Our two parameter formulation of growth has the form

\[
P(t) = \begin{cases} 
P_0 e^{\alpha t}, & \beta = 1 \\
\frac{P_0^{\beta-1}}{1 - P_0^{\beta-1}(\beta - 1)\alpha t} & , \beta \neq 1
\end{cases}
\]  \quad (1)

where \( P \) is the initial population size in haplotypes, \( t \) is time, and the model parameters \( \alpha \) and \( \beta \) are constants. For \( \beta \neq 1 \) population size approaches infinity as the denominator of the solution approaches 0 which occurs as

\[
P_0^{\beta-1}(\beta - 1)\alpha t \to 1.
\]  \quad (2)

This can be rewritten as

\[
t \to \frac{1}{\alpha P_0^{\beta-1}(\beta - 1)},
\]  \quad (3)

implying there is a finite time at which a population growing according to this model approaches infinite size. Assuming \( t, \alpha, \) and \( P_0 \) are all positive, \( \beta > 1 \) results in infinite size in finite time. In coalescent simulation, the parameters \( \alpha \) and \( \beta \) can be selected with equation (3) to make sure population size in the present is finite.
A comparison of sample size and current effective population size

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].


Figure S1  Ratio of sample size to average population size across a range of $\beta$ values. Results are based on average values from 1,000 independent samples, each of 20,000 haplotypes simulated from an ancestral population of 20,000 haplotypes growing to 8,000,000 haplotypes over 500 generations. The figure also gives a sense of the average number of haplotypes remaining at the end of growth for each model, corresponding to the beginning of linear decline and particularly noticeable for $\beta = 0.1$ and 0.5.
Measuring linkage disequilibrium decay

We measure the decay of linkage disequilibrium in our samples with two commonly used pairwise statistics: \( r^2 \) and \( D' \).

For two variants, A and B, let \( p_A \) and \( p_B \) be their respective minor allele frequencies and \( p_{AB} \) the frequency of the haplotype with both minor alleles, then we can define the disequilibrium coefficient \( D_{AB} = p_{AB} - p_A p_B \). Using \( D \), \( r^2 \) is the correlation coefficient between the loci and is defined as

\[
r_{AB}^2 = \frac{D_{AB}^2}{p_A (1 - p_A) p_B (1 - p_B)}.
\]

\( D' \) is defined as

\[
D'_{AB} = \begin{cases} 
\frac{D_{AB}}{\min(p_A p_B, (1 - p_A)(1 - p_B))} & D_{AB} \leq 0 \\
\frac{D_{AB}}{\min((1 - p_A) p_B, p_A (1 - p_B))} & D_{AB} \geq 0 
\end{cases}
\]

In our research we are interested in the magnitude but not the sign of \( D' \), and consequently, for easier comparisons, we present the absolute value of the statistic.
Figure S2  The amount of very rare non-singleton variation in samples of 10,000 haplotypes with an initial growth rate of $\alpha=100$ and 500 generations of growth.
The average number of variants per kilobase (kb) in samples of 10,000 haplotypes with an initial growth rate of $\alpha=20$. Each colored line represents a different growth duration, with dashed lines giving the values for all variants per kb, and solid lines only the values for singleton variants per kb. The shorter the duration of growth, the less impact accelerating growth has on the values. And for any given acceleration value $\beta$ the longer the duration the greater the number of variants and singletons present in the samples.

Figure S3
Figure S4  The current population sizes presented in figure 4 correspond to distinct $\alpha, \beta$ pairs, plotted here. These results are based on the average number of (A) singletons and (B) all variants in 30 kb of simulated sequence in 10,000 haplotype samples drawn from a population growing from an ancestral size of 20,000 haplotypes over 500 generations. Each colored line represents a different number of singletons or total variants. Contour lines were generated using linear interpolation over a fine-scale range of $\alpha$ determined using a grid search.
Figure S5  Pairwise linkage disequilibrium decay measured by $r^2$ and $D'$ in samples of 10,000 haplotypes from populations with an initial growth rate of $\alpha=100$ and 500 generations of growth. The panels show LD decay in variants with a sample minor allele frequency of (A) 10% to 15%; (B) 1% to 2%; (C) a minor allele count of 10. Each colored line represents a different growth model.