

1 **Coalescent inferences in conservation genetics: should the exception become the rule?**

2 Valeria Montano¹⁺

3 ¹School of Life Sciences, Ecole Polytechnique Fédérale de Lausanne, CH-1015 Lausanne,

4 Switzerland

5 ⁺ email: montano@epfl.ch

6 **Abstract**

7 Genetic estimates of effective population size (N_e) are an established means to develop informed
8 conservation policies. Another key goal to pursue the conservation of endangered species is keeping
9 the connectivity across fragmented environments, to which genetic inferences of gene flow and
10 dispersal greatly contribute. Most current statistical tools for estimating such population demographic
11 parameters are based on Kingman's coalescent (KC). However, KC is inappropriate for taxa
12 displaying skewed reproductive variance, a property widely observed in natural species. Coalescent
13 models that consider skewed reproductive success – called multiple merger coalescent (MMCs) –
14 have been shown to substantially improve estimates of N_e when the distribution of offspring per
15 capita is highly skewed. MMCs predictions of standard population genetic parameters, including the
16 rate of loss of genetic variation and the fixation probability of strongly selected alleles, substantially
17 depart from KC predictions. These extended models also allow studying gene genealogies in a spatial
18 continuum, providing a novel theoretical framework to investigate spatial connectivity. Therefore,
19 development of statistical tools based on MMC's should substantially improve estimates of
20 population demographic parameters with major conservation implications.

21 **Recent developments in coalescent theory**

22 Estimates of effective population size (N_e), defined by Wright as the number of reproducing
23 lineages in an idealized population [1], are among the parameters used by the International Union for
24 the Conservation of Nature (IUCN) to classify endangered species and to identify the minimum
25 viable population size preventing extinction [2-3]. It has been suggested that IUCN thresholds of N_e
26 recommended to avoid inbreeding depression and maintain evolutionary potential should be revised,
27 as theoretical predictions often fail to match empirical observations [3]. However, a theoretical
28 revision of N_e thresholds will be ineffective to improve conservation recommendations if it is based
29 on inappropriate evolutionary models.

30 Most methods applied in molecular ecology to infer demographic parameters from genetic
31 data (e.g., Beast, Splatche, Ima, $\delta a \delta i$, FastSimcoal2, [4-8]) rely on Kingman's coalescent (KC; [9]) or
32 its forward dual, the Wright-Fisher model (WF; [10]). Although KC has proven robust to violations
33 of most of its assumptions, it drastically fails to approximate the genealogies of species with high
34 reproductive skew [11], whereby few individuals contribute most of the offspring to the next
35 generation (Sweepstakes Reproductive Success, or SRS [12]). Skewed distribution of per-capita
36 reproductive success is widely observed among both marine and terrestrial species, from plants to
37 parasites, but also among social birds and mammals [13]. SRS generally characterizes clonally
38 reproducing organisms as much as species with high fecundity and low investment in parental care
39 and thus applies to many endangered species, for instance amphibians and commercial fish.
40 Moreover, skewed individual reproductive success is not only due to intrinsic reproductive properties
41 of a species, but can happen during strong population bottlenecks where only few individuals survive
42 (e.g., a virus infecting a new host), during rapid population expansions [14], and during non-neutral
43 processes, such as the appearance of a strongly beneficial allele which can drag a genome to replace
44 an important fraction of the population within a few generations [15] (Figure 1).

45 KC model neglects the probability of more than two lineages to merge at each coalescent
46 event, but when the offspring of a few individuals replaces a large fraction of the population at each
47 reproductive event, the probability of multiple lineages merging in backward time becomes high.
48 Hence, under skewed reproductive success, KC forces lineages involved in multiple and/or
49 simultaneous merges to coalesce pairwise producing genealogical trees with misleading branch
50 lengths and shape [11,14,15]. KC is a limit case of more complex coalescent processes, called

51 multiple merger coalescences (MMCs), addressed in several recent studies, e.g., [11,12,14-18] and
52 excellently reviewed in [18]. MMCs cover comprehensive scenarios, spanning from multiple
53 lineages merging into one at each coalescent event (Λ - coalescent and its limit cases – β coalescent
54 and Bolthausen-Sznitman coalescent [18]) to simultaneous multiple merging of multiple lineages at
55 each coalescent event (Ξ -coalescent [18]). In MMC models, time-dependent changes in allele
56 frequencies depart from KC predictions, consequently, probability of and time to fixation of both
57 neutral and beneficial alleles, and, thus, the expected number of segregating sites dramatically change
58 [19,20]. All of these measures are important to evaluate the health status of endangered species and
59 their potential for adaptation to challenging environments [3].

60 When reproduction is highly skewed, few lineages substantially contribute to the next
61 generation which means that the value of N_e , expressed by the parameter θ ($2N_e\mu$), is expected to be
62 very low. However, under MMCs, alleles can persist at the same frequency for longer time than
63 under KC before changing state, implying a reduced probability of loss or fixation for very low or
64 high frequency alleles, respectively [19,20]. In contrast, when offspring variance and N_e are small,
65 alleles at low frequencies are more likely to be lost by drift. Hence, under MMCs, the number of
66 segregating sites and the number of singletons are predicted and empirically observed to assume
67 close values, while under KC predicted number of singletons is usually much lower than number of
68 segregating sites [11,16-18,21]. As a consequence, new beneficial mutations also show a higher
69 chance to get lost under KC than under MMCs [19,20]. When few individuals contribute most of the
70 offspring to the next generation, the frequency of few genotypes can increase substantially more than
71 predicted by neutral KC. We can think of this scenario in terms of single lineages' rapid expansion,
72 from which it follows that a high number of singletons can appear as the local genealogies become
73 star-like. However, this scenario does not imply an expansion of the population size which can
74 remain constant.

75 These differences between the KC and MMCs predictions explain two important results.
76 First, MMCs estimates of N_e in marine species point to much lower values than KC estimates. In [11],
77 the value of θ calculated for a population of oysters is 50 under KC and 0.031 under MMCs. From a
78 conservation perspective, this result implies that high genetic variability can be generated by a very
79 low number of lineages and thus an actual population might decline substantially without evident loss
80 of genetic variation. At the same time, the ability of few individuals to quickly regenerate

81 considerable genetic variation and the chance of new beneficial mutations to persist might result in
82 high potential for rapid adaptation. Second, under MMCs and constant population size, a low θ value
83 can recover both the observed number of segregating sites and singletons, while KC estimates fail to
84 do so [11,21]. Therefore, conclusions pointing to population expansion based on excess of singletons
85 – negative values of Tajima's D – should be carefully evaluated in molecular ecology studies.

86

87 **Spatial connectivity and continuous space evolution**

88 Another theoretical advance of MMCs is the possibility to model continuous space evolution
89 overcoming historical limitations. Indeed, models based on KC fail to control local population
90 growth in continuous space, with the consequence that parts of the space grow unlimitedly and others
91 become completely empty (a dynamic known as pain in the torus; [22,23]). As maintaining
92 connectivity across habitats is indicated as a conservation priority [24], approaches to estimate
93 connectivity in continuous landscapes based on circuit theory were developed as alternative to
94 coalescent-based models [24,25]. Explicit spatial coalescent simulators based on KC (e.g,[5]) are still
95 hampered by the use of discrete units which force coalescent events in non-contiguous populations
96 [25] thus limiting their usefulness compared to alternative approaches [24,25]. In species with long
97 distance dispersal ability and skewed reproductive success, local populations show low values of N_e
98 associated to higher pairwise F_{ST} between closer than more distant populations [26]. This pattern can
99 be explained by local bottlenecks due to few individuals reproducing and long distance dispersal
100 events.

101 A forward model based on extinction-recolonization events (Λ -Fleming-Viot) allows to model
102 evolution in spatial continuum using stochastic regulation of local size by randomly drawing the
103 number of individuals destined for extinction (extinction event) and the number that will repopulate
104 the same area from local or external parental lineages (recolonization event) [27,28]. The multiple
105 merging spatial- Λ -coalescent is the backward dual of the forward Λ -Fleming-Viot processes [27,28].
106 Indeed, when lineages disappear backwardly during a recolonization event, multiple lineages will
107 merge into the same or more parental individuals depending on how many parental lineages are
108 responsible for the recolonization. When a parental lineage immigrates into a new area, the position
109 of the descendant coalescing lineage will be spatially tracked back to a different part of the lattice
110 corresponding to the origin of the parental lineage, such that the coalescing lineage is said to “jump”

111 [27]. Allowing for local bottlenecks and long distance jumps, the spatial- Λ -coalescent can recover
112 both small local N_e and long-distance correlated genealogies deriving from long distance dispersal
113 events [27,28]. Without needing to assume discrete demes or homogeneous population distribution,
114 this new framework has been shown to predict very well local and global N_e values when classic F_{ST}
115 measures otherwise largely uncorrelate to observed values [26-29].

116

117 **Available statistical tools based on MMCs**

118 Given the wide relevance of MMCs models to describe the demographic histories of natural
119 populations (e.g., SRS, bottlenecks, expansions, positive selection), it is important to compare the fit
120 of KC versus MMCs to describe a population demographic history, before a parameter of interest is
121 estimated from empirical genetic data. While in species with highly skewed reproductive success
122 MMCs can be assumed to outperform KC, in less trivial cases, e.g., human rapid population
123 expansion [14], a model comparison is needed to accept or reject KC.

124 At the state of the art, some MMCs maximum likelihood estimators have been developed and
125 are available to infer the effective population size and skewness of the offspring distribution of
126 marine species [11,25,30], such as Metagenetree [17] (Table 1). A recent software based on spatial- Λ -
127 coalescent (*phyrex*) by [29] estimates global N_e values in continuous space as an alternative to classic
128 F_{ST} estimates. Moreover, two MMCs simulators are currently available: algorithms by Kelleher et al
129 for continuous space evolution [29] and Hybrid-Lambda for species evolution [31], which could be
130 used to fit evolutionary hypotheses to observations using simulation approaches (Table1). Indeed,
131 Joseph et al 2016 [32] developed an ABC pipeline based on the simulator presented in [29] (Table1).
132 At the same time, empirical conservation biologists will benefit from being aware of the biological
133 relevance of MMCs and when and why they should be applied.

134

135 **Acknowledgements**

136 I am grateful to Mauricio Gonzalez-Forero, Jeffrey Jensen, Sebastian Matuszewski, Stefan
137 Laurent, Oscar Gaggiotti, Chiara Batini and two anonymous reviewers for helpful comments.

138

139 **References**

1. Wright S. Evolution in Mendelian populations. *Genetics*. 1931 Mar;16(2):97–159.
2. Luikart G, Ryman N, Tallmon DA, Schwartz MK, Allendorf FW. Estimation of census and effective population sizes: the increasing usefulness of DNA-based approaches. *Conserv Genet*. 2010 Feb 27;11(2):355–73.
3. Frankham R, Bradshaw CJA, Brook BW. Genetics in conservation management: Revised recommendations for the 50/500 rules, Red List criteria and population viability analyses. *Biol Conserv*. 2014 Feb;170:56–63.
4. Drummond AJ, Rambaut A, Shapiro B, Pybus OG. Bayesian Coalescent Inference of Past Population Dynamics from Molecular Sequences. *Mol Biol Evol*. 2005 May 1;22(5):1185–92.
5. Currat M, Ray N, Excoffier L. splatc: a program to simulate genetic diversity taking into account environmental heterogeneity. *Mol Ecol Notes*. 2004 Mar 1;4(1):139–42.
6. Hey J. Isolation with Migration Models for More Than Two Populations. *Mol Biol Evol*. 2010 Apr 1;27(4):905–20.
7. Gutenkunst RN, Hernandez RD, Williamson SH, Bustamante CD. Inferring the Joint Demographic History of Multiple Populations from Multidimensional SNP Frequency Data. *PLoS Genet*. 2009 Oct 23;5(10):e1000695.
8. Excoffier L, Dupanloup I, Huerta-Sánchez E, Sousa VC, Foll M. Robust Demographic Inference from Genomic and SNP Data. *PLoS Genet*. 2013 Oct 24;9(10):e1003905.
9. Kingman JFC. The coalescent. *Stoch Process Their Appl*. 1982 Sep;13(3):235–48.
10. Ewens WJ. *Mathematical Population Genetics*. New York, NY: Springer New York; 2004.
11. Eldon B, Wakeley J. Coalescent Processes When the Distribution of Offspring Number Among Individuals Is Highly Skewed. *Genetics*. 2006 Apr 1;172(4):2621–33.
12. Hedgecock D, Pudovkin AI. Sweepstakes Reproductive Success in Highly Fecund Marine Fish and Shellfish: A Review and Commentary. *Bull Mar Sci*. 2011 Oct 1;87(4):971–1002.

13. Rubenstein DR, Lovette IJ. Reproductive skew and selection on female ornamentation in social species. *Nature*. 2009 Dec 10;462(7274):786–9.
14. Bhaskar A, Clark AG, Song YS. Distortion of genealogical properties when the sample is very large. *Proc Natl Acad Sci*. 2014 Feb 11;111(6):2385–90.
15. Neher RA, Hallatschek O. Genealogies of rapidly adapting populations. *Proc Natl Acad Sci*. 2013 Jan 8;110(2):437–42.
16. Eldon B. Estimation of parameters in large offspring number models and ratios of coalescence times. *Theor Popul Biol*. 2011 Aug;80(1):16–28.
17. Birkner M, Blath J, Steinrücken M. Importance sampling for Lambda-coalescents in the infinitely many sites model. *Theor Popul Biol*. 2011 Jun;79(4):155–73.
18. Tellier A, Lemaire C. Coalescence 2.0: a multiple branching of recent theoretical developments and their applications. *Mol Ecol*. 2014 Jun;23(11):2637–52.
19. Der R, Epstein CL, Plotkin JB. Generalized population models and the nature of genetic drift. *Theor Popul Biol*. 2011 Sep;80(2):80–99.
20. Der R, Epstein C, Plotkin JB. Dynamics of Neutral and Selected Alleles When the Offspring Distribution Is Skewed. *Genetics*. 2012 Aug 1;191(4):1331–44.
21. Sargsyan O, Wakeley J. A coalescent process with simultaneous multiple mergers for approximating the gene genealogies of many marine organisms. *Theor Popul Biol*. 2008 Aug;74(1):104–14.
22. Felsenstein J. A Pain in the Torus: Some Difficulties with Models of Isolation by Distance. *Am Nat*. 1975 May 1;109(967):359–68.
23. Barton NH, Etheridge AM, Véber A. A New Model for Evolution in a Spatial Continuum. *Electron J Probab*. 2010 Feb 3;15(0).
24. McRae BH. Isolation by resistance. *Evol Int J Org Evol*. 2006 Aug;60(8):1551–61.

25. Dupas S, le Ru B, Branca A, Faure N, Gigot G, Campagne P, et al. Phylogeography in continuous space: coupling species distribution models and circuit theory to assess the effect of contiguous migration at different climatic periods on genetic differentiation in *Busseola fusca* (Lepidoptera: Noctuidae). *Mol Ecol*. 2014 May 1;23(9):2313–25.
26. Eldon B, Wakeley J. Coalescence Times and FST Under a Skewed Offspring Distribution Among Individuals in a Population. *Genetics*. 2009 Feb 1;181(2):615–29.
27. Barton NH, Etheridge AM, Kelleher J, Véber A. Inference in two dimensions: Allele frequencies versus lengths of shared sequence blocks. *Theor Popul Biol*. 2013 Aug;87:105–19.
28. Kelleher J, Barton NH, Etheridge AM. Coalescent simulation in continuous space. *Bioinformatics*. 2013 Apr 1;29(7):955–6.
29. Guindon S, Guo H, Welch D. Demographic inference under the coalescent in a spatial continuum. *bioRxiv*. 2016 Mar 2;042135.
30. Árnason E, Halldórsdóttir K. Nucleotide variation and balancing selection at the *Ckma* gene in Atlantic cod: analysis with multiple merger coalescent models. *PeerJ*. 2015 Feb 24;3:e786.
31. Zhu S, Degnan JH, Goldstien SJ, Eldon B. Hybrid-Lambda: simulation of multiple merger and Kingman gene genealogies in species networks and species trees. *BMC Bioinformatics*. 2015 16:292.
32. Joseph TA, Hickerson MJ, Alvarado-Serrano DF. Demographic inference under a spatially continuous coalescent model. *Heredity*. 2016 April doi: 10.1038/hdy.2016.28.

Figure 1. Examples of haploid genealogies presenting skewed reproductive success in forward and thus multiple merging in backward. Red edges indicate the sampled lineages. The yellow arrows represent the generation at which multiple merges occur and the blue arrows represent the generation at which the demographic event occurs. In A) SRS always leads to skewed offspring variance and thus multiple mergers can be observed at each generation, even when population size remains constant. In B) population expansion happens at the last generation with low reproductive variance and number of pre-capita offspring, hence the multiple mergers take place at the previous generation; in C) the population bottleneck and the multiple merging events occur at the same generation. In D) a selective sweep drags one genome to replace part of the population, thus the demographic event and the multiple merges co-occur.

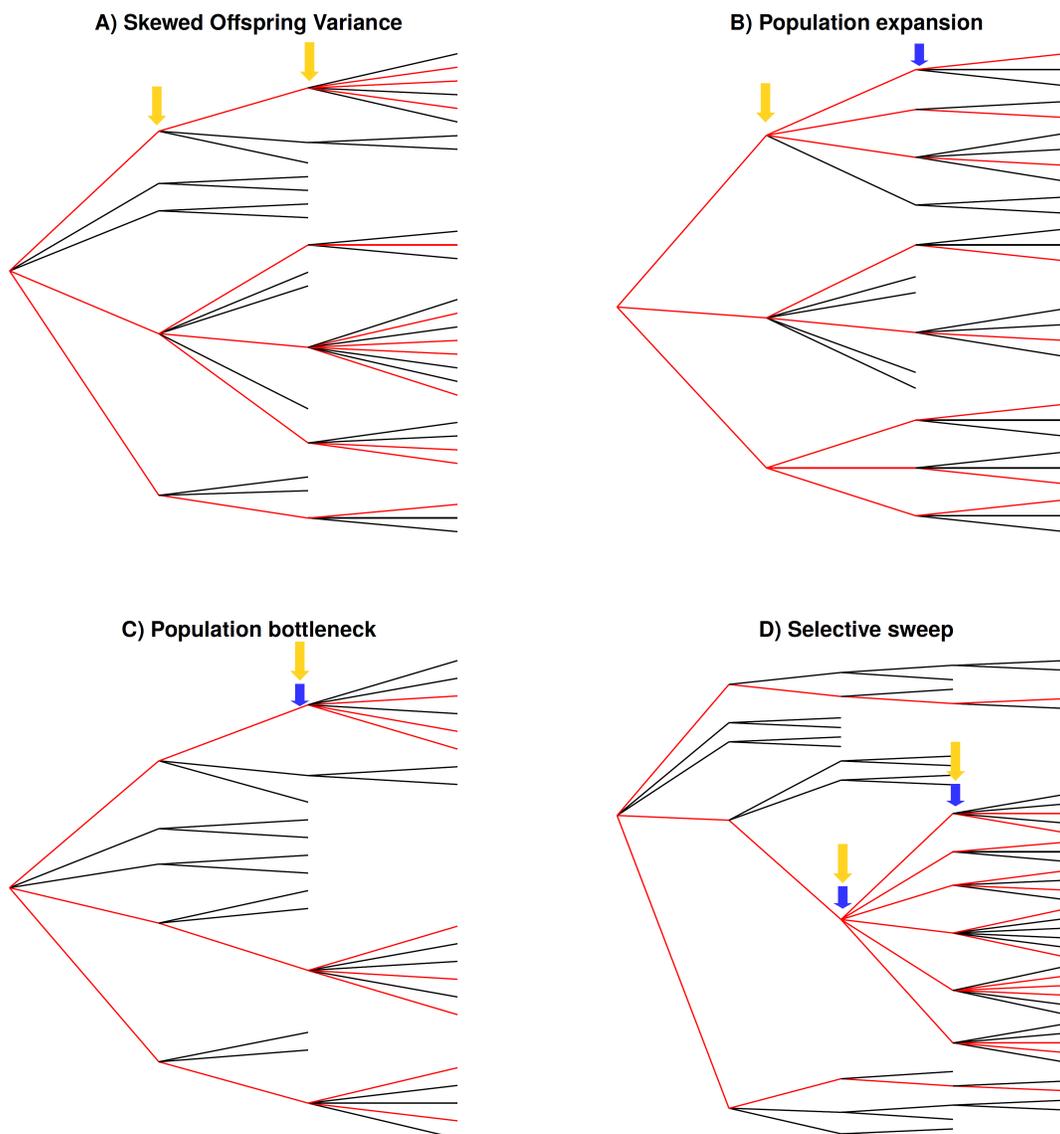


Table 1. Available statistical tools based on MMC models.

MMC tools					
<i>Name</i>	<i>Type</i>	<i>Model</i>	<i>Spatially explicit</i>	<i>Reference</i>	<i>Source</i>
Eldon & Wakeley	Estimator	Λ -coalescent	No	Eldon and Wakeley 2006	Available under request to authors
Metagenetree	Estimator	Λ -coalescent	No	Birkner et al 2011	http://metagenetree.sourceforge.net/
Phyrex	Estimator	Spatial- Λ -coalescent	Yes	Guindon et al 2016	https://github.com/stephaneguindon/phyml
Hybrid-Lambda	Simulator	B and Λ -coalescent	No	Zhu et al 2015	https://github.com/hybridLambda/hybrid-Lambda
ABC-Discsim	Simulator and Estimator	Spatial- Λ -coalescent	Yes	Kelleher et al 2014; Joseph et al 2016	https://github.com/tyjo/ABC-Discsim