On the importance of skewed offspring distributions and background selection in viral population genetics

Kristen K. Irwin[1,2], Stefan Laurent[1,2], Sebastian Matuszewski [1,2], Séverine Vuilleumier[1,2], Louise Ormond[1,2], Hyunjin Shim [1,2], Claudia Bank[1,2,3], and Jeffrey D. Jensen[1,2,4]

[1] – École Polytechnique Fédérale de Lausanne (EPFL), School of Life Sciences, Lausanne, Switzerland
[2] – Swiss Institute of Bioinformatics (SIB), Lausanne, Switzerland
[3] – Instituto Gulbenkian de Ciência (IGC), Oeiras, Portugal
[4] – Arizona State University (ASU), School of Life Sciences, Center for Evolution & Medicine, Phoenix, USA

Word Count: 3400

Keywords: Virus, Background Selection, Multiple Merger Coalescent, Skewed Offspring Distribution

Abstract

Many features of viral populations make them excellent candidates for population genetic study, including a very high rate of mutation, high levels of nucleotide diversity, exceptionally large census population sizes, and frequent positive selection. However, these attributes also mean that special care must be taken in population genetic inference. For example, highly skewed offspring distributions, frequent and severe population bottleneck events associated with infection and compartmentalization, and strong purifying selection all affect the distribution of genetic variation but are often not taken in to account. Here, we draw particular attention to multiple-merger coalescent events and background selection, discuss potential mis-inference associated with these processes, and highlight potential avenues for better incorporating them in to future population genetic analyses.

Introduction

Viruses appear to be excellent candidates for studying evolution in real time; they have short generation times, high levels of diversity often driven by very large mutation rates and population sizes, and they experience frequent positive selection in response to host immunity or antiviral treatment. However, despite these desired attributes, standard population genetic models must be used with caution when making evolutionary inference.

Firstly, population genetic inference is usually based on a coalescence model of the Kingman type, under which only two lineages may coalesce at a time. This assumption results in Poisson-shaped offspring distributions. In contrast, viruses have highly variable reproductive rates, taken as rates of replication; these may vary based on cell or tissue type, level of cellular differentiation, or stage in the lytic/lysogenic cycle (Knipe and Howley, 2007), resulting in highly skewed offspring distributions. This model violation is further intensified by the strong bottlenecks associated with infection and by strong positive selection (Neher and Hallatschek, 2013). Therefore, virus genealogies may be best characterized by *multiple merger* coalescent (MMC) models (e.g, (Eldon and Wakeley, 2008) instead of the Kingman coalescent.

Secondly, the mutation rates of many viruses, particularly RNA viruses, are among the highest observed across taxa (Lauring *et al*, 2013; Cuevas *et al*, 2015) Though these high rates of mutation are what enable new beneficial mutations to arise, potentially allowing for rapid resistance to host immunity or antiviral drugs, they also render high mutational loads (Sanjuán, 2010; Lauring *et al*, 2013). Specifically, the distribution of fitness effects (DFE) has now been described across taxa – demonstrating that the input of deleterious mutations far outnumbers the input of beneficial mutations (Acevedo *et al*, 2014; Bank *et al*, 2014; Bernet and Elena, 2015; Jiang *et al*, 2016). The purging of these deleterious mutants through purifying selection can affect other areas in the genome through a process known as background selection (BGS)

(Charlesworth *et al*, 1993). Accounting for these effects is important for accurate evolutionary inference in general (Ewing and Jensen, 2016), but essential for the study of viruses due to their particularly high rates of mutation and compact genomes (Renzette *et al*, 2016).

Given these distinctive features of viral populations and the increasing use of population genetic inference in this area, it is crucial to account for these processes that are shaping the amount and distribution of variation across their genomes. We aim here to draw particular attention to multiple-merger coalescent events and background selection, and the repercussions of ignoring them in population genetic inference, highlighting particular applications to viruses. We conclude with general recommendations for how best to address these topics in the future.

Skewed Offspring Distributions and the Multiple Merger Coalescent

Inferring evolutionary history using the Wright-Fisher model: benefits and shortcomings

Many population genetic statistics and subsequent inference are based on the Kingman coalescent and the Wright-Fisher (WF) model (Wright, 1931; Kingman, 1982). With increasing computational power, the WF model has also been implemented in forward-time methods for inferring population genetic parameters such as selection coefficients and effective population sizes (N_e) from time-sampled data (Ewens, 1979; Williamson and Slatkin, 1999; Malaspinas *et al*, 2012; Foll *et al*, 2014; Foll *et al*, 2015; Ferrer-Admetlla *et al*, *in press*). These methods are robust to some violations of WF model assumptions, such as constant population size, panmixia, and non-overlapping generations, and also have been extended to accommodate selection, migration and population structure (Neuhauser and Krone, 1997; Nordborg, 1997).

However, it has been suggested that violation of the assumption of a small variance in offspring number in the WF model leads to erroneous inference of population genetic parameters (Eldon and Wakeley, 2006). Biological factors such as sweepstake reproductive events, population bottlenecks, and recurrent positive selection may lead to skewed distributions in offspring number (Eldon and Wakeley, 2006; Li *et al*, 2014); examples include various prokaryotes, fungi, plants, marine organisms, and viruses (reviewed Tellier and Lemaire, 2014). These skewed offspring distributions can also result in elevated linkage disequilibrium (LD) despite frequent recombination (Eldon and Wakeley, 2008; Birkner *et al*, 2013). They may also skew estimates of F_{ST} from those expected under WF models, as there is a high probability of alleles being *identical by descent* in subpopulations, where the expectation of coalescent times within populations is less than that between populations

regardless of the timescale or magnitude of gene flow (Eldon and Wakeley, 2009).

Beyond WF assumptions: the Multiple Merger Coalescent

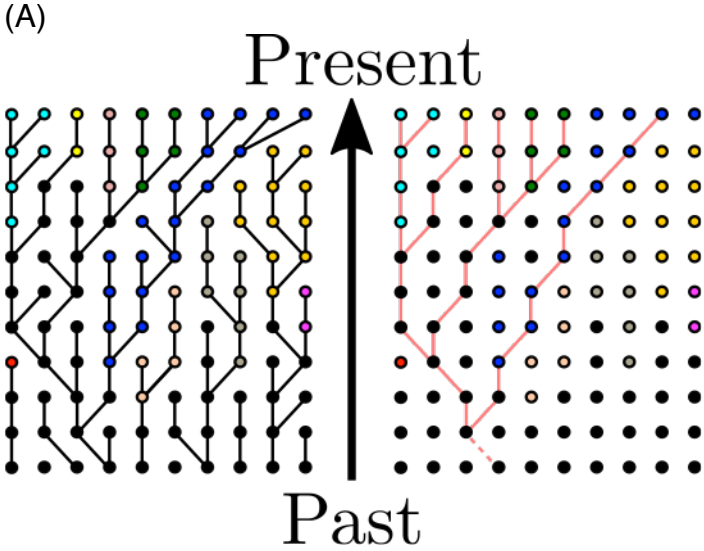
A more general coalescent class of models, summarized as the MMC class, can account for these violations, particularly for (non-Poisson) skewed offspring distributions, by allowing more than two lineages to coalesce at a time (Table 1). These are often derived from Moran models, (Moran, 1958), generalized to allow multiple offspring per individual. In contrast to the WF model (for which $P(k > 2) = 0$), a probability distribution for k-merger events determines coalescence.

The parameters inferred under the MMC differ notably from those inferred under the Kingman coalescent in several respects. In a Kingman coalescent, effective size N_e scales linearly with census size N, whereas for the MMC it does not (Huillet and Möhle, 2011). Thus genetic diversity is only weakly related to population size. Coalescent trees under the MMC also have more pronounced star-like genealogies (Figure 1), and their site frequency spectra (SFSs) are skewed toward an excess of low frequency and high frequency variants, generating a more negative Tajima's D (Birkner *et al*, 2013). With similar migration and population size, alleles fix at a higher rate per population in the MMC than under the Kingman coalescent, and thus higher F_{ST} is expected between subpopulations (Eldon and Wakeley, 2009). Further, the efficacy of selection increases, as selection acts almost deterministically between multiple merger events (Tellier and Lemaire, 2014). Theoretical analyses demonstrate that the fixation probability of a new mutant with a positive selection coefficient approaches 1 as the population size increases, in stark contrast with traditional expectations under the standard Wright-Fisher model (Der *et al*, 2011).

Not accounting for skewed offspring distributions can lead to mis-inference. For instance, Eldon and Wakeley (2006) showed that for Pacific oysters, which have been shown to occasionally undergo sweepstake-like reproductive events, the estimated population-wide mutation rate θ inferred under the Kingman coalescent is two orders of magnitude larger than that obtained from the ψ-coalescent (see below) - 9 vs 0.0308, respectively - and, indeed, does not well fit the data.

Figure 1: Multiple-Merger and Kingman Coalescent Realizations

(A)



(B)

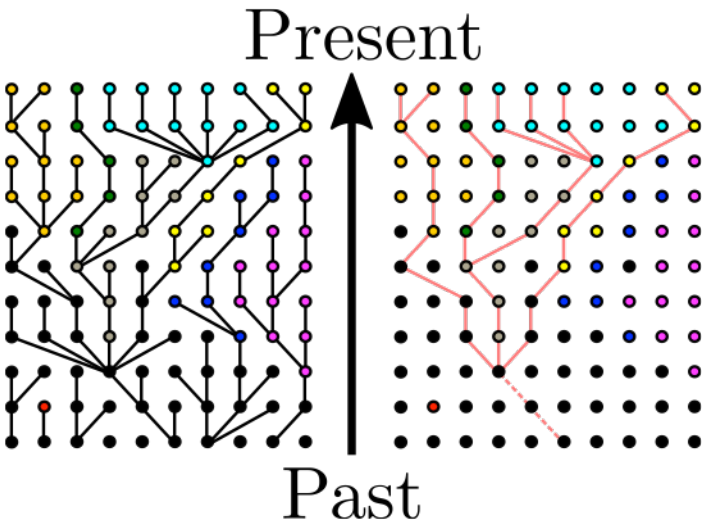


Figure 1: Example genealogies and samples from (A) the Kingman coalescent and (B) a multiple-merger coalescent. Panels on the left show the evolutionary process of the whole population, whereas those on the right show a possible sampling and its resulting genealogy. Colors correspond to different (neutral) derived allelic states, where black denotes the wild type.

172 Table 1: Hierarchy of coalescent models, in decreasing order of generality

173

Coalescent model	Allows MMs?	Allows simult-aneous MMs?	Distribution and parameters	References
Ξ-coalescent	Yes	Yes	MMC events occur with the probability Λ, where participating lineages are randomly grouped into M simultaneous mergers with probability 1/M	Schweinsberg (2000); Möhle and Sagitov (2001)
Λ-coalescent	Yes	No	MMC events occur with the probability Λ (but ≤1 event/gen)	Donnelly and Kurtz (1999); Pitman (1999); Sagitov (1999)
ψ-coalescent	Yes	No	Λ follows a distribution where a fraction Ψ of the population is replaced by the offspring of a single individual	Eldon and Wakeley (2006); Eldon and Wakeley (2008); Eldon and Wakeley (2009); Eldon and Degnan (2012)
β-coalescent	Yes	No	Λ follows β-distribution: beta(α, 2-α) with $1 \leq \alpha < 2$	Schweinsberg (2003); Berestycki *et al* (2007); Berestycki *et al* (2008); Birkner and Blath (2008); Birkner *et al* (2013); Steinrücken *et al* (2013)
Bolthausen-Sznitman	Yes	No	Λ follows β-distribution with α=1: beta(1,1) = uniform on [0,1]	Bolthausen and Snznitman (1998); Basdevant and Goldschmidt (2008); Neher and Hallatschek (2013)
Kingman coalescent	No	No	Λ follows β-distribution with α=2: only two lineages are allowed to merge at a time	Kingman (1982)

174

The ψ-coalescent

Introduced by Eldon and Wakeley (2006), the ψ-coalescent (also called the 'Dirac-coalescent') differentiates two possible reproductive events (Figure 2) (Eldon *et al*, 2015) . Either a standard Moran model reproduction event occurs (with probability 1-ε), where a single individual is randomly chosen to reproduce and the (single) offspring replaces one randomly chosen non-parental individual; all other individuals, including the parent, persist. Alternatively, a 'sweepstake' reproductive event occurs (with probability ε) (Hedgecock, 1994)), where a single parent replaces ψ*N individuals. Consequently, an individual may have many offspring and the underlying coalescent process will be characterized by MM events, or in the case of the Ξ-coalescent, multiple MM events. However, in contrast to other MMC models (*e.g.*, λ- or Ξ-coalescent), the parameter ψ has a clear biological interpretation as the fraction of the population that is replaced in each sweepstake reproductive event. Though the assumption of a fixed ψ (as in the normal ψ-coalescent) seems biologically unrealistic, it can be avoided by treating ψ as a Poisson parameter.

Figure 2: Depiction of the ψ coalescent

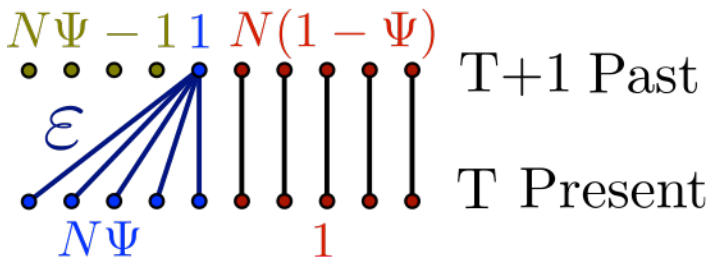


Figure 2: Lineages between the present and one generation in the past, where *N* is the population size, ε is the probability of a sweepstake event, and ψ is the fraction of the population that is replaced in each such event. Labels in the top row give the number of parental individuals reproducing in a given manner (represented by color), whereas labels in the bottom row give the number of corresponding offspring per parent. Note that time is running backwards in the coalescent framework.

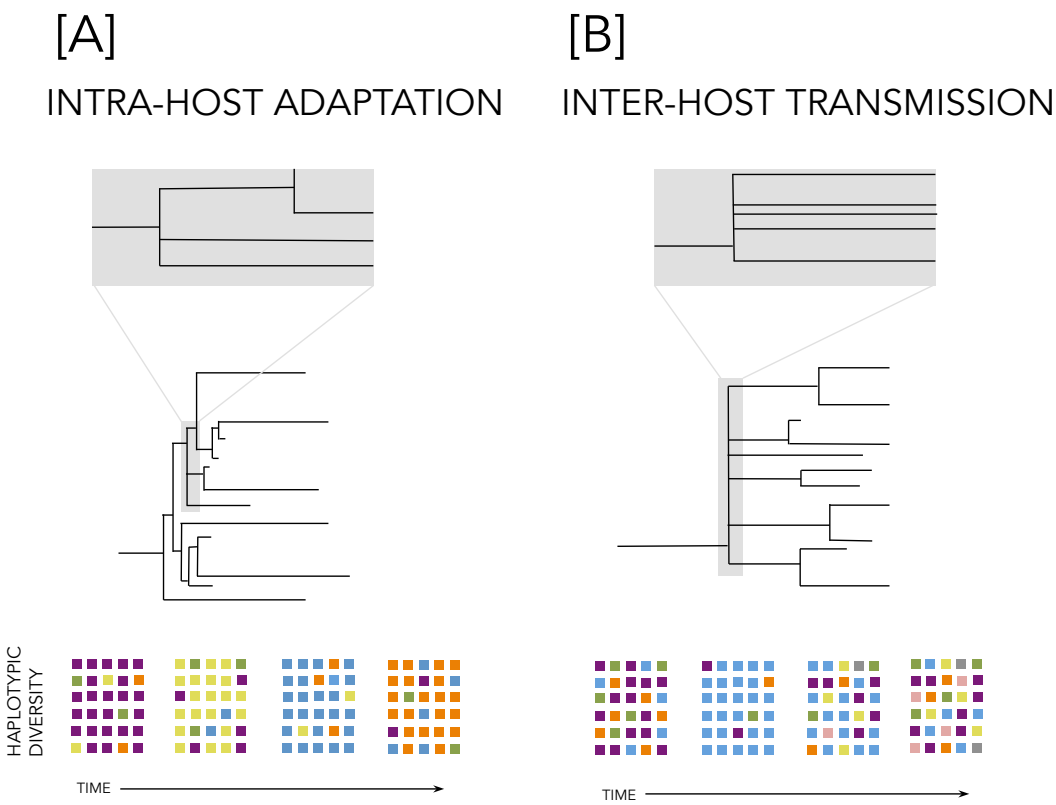
Application to Viruses

There are several reasons why a modified Moran model may better capture viral evolution than models converging to the Kingman coalescent, although it does not account for fitness differences between individuals. First, viral evolution is driven by strong bottlenecks during host transmission and intrahost selection processes, which likely result in skewed offspring distributions (Figure 3) (Gutiérrez *et al*, 2012; Tellier and Lemaire, 2014). Further, viruses display the MMC-typical low N_e/N ratio (Pennings *et al*, 2014; Tellier and Lemaire, 2014), can adapt rapidly (Neher and Hallatschek, 2013), and may have sweepstake-like reproductive events in which a single virion can propagate a large fraction of the entire population (Grenfell *et al*, 2004; Pybus and Rambaut, 2009). For example, the influenza virus hemagglutinin (HA) segment appears to be under strong directional selection imposed by host immunity (and sometimes drug treatment), resulting in a ladder-like genealogy (Grenfell *et al*, 2004), suggesting that only a few viruses seed the entire next generation.

The processes that make viruses ideal candidates for MMCs can, however, differ by scale (see Figure 3); for example, following transmission events, there are severe founder events and potentially high recombination within the host (*e.g.*, HIV, HCMV). Subsequent compartmentalization may introduce intra-host population structure through bottlenecks, colonization events, and extinction events (Renzette *et al*, 2013). To date, it remains unclear how often MMCs fit the patterns of variation observed in intra-host relative to inter-host viral populations – but such comparisons are increasingly feasible. Finally, periods of latency - temporary viral inactivation with cessation of reproduction - should be incorporated in such modeling. Thus, multiple MMC models are a necessary but not final step towards addressing the various patterns observed at different scales of virus evolution (Table 1).

The large data sets often generated from viruses may also prove impractical for the likelihood-based methods commonly employed for MMCs. This limitation has partially been overcome by Eldon *et al.* (2015), who proposed an approximate likelihood method along with an Approximate Bayesian Computation (ABC) approach based on the SFS to distinguish between the MMC and exponential population growth. Although both effects are expected to result in very similar SFSs, characterized by an excess of singletons as compared to the Kingman coalescent, the bulk and tail of the SFS (i.e., the higher-order frequency classes) typically differ, which can be assessed by approximate likelihood-ratio tests and Approximate Bayes Factors (Eldon *et al*, 2015).

249 Figure 3: Example Processes Spurring MM Events in Viral Populations



250

251 Figure 3: Examples include (A) intra-host adaptation (a selective process) and
252 (B) inter-host transmission (a demographic process). The tree in (A)
253 characterizes, for example, NA or HA evolution in the influenza A virus, driven
254 by positive selection; selection by host immunity is ongoing, while that from
255 drug treatment may be intermittent. The tree in (B) represents inter-host
256 transmission and its associated bottleneck; for viruses that compartmentalize
257 (such as HCMV and HIV), similar patterns follow transmission to new
258 compartments. The colored squares below the trees roughly indicate the
259 diversity of the population through time. Intra-host adaptation may temporally
260 decrease diversity owing to genetic hitchhiking, though single snapshots may
261 not reflect varying temporal levels of diversity. During inter-host transmission,
262 diversity decreases owing to the associated bottleneck but then may quickly
263 recover in the new host.

264

[BOX 1: Future challenges in MMC models]

In order to make MMC models biologically relevant for viruses, a number of important tasks remain:

1. Describe summary statistics that capture demographic features and processes when offspring distributions are highly skewed; such patterns will be required for large-scale inference in a computationally efficient (e.g., Approximate Bayesian) framework.
2. Better understand the behavior of commonly used summary statistics under such models, as done for F_{ST} by Eldon and Wakeley (2009), for commonly used divergence, SFS, and LD-based statistics.
3. Determine which MMCs are best suited for different scales of viral evolution (i.e., intra-host, inter-host, global); develop novel models if necessary.
4. Investigate the effect of violations of MMC assumptions (e.g., overlapping generations, number of multiple merger events) on inference.

[END BOX 1]

Purifying Selection and Linkage in Viral Populations

Modeling Background Selection

The joint modeling of the effects of genetic drift and positive selection, including in experimental evolution studies of viral populations, has improved our ability to distinguish adaptive from neutral mutations by minimizing the chance that the rapid fixation of a neutral allele is incorrectly interpreted as strong positive selection (Li *et al*, 2012; Foll *et al*, 2014). However, there is another process that must be incorporated if we are to fully understand mutation trajectories in viral populations: background selection (BGS).

BGS was originally proposed to explain patterns of reduced diversity in regions of low recombination – patterns that were previously suggested to be the signature of genetic hitchhiking (HH) around strongly beneficial mutations (see Begun and Aquadro 1992 and Charlesworth *et al* 1993). It was argued that only neutral mutations present on the "least-loaded" chromosomes – that is, those with the fewest deleterious mutations – have appreciable probabilities of reaching high frequencies or fixation. Kimura and Maruyama (1966) showed that the proportion of individuals belonging to the least-loaded class is

$$f_0 = \exp\left(-\frac{U}{2hs}\right) \quad , \tag{1}$$

where U is the rate of mutation to a deleterious state, s is the selection coefficient against homozygous mutations, and h is the dominance coefficient.

The least-loaded class, and thus genetic diversity in the presence of BGS, is therefore dependent on the balance between the influx of deleterious mutations (occurring at rate U) and their removal by natural selection (according to the product hs). Assuming that offspring exclusively originate

from the least-loaded class of individuals, Charlesworth *et al* (1993) expressed the expected neutral diversity due to background selection as

$$\pi = 4\, f_o N_e\, \mu\ , \tag{2}$$

where N_e is the effective population size and μ is the mutation rate. As BGS reduces the number of reproducing individuals, genetic drift increases, thus reducing genetic diversity and increasing stochasticity in allele trajectories. Further, since only the genetic diversity segregating in the least-loaded class can be observed, population size inferred from measures of genetic diversity may be underestimated if BGS is not properly taken into account (Ewing and Jensen, 2016).

In the BGS model described above, strongly deleterious mutations are maintained in mutation-selection balance such that no skew in the SFS is expected, as rare variants are rapidly purged. Thus, a simple re-scaling of N_e is often used as a proxy for the effects of BGS (*e.g.*, Hudson and Kaplan, 1995; Zeng and Charlesworth, 2011; Prüfer *et al*, 2012; Zeng, 2013). However, recent work has demonstrated that, while this re-scaling is appropriate for strongly deleterious mutations, it is largely inappropriate for weakly deleterious mutations that may segregate in the population. Experimental work on the shape of the distribution of fitness effects (DFE) in many organisms indicates that weakly deleterious mutations represent an important class (*e.g.*, Eyre-Walker and Keigntley, 2007; Bank *et al*, 2014). These mutations may act to skew the SFS towards rare alleles as they decrease the expected frequency of linked neutral mutations relative to neutral expectations. As subsequent demographic inference is based on the shape of this SFS, this effect should be properly accounted for by directly simulating weakly deleterious mutations rather than implementing a simple rescaling, as is common practice. Figure 4 shows the skew in estimates of population size and migration rates obtained using an ABC approach when

BGS is prevalent for two populations A and B that have split at time τ=2N_e generations (reproduced from Ewing and Jensen, 2016).

Figure 4: Bias in parameter inference at intermediate levels of BGS

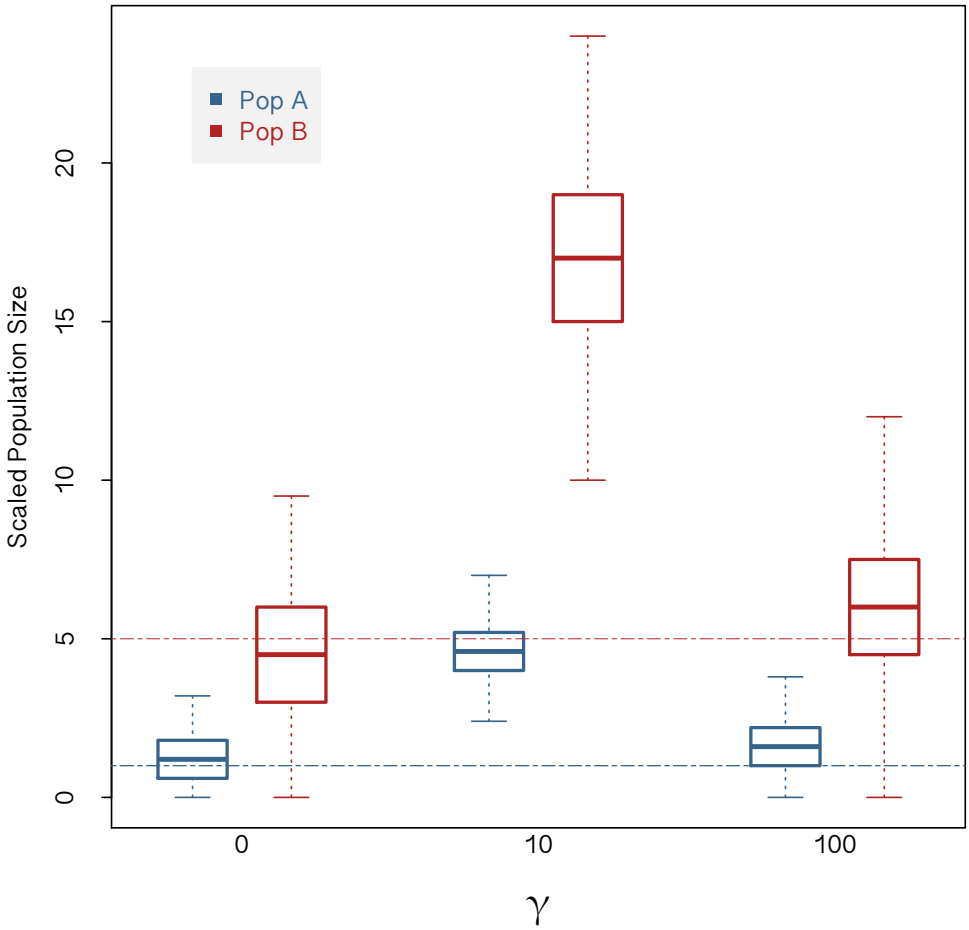


Figure 4: Bias in parameter inference for different levels of BGS, redrawn from Ewing & Jensen (2016). Posterior densities from ABC inference for population size are shown. Selection strength is given as γ, where $\gamma = 2N_e s$. Population A has a true scaled size of 1 (blue line), and population B a true scaled size of 5 (red line). As shown, the greatest mis-inference occurs in the presence of weakly deleterious mutations.

The Effects of Background Selection on Inference for Viral Populations

Efforts to estimate the impact of BGS in non-viral organisms have been well reported. One of the most notable examples is that of Comeron (2014), who estimated levels of BGS in *Drosophila melanogaster* based on the results of Hudson and Kaplan (1995) and Nordborg *et al* (1996) using a high-definition recombination map, with results indicating strong effects across the genome. For viruses, similar efforts are in their infancy, with the first attempt at such estimation in a virus reported recently by Renzette *et al* (2016), where they utilized the theoretical predictions of Innan and Stephan (2003). Interestingly, the full spectrum of recombination frequencies is available in viral systems – from non-recombining (*e.g.,* most negative-sense RNA viruses), to re-assorting (*e.g.*, Influenza virus), to rarely recombining (*e.g.,* Hepatitis C and West Nile viruses), to frequently recombining (*e.g.*, HIV), offering a highly promising framework for comparative analyses investigating the pervasiveness of BGS effects (Chare *et al*, 2003; Simon-Loriere and Holmes, 2011). Further, given the high mutation rates and compact genomes of many viruses, evolutionary theory suggests effects at least equal to those seen in *Drosophila*.

In order to accomplish such inference, improved recombination maps for viral genomes will be important. With such maps in hand, and given the amenability of viruses to experimental perturbation, it may indeed be feasible to understand and account for BGS in models of viral evolution.

[BOX 2: Future challenges in identifying the effects of BGS]

As BGS almost certainly impacts inference in viral populations, accounting for its effects is critical. Some future challenges include:

1. Accounting for BGS effects on the SFS by directly simulating weakly deleterious mutations, rather than by rescaling N_e.
2. Improving recombination maps for viral genomes.
3. Developing models combining the effects of non-equilibrium demography, positive selection, and BGS, ideally to allow for the joint estimation of all associated parameters.
4. Extending methods applied to other taxa to virus populations; for example, establishing a baseline of variation for use as a null expectation to estimate BGS levels across the genome, as done for *Drosophila*.

[END BOX 2]

Future Directions

Given that skewed offspring distributions and pervasive linked selection are likely important factors influencing the inference of viral population parameters, it is important to note that multiple backward and forward simulation programs have recently been developed which make the modeling of these processes feasible (Hernandez, 2008; Messer, 2013; Thornton, 2014; Eldon *et al*, 2015; Zhu *et al*, 2015). This will allow researchers to directly simulate from parameter ranges that may be relevant for their population of interest, developing a better intuition for the importance of these

processes in shaping observed genomic diversity. More concretely, the ability to now simulate in a computationally efficient framework opens the possibility of directly implementing ABC inference approaches under these models. Thus, by drawing mutations from a biologically realistic distribution of fitness effects and allowing offspring distributions to appropriately vary, it is now possible to re-implement common demographic estimation or genome scan approaches; such modified approaches would be based on more appropriate null expectations of the shape of the SFS, the extent of linkage disequilibrium, and the degree of population divergence.

Acknowledgements

We would like to thank Bjarki Eldon for helpful suggestions during the early stages of this manuscript. This work was funded by a European Research Council (ERC) Starting Grant to JDJ, as well as Swiss National Science Foundation (FNS) grants to JDJ (31003A_159835) and SV (PMPDP3_158381).

Conflict of Interest

The authors declare no conflict of interest.

Data Archiving

As a review article, no new data was processed, analyzed, or used directly.

References

Acevedo A, Brodsky L, Andino R (2014). Mutational and fitness landscapes of an RNA virus revealed through population sequencing. *Nature* **505:** 686-690.

Bank C, Hietpas RT, Wong A, Bolon DN, Jensen JD (2014). A Bayesian MCMC approach to assess the complete distribution of fitness effects of new mutations: Uncovering the potential for adaptive walks in challenging environments. *Genetics* **196:** 841-852.

Basdevant A, Goldschmidt C (2008). Asymptotics of the allele frequency spectrum associated with the Bolthausen-Sznitmzn coalescent. *Electronic Journal of Probability* **13**(17)**:** 486-512.

Begun DJ, Aquadro CF (1992). Levels of naturally occurring DNA polymorphism correlate with recombination rates in *D. melanogaster*. *Nature* **356:** 519-520.

Berestycki J, Berestycki N, Schweinsberg J (2007). Beta-coalescents and cintinuous stable random trees. *The Annals of Probability* **35**(5)**:** 1835-1887.

Berestycki J, Berestycki N, Schweinsberg J (2008). Small-time behavior of beta coalescents. *Annales de l'Institut Henri Poincaré - Probabilitiés et Stastiques* **44**(2)**:** 214-238.

Bernet GP, Elena SF (2015). Distribution of mutational fitness effects and of epistasis in the 5' untranslated region of a plant RNA virus. *BMC Evolutionary Biology* **15:** 274-287.

Birkner M, Blath J (2008). Computing likelihoods for coalescents with multiple collisions in the infinitely many sites model. *Journal of Mathematical Biology* **57**(3)**:** 435-465.

Birkner M, Blath J, Eldon B (2013). An ancestral recombination graph for diploid populations with skewed offspring distribution. *Genetics* **193:** 255-290.

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

Chare ER, Gould EA, Holmes EC (2003). Phylogenetic analysis reveals a low rate of homologous recombination in negative-sense RNA viruses. *Journal of General Virology* **84:** 2691-2703.

Charlesworth B, Morgan MT, Charlesworth D (1993). The effect of deleterious mutations on neutral molecular variation. *Genetics* **134:** 1289-1303.

Comeron JM (2014). Background selection as a baseline for nucleotide variation across the *Drosophila* genome. *PLoS Genetics* **10**(6)**:** e1004434.

Cuevas JM, Geller R, Garijo R, López-Aldeguer J, Sanjuán R (2015). Extremely high mutation rate of HIV-1 in vivo. *PLoS Biology* **13**(9)**:** e1002251.

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

Donnelly P, Kurtz TG (1999). Particle representations for measure-valued population models. *The Annals of Probability* **27**(1)**:** 166-205.

Eldon B, Birkner M, Blath J, Freund F (2015). Can the site-frequency spectrum distinguish exponential population growth from multiple-merger coalescents? *Genetics* **199:** 841-856.

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). Coalesent processes when the distribution of offspring number among individuals is highly skewed. *Genetics* **172:** 2621-2633.

Eldon B, Wakeley J (2008). Linkage disequlibrium 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 distirbution among individuals in a population. *Genetics* **181:** 615-629.

Ewens WJ (1979). Testing the generalized neutrality hypothesis. *Theoretical Population Biology* **15**(2)**:** 205-216.

Ewing GB, Jensen JD (2016). The consequences of not accounting for background selection in demographic inference. *Molecular Ecology* **25:** 135-141.

Eyre-Walker A, Keightley PD (2007). The distirbution of fitness effects of new mutations. *Nature Reviews Genetics* **8:** 610-618.

Ferrer-Admetlla A, Leuenberger C, Jensen JD, Wegmann D (*in press*). An Approximate Markov Model for the Wright-Fisher Diffusion and its Application to Time Series Data. *Genetics*.

539 Foll M, Poh Y, Renzette N, Ferrer-Admetlla A, Bank C, Shim H *et al* (2014).
540 Influenza virus drug resistance: a time-sampled population genetic
541 perspective. *PLoS Genetics* **10**(2)**:** e1004185.
542
543 Foll M, Shim H, Jensen JD (2015). WFABC: a Wright-Fisher ABC-based
544 approach for inferring effective population sizes and selection coefficients
545 from time-sampled data. *Molecular Ecology Resources* **15**(1)**:** 87-98.
546
547 Grenfell BT, Pybus OG, Gog JR, Wood JLN, Daly JM, Mumford JA *et al*
548 (2004). Unifying the epidemiological and evolutionary dynamics of pathogens.
549 *Science* **303:** 327-332.
550
551 Gutiérrez S, Michalakis Y, Blanc S (2012). VIrus population bottlenecks during
552 within-host progression and host-to-host transmission. *Current Opinion in*
553 *Virology* **2:** 546-555.
554
555 Hedgecock D (1994). Population genetics of marine organisms. *US Globec*
556 *News* **6**(11)**:** 1-8.
557
558 Hernandez R (2008). A flexible forward simulator for populations subject to
559 selection and demography. *Bioinformatics* **24**(23)**:** 2786-2787.
560
561 Hudson RR, Kaplan NL (1995). Deleterious background selection with
562 recombination. *Genetics* **141:** 1605-1617.
563
564 Huillet T, Möhle M (2011). Population genetics models with skewed fertilities:
565 a forward and backward analysis. *Stochastic Models* **27:** 521-554.
566
567 Innan H, Stephan W (2003). Distinguishing the hitchhiking and background
568 selection models. *Genetics* **165:** 2307-2312.
569
570 Jiang L, Liu P, Bank C, Renzette N, Prachanronarong K, Yilmaz LS *et al*
571 (2016). A balance between inhibitor binding and substrate processing confers
572 influenza drug resistance. *Journal of Molecular Biology* **428:** 538-523.
573
574 Kimura M, Maruyama T (1966). The mutational load with epistatic gene
575 interactions in fitness. *Genetics* **54**(6)**:** 1337-1351.
576
577 Kingman JFC (1982). The coalescent. *Stochastic Processes and their*
578 *Applications* **13:** 235-248.
579
580 Knipe DM, Howley PM (2007). *Fields Virology*, Vol 1. Lippincott Williams &
581 Wilkins: Philadelphia.
582
583 Lauring AS, Frydman J, Andino R (2013). The role of mutational robustness in
584 RNA virus evolution. *Nature Reviews Genetics* **11:** 327-336.
585

Li J, Li H, Jakobsson M, Li S, Sjödin P, Lascoux M (2012). Joint analysis of demography and selection in poulation genetics: where do we stand and where could we go? *Molecular Ecology* **21:** 28-44.

Li LM, Grassly NC, Fraser C (2014). Genomic analysis of emerging pathogens: methods, application and future trends. *Genome Biology* **15:** 541-550.

Malaspinas A, Malaspinas O, Evans SN, Slatkin M (2012). Estimating allele age and selection coefficient from time-serial data. *Genetics* **192:** 599-607.

Messer PW (2013). SLiM: Simulating evolution with selection and linkage. *Genetics* **194:** 1037-1039.

Möhle M, Sagitov S (2001). A classification of coalescent processes for haploid exchangable population models. *The Annals of Probability* **29**(4)**:** 1547-1562.

Moran PAP (1958). Random processes in genetics. *Mathematical Proceedings of the Cambridge Philosophical Society* **54**(1)**:** 60-71.

Neher RA, Hallatschek O (2013). Genealogies of rapidly adapting populations. *Proceedings of the National Academy of Sciences* **110**(2)**:** 437-442.

Neuhauser C, Krone SM (1997). The genealogy of samples in models with selection. *Genetics* **145:** 519-534.

Nordborg M (1997). Structured coalescent processes on different time scales. *Genetics* **146:** 1501-1514.

Nordborg M, Charlesworth B, Charlesworth D (1996). The effect of recombination on background selection. *Genetical Reserach* **67**(2)**:** 159-174.

Pennings PS, Kryazhimskiy S, Wakeley J (2014). Loss and recovery of genetic diversity in adapting populations of HIV. *PLoS Genetics* **10**(1)**:** e1004000.

Pitman J (1999). Coalescents with multiple collisions. *Journal of Applied Probability* **27:** 1870-1902.

Prüfer K, Munch K, Hellmann I, Akagi K, Miller JR, Walenz B *et al* (2012). The bonobo genome compared with the chimpanzee and human genomes. *Nature* **486:** 527-531.

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

Renzette N, GIbson L, Bhattacharjee B, Fisher D, Schleiss MR, Jensen JD *et al* (2013). Rapid intrahost evolution of human cytomegalovirus is shaped by demography and positive selection. *PLoS Genetics* **9**(9)**:** e1003735.

Renzette N, Kowalik TF, Jensen JD (2016). On the relative rules of background selection and geneic hitchhiking in shaping human cytometgalovirus genetic diversity. *Molecular Ecology* **25**(1)**:** 403-413.

Sagitov S (1999). The general coalescent with asynchronous mergers of ancestral lines. *Journal of Applied Probability* **36:** 1116-1125.

Sanjuán R (2010). Mutational fitness effects in RNA and single-stranded DNA viruses: common patterns revealed by site-directed mutagenesis studies. *Philosophical Transactions of the Royal Society B* **365:** 1975-1982.

Schweinsberg J (2000). Coalescents with simultaneous multiple collisions. *Electronic Journal of Probability* **5**(12)**:** 1-50.

Schweinsberg J (2003). Coalescent processes obtained from supercritical Galton-Watson processes. *Stochastic processes and their Applications* **106:** 107-139.

Simon-Loriere E, Holmes EC (2011). Why do RNA viruses recombine? *Nature Reviews Microbiology* **9:** 617-626.

Steinrücken M, Birkner M, Blath J (2013). Analysis of DNA sequence variation within marine species using Beta-coalescents. *Theoretical Population Biology* **87:** 15-24.

Tellier A, Lemaire C (2014). Coalescence 2.0: a multiple branching of recent theoretical developments and their applications. *Molecular Ecology* **23:** 2637-2652.

Thornton KR (2014). A C++ template library for efficient forward-time population genetic simulation of large populations. *Genetics* **198:** 157-166.

Williamson EG, Slatkin M (1999). Using maximum likelihood to estimate population size from temporal change in allele frequencies. *Genetics* **152:** 755-761.

Wright S (1931). Evolution in Mendelian populations. *Genetics* **16:** 97-159.

Zeng K (2013). A coalescent model of background selection with recombination, demography and variation in selection coefficients. *Heredity* **100:** 363-371.

Zeng K, Charlesworth B (2011). The joint effects of background selection and genetic recombination on local gene genealogies. *Genetics* **189:** 251-266.

Zhu S, Degnan JH, Goldstien SJ, Eldon B (2015). Hybrid-Lambda: simulation of multiple merger and Kingman gene genealogies in species networks and species trees. *BMC Bioinformatics* **16:** 292-298.