plant disease. *Science*, **297**, 537–541.

Brown JKM, Hovmøller MS, Wyand RA, Yu DZ (2002) Oases in the desert: dispersal and host specialization of biotrophic fungal pathogens of plants. In: *Dispersal Ecology* (eds Bullock JM, Kenward RE & Hails RS), pp. 395–409. Blackwell Science, Oxford, UK.

Brunet É, Derrida B (2012) How genealogies are affected by the speed of evolution. *Philosophical Magazine*, **92**, 255–271.

Brunet É, Derrida B, Mueller AH, Munier S (2007) Effect of selection on ancestry: an exactly soluble case and its phenomenological generalization. *Physical Review E*, **76**, 041104.

Brunner PC, Torriani SFF, Croll D, Stukenbrock EH, McDonald BA (2013) Coevolution and life cycle specialization of plant cell wall degrading enzymes in a hemibiotrophic pathogen. *Molecular Biology and Evolution*, **30**, 1337–1347.

Cannings C (1974) The latent roots of certain Markov chains arising in genetics: a new approach, I. Haploid models. *Advances in Applied Probability*, **6**, 260–290.

Cenik C, Wakeley J (2010) Pacific Salmon and the coalescent effective population size. *PLoS One*, **5**, e13019.

Charlesworth B, Charlesworth D, Barton NH (2003) The effects of genetic and geographic structure on neutral variation. *Annual Review of Ecology, Evolution, and Systematics*, **34**, 99–125.

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

Coventry A, Bull-Otterson LM, Liu X *et al.* (2010) Deep resequencing reveals excess rare recent variants consistent with explosive population growth. *Nature Communications*, **1**, 131.

Der R, Epstein CL, Plotkin JB (2011) Generalized population models and the nature of genetic drift. *Theoretical Population Biology*, **80**, 80–99.

Der R, Epstein C, Plotkin JB (2012) Dynamics of neutral and selected alleles when the Offspring distribution is skewed. *Genetics*, **191**, 1331–1344.

Desai MM, Walczak AM, Fisher DS (2013) Genetic diversity and the structure of genealogies in rapidly adapting populations. *Genetics*, **193**, 565–585.

Donnelly P, Kurtz TG (1999) Particle representations for measure-valued population models. *The Annals of Probability*, **27**, 166–205.

Duan X, Tellier A, Wan A *et al.* (2009) *Puccinia striiformis* f.sp. *tritici* presents high diversity and recombination in the over-summering zone of Gansu, China. *Mycologia*, **102**, 44–53.

Durrett R, Schweinsberg J (2005) A coalescent model for the effect of advantageous mutations on the genealogy of a population. *Stochastic Processes and their Applications*, **115**, 1628–1657.

Eldon B (2009) Structured coalescent processes from a modified Moran model with large offspring numbers. *Theoretical Population Biology*, **76**, 92–104.

Eldon B (2011) Estimation of parameters in large offspring number models and ratios of coalescence times. *Theoretical Population Biology*, **80**, 16–28.

Eldon B, Degnan JH (2012) Multiple merger gene genealogies in two species: monophyly, paraphyly, and polyphyly for two examples of Lambda coalescents. *Theoretical Population Biology*, **82**, 117–130.

Eldon B, Wakeley J (2006) Coalescent processes when the distribution of Offspring number among individuals is highly skewed. *Genetics*, **172**, 2621–2633.

Eldon B, Wakeley J (2008) Linkage disequilibrium under skewed Offspring distribution among individuals in a population. *Genetics*, **178**, 1517–1532.

Eldon B, Wakeley J (2009) Coalescence times and FST under a skewed Offspring distribution Among individuals in a population. *Genetics*, **181**, 615–629.

Etheridge AM (2011) *Some Mathematical Models from Population Genetics. École d'Été de Probabilités de Saint-Flour XXXIX-2009*. Springer, Berlin, Heidelberg.

Fisher RA (1930) *The Genetical Theory of Natural Selection*. Clarendon Press, Oxford, UK.

Freckleton RP, Watkinson AR (2002) Large-scale spatial dynamics of plants: metapopulations, regional ensembles and patchy populations. *Journal of Ecology*, **90**, 419–434.

Gutiérrez S, Michalakis Y, Blanc S (2012) Virus population bottlenecks during within-host progression and host-to-host transmission. *Current Opinion in Virology*, **2**, 546–555.

Haag CR, Riek M, Hottinger JW, Pajunen VI, Ebert D (2005) Genetic diversity and genetic differentiation in *Daphnia* metapopulations with subpopulations of known age. *Genetics*, **170**, 1809–1820.

Hanski I (1998) Metapopulation dynamics. *Nature*, **396**, 41–49.

Harrang E, Lapègue S, Morga B, Bierne N (2013) A high load of non-neutral amino-acid polymorphisms explains high protein diversity despite moderate effective population size in a marine bivalve with sweepstakes reproduction. *Genes, Genomes, Genetics*, **3**, 333–341.

Hedgecock D (1994) Does variance in reproductive success limit effective population size of marine organisms? In: *Genetics and Evolution of Aquatic Organisms* (ed. Beaumont A), pp. 122–134. Chapman & Hall, London, UK.

Hedgecock D, Pudovkin AI (2011) Sweepstakes reproductive success in highly fecund marine fish and shellfish: a review and commentary. *Bulletin of Marine Science*, **87**, 971–1002.

Hernandez RD, Kelley JL, Elyashiv E *et al.* (2011) Classic selective sweeps were rare in recent human evolution. *Science*, **331**, 920–924.

Heuer B, Sturm A (2013) On spatial coalescents with multiple mergers in two dimensions. *Theoretical Population Biology*, **87**, 90–104.

Hoban S, Bertorelle G, Gaggiotti OE (2012) Computer simulations: tools for population and evolutionary genetics. *Nature Reviews Genetics*, **13**, 110.

Hovmøller MS, Sørensen CK, Walter S, Justesen AF (2011) Diversity of *Puccinia striiformis* on cereals and grasses. *Annual Review of Phytopathology*, **49**, 197–217.

Hudson RR (1983) Properties of a neutral allele model with intragenic recombination. *Theoretical Population Biology*, **23**, 183–201.

Huillet T, Möhle M (2011) Population genetics models with skewed fertilities: a forward and backward analysis. *Stochastic Models*, **27**, 521–554.

Ingvarson PK (2010) Nucleotide polymorphism, linkage disequilibrium and complex trait dissection in populus. In: *Genetics and Genomics of Populus Plant Genetics and Genomics: Crops and Models* (eds Jansson S, Bhalerao R & Groover A), pp. 91–111. Springer, New York.

Kaj I, Krone SM (2003) The coalescent process in a population with stochastically varying size. *Journal of Applied Probability*, **40**, 33–48.

Kimura M (1954) Process leading to quasi-fixation of genes in natural populations due to random fluctuation of selection intensities. *Genetics*, **39**, 280–295.

Kingman JFC (1982) The coalescent. *Stochastic Processes and their Applications*, **13**, 235–248.

Krone SM, Neuhauser C (1997) Ancestral processes with selection. *Theoretical Population Biology*, **51**, 210–237.

Li G, Hedgecock D (1998) Genetic heterogeneity, detected by PCR-SSCP, among samples of larval Pacific oysters (*Crassostrea gigas*) supports the hypothesis of large variance in reproductive success. *Canadian Journal of Fisheries and Aquatic Sciences*, **55**, 1025–1033.

Limic V, Sturm A (2006) The spatial Λ-coalescent. *Electronic Journal of Probability*, **11**, 363–393.

Malecot G (1941) Etude mathématiques des populations "mendéliennes". *Annales de l'Université de Lyon, Sciences A*, **4**, 45–60.

Maynard-Smith J, Haigh J (1974) The hitch-hiking effect of a favourable gene. *Genetics Research*, **23**, 23–35.

Möhle M, Sagitov S (2001) A classification of coalescent processes for haploid exchangeable population models. *The Annals of Probability*, **29**, 1547–1562.

Moran PAP (1958) Random processes in genetics. *Mathematical Proceedings of the Cambridge Philosophical Society*, **1**, 60–71.

Morelli G, Song Y, Mazzoni CJ *et al.* (2010) *Yersinia pestis* genome sequencing identifies patterns of global phylogenetic diversity. *Nature Genetics*, **42**, 1140–1143.

Moury B, Fabre F, Senoussi R (2007) Estimation of the number of virus particles transmitted by an insect vector. *Proceedings of the National Academy of Sciences*, **104**, 17891–17896.

Neher RA (2013) Genetic draft, selective interference, and population genetics of rapid adaptation. *Annual Review of Ecology, Evolution, and Systematics*, **44**, 195–215.

Neher RA, Hallatschek O (2013) Genealogies of rapidly adapting populations. *Proceedings of the National Academy of Sciences USA*, **110**, 437–442.

Neher RA, Shraiman BI (2011) Genetic draft and quasi-neutrality in large facultatively sexual populations. *Genetics*, **188**, 975–996.

Neher RA, Kessinger TA, Shraiman BI (2013) Coalescence and genetic diversity in sexual populations under selection. *Proceedings of the National Academy of Sciences USA*, **110**, 15836–15841.

Nelson MR, Wegmann D, Ehm MG *et al.* (2012) An abundance of rare functional variants in 202 drug target genes sequenced in 14 002 people. *Science*, **337**, 100–104.

Nunney L (1995) Measuring the ratio of effective population size to adult numbers using genetic and ecological data. *Evolution*, **49**, 389–392.

Pitman J (1999) Coalescents with multiple collisions. *The Annals of Probability*, **27**, 1870–1902.

Pybus OG, Rambaut A (2009) Evolutionary analysis of the dynamics of viral infectious disease. *Nature Reviews Genetics*, **10**, 540–550.

Rosenberg NA, Nordborg M (2002) Genealogical trees, coalescent theory and the analysis of genetic polymorphisms. *Nature Reviews Genetics*, **3**, 380–390.

Sagitov S (1999) The general coalescent with asynchronous mergers of ancestral lines. *Journal of Applied Probability*, **36**, 113–1125.

Sargsyan O, Wakeley J (2008) A coalescent process with simultaneous multiple mergers for approximating the gene genealogies of many marine organisms. *Theoretical Population Biology*, **74**, 104–114.

Savolainen O, Lascoux M, Merilä J (2013) Ecological genomics of local adaptation. *Nature Reviews Genetics*, **14**, 807–820.
Schweinsberg J (2000) Coalescents with simultaneous multiple collisions. *Electronic Journal of Probability*, **5**, 1–50.
Schweinsberg J (2003) Coalescent processes obtained from supercritical Galton-Watson processes. *Stochastic Processes and their Applications*, **106**, 107–139.
Schweinsberg J, Durrett R (2005) Random partitions approximating the coalescence of lineages during a selective sweep. *The Annals of Applied Probability*, **15**, 1591–1651.
Simon M, Cordo C (1997) Inheritance of partial resistance to *Septoria tritici* in wheat (*Triticum aestivum*): limitation of pycnidia and spore production. *Agronomie*, **17**, 343–347.
Sousa V, Hey J (2013) Understanding the origin of species with genome-scale data: modelling gene flow. *Nature Reviews Genetics*, **14**, 404–414.
Städler T, Haubold B, Merino C, Stephan W, Pfaffelhuber P (2009) The impact of sampling schemes on the site frequency spectrum in nonequilibrium subdivided populations. *Genetics*, **182**, 205–216.
Steinrücken M, Birkner M, Blath J (2013) Analysis of DNA sequence variation within marine species using Beta-coalescents. *Theoretical Population Biology*, **87**, 15–24.
Stukenbrock EH, McDonald BA (2008) The origins of plant pathogens in agro-ecosystems. *Annual Review of Phytopathology*, **46**, 75–100.
Stukenbrock EH, Bataillon T, Dutheil JY *et al.* (2011) The making of a new pathogen: insights from comparative population genomics of the domesticated wheat pathogen *Mycosphaerella graminicola* and its wild sister species. *Genome Research*, **21**, 2157–2166.
Stukenbrock EH, Christiansen FB, Hansen TT, Dutheil JY, Schierup MH (2012) Fusion of two divergent fungal individuals led to the recent emergence of a unique widespread pathogen species. *Proceedings of the National Academy of Sciences USA*, **109**, 10954–10959.
Tajima F (1989) Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics*, **123**, 585–595.
Taylor JE, Véber A (2009) Coalescent processes in subdivided populations subject to recurrent mass extinctions. *Electronic Journal of Probability*, **14**, 242–288.
Tellier A, Laurent SJY, Lainer H, Pavlidis P, Stephan W (2011) Inference of seed bank parameters in two wild tomato species using ecological and genetic data. *Proceedings of the National Academy of Sciences USA*, **108**, 17052–17057.
Wakeley J (2008) *Coalescent Theory: An Introduction*. Roberts & Company Publishers, Greenwood Village, Colorado.
Wakeley J (2013) Coalescent theory has many new branches. *Theoretical Population Biology*, **87**, 1–4.
Wakeley J, Aliacar N (2001) Gene genealogies in a metapopulation. *Genetics*, **159**, 893–905.
Wakeley J, Hey J (1997) Estimating ancestral population parameters. *Genetics*, **145**, 847–855.
Wakeley J, Takahashi T (2003) Gene genealogies when the sample size exceeds the effective size of the population. *Molecular Biology and Evolution*, **20**, 208–213.
Wallner WE (1987) Factors affecting insect population dynamics: differences between outbreak and non-outbreak species. *Annual Review of Entomology*, **32**, 317–340.
Watterson GA (1984) Allele frequencies after a bottleneck. *Theoretical Population Biology*, **26**, 387–407.
Wicker T, Oberhaensli S, Parlange F *et al.* (2013) The wheat powdery mildew genome shows the unique evolution of an obligate biotroph. *Nature Genetics*, **45**, 1092–1096.
Wright S (1931) Evolution in Mendelian populations. *Genetics*, **16**, 0097–0159.

A.T. and C.L. wrote the manuscript. A.T. performed the simulations for Fig. 2.

Appendix 1

We indicate in this section the general formulae for different types of coalescent models describing the rates of coalescence of k lineages when there are n active ancestral lineages in total. For a sample size n, each k-tuple of lineages is merging to form a single lineage at rate $\lambda_{n,k}$, and no other transitions are possible.

$$\lambda_{n,k} = \int_0^1 x^{k-2}(1-x)^{n-k}\Lambda(dx) \text{ with } 2 \leq k \leq n \quad \text{(A1)}$$

where Λ is a probability measure on [0,1] with which events of a certain size happen in the total population.

Kingman coalescent

The Kingman coalescent is obtained from eqn A1 with:

$$\Lambda(dx) = \delta_0 dx \quad \text{(A2)}$$

which is the Dirac point mass function at 0. In other words, the Kingman coalescent is defined with the rate of coalescence being $\lambda_{n,k} = 0$ except if $k = 2$ where $\lambda_{n,2} = 1$. Only binary collisions, that is, coalescence of two lineages, are thus possible.

Λ-Coalescent

The Λ-coalescent is generally defined in eqn A1 where Λ specifies the size and probability of the coalescent events following given specific Dirac mass functions with peculiar interpretations. The function

$$\Lambda(dx) = c\delta_y dx\,(c > 0) \quad \text{(A3)}$$

defines a model where a fraction y of the population is replaced by the offsprings of a single ancestor at rate c. This macroscopic reproduction event is called an 'extreme reproductive behaviour' and leads to multiple collisions in the genealogy (coalescence of more than two lineages).

Another special extreme case occurs when choosing $\Lambda = \delta_1 dx$ where the Dirac mass function is on one, which generates a star-like-genealogy where all n lineages merge to a single ancestor.

Beta-coalescent

This class of coalescent is a special case of Λ-coalescents for which Λ has a density defined as a function beta (α, 2- α) (with $1 \leq \alpha < 2$). Note that the special case where $\alpha = 2$ corresponds to the classic Kingman coalescent above.

$$\Lambda(dx) = \frac{1}{\Gamma(2-\alpha)\Gamma(\alpha)} x^{1-\alpha}(1-x)^{\alpha-1}dx \qquad \text{(A4)}$$

As for the Λ-coalescent, the 'extreme reproductive behaviour' in eqn A4 leads to multiple collisions in the genealogy with coalescence of more than two lineages.

Bolthausen–Sznitman

The Bolthausen–Sznitman coalescent is obtained as a specific case of the Λ-coalescent where $\Lambda(dx)$ = beta(1,1), or in other words, a beta-coalescent with $\alpha = 1$. In this case, the Dirac function has a uniform distribution of mass on [0,1].

ψ-coalescent

Eldon and Wakeley (2006) have developed a model where a single parent chosen at random from the population of size N contributes either 1) one offspring with probability $1-\varepsilon$ as in the classic Wright–Fisher model, or 2) ψN offsprings with probability ε. The latest case generates a large reproduction event, and a multiple merging of several lineages. The case where $\varepsilon = 1/N^{\gamma}$ is considered so that if $0 < \gamma < 2$. The function Λ from eqn A1 is:

$$\Lambda(dx) = \left(\delta_0 + c\Psi^2\delta_\Psi\right)dx$$

This function defines a model where a fraction ψ of the population is replaced by the offsprings of a single ancestor at rate c.

Eldon and Wakeley rewrote eqn A1 to define the rate of multiple merger of k lineages among n under their model:

$$\lambda_{n,k} = \binom{n}{k}\Psi^k(1-\Psi)^{n-k}, \text{ with } 0 < \Psi < 1$$

.