An underlying assumption of the coalescent model is that the sample size is much smaller than the overall population size. When this assumption does not hold, the simplifying assumptions that there is at most one coalescent event per generation and that no more than two samples coalesce to the same common ancestor in a given event become unrealistic. Wakeley and Takahashi (2003) performed a thorough study of how violations of the assumption \( n \ll N \) alter a coalescent sample’s frequency spectrum and conclude that while \( n \leq N \) the effects are “surprisingly mild”. Of concern in our work is that FTE growth is so rapid that looking backwards in time the underlying population shrinks at a faster rate than our sample, resulting in a situation where the sample size and population size are of comparable magnitude and consequently the simplifying assumptions of the basic coalescent no longer hold.

To assess this possibility, we perform a series of simulations with a sample size of 20,000 haplotypes, a current population size of 8,000,000 haplotypes and an ancestral size of 20,000 haplotypes during which we calculate the ratio of sample size to average overall population size during growth. We find that for \( \beta \) between 0.1 and 3.5, the range investigated for this study, the average sample size to population size ratio does not exceed 0.177 (Figure S1). In particular, for \( \beta \) values close to exponential growth, those between 0.5 and 1.5, the ratio does not exceed 0.05. In light of this and the values observed in our models, our work does not appear to be violating simplifying assumptions of the basic coalescent substantially, and using our growth model within this framework is reasonable. This finding is in line with the work of Bhaskar et al. (2013), who compared the results from the coalescent to a discrete-time Wright Fisher model which allowed multiple- and simultaneous-mergers. In their work they found that under the demographic model of Tennessen et al (2012), with recent accelerating growth to a current diploid population size of \( \sim 10^6 \), the quantities of singletons and doubletons in a sample size of 10,000 haploids differed between models by <0.3%.

In our paper, we also present a constant population size model and a model of instantaneous growth as bounds on many of the results. While they are not present on figures S1, the constant population size model begins with a ratio of sample size to population size of 1, and spends a good portion of its history close to this value. Likewise, at the time where the transition between current and ancestral size occurs in the instantaneous growth model sample size is still very large relative to population size, achieving a maximum ratio of 0.61. In both these cases the simplifying assumptions of the basic coalescent are likely more substantially violated than in our FTE models. Wakeley and Takahashi (2003) report that the major effect of violating this assumption is a deficiency in singleton variation, so results for these models may under-represent singletons. We
include these models in this study for comparison only. If drawing accurate inferences from these models were of real interest, it would be important to correct for this issue, and it would likely require simulations using a more complex coalescent model.

References

BHASKAR, A., A. G. CLARK and Y. S. SONG, 2013 Distortion of genealogical properties when the sample is very large. arXiv:1308.0091 [q-bio.PE].
