The Impact of Accelerating Faster than Exponential Population Growth on Genetic Variation

Mark Reppell,^{*,1} Michael Boehnke,^{*} and Sebastian Zöllner^{*,1,†}

^{*}Department of Biostatistics and Center for Statistical Genetics, and [†]Department of Psychiatry, University of Michigan, Ann Arbor, Michigan 48109

ABSTRACT Current human sequencing projects observe an abundance of extremely rare genetic variation, suggesting recent acceleration of population growth. To better understand the impact of such accelerating growth on the quantity and nature of genetic variation, we present a new class of models capable of incorporating faster than exponential growth in a coalescent framework. Our work shows that such accelerated growth affects only the population size in the recent past and thus large samples are required to detect the models' effects on patterns of variation. When we compare models with fixed initial growth rate, models with accelerating growth achieve very large current population sizes and large samples from these populations contain more variation than samples from populations with constant growth. This increase is driven almost entirely by an increase in singleton variation. Moreover, linkage disequilibrium decays faster in populations with accelerating growth. When we instead condition on current population size, models with accelerating growth result in less overall variation and slower linkage disequilibrium decay compared to models with exponential growth. We also find that pairwise linkage disequilibrium of very rare variants contains information about growth rates in the recent past. Finally, we demonstrate that models of accelerating growth may substantially change estimates of present-day effective population sizes and growth times.

LARGE-scale, deep sequencing studies have revealed a previously unknown abundance of genetic variation in the human genome, the majority of which is very rare. Nelson *et al.* (2012) sequenced 202 drug target genes in 14,001 individuals and found an alternate allele at 1 in 17 sequenced sites. These variants were overwhelmingly rare, with >60% of the observed single nucleotide variants (SNVs) being singletons, present in just one individual. The Exome Sequencing Project (ESP) sequenced the exome of 2440 individuals and found >500,000 different variant sites, with 57% of these variants singletons (Tennessen *et al.* 2012). Coventry *et al.* (2010) sequenced two genes in 13,715 individuals and estimated 579 sites with alternate alleles in 12.3 kb; singletons made up >28% of variants at *HHEX* and >17% of variants at *KCNJ11*. This newly

discovered abundance of variation in humans, composed primarily of very rare variants, is consistent with recent massive population growth (Tajima 1989). This result is in stark contrast to previous studies in which weak signals of population growth were observed, and the precise model of population growth (instantaneous or exponential growth) had little impact on the expected pattern of diversity (Adams and Hudson 2004; Marth *et al.* 2004; Williamson *et al.* 2005; Keinan *et al.* 2007; Gutenkunst *et al.* 2009; Wall *et al.* 2009; Tennessen *et al.* 2012). Keinan and Clark (2012) explained that many studies (Schaffner *et al.* 2005; Gutenkunst *et al.* 2009; Gravel *et al.* 2011) were based on genotype or small-scale resequencing data and thus were unable to assay the rarest portions of the frequency spectrum. As a result, models fit with these data failed to account for very recent history and arrived at low estimates of human growth rates and small current effective population sizes. Coventry *et al.* (2010), with a much larger sample size, arrived at a faster estimated exponential growth rate in their European origin samples than previous studies. However, when they generated simulated samples using their best estimate growth rate and compared the results to the site frequency spectrum observed at *HHEX* they found the exponential model was still lacking: the number of singleton variants

Copyright © 2014 by the Genetics Society of America

doi: 10.1534/genetics.113.158675

Manuscript received October 13, 2013; accepted for publication December 24, 2013; published Early Online December 30, 2013.

Supporting information is available online at <http://www.genetics.org/lookup/suppl/doi:10.1534/genetics.113.158675/-DC1>.

¹Corresponding authors: Department of Biostatistics, University of Michigan, 1415 Washington Heights, Ann Arbor, MI 48109. E-mail: mreppell@umich.edu; and Department of Biostatistics, University of Michigan, 1415 Washington Heights, Ann Arbor, MI 48109. E-mail: szollne@umich.edu

in their sequenced samples significantly exceeded model predictions. Such an abundance of singletons suggests a very large recent population size, but the observed frequencies of less rare variants are consistent with exponential growth rates that fail to achieve a sufficient population size to explain the quantity of singletons. Coventry *et al.* (2010) hypothesized that a recent acceleration in growth could address the discrepancies between their models and their data. To address this issue in their data, Tennessen *et al.* (2012) fit a piecewise growth model with two periods of growth to model a recent acceleration, improving the fit of the simulated frequency spectrum to the observed data. A technical weakness of modeling accelerating growth by piecewise exponential functions is that while population size over time is continuous, the resulting growth is not continuous. Moreover, the parameter space data increase with every change in growth rate and the correct number of changes is *a priori* unknown. Additionally, the exponential models fit by Coventry *et al.* (2010), Tennessen *et al.* (2012), and Nelson *et al.* (2012) all arrived at current effective population size estimates for Europe of <5 million individuals, a seemingly small number compared to the recent census estimates of a continental population >738 million (United Nations Department of Economic and Social Affairs Population Division 2011). The potential for continuous models incorporating faster than exponential (FTE) growth to accurately reflect recent acceleration in population growth, explain the abundance of singleton variation discovered in human samples, maintain accurate prediction of the observed patterns of common variants, and avoid a drastic expansion of the growth parameter space is an important open question. Such FTE models may also result in estimates of present-day effective population size much larger than the numbers obtained by Coventry *et al.* (2010), Tennessen *et al.* (2012), and Nelson *et al.* (2012) and thus conveniently provide a way to reduce the gap between census population sizes and effective population size estimates.

To address this question, we propose a two-parameter class of continuous population growth functions which allow accelerating, FTE growth. We integrated this model into the coalescent framework (Repell *et al.* 2012), a widely used stochastic model that traces the ancestry of a sample backward through time to its most recent common ancestor (Kingman 1982a,b; Hudson 1983). The original coalescent assumed a constant population size; Donnelly and Tavaré (1995) building on previous work (Slatkin and Hudson 1991; Griffiths and Tavaré 1994) extended the model to allow for populations with size varying deterministically over time; we adapted their work to our two-parameter model of FTE growth. Using our model, we study rare variant frequency spectra and population genetic summary statistics under a wide range of model parameters. In models with the same initial growth rate, where faster acceleration in FTE models results in larger current population sizes, we find that FTE growth produces an abundance of singleton variation in samples without a subsequent increase in the

quantity of more common variants. We also find that when time of growth, ancestral size, and current population size are fixed, FTE growth actually results in fewer total variants, fewer rare variants, and slower decay of LD than exponential growth. With a fixed current size we find that pairwise linkage disequilibrium between very rare variants contains information about recent growth rates. Our work highlights the importance of sample size in distinguishing between growth models and explores the impact of the duration of growth. We find that the addition of the acceleration parameter to growth time and present population size in our models creates a parameter space where sequence summary statistics are consistent with multiple growth times and current sizes and in particular samples from populations that underwent accelerating growth to larger current sizes share characteristics with samples from exponentially growing populations with smaller current sizes.

Methods

Model of faster than exponential growth

We model faster than exponential growth with a set of functions that fit well into the coalescent framework and are motivated by the differential equation suggested by Tolle (2003):

$$\frac{dP}{dt} = \alpha P_t^\beta. \quad (1)$$

Here, P is population size in number of haplotypes, t is time in generations, and the model parameters α and β are constants. When $\beta = 1$, the solution to this equation is an exponential growth function, for $\beta < 1$ the solution results in slower than exponential growth, while for $\beta > 1$, the solution results in FTE growth where not just the population size but also the rate of growth increases with time. If we solve Equation 1 with initial population size P_0

$$P_t = \begin{cases} P_0 e^{\alpha t}, & \beta = 1 \\ \left[\frac{P_0^{\beta-1}}{1 - P_0^{\beta-1}(\beta-1)\alpha t} \right]^{1/(\beta-1)}, & \beta \neq 1. \end{cases} \quad (2)$$

With this parameterization, we interpret α as the exponential growth constant and β as an “acceleration” parameter (Figure 1). Note that under this model, populations can achieve infinite population size in finite time (Supporting Information, File S1). As coalescent frameworks condition on a finite present-day population size, this property does not affect our model.

FTE growth in the coalescent

In the coalescent model we begin with a sample of haplotypes drawn from a current population. Moving backward in time, coalescent events occur where two sample lineages coalesce into a single lineage through a common ancestor, decreasing the number of distinct lineages by one. Coalescent events are observed until the most recent common ancestor of the entire

Figure 1 (A) When conditioning on growth time, ancestral population size, and current population size, and comparing to exponential growth ($\beta = 1$), a population where $\beta > 1$ was smaller in size during its past, while growth with $\beta < 1$ results in a population that was larger. (B) When current population size is allowed to vary, and the exponential growth parameter α is fixed, values of $\beta > 1$ result in current population larger than an exponentially growing population ($\beta = 1$), while values of $\beta < 1$ result in smaller current population sizes.

sample has been reached, and only a single lineage remains (Kingman 1982a). The distribution of coalescent event times depends on the population size and the number of distinct sample lineages remaining. Donnelly and Tavaré (1995) showed that in a population with deterministically varying past size, given an original sample of n haplotypes with j distinct lineages remaining and previous times between coalescent events T_i ($i = n, n - 1, \dots, j + 1$), the probability of the time to the next coalescent event T_j is defined by

$$\begin{aligned} \text{Prob}(T_j > k | T_n + \dots + T_{j+1} = S) \\ = \exp\left(-\binom{j}{2} [\Lambda(k + S) - \Lambda(S)]\right). \end{aligned} \quad (3)$$

Here $\Lambda(t)$ determines how the population size changes over time and is defined as

$$\Lambda(t) = \int_0^t \frac{1}{\lambda(s)} ds, \quad (4)$$

where

$$\lambda(s) = \frac{P_s}{P_0} \quad (5)$$

is the ratio of population size at time s and at the present time. Once $\Lambda(t)$ is specified, it can be substituted into Equation 3 and the time between coalescent events, T_i , $i = n, n - 1, \dots, 2$, drawn. For our model, because the coalescent models populations backward in time, Equation 1 must be changed to

$$\frac{dP}{dt} = -\alpha P_t^\beta \quad (6)$$

before it can be transformed into the corresponding $\Lambda(t)$. Then we can solve, when $\beta \neq 1$:

$$\begin{aligned} \Lambda(t) &= \int_0^t (1 + \alpha s(\beta - 1)P_0^{\beta-1})^{1/(\beta-1)} ds \\ &= \frac{(1 + \alpha t(\beta - 1)P_0^{\beta-1})^{\beta/(\beta-1)} - 1}{\alpha \beta P_0^{\beta-1}}. \end{aligned} \quad (7)$$

And when $\beta = 1$:

$$\Lambda(t) = \int_0^t \frac{1}{e^{-\alpha s}} ds = \frac{e^{\alpha t} - 1}{\alpha}. \quad (8)$$

By substituting $\Lambda(t)$ into Equation 3 we can generate coalescent event times for a population that has been growing or contracting according to our two-parameter model.

Simulation parameters

In our simulations, we assume an ancestral population of 20,000 haplotypes, which grows over the most recent 100 to 3000 generations. We simulate 30 kb from samples of between 100 and 20,000 haplotypes and assume an ancestral mutation parameter $\theta = 16.8 = 2N_e m$, where N_e is the effective ancestral population size in haplotypes and m is the product of per-base mutation rate and the number of bases analyzed. We also assume a uniform ancestral recombination parameter $\rho = 12 = 2N_e r$. Here, N_e is again the effective ancestral population size in haplotypes and r is the product of per-base recombination rate and number of bases analyzed. These parameter values correspond to a per-base mutation rate of 1.4×10^{-8} (Campbell *et al.* 2012; Nelson *et al.* 2012) and a recombination rate of 1 cM/Mb (Kong *et al.* 2002). In our initial analyses we assume a current population size of 8×10^6 haplotypes and 500 generations of growth, roughly consistent with the data of Coventry *et al.* (2010). For comparison we also simulate samples drawn from (1) populations that maintain a constant size of 20,000 haplotypes throughout history and (2) populations that grow from the ancestral 20,000 haplotypes to a size of 8×10^6 haplotypes instantaneously 500 generations in the past. For each pair of α and β values, we report results for 1000 independent simulation replicates. We also use simulation to verify that for the range of β -values investigated, the coalescent assumption that sample size remains much smaller than effective population size is not violated in a way that changes our findings or conclusions (File S2 and Figure S1).

To capture the scale of patterns in linkage disequilibrium (LD) we simulate longer sequences of 100 kb and calculate pairwise r^2 and the absolute value of D' (File S3). We bin

variants based on minor allele count and sample a subset for which we calculate all pairwise statistics. Binning allows us to compare LD across models without having the results determined solely by differences in the abundance of very rare variants.

Results

Using coalescent simulations, we generate samples from populations that grow according to our two-parameter model in which α is an exponential-like growth parameter and β an acceleration parameter. We then quantify how changes in growth patterns affect different portions of the sample allele frequency spectrum and linkage disequilibrium. We also demonstrate how varying sample sizes and growth times alter the quantity and nature of genetic diversity within our growth models.

Accelerating growth with fixed current population size

To isolate the trajectory of the population size under a model of accelerating growth we first simulate samples under a simple model where ancestral size, growth time, and current population size are fixed while β varies between 0.1 and 3.5. We use an ancestral size of 20,000 haplotypes, which grows over a period of 500 generations to a current effective size of 8×10^6 haplotypes. We also consider models in which the population remains at its ancestral size throughout and where the population instantaneously grows to 8,000,000 haplotypes 500 generations in the past.

When conditioning on a fixed current size, a population with accelerating growth initially grows more slowly than a population with constant growth, accelerating to exceed the constant rate only during the very recent past (Figure 1). With increasing β , large changes of population size occur more recently and the population size is closer to its ancestral size for longer. Hence the total variation decreases as β increases (Figure 2A). For example, samples of 10,000 haplotypes drawn from a model where $\beta = 0.5$ have on average 54.0 variants/kb, 1.4 times greater than the average of 39.6 variants/kb under exponential growth, and 5.9 times greater than the average of 9.1 variants/kb under accelerating growth with $\beta = 3.5$.

Rare variants drive the divergence in the total number of variant sites between models (Figure 2B). In a sample of 10,000 haplotypes, variants with minor allele count ≤ 10 account for 99% of the reduction in variable sites between models with $\beta = 1$ and $\beta = 2$ and 99.8% of the reduction between models with $\beta = 1$ and $\beta = 3.5$. Correspondingly, the proportion of variants that are very rare also decreases with increasing β (Figure 2D). The proportion of variants that are singletons drops from 0.66 to 0.25 between models where $\beta = 1$ and $\beta = 3.5$.

Models with constant population size and models with instantaneous growth bound the family of FTE models with a fixed current population size (Figure 2). As β increases, the average frequency spectrum more closely resembles the

average frequency spectrum of a model without growth. Conversely, as β approaches 0, the average frequency spectrum more closely resembles the average frequency spectrum from a model with instantaneous growth. Larger values of β correspond to histories where population size is very close to its ancestral value for their entirety. Consequently, genetic sequences simulated under these models share many similarities to those generated under a constant population size model. As β decreases, achieving the current size requires initially fast growth that is slowing over time and sequences simulated under these models bear resemblance to those from instantaneous growth models. Differences between growth models become apparent only with increasing sample size (Figure 2, A and B). In a sample of 100 haplotypes, the most extreme models we simulate differ by <1 variant/kb: a model with no growth averages 2.9 variants/kb compared with 3.6 variants/kb with instantaneous growth. And samples with $\beta = 1$ contain <0.1 more variants/kb on average than $\beta = 1.1$, both resulting in an average of 3.4 variants/kb. When sample size expands to 20,000 haplotypes, instantaneous growth averages 108.5 more variants/kb than a model without growth (114.4 variants/kb vs. 5.9 variants/kb), and the differences between $\beta = 1$ with 62.1 variants/kb and $\beta = 1.1$ with 56.1 variants/kb become apparent.

Accelerating growth with variable current population size

In practice, accelerating growth should lead to a larger present-day population size than growth at a constant rate. To examine this, we sample 100 to 20,000 haplotypes from populations where the initial growth rate α has been fixed at a value of 100 and β is allowed to vary between exponential growth ($\beta = 1$) and 1.1. Growth accelerates very quickly with increasing β . When we fix the time of growth at 500 generations our parameter settings result in growth from an ancestral size of 20,000 haplotypes to a current size of between 2.4×10^5 and 1.4×10^9 haplotypes, with most of the change in population size again occurring in the very recent past.

As growth accelerates, the total quantity of genetic variation in samples quickly increases (Figure 3A). For example, in a sample of 10,000 haplotypes there are on average 15.5 variants/kb with exponential growth, 25.5 variants/kb when growth accelerates with $\beta = 1.04$, and 50.8 variants/kb when $\beta = 1.1$. This increase in variation is driven almost entirely by an increase in the quantity of singleton variants (Figure 3, B and C), which proportionally come to dominate the frequency spectrum (Figure 3D). Ten thousand haplotypes contain on average 4.9 singletons/kb with exponential growth, 12.3 singletons/kb when growth accelerates with $\beta = 1.04$, and 40.2 singletons/kb when $\beta = 1.1$. Between $\beta = 1$ and $\beta = 1.1$ the proportion of all variants that are singletons rises from 0.32 to 0.79. For non-singleton rare variants, the effect of accelerating growth is relatively small, and in fact the average number observed in

Figure 2 Sequence properties for growth models with a current size fixed at 8×10^6 haplotypes reached after 500 generations of growth. (A) Average number of variant sites per kilobase by sample size. Each colored line represents a different growth model. (B) Average number of singleton vs. nonsingleton variants per kilobase under different growth models as sample size increases. (C) Site frequency spectrum showing the average number of variants with a given minor allele count per kilobase of sequence in a sample of 10,000 haplotypes. (D) The change in the proportion of all variants that have the given allele counts as growth accelerates in a sample of 10,000 haplotypes.

a sample can be smaller under faster accelerating models (Figure 3C and Figure S2). In samples of 10,000 haplotypes the number of variants per kilobase with minor allele counts between 4 and 10 drops from 2.6 to 2.0 between $\beta = 1$ and $\beta = 1.1$. Additionally, the number of variants with a given-minor allele count does not follow a monotonic trend as growth accelerates. For example, with 10,000 haplotypes, the number of doubletons increases between $\beta = 1$ and $\beta = 1.04$ from 2 to 3.7/kb before decreasing to 3.4/kb when $\beta = 1.1$, and the proportion of all variants that are doubletons follows the same trend ($\beta_1 = 0.14$, $\beta_{1.04} = 0.15$, $\beta_{1.1} = 0.07$) (Figure 3D).

As with a fixed current size, differences between growth models become apparent only as sample size increases (Figure 3A). With samples of 100 haplotypes, a model where $\beta = 1$ has an average of 3.3 variants/kb, only 0.2 variants/kb smaller than when $\beta = 1.1$. This contrasts sharply with 10,000 haplotypes, where the difference grows to 35.3 more variants/kb ($\beta_1 = 15.5$ variants/kb, $\beta_{1.1} = 50.8$ variants/kb).

The duration of accelerating growth

To quantify how the duration of population growth affects our results we perform simulations with samples of 10,000 haplotypes where the initial growth rate α is fixed at 20, β is

allowed to vary between $\beta = 1$ and $\beta = 1.1$, and growth duration varies between 100 and 3,000 generations. When we calculate the site frequency spectra for these samples we observe that for all growth durations, the patterns observed in the section above hold: singleton variation increases with accelerating growth, common variation is relatively unaffected, and nonsingleton rare variation follows nonmonotonic patterns where maxima occur at intermediate growth rates. The shorter the duration of growth the less pronounced these patterns are, and the smaller the differences between the growth models. When we compare models with the same rate of acceleration but different growth durations, we find that shorter durations result in frequency spectra with fewer of all variant types and proportionally fewer singletons and other rare variants (Figure S3).

Growth time, present population size, and acceleration create a parameter space where sequence characteristics are consistent with multiple sets of growth parameters. With exponential growth, once growth time and growth rate are defined, a population's current size is fixed, and the frequency spectrum can be predicted from these parameters. The addition of an acceleration parameter changes this determinism; the β parameter allows a model where growth time and α have been defined to grow to any current size. Therefore, different combinations of acceleration and

Figure 3 Sequence properties for growth models with same initial growth rate $\alpha = 100$ and 500 generations of growth. (A) Average number of variant sites per kilobase by sample size. Each colored line represents a different growth model. (B) Average number of singleton vs. nonsingleton variants per kilobase as sample size increases for an exponential growth model and a model where $\beta = 1.1$. (C) Site frequency spectrum showing the average number of variants with a given minor allele count per kilobase of sequence in a sample of 10,000 haplotypes. (D) The change in the proportion of all variants that have the given allele counts as growth accelerates in a sample of 10,000 haplotypes.

current population size can generate the same values of a sequence summary statistic (Figure 4). In particular, in a model of accelerated growth the same summary statistic can be generated by a much larger present-day population size compared to a model of constant growth. Assuming an ancestral size of 20,000 haplotypes, 500 generations of growth, and a sample of 10,000 haplotypes, an average of 40 variants/kb of sequence is consistent with both a population that grew to a current size of 8.4×10^6 haplotypes at a constant exponential rate or 18.9×10^6 haplotypes following $\beta = 1.1$. In this example the initial growth rate is slower under accelerating growth (Figure S4) but the β parameter results in a rapid acceleration and much larger current population size. Note that even though growth follows very different trajectories between these models, the ancestral trees have the same mean total branch length; consequently the same average number of variants is found in the resulting samples.

Similarly, different combinations of acceleration and growth time can generate the same values of a sequence summary statistics (Figure 5). For example, assuming an ancestral size of 20,000 haplotypes and 30 kb sequenced in a samples of 10,000 haplotypes, we expect about 750 variant sites both from a model of constant growth with rate of $\alpha = 20$ for 2589 generations and a model where $\alpha = 20$

but growth accelerates with $\beta = 1.1$ for 1064 generations. Likewise, under the same scenario, we expect 200 singletons both from a model of exponential growth for 2605 generations and from a model with accelerating growth ($\beta = 1.1$) for 903 generations.

Linkage disequilibrium

To understand the impact of growth on LD decay, we simulate 10,000 haplotypes of length 0.1 cM for $\alpha = 100$ and 500 generations of growth and calculate pairwise r^2 and the absolute value of D' . Even though these growth models result in current population sizes that differ by three orders of magnitude, historically their sizes are very similar, and as a result the pattern of LD decay with faster acceleration is subtle. For variants with minor allele count >50 , the rate of pairwise LD decay increases as β increases and accelerated growth results in larger current population sizes. Between variants with minor allele frequencies between 0.1 and 0.15, at a distance of 40 kb r^2 is on average 0.046 in an exponential growth model, 0.045 for $\beta = 1.04$, and 0.041 when $\beta = 1.1$ (Figure S5). In the same frequency bin, at 40 kb D' is on average 0.598 in populations with growth at $\beta = 1$ compared to 0.597 for populations with growth at $\beta = 1.04$ and 0.569 when $\beta = 1.1$. In rarer variants, independent of growth model, at a given physical distance the average r^2

Figure 4 Contour plots showing that under different growth parameters summary statistics are consistent with multiple current population sizes. In 30 kilobases of sequence from samples of 10,000 haplotypes, after 500 generations of growth the average quantities of (A) singletons and (B) all variants from models with larger current population sizes that grew at faster than exponential rates are the same as the average quantities in models with smaller current sizes where the population grew at a slower rate. 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 α determined using a grid search. For each α in the grid, average parameter values were calculated from 100 independent simulation replicates.

is lower and D' is higher (Devlin and Risch 1995; VanLiere and Rosenberg 2008). For example, both with $\beta = 1$ and $\beta = 1.1$ the average pairwise r^2 decays <0.1 in 2 kb for variants with minor allele frequencies between 0.01 and 0.02, compared with 22 kb in variants with minor allele frequencies between 0.1 and 0.15. However, when we look at the relationship between LD decay and acceleration of population growth, it is qualitatively the same across allele counts. Pairwise r^2 in variants with minor allele frequencies between 0.01 and 0.02 from a sample of 10,000 haplotypes and 500 generations of growth with $\alpha = 100$, at 40 kb is on average 0.026 when $\beta = 1$ vs. 0.023 when $\beta = 1.1$, and for the same comparison with D' the values are 0.944 for $\beta = 1$ and 0.852 for $\beta = 1.1$. The pattern of faster decay as β increases holds true in smaller sample sizes, but is less pronounced. In samples of 1000 haplotypes, at 40 kb, variants with minor allele frequencies between 0.01 and 0.02 have an average pairwise r^2 of 0.024 when $\beta = 1$ compared with 0.023 for $\beta = 1.1$, and average D' is 0.018 greater ($D'_{\{\beta=1\}} = 0.973$, $D'_{\{\beta=1.1\}} = 0.955$).

Varying growth duration between 250 and 750 generations, we find that for any given duration and α , LD decays faster in models where β is greater. When we compare models with the same α and β values but different growth durations we find that shorter growth times, with smaller current sizes, have slower LD decay. For example, in samples of 10,000 haplotypes from models with $\alpha = 100$ and $\beta = 1.06$, the average pairwise r^2 of variants with minor allele frequencies between 0.01 and 0.02 decays <0.03 at distances of 42, 28, and 24 kb in models with 250, 500, and 750 generations of growth, respectively.

By returning to models with fixed current size we are able to compare pairwise LD decay over a larger range of accelerating growth models. As was described in the section above, with a fixed current size, larger β and hence faster acceleration of growth means that a population has been large for a shorter period of time. We find the same pattern of faster LD decay in larger populations with these models;

only their relationship with β is reversed. Lower values of β correspond to larger recent population sizes and consequently faster decay in LD (Figure 6). In samples of 10,000 haplotypes from populations that expand from an ancestral size of 20,000 haplotypes to a current size of 8×10^6 haplotypes over 500 generations in variants with minor allele frequencies between 0.1 and 0.15 average pairwise r^2 decays <0.1 at a distance of 18, 20, and 21 kb for models where $\beta = 0.5$, 1, or 3.5 respectively, and D' falls <0.75 at distances of 17, 18, and 23 kb. As above, the trends in LD decay are qualitatively the same across variant frequencies, but variants with lower frequencies have lower average pairwise r^2 and the D' between them decays over greater distances. In variants with minor allele frequencies between 0.01 and 0.02, at a distance of 40 kb, pairwise r^2 is 27% greater in a model where $\beta = 3.5$ than when $\beta = 0.5$ ($r^2_{\{\beta=0.5\}} = 0.022$, $r^2_{\{\beta=3.5\}} = 0.028$), and likewise D' is 16% greater ($D'_{\{\beta=0.5\}} = 0.835$, $D'_{\{\beta=3.5\}} = 0.968$).

The range of models we are able to consider with a fixed current size allows us to observe an interesting pattern in variants with minor allele count between 2 and 20. For these variants, D' maintains the same pattern of decay occurring over increasingly long genetic distances, but faster decay in models with smaller β and larger recent population sizes. In contrast, average pairwise r^2 for the rarest variants is lower in models with intermediate rates of growth (Figure 6C). For example, at a minor allele count of 10, average r^2 for variants 20 kb apart where $\beta = 1$ is 38% lower than when $\beta = 0.5$ and 70% lower than for $\beta = 3.5$ ($r^2_{\{\beta=0.5\}} = 0.013$, $r^2_{\{\beta=1\}} = 0.008$, $r^2_{\{\beta=3.5\}} = 0.027$).

Discussion

Recent sequencing studies suggest an acceleration of growth in human populations, as exponential growth models cannot capture the observed excess of singleton variants (Coventry *et al.* 2010; Tennessen *et al.* 2012). The frequency spectra in these studies can be better modeled by defining growth as

Figure 5 Contour plots showing that under different growth parameters summary statistics are consistent with multiple growth durations. In 30 kilobases of sequence from samples of 10,000 haplotypes, after beginning with the same initial growth rate of $\alpha = 20$, the average quantities of (A) singletons and (B) all variants from models that grow for a shorter amount of time at an accelerated rate are the same as those from models that grow for longer at a slower rate. 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 α determined using a grid search. Average parameter values were calculated from 100 independent simulation replicates at each grid point.

a discontinuous piecewise function with an arbitrarily selected number of segments. However, even the simplest of such models require several additional parameters. Instead, we provide an approach for including an acceleration parameter β into the population growth model, obtaining a continuous model that can allow for a wide range of growth trajectories.

When modeling accelerating growth while conditioning on current population size, we observed that the total amount of variation and the amount of rare variation both decrease with increasing β . With a fixed current size, when $\beta < 1$ the initial growth parameter α must be large, and as β increases α correspondingly decreases. Large values of α result in populations that quickly expand at the onset of growth and have a larger size for much of their history. Small α values yield populations with slow initial growth, which achieve only sizes significantly larger than their ancestral size in the recent past. In smaller populations common ancestors are found more quickly, coalescent trees are smaller, and fewer variants are found in samples, explaining our results as β increases. For large β the population history closely resembles a model without growth for all but a small fraction of its history, explaining why models without growth bound our FTE models as $\beta \rightarrow \infty$. Likewise, the closer β is to 0, the faster a population's size diverges from its ancestral state, and so more closely resembles a model of instantaneous growth. In this study we assume an ancestral size of 20,000 haplotypes and a period of growth lasting 500 generations; however, the manner in which the demographic history of a population is altered by our growth model, as described in this paragraph, is completely independent of the exact parameter values.

Under an alternate set of conditions, where current size is not fixed and accelerating growth leads to a greater current population size, we observe that the total amount of variation increases with increasing β and that this increase is driven almost entirely by the amount of singleton variation present in samples. Under these conditions, when we compare a population that grows exponentially to one that

grows with $\beta > 1$, initially, growth and population size are determined by the initial growth rate α . Acceleration only manifests itself in the most recent generations. Consequently, for both common and nonsingleton rare variants, the frequency spectra from FTE models closely resemble those from exponentially growing models; however, the very recent acceleration to a much larger current size gives these models an abundance of singletons unmatched by simple exponential models.

Our work shows that the quantity of nonsingleton rare variants changes in a nonmonotonic fashion as β increases. This is best explained by looking at the average population size at the time these variants arise. The number of alleles with a given allele count is a function of the total length of coalescent branches with that number of descendants, a value dependent on population size. In a large population branches are longer; however, coalescent events are also less common, so branches with a given number of descendants occur further in the past than they would in a smaller population. As β increases the population size and mutation age change at different rates. As β increases from low to intermediate values, population size at the times nonsingleton rare variants arise increases even as average mutation age decreases. However, as β increases further, even as average mutation age continues to decrease, the population size at the time nonsingleton rare variants arise decreases more rapidly and is smaller than in the intermediate growth models. Thus, the total length of branches with a given number of descendants is larger at intermediate growth rates and the quantity of variants changes in the observed nonmonotonic manner.

In both the fixed and variable current size contexts we illustrate the importance of sample size for detecting differences between growth models. Small samples contain a limited amount of genetic variation for use in inference. And the inability to distinguish between variants with allele frequencies $< 1/(\text{sample size})$ means that small samples lack resolution for studying rare and recently arisen variants that contain information about recent demographics. This

Figure 6 Pairwise linkage disequilibrium decay measured by r^2 and D' in samples of 10,000 haplotypes from populations with current population sizes of 8×10^6 haplotypes reached after 500 generations of growth. LD decay in variants with a sample minor allele frequency is shown: (A) 10–15%; (B) 1–2%; (C) a minor allele count of 10. Each colored line represents a different growth model.

finding supports the work of Keinan and Clark (2012) in their explanation of why earlier, smaller studies failed to detect evidence of recent massive population growth. It also underpins why more accurate models are needed as study sample sizes increase: in 100 haplotypes, failing to include any growth has limited consequences on the site frequency spectrum and resulting inferences; in 20,000 haplotypes even models with $\beta = 1$ and $\beta = 1.02$ result in substantially different frequency spectra.

The addition of the acceleration parameter allows better modeling of very rare variants; the resulting model can “bend” growth trajectories between any ancestral population size and current population size, over any amount of time, flexibility impossible with exponential growth alone. Our work shows how samples from populations that have undergone accelerating growth share the summary statistics of samples from populations with much smaller current sizes achieved via slower growth. Similarly, samples from populations that have grown at an accelerating FTE rate for a shorter length of time share summary characteristics with samples that grew at a slower rate for a longer period of time. Thus, estimates of current population size and the duration of growth estimated from genetic data may differ greatly if accelerating growth is considered.

Our linkage disequilibrium results are also a direct result of how demographic history changes with β . When we look at our models where current population size is allowed to expand with accelerating growth we find that both r^2 and D' decay more quickly in models with larger β , but the effect is small. In larger populations the average time since samples have shared a common ancestor is longer, increasing the

likelihood of recombination events that break down D' and r^2 . In our models with accelerating growth, it is only over a very brief recent time that the models with larger β have been substantially larger than those with smaller β , and consequently while they show faster decay, it is a very modest difference. However, differences between models are detectable, and this suggests that in the future, pairwise LD could be used to help assess the fit of growth models.

The effects of accelerated growth on LD are larger in our models with fixed current size, as these populations substantially differ in size for much longer. We observe that as β increases LD decays more slowly. This pattern holds across the full frequency spectrum for D' , and the majority of the frequency spectrum for r^2 ; however, as minor allele count approaches 1 the pattern for r^2 changes. When we look at the ancestry of a sample, extremely rare variants likely arose only recently and are carried by very few lineages, so the probability of a recombination event occurring between them is small. For these variants every pair of variants has an r^2 of either 1 if they arose on the same lineage or near 0 if they did not. Therefore, average pairwise r^2 becomes equivalent to the probability that the two variants arose on the same lineage. The total quantity of variants with a given minor allele count in a sample is a function of the total length of all lineage branches with exactly that number of descendants, which is itself a function of population size. The probability that two variants with a given minor allele count arise on the same branch is a function of the variability of the branch lengths with exactly that number of descendants. As variability decreases, the probability that two variants arose on the same branch also decreases. The distribution of

branch lengths for a given minor allele count reflects how the population size changes while ancestral lineages with that number of descendants exist. For rare variants, what we observe is that in both models with small β and those with large β the majority of lengths are very short, and the variance is lower than under models with intermediate β . For large β , the population size shrinks so fast that branch lengths corresponding to every variant count are predominantly short. For small β the population size has been large for some time, so lineages rarely find common ancestors during the period of growth, and nonsingleton variants arise mostly in the ancestral population, where all branch lengths are again short. Intermediate β yield a Goldilocks situation, where population sizes result in lineages that have found enough common ancestors that mutations are found in multiple present-day samples, but branch lengths remain relatively long because the time between common ancestors is longer than it would be at the ancestral population size. This observation gives rise to the idea, to be explored in a future study, that pairwise r^2 between very rare variants could be used as a tool to help estimate recent growth rates.

Presently, the amount of high-quality deep coverage sequencing data available to the research community is increasing by the day. These data offer not only the potential for a better understanding of the genetic underpinnings of a multitude of diseases and traits, but also important insight into the demographic history of our species. The accuracy of these insights will be directly dependent on the models they are based on. The general two-parameter models introduced by this study can substantially revise estimates of current population sizes while simultaneously modeling the excess of very rare variants observed in large human resequencing studies, providing a valuable tool for modeling the complex history of humanity.

Acknowledgments

The authors thank Mary-Kate Wing and Shyam Gopalakrishnan for help with optimization and distribution of the FTEC program. This work was supported by National Institutes of Health grants HG000376 (to M.B.) and HG005855 (to S.Z.) and National Institutes of Health training grant HG000040.

Literature Cited

- Adams, A. M., and R. R. Hudson, 2004 Maximum-likelihood estimation of demographic parameters using the frequency spectrum of unlinked single-nucleotide polymorphisms. *Genetics* 168: 1699–1712.
- Campbell, C. D., J. X. Chong, M. Malig, A. Ko, B. L. Dumont *et al.*, 2012 Estimating the human mutation rate using autozygosity in a founder population. *Nat. Genet.* 44: 1277–1281.
- Coventry, A., L. M. Bull-Otterson, X. Liu, A. G. Clark, T. J. Maxwell *et al.*, 2010 Deep resequencing reveals excess rare recent variants consistent with explosive population growth. *Nat. Commun.* 1: 131.
- Devlin, B., and N. Risch, 1995 A comparison of linkage disequilibrium measures for fine-scale mapping. *Genomics* 29: 311–322.
- Donnelly, P., and S. Tavaré, 1995 Coalescents and genealogical structure under neutrality. *Annu. Rev. Genet.* 29: 401–421.
- Gravel, S., B. M. Henn, R. N. Gutenkunst, A. R. Indap, G. T. Marth *et al.*, 2011 Demographic history and rare allele sharing among human populations. *Proc. Natl. Acad. Sci. USA* 108: 11983–11988.
- Griffiths, R. C., and S. Tavaré, 1994 Sampling theory for neutral alleles in a varying environment. *Philos. Trans. R. Soc. B Biol. Sci.* 344: 403–410.
- Gutenkunst, R. N., R. D. Hernandez, S. H. Williamson, and C. D. Bustamante, 2009 Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. *PLoS Genet.* 5: e1000695.
- Hudson, R. R., 1983 Testing the constant-rate neutral allele model with protein sequence data. *Evolution* 37: 203–217.
- Keinan, A., and A. G. Clark, 2012 Recent explosive human population growth has resulted in an excess of rare genetic variants. *Science* 336: 740–743.
- Keinan, A., J. C. Mullikin, N. Patterson, and D. Reich, 2007 Measurement of the human allele frequency spectrum demonstrates greater genetic drift in East Asians than in Europeans. *Nat. Genet.* 39: 1251–1255.
- Kingman, J. F. C., 1982a The coalescent. *Stochastic Process. Appl.* 13: 235–248.
- Kingman, J. F. C., 1982b On the genealogy of large populations. *J. Appl. Probab.* 19: 27–43.
- Kong, A., D. F. Gudbjartsson, J. Sainz, G. M. Jonsdottir, S. A. Gudjonsson *et al.*, 2002 A high-resolution recombination map of the human genome. *Nat. Genet.* 31: 241–247.
- Marth, G. T., E. Czabarka, J. Murvai, and S. T. Sherry, 2004 The allele frequency spectrum in genome-wide human variation data reveals signals of differential demographic history in three large world populations. *Genetics* 166: 351–372.
- Nelson, M. R., D. Wegmann, M. G. Ehm, D. Kessner, P. St Jean *et al.*, 2012 An abundance of rare functional variants in 202 drug target genes sequenced in 14,002 people. *Science* 337: 100–104.
- Reppell, M., M. Boehnke, and S. Zollner, 2012 FTEC: a coalescent simulator for modeling faster than exponential growth. *Bioinformatics* 28: 1282–1283.
- Schaffner, S. F., C. Foo, S. Gabriel, D. Reich, M. J. Daly *et al.*, 2005 Calibrating a coalescent simulation of human genome sequence variation. *Genome Res.* 15: 1576–1583.
- Slatkin, M., and R. R. Hudson, 1991 Pairwise comparisons of mitochondrial DNA sequences in stable and exponentially growing populations. *Genetics* 129: 555–562.
- Tajima, F., 1989 The effect of change in population size on DNA polymorphism. *Genetics* 123: 597–601.
- Tennessen, J. A., A. W. Bigham, T. D. O'Connor, W. Fu, E. E. Kenny *et al.*, 2012 Evolution and functional impact of rare coding variation from deep sequencing of human exomes. *Science* 337: 64–69.
- Tolle, J., 2003 Can growth be faster than exponential, and just how slow is the logarithm? *Math. Gazette* 87: 522–525.
- VanLiere, J. M., and N. A. Rosenberg, 2008 Mathematical properties of the r^2 measure of linkage disequilibrium. *Theor. Popul. Biol.* 74: 130–137.
- Wall, J. D., K. E. Lohmueller, and V. Plagnol, 2009 Detecting ancient admixture and estimating demographic parameters in multiple human populations. *Mol. Biol. Evol.* 26: 1823–1827.
- Williamson, S. H., R. Hernandez, A. Fledel-Alon, L. Zhu, R. Nielsen *et al.*, 2005 Simultaneous inference of selection and population growth from patterns of variation in the human genome. *Proc. Natl. Acad. Sci. USA* 102: 7882–7887.

Communicating editor: Y. S. Song.

GENETICS

Supporting Information

<http://www.genetics.org/lookup/suppl/doi:10.1534/genetics.113.158675/-/DC1>

The Impact of Accelerating Faster than Exponential Population Growth on Genetic Variation

Mark Reppell, Michael Boehnke, and Sebastian Zöllner

File S1

Achieving infinite population size in finite time

Our two parameter formulation of growth has the form

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

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

$$P_0^{\beta-1}(\beta-1)\alpha t \rightarrow 1. \quad (2)$$

This can be rewritten as

$$t \rightarrow \frac{1}{\alpha P_0^{\beta-1}(\beta-1)}, \quad (3)$$

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

File S2

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 β 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 β 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].
- TENNESSEN, J. A., A. W. BIGHAM, T. D. O'CONNOR, W. FU, E. E. KENNY *et al.*, 2012 Evolution and functional impact of rare coding variation from deep sequencing of human exomes. *Science* **337**: 64-69.
- WAKELEY, J., and T. TAKAHASHI, 2003 Gene genealogies when the sample size exceeds the effective size of the population. *Mol Biol Evol* **20**: 208-213.

Figure S1 Ratio of sample size to average population size across a range of β 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.

File S3

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.

Figure S3 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 β the longer the duration the greater the number of variants and singletons present in the samples.

Figure S4 The current population sizes presented in figure 4 correspond to distinct α , β 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 α 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.