Theoretical Population Biology 74 (2008) 182–190

Contents lists available at ScienceDirect

Theoretical Population Biology

journal homepage: www.elsevier.com/locate/tpb

# To what extent does genealogical ancestry imply genetic ancestry?

Frederick A. Matsen *, Steven N. Evans

*Department of Statistics, University of California at Berkeley, 367 Evans Hall, Berkeley, CA 94720-3860, USA*

ARTICLE INFO

*Article history:*
Received 14 May 2008
Available online 26 June 2008

*Keywords:*
Most recent common ancestor
Diploid
Coalescent
Branching process
Total positivity
Monotone likelihood ratio

ABSTRACT

Recent statistical and computational analyses have shown that a genealogical most recent common ancestor (MRCA) may have lived in the recent past [Chang, J.T., 1999. Recent common ancestors of all present-day individuals. Adv. Appl. Probab. 31, 1002–1026. 1027–1038; Rohde, D.L.T., Olson, S., Chang, J.T., 2004. Modelling the recent common ancestry of all living humans. Nature 431, 562–566]. However, coalescent-based approaches show that genetic most recent common ancestors for a given non-recombining locus are typically much more ancient [Kingman, J.F.C., 1982a. The coalescent. Stochastic Process Appl. 13, 235–248; Kingman, J.F.C., 1982b. On the geneology of large populations. J. Appl. Probab. 19A, 27–43]. It is not immediately clear how these two perspectives interact. This paper investigates relationships between the number of descendant alleles of an ancestor allele and the number of genealogical descendants of the individual who possessed that allele for a simple diploid genetic model extending the genealogical model of [Chang, J.T., 1999. Recent common ancestors of all present-day individuals. Adv. Appl. Probab. 31, 1002–1026. 1027–1038].

© 2008 Elsevier Inc. All rights reserved.

## 1. Introduction and model

Joseph Chang's 1999 paper (Chang, 1999) showed that a well-mixed closed diploid population of $n$ individuals will have a genealogical common ancestor in the recent past. Specifically, the paper showed that if $T_n$ is the number of generations back to the most recent common ancestor (MRCA) of the population, then $T_n$ divided by $\log_2 n$ converges to one in probability as $n$ goes to infinity. A modification of this result was applied by Rohde et al. (2004) to estimate $T_n$ for the human world population.

Chang (1999) initiated a discussion in which many of the leading figures of population genetics expressed interest in the relationship between the genealogical and genetic perspectives for such models (Donnelly et al., 1999). For example, Peter Donnelly wrote "[r]esults on the extent to which common ancestors, in the sense of [Chang's] paper, are ancestors in the genetic sense…would also be of great interest" (Donnelly et al., 1999). Every other discussant also either discussed the relationship of Chang's work to genetics or expressed interest in doing so.

Given this interest, surprisingly little work has been done specifically about the interplay between the two perspectives. Wiuf and Hein, in their reply, wrote three paragraphs containing some simple initial observations (Donnelly et al., 1999). Some simulation work has been done by Murphy (2004) with a more realistic population model. In a related though different vein, Möhle and Sagitov (2003) derived limiting results for the diploid coalescent, in the classical setting of a small sample from a large population.

In an interesting series of papers, Derrida, Manrubia, Zanette, and collaborators (Derrida et al., 1999, 2000a,b; Manrubia et al., 2003) have investigated the distribution of the number of repetitions of ancestors in a genealogical tree, as well as the degree of concordance between the genealogical trees for two distinct individuals. Our paper, on the other hand, is concerned with correlations between the number of genealogical descendants of an individual and the number of descendant alleles of that individual. The interesting time-frame in our paper is different than theirs: they focus on the period substantially after $T_n$, while for us any interesting correlation is erased with high probability after time about $1.77T_n$.

Our paper attempts to connect the genealogical and genetic points of view by investigating several different questions concerning the interaction of genealogical ancestry and genetic ancestry in a diploid model incorporating Chang's model. In the classical Wright–Fisher fashion, we consider $2n$ alleles contained in $n$ diploid individuals. Each discrete generation forward in time, every individual selects two alleles from the previous generation independently and uniformly to "inherit". If an individual $X$ at time $t$ inherits genetic information from an individual $Y$ at time $t-1$, then we consider $Y$ to be a "parent" of $X$ in the genealogical sense. As with Chang's model, the two parents are permitted to be the same individual and each allele of a child may descend from the same parent allele. We illustrate the basic operation of the model in

* Corresponding author.
*E-mail addresses:* matsen@berkeley.edu (F.A. Matsen), evans@stat.Berkeley.EDU (S.N. Evans).
*URLs:* http://www.stat.berkeley.edu/users/matsen/ (F.A. Matsen), http://www.stat.berkeley.edu/users/evans/ (S.N. Evans).

0040-5809/$ – see front matter © 2008 Elsevier Inc. All rights reserved.
doi:10.1016/j.tpb.2008.06.003

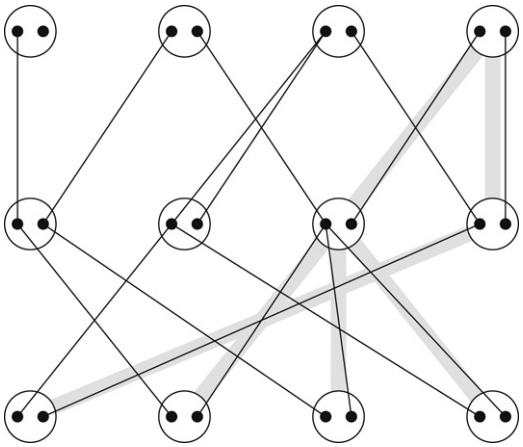

**Fig. 1.** An example instance of our model with four individuals and three generations. Time increases moving down the diagram. The two alleles of each individual are depicted as two dots within the larger circles; a thin black line indicates genetic inheritance, i.e. the lower allele is descended from the upper allele. This sample genealogy demonstrates that the genealogical MRCA need not have any genetic relation to present-day individuals. The individual at the far right on the top row is in this case the (unique) MRCA as demonstrated by the thick gray lines, however none of its genetic material is passed onto the present day.

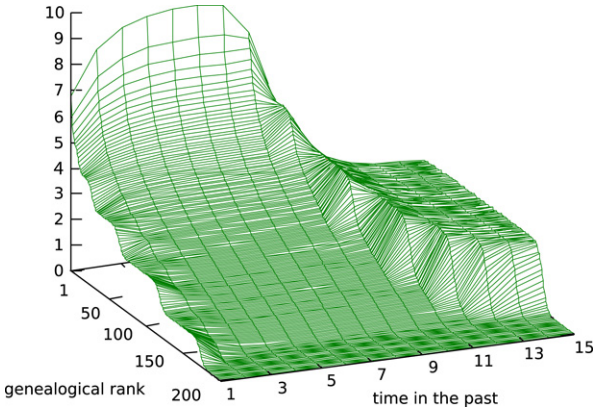


**Fig. 2.** The expected number of descendant alleles from historical individuals sorted by number of genealogical descendants. Results by simulation of a population of size 200. For example, the value at "genealogical rank" 50 and time 2 is the expected number of alleles in the current population which descend from one or other of the two alleles present in the individual two generations ago who had no more genealogical descendants in the present population than did 49 other individuals in the population two generations ago. As described in the text, this curve attains an interesting characteristic shape around generation 3 that lasts until generation 8. We investigate that shape in Fig. 3 and Proposition 2.

Fig. 1. Each individual is represented as a circle, and each of a given individual's alleles are represented as dots within the circle. Time increases down the figure and inheritance of alleles is represented by lines connecting them.

We have chosen notation in order to fit with Chang's original article. The initial generation will be denoted $t = 0$ and other generations will be counted forwards in time; thus the parents of the $t = 1$ generation will be in the $t = 0$ generation, and so on. The $n$ individuals of generation $t$ will be denoted $I_{t,1}, \ldots, I_{t,n}$. The two alleles present at a given locus of individual $I_{t,i}$ will be labeled $A_{t,i,1}$ and $A_{t,i,2}$. Using this notation, each allele $A_{t,i,c}$ of generation $t$ selects an allele $A_{t-1,j,d}$ uniformly and independently from all of the alleles of the previous generation; given such a choice we say that allele $A_{t,i,c}$ is *descended genetically* from allele $A_{t-1,j,d}$. We define more distant ancestry recursively: allele $A_{t,i,c}$ is descended from allele $A_{t',j,d}$ if $t > t'$ and there exists a $k$ and $e$ such that allele $A_{t,i,c}$ is descended genetically from allele $A_{t-1,k,e}$ and allele $A_{t-1,k,e}$ is descended from or is the same as allele $A_{t',j,d}$.

One can make a similar recursive definition of genealogical ancestry that matches Chang's notion of ancestry: individual $I_{t,i}$ is *descended genealogically* from individual $I_{t',j}$ if $t > t'$ and there exists a $k$ such that individual $I_{t,i}$ is a parent of individual $I_{t-1,k}$ and individual $I_{t-1,k}$ is descended from or is the same as individual $I_{t',j}$.

Define $\mathcal{Q}_t^i$ to be the alleles that are genetic descendants at time $t$ of the two alleles present in individual $I_{0,i}$, and let $Q_t^i$ be the number of such alleles. We will call the elements of $\mathcal{Q}_t^i$ the *descendant alleles* of individual $I_{0,i}$. Define $\mathcal{G}_t^i$ to be the genealogical descendants at time $t$ of the individual $I_{0,i}$, and let $G_t^i$ be the number of such individuals. We will say that a (genealogical) most recent common ancestor (MRCA) first appears at time $t$ if there is an individual $I_{0,i}$ in the population at time 0 such that $G_t^i = n$ and $G_s^j < n$ for all $j$ and $s < t$; that is, individual $i$ in generation 0 is a genealogical ancestor of all individuals in generation $t$, but there is no individual in generation 0 that is a a genealogical ancestor of all individuals in any generation previous to generation $t$. Let $T_n$ denote the generation number at which the MRCA first appears. The main conclusion of Chang's 1999 paper is that the ratio $T_n/\log_2 n$ converges to one in probability as $n$ tends to infinity.

Our intent is to investigate the degree to which genealogical ancestry implies genetic ancestry. Unsurprisingly, historical individuals with more genealogical descendants will have more descendant alleles in expectation: in Proposition 1 we show that $\mathbb{E}[Q_t^i \mid G_t^i = k]$ is a super-linearly increasing function in $k$. However, in any realization of the stochastic process, individuals with more genealogical descendants need not have more descendant alleles. For example, in Fig. 1 we show a case where the MRCA has no genetic relationship to any present day individuals. In the above notation, $G_2^4 = n = 4$ and yet $Q_2^4 = 0$.

Another approach is based on the rank of $G_t^i$. Loosely speaking, we are interested in the number of descendant alleles of the generation-$t$ individual with the $x$th most genealogical descendants. More rigorously, we consider the renumbering (opposite to the way rank is typically defined in statistics) $F(t, 1), \ldots, F(t, n)$ of the indices $1, \ldots, n$ such that

$$G_t^{F(t,1)} \geq \cdots \geq G_t^{F(t,n)}$$

and if $G_t^{F(t,i)} = G_t^{F(t,j)}$ then fix $F(t, i) < F(t, j)$ when $i < j$. We then investigate $\left\{Q^{F(t,k)} : 1 \leq k \leq n\right\}$. These quantities give us concrete information about our main question in a relative sense: how much do individuals with many genealogical descendants contribute to the genetic makeup of present-day individuals compared to those with only a few? In Fig. 2 we simulate our process 10 000 times and then take an average for each time step, approximating $\mathbb{E}\left[Q_t^{F(t,k)}\right]$.

After several generations, the curve depicting $\mathbb{E}\left[Q_t^{F(t,k)}\right]$ acquires a characteristic shape which persists for some time, in this figure between time 3 and time 8. In order to explain what this curve is, we need to introduce some elementary facts about branching processes.

Recall that a branching process is a discrete time Markov process that tracks the population size of an idealized population (Athreya and Ney, 1972; Grimmett and Stirzaker, 2001). Each individual of generation $t$ produces an independent random number of offspring in generation $t + 1$ according to some fixed probability distribution (the *offspring distribution*). This distribution is the same across all individuals. We will use the Poisson(2) branching process where the offspring distribution is Poisson with mean 2 and write $B_t$ for the number of individuals in the $t$th generation starting with one individual at time $t = 0$. It is a standard fact that the random variables $W_t = B_t/2^t$ converges almost surely as $t \to \infty$ to a random variable $W$ that is strictly positive on the event that the branching process doesn't

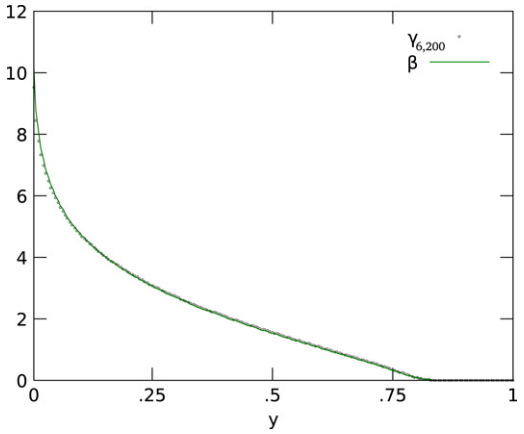


**Fig. 3.** A plot of $\gamma_{6,200}$ and $\beta$, showing experimentally that the characteristic shape in Fig. 2 is very close to the "tail-quantile" curve of a normalized Poisson(2) branching process. The curve for $\gamma_{6,200}$ was taken from Fig. 2. To construct the curve for $\beta$, we wrote a subroutine that simulated 200 Poisson(2) branching processes simultaneously, then sorted the normalized results after 10 generations. This subroutine was run 10 000 times and the average was taken. Note that the distribution had stabilized after 10 generations.

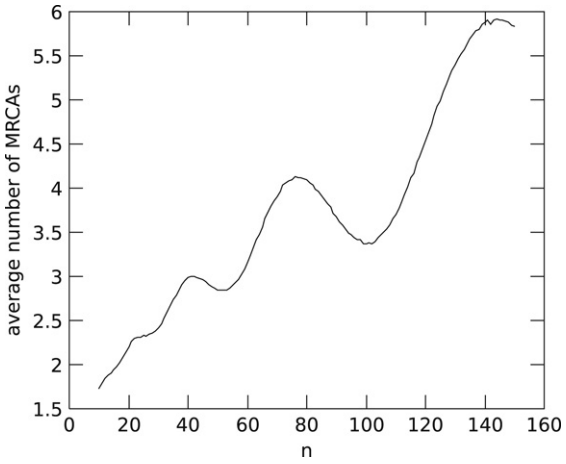


**Fig. 4.** The dependence of the expected number of MRCAs on population size. Average of 10 000 simulations.

die out (that is, on the event that $B_t$ is strictly positive for all $t \geq 0$)—cf. Theorem 8.1 of Athreya and Ney (1972). Denote by $R$ the distribution of the limit random variable $W$. The probability measure $R$ is diffuse except for an atom at 0 (that is, 0 is the only point to which $R$ assigns non-zero mass). Also, the support of $R$ is the whole of $\mathbb{R}_+$ (that is, every open sub-interval of $\mathbb{R}_+$ is assigned strictly positive mass by $R$).

Returning to our discussion of $\mathbb{E}\left[Q_t^{F(t,k)}\right]$, define a non-increasing function $\gamma_{t,n}(c) : (0, 1) \to \mathbb{R}_+$ by

$$\gamma_{t,n}(c) = \mathbb{E}\left[Q_t^{F(t,\lfloor cn \rfloor)}\right],$$

and define a non-increasing, continuous function $\beta : (0, 1) \to \mathbb{R}_+$ by

$$\begin{aligned}\beta(c) &= \min\{r \geq 0 : R((r/2, \infty)) \leq c\} \\ &= \min\{r \geq 0 : R([0, r/2]) \geq 1 - c\}. \end{aligned} \tag{1}$$

That is, $\beta(c)$ is the $(1 - c)^{\text{th}}$ quantile of $2W$, where the random variable $W$ is the limit of the normalized Poisson(2) branching process introduced above. Note that the function $\beta$ is strictly decreasing on the interval $(0, 1-R(\{0\}))$; that is, $\beta(c)$ is the **unique** value $r$ for which $R((r/2, \infty)) = c$ when $0 < c < 1 - R(\{0\})$. We see experimentally that $\gamma_{6,200}$ is quite close to $\beta$ in Fig. 3, and establish a convergence result in Proposition 2. Although a closed-form expression for the distribution $R$ is not available, there is a considerable amount known about this classical object (Van Mieghem, 2005). Note that the long-time behavior in Fig. 2 is easily explained: it is simply the uniform distribution across only the common ancestors, that form $1 - e^{-2} \simeq 0.864$ of the population.

Thus far we have examined the connection between genealogical ancestry and genetic ancestry in the population as a whole; one may wonder about the number of descendants of the MRCA itself. Unfortunately, the story there is not as simple as could be desired. For example, there are usually multiple MRCAs appearing (by definition) in the same generation, and the expected number depends on $n$ in a surprising way (see Fig. 4). We investigate this genealogical issue and related genetic questions in Section 4.

## 2. Monotonicity of the number of descendant alleles in terms of genealogy

In this section we prove the following result.

**Proposition 1.** *For each time $t \geq 2$, the function $k \mapsto k^{-1}\mathbb{E}[Q_t^i \mid G_t^i = k]$, $0 \leq k \leq n$, is strictly increasing.*

The key observation in the proof of Proposition 1 will be that the random variables $G_t^i$ and $G_{t+1}^i$ enjoy the property of *total positivity* investigated extensively in the statistical literature following Karlin (1968) (see, for example, Brown et al. (1981)).

**Definition 1.** A pair of random variables $(X, Y)$ has a *strict* TP(2) joint distribution if

$$\begin{aligned}&\mathbb{P}\{X = x, Y = y\}\,\mathbb{P}\{X = x', Y = y'\} \\ &\quad > \mathbb{P}\{X = x, Y = y'\}\,\mathbb{P}\{X = x', Y = y\}\end{aligned}$$

for all $x < x'$ and $y < y'$ such that the left-hand side is strictly positive.

The proof of the next result is clear.

**Lemma 2.** *The following are equivalent to strict TP(2) for $x < x'$ and $y < y'$:*

$$\frac{\mathbb{P}\{Y = y' \mid X = x'\}}{\mathbb{P}\{Y = y' \mid X = x\}} > \frac{\mathbb{P}\{Y = y \mid X = x'\}}{\mathbb{P}\{Y = y \mid X = x\}} \tag{2}$$

$$\frac{\mathbb{P}\{X = x' \mid Y = y'\}}{\mathbb{P}\{X = x' \mid Y = y\}} > \frac{\mathbb{P}\{X = x \mid Y = y'\}}{\mathbb{P}\{X = x \mid Y = y\}}. \tag{3}$$

**Lemma 3.** *The pair $(G_t^i, G_{t+1}^i)$ has a strict TP(2) joint distribution.*

**Proof.** We will show condition (2). By definition of our model, the number of genealogical descendants in generation $t + 1$ has a conditional binomial distribution as follows:

$$\begin{aligned}&\mathbb{P}\{G_{t+1}^i = k \mid G_t^i = r\} \\ &\qquad = \binom{n}{k}\left(2r/n - (r/n)^2\right)^k\left(1 - 2r/n + (r/n)^2\right)^{n-k}.\end{aligned}$$

Set $x(r) = 2r/n - (r/n)^2$, a function that is strictly increasing in $r$ for $0 \leq r \leq n$. Then

$$\frac{\mathbb{P}\{G_{t+1}^i = k + 1 \mid G_t^i = r\}}{\mathbb{P}\{G_{t+1}^i = k \mid G_t^i = r\}} = \frac{n - k}{k + 1} \cdot \frac{x(r)}{1 - x(r)},$$

a function that is a strictly increasing function of $r$ for $1 \leq r \leq n$. ■

The following definition is well known to statisticians (Lehmann, 1986).

**Definition 4.** Consider a reference measure $\mu$ on some space $\mathcal{X}$ and a parameterized family $\{p_\theta : \theta \in \Theta\}$ of probability densities with respect to $\mu$, where $\Theta$ is a subset of $\mathbb{R}$. Let $T$ be a real-valued function defined on $\mathcal{X}$. The family of densities has the *monotone likelihood ratio* property in $T$ with respect to the parameter $\theta$ if for any $\theta' < \theta''$ the densities $p_{\theta'}$ and $p_{\theta''}$ are distinct and $x \mapsto p_{\theta''}(x)/p_{\theta'}(x)$ is a nondecreasing function of $T(x)$.

**Lemma 5.** *Fix a time $t \geq 0$. If the function $f : \mathbb{R} \to \mathbb{R}$ is strictly increasing, then the function $k \mapsto \mathbb{E}[f(G_t^i) \mid G_{t+1}^i = k]$, $1 \leq k \leq n$, is strictly increasing.*

**Proof.** By Lemma 3 and inequality (3), the family of probability densities (with respect to counting measure) $\mathbb{P}\left\{G_t^i = r \mid G_{t+1}^i = k\right\}$ parameterized by $k$ has monotone likelihood ratios in $r$ with respect to $k$. Now apply Lemma 2(i) of Lehmann (1986). ■

**Proof of Proposition 1.** For $t \geq 1$, let $\alpha_t(k) = k^{-1}\mathbb{E}[Q_t^i \mid G_t^i = k]$.

First note that each individual of $\mathcal{G}_1^i$ has a $1/n$ chance of choosing $I_{0,i}$ as a parent twice, thus

$$\alpha_1(k) = (1 + 1/n).$$

The result will thus follow by induction on $t$ if we can show for $t \geq 1$ that the function $k \mapsto \alpha_{t+1}(k)$ is strictly increasing whenever the function $k \mapsto \alpha_t(k)$ is non-decreasing. Therefore, fix $t \geq 1$ and suppose that the function $k \mapsto \alpha_t(k)$ is non-decreasing.

We first claim that

$$\begin{aligned} k^{-1}\mathbb{E}[Q_{t+1}^i \mid G_t^i = r, G_{t+1}^i = k] &= (r^{-1} + n^{-1})\mathbb{E}[Q_t^i \mid G_t^i = r] \\ &= f(r), \end{aligned} \tag{4}$$

where

$$f(r) = \left(r^{-1} + n^{-1}\right)\mathbb{E}[Q_t^i \mid G_t^i = r] = \left(1 + \frac{r}{n}\right)\alpha_t(r).$$

The proof of this claim is as follows.

Recall that $\mathcal{G}_t^i$ is the set of generation $t$ individuals descended from $I_{0,i}$, so that $\mathcal{G}_t^i$ has $G_t^i$ elements. Suppose that $G_t^i = r$ and number the elements of $\mathcal{G}_t^i$ as $1, \ldots, r$. Let $V_{j,c}$ be the indicator random variable for the event that the allele $A_{t,j,c}$ is descended from one of the alleles of $I_{0,i}$ for $1 \leq j \leq r$. By definition, the sum of the $V_{j,c}$ is equal to $Q_t^i$. Note that any individual in $\mathcal{G}_{t+1}^i$ has one parent uniformly selected from $\mathcal{G}_t^i$ and the other uniformly selected from the population as a whole. Selections for different individuals are independent. Therefore,

$$\begin{aligned} &\mathbb{E}[Q_{t+1}^i \mid G_t^i = r, G_{t+1}^i = k, V_{1,1}, V_{1,2}, \ldots, V_{r,1}, V_{r,2}] \\ &= k\left[\frac{1}{r}\sum_{\ell=1}^{r}(V_{\ell,1} + V_{\ell,2}) + \frac{1}{n}\left(\sum_{\ell=1}^{r}(V_{\ell,1} + V_{\ell,2}) + \sum_{\ell=r+1}^{n} 0\right)\right] \\ &= k(r^{-1} + n^{-1})Q_t^i. \end{aligned}$$

By the tower property of conditional expectation,

$$\begin{aligned} \mathbb{E}[Q_{t+1}^i \mid G_t^i = r, G_{t+1}^i = k] &= k(r^{-1} + n^{-1}) \\ &\quad \times \mathbb{E}[Q_t^i \mid G_t^i = r, G_{t+1}^i = k]. \end{aligned}$$

An application of the Markov property now establishes our claim (4). Thus

$$\begin{aligned} k^{-1}\mathbb{E}[Q_{t+1}^i \mid G_{t+1}^i = k] &= \sum_{r=1}^{n} k^{-1}\mathbb{E}[Q_{t+1}^i \mid G_t^i = r, G_{t+1}^i = k] \\ &\quad \times \mathbb{P}\{G_t^i = r \mid G_{t+1}^i = k\} \\ &= \sum_{r=1}^{n} f(r)\mathbb{P}\{G_t^i = r \mid G_{t+1}^i = k\} \\ &= \mathbb{E}[f(G_t^i) \mid G_{t+1}^i = k]. \end{aligned}$$

This is strictly increasing in $k$ by Lemma 5 and the observation that $f$ is strictly increasing. ■

## 3. The mysterious shape in Fig. 2

In this section we investigate the shape of the curve relating the number of descendant alleles to genealogical rank. As shown in Fig. 2, this curve attains a characteristic shape after several generations; the shape is maintained for a period prior to the time when the genealogical MRCA appears. We show that this curve is essentially the limiting "tail-quantile" of a normalized Poisson(2) branching process.

An important component of our analysis will be a multigraph representing ancestry that we will call the *genealogy*. A multigraph is similar to a graph except that multiple edges between pairs of nodes are allowed. Specifically, a multigraph is an ordered pair $(V, E)$ where $V$ is a set of nodes and $E$ is a multiset of unordered pairs of nodes.

**Definition 6.** Define the time $t$ *ancestry multigraph* $\mathcal{G}_t$ as follows. The nodes of this multigraph are the set of all individuals of generations zero through $t$; for any $0 < t' \leq t$ connect an $I_{t',k}$ to $I_{t'-1,j}$ if $I_{t',k}$ is descended from $I_{t'-1,j}$. If both parents of $I_{t',k}$ are $I_{t'-1,j}$, then add an additional edge connecting $I_{t',k}$ and $I_{t'-1,j}$. Define the time $t$ *genealogy* $\mathcal{G}_t^i$ to be the subgraph of $\mathcal{G}_t$ consisting of $I_{0,i}$ and all of its descendants $\bigcup_{t'=0}^{t} \mathcal{G}_{t'}^i$ up to time $t$.

**Definition 7.** We define an ancestry path in $\mathcal{G}_t^i$ to be a sequence of individuals $I_{0,i(0)}, I_{1,i(1)}, \ldots, I_{t,i(t)}$ with $i(0) = i$ where for each $0 < t' \leq t$, $I_{t'-1,i(t'-1)}$ is a parent of $I_{t',i(t')}$. Let $P_t^i$ be the number of ancestry paths in $\mathcal{G}_t^i$.

We emphasize that a parent being selected twice by a single individual results in a "doubled" edge; paths that differ only in their choice of what edge to traverse between parent to child are considered distinct. Thus, each such doubled edge doubles the number of ancestry paths that contain the corresponding parent-child pair.

Our result concerning the connection between the curve in Fig. 2 and the Poisson(2) branching process can be stated as follows. Define a random probability measure on the positive quadrant that puts mass $1/n$ at each of the points $(\mathbb{E}[Q_t^i|\mathcal{G}_t], 2^{1-t}G_t^i)$. We show below that this random probability measure converges in probability to a deterministic probability measure concentrated on the diagonal and has projections onto either axis given by the limiting distribution of $2^{1-t}B_t$ as $t \to \infty$.

We may describe the convergence more concretely by using the idea of "sorting by the number of genealogical descendants" as in the introduction; using the notation introduced there, let the random variable $F(t, k)$ denote the index of the individual in generation 0 with the $k$th greatest number of genealogical descendants at time $t$. Recall the non-increasing, continuous function $\beta : (0, 1) \to \mathbb{R}_+$ defined in Eq. (1).

**Proposition 2.** *Suppose that $0 < a < b < 1 - R(\{0\})$, so that $\infty > \beta(a) > \beta(b) > 0$. Then*

$$\frac{1}{n(b-a)} \cdot \#\left\{an \leq k \leq bn : \mathbb{E}\left[Q_t^{F(t,k)} \,\middle|\, \mathcal{G}_t\right] \in [\beta(b), \beta(a)]\right\}$$

*converges to 1 in probability as $t = t_n$ and $n$ go to infinity in such a way that $2^{2t_n}/n \to 0$.*

Note that the condition $2^{2t_n}/n \to 0$ is satisfied, for example, when $t_n = \tau \log_2 n$ for $\tau < 1/2$.

The proof of Proposition 2 formalizes the following three common-sense notions about the ancestry process.

Note that for $t > 1$, the genealogy will not necessarily be a tree: it may be possible to follow two different ancestry paths through $\mathcal{G}_t^i$ to a given time-$t$ individual. However, our first intuition is that

this possibility is rare when $n$ is large and $t$ is small relative to $n$, and such events do not affect the values of $G_t^i$ and $Q_t^i$ in the limit.

Second, the fact that each of the above genealogies is usually a tree suggests that we may be able to relate the ancestry process to a branching process. In our case, the number of immediate descendants for an individual $I_{t'-1,j}$ is the number of times a individual of generation $t'$ chooses $I_{t'-1,j}$ as a parent. These numbers are not exactly independent: for example, if all of the individuals of generation $t'$ descend only from a single individual of generation $t'-1$, then the number of descendants of the other individuals is exactly zero. However, we will show that these numbers are close to independent when $n$ becomes large. Also, note that the marginal distribution of the number of next-generation descendants of a single individual is binomial: there are $2n$ trials each with probability $1/n$. As $n$ goes to infinity, this is approximately a Poisson(2) random variable. In summary, we will show that the genealogy of an individual is close to that of a Poisson(2) branching process for short times relative to the population size.

Third, we note that there is a simple relationship between the number of paths $P_t^i$ and the expected number of descendant alleles $Q_t^i$:

**Lemma 8.** $\mathbb{E}\left[Q_t^i \mid \mathfrak{G}_t^i\right] = 2^{1-t}P_t^i$.

**Proof.** Consider an arbitrary path in the ancestry graph $\mathfrak{G}_t^i$ and pick an arbitrary edge in that path. Suppose the edge connects $I_{t'-1,j}$ to $I_{t',i}$. By the definition of the model, $I_{t',i}$ has probability $1/2$ of inheriting any fixed allele of $I_{t'-1,j}$. Thus, the contribution of any single allele of $I_{0,i}$ and given path in $\mathfrak{G}_t^i$ to the expectation of $Q_t^i$ is $2^{-t}$. The contribution of both alleles of $I_{0,i}$ is $2^{1-t}$. The total number of alleles descended from the alleles of $I_{0,i}$ is the sum over the contributions of all paths, and the expectation of this sum is the sum of the expectations. ■

We will use the probabilistic method of coupling to formalize the connection between the genealogical process and the branching process. A coupling of random variables $X$ and $Y$ that are not necessarily defined on the same probability space is a pair of random variables $X'$ and $Y'$ defined on a single probability space such that the marginal distributions of $X$ and $X'$ (respectively, $Y'$ and $Y$) are the same. A simple example of coupling is "Poisson thinning", a coupling between an $X \sim \text{Poisson}(\lambda_1)$ and a $Y \sim \text{Poisson}(\lambda_2)$ where $\lambda_1 \geq \lambda_2$. To construct the pair $(X', Y')$, one first gains a sample for $X'$ by simply sampling from $X$. The sample from $Y'$ is then gained by "throwing away" points from the sample for $X'$ with probability $\lambda_2/\lambda_1$; i.e. the distribution for $Y'$ conditioned on the value $x$ for $X'$ is just Binomial$(x, 1-\lambda_2/\lambda_1)$.

We note that coupling is a popular tool for questions with a flavor similar to ours. Recently Barbour (2007) has coupled an epidemics model to a branching process and Durrett et al. (submitted for publication) have used coupling to analyze a model of carcinogenesis.

Recall that we defined $W_t = B_t/2^t$, where $B_t$ is a Poisson(2) branching processes started at time $t = 0$ from a single individual, and we observed that the sequence of random variables $W_t$ converges almost surely to a random variable $W$ with distribution $R$. The following lemma is the coupling result that will give the convergence of the sampling distribution of the $P_t^i$ and $G_t^i$ to $R$ in Lemma 10 below.

**Lemma 9.** *There is a coupling between $P_t^i, G_t^i$, and $B_t^i$, where $B_t^1, B_t^2, \ldots$ is a sequence of independent Poisson(2) branching processes, such that for a fixed positive integer $\ell$ the probability*

$$\mathbb{P}\{P_t^i = G_t^i = B_t^i,\ 1 \leq i \leq \ell\}$$

*converges to one as n goes to infinity with $t = t_n$ satisfying $2^{2t_n}/n \to 0$.*

**Proof.** We introduce the coupling between the ancestral process and the branching process by looking first at the transition from generation 0 to generation 1. Suppose that we designate a set $S$ of $k$ individuals in generation 0 and write $G$ for the number of descendants these $k$ individuals have in generation 1.

The probability that there is an individual in generation 1 who picks both of its parents from the $k$ designated individuals is

$$1 - \left(1 - (k/n)^2\right)^n \leq k^2/n.$$

Couple the random variable $G$ with a random variable $P$ that is the same as $G$ except that we (potentially repeatedly) re-sample any generation 1 individual who chooses two parents from $S$ until it has at least one parent not belonging to $S$. The random variable $P$ will have a binomial distribution with number of trials $n$ and success probability

$$\frac{2\frac{k}{n}\left(1-\frac{k}{n}\right)}{1-\frac{k^2}{n^2}} = \frac{2\frac{k}{n}}{1+\frac{k}{n}}, \tag{5}$$

which is simply the probability of an individual selecting exactly one parent from the set of $k$ given that it does not select two. By the above,

$$\mathbb{P}\{G \neq P\} \leq \frac{k^2}{n}.$$

By a special case of Le Cam's Poisson approximation result (Grimmett and Stirzaker, 2001; Le Cam, 1960), we can couple the random variable $P$ to a random variable $Y$ that is Poisson distributed with mean

$$n\frac{2\frac{k}{n}}{1+\frac{k}{n}} = \frac{2k}{1+\frac{k}{n}}$$

in such a way that

$$\mathbb{P}\{P \neq Y\} \leq n\left(\frac{2\frac{k}{n}}{1+\frac{k}{n}}\right)^2 \leq 4\frac{k^2}{n}.$$

Moreover, a straightforward argument using Poisson thinning shows that we can couple the random variable $Y$ with a random variable $B$ that is Poisson distributed with mean $2k$ such that

$$\mathbb{P}\{Y \neq B\} \leq \left|2k - \frac{2k}{1+\frac{k}{n}}\right| \leq 2\frac{k^2}{n}.$$

Putting this all together, we see that we can couple the random variables $G$, $P$, and $B$ together in such a way that

$$\mathbb{P}(\neg\{G = P = B\}) \leq 8\frac{k^2}{n}$$

where $\neg$ denotes complement. Note that $B$ may be thought of as the sum of $k$ independent random variables, each having a Poisson distribution with mean 2.

Fix an index $i$ with $1 \leq i \leq n$. Returning to the notation used in the rest of the paper, the above triple $(G, P, B)$ correspond to $(G_t^i, P_t^i, B_t^i)$, and $k$ plays the role of $G_{t-1}^i$. Now suppose we start with one designated individual $i$ in the population at generation 0. Let $S_t$ denote the event

$$\{P_t^i = G_t^i = B_t^i\}.$$

The above argument shows that we can couple the process $P^i$ with the branching process $B^i$ in such a way that

$$\begin{aligned}\mathbb{P}\{\neg S_t\} &\leq \mathbb{P}\{\neg S_{t-1}\} + \mathbb{P}\{\neg S_t, S_{t-1}\} \\ &\leq \mathbb{P}\{\neg S_{t-1}\} + \mathbb{E}\left[8\frac{B_{t-1}^2}{n}\right] \\ &\leq \mathbb{E}\{\neg S_{t-1}\} + \frac{c2^{2(t-1)}}{n}\end{aligned}$$

for a suitable constant $c$ (using standard formulae for moments of branching processes). Iterating this bound gives

$$\mathbb{P}\{\neg S_t\} \leq \frac{c' 2^{2t}}{n}$$

for a suitable constant $c'$.

This tells us that when $n$ is large, the random variable $P_t^i$ is close to the random variable $B_t^i$ not just for fixed times but more generally for times $t$ such that $2^{2t}/n \to 0$. As mentioned above, this condition is satisfied when $t = \tau \log_2 n$ for $\tau < 1/2$.

Next, we elaborate the above argument to handle the descendants of $\ell$ individuals. Let $S_t^\ell$ denote the event that the $\ell$ coupled triples of random variables are equal, that is,

$$\{P_t^1 = G_t^1 = B_t^1, P_t^2 = G_t^2 = B_t^2, \ldots, P_t^\ell = G_t^\ell = B_t^\ell\},$$

where $B_t^i$ is the branching process coupled to $P_t^i$ and $G_t^i$. By mimicking the above argument, we can show that

$$\mathbb{P}\{\neg S_t^\ell\} \leq \mathbb{P}\{\neg S_{t-1}^\ell\} + c\ell 2^{2(t-1)}/n.$$

Again, iterating this bound gets

$$\mathbb{P}\{\neg S_t^\ell\} \leq c'\ell 2^{2t}/n$$

for some $c'$. ■

For any Borel subset $C$ of $\mathbb{R}_+^2$, let $\eta_{t,n}(C)$ denote the joint empirical distribution of the normalized $P_t^i$ and the normalized $G_t^i$ at time $t$, i.e.

$$\eta_{t,n}(C) = \frac{1}{n} \cdot \#\{1 \leq i \leq n : (2^{-t}P_t^i, 2^{-t}G_t^i) \in C\}.$$

In Lemma 10 we demonstrate that the $\eta_{t,n}$ converge in probability to the deterministic probability measure $\eta(dx, dy) = R(dx)\delta_x(dy) = \delta_y(dx)R(dy)$ concentrated on the diagonal, where $\delta_z$ denotes the unit point mass at $z$.

The mode of convergence may require a bit of explanation. When we say that a real-valued random variable converges in probability to a fixed quantity, there is an implicit and commonly understood notion of convergence of a sequence of real numbers. However, here the random quantities are probability measures, and the underlying notion we use for convergence of measures is that of weak convergence. Recall that a sequence of probability measures $\mu_n$ on $\mathbb{R}_+^2$ is said to converge to $\mu$ weakly if $\int f d\mu_n$ converges to $\int f d\mu$ for all bounded continuous functions $f : \mathbb{R}_+^2 \to \mathbb{R}$. The following are equivalent conditions for sequence of probability measures $\mu_n$ to converge to $\mu$ weakly: (i) $\limsup_n \mu_n(F) \leq \mu(F)$ for all closed sets $F \subseteq \mathbb{R}_+^2$, (ii) $\liminf_n \mu_n(G) \geq \mu(G)$ for all open sets $G \subseteq \mathbb{R}_+^2$, (iii) $\lim_n \mu_n(A) = \mu(A)$ for all Borel sets $A \subseteq \mathbb{R}_+^2$ such that $\mu(\partial A) = 0$, where $\partial A$ is the boundary of $A$.

**Lemma 10.** *Suppose that $t = t_n$ converges to infinity as $n$ goes to infinity in such a way that $\lim_{n\to\infty} 2^{2t_n}/n \to 0$. Then the sequence of random measures $\eta_{t,n}$ converges in probability as $n \to \infty$ to the deterministic probability measure $\eta$ on $\mathbb{R}_+^2$ that assigns mass $R(A \cap B)$ to sets of the form $A \times B$.*

**Proof.** For brevity, let $H_t^i$ denote the pair $(2^{-t}P_t^i, 2^{-t}G_t^i)$. Fix a bounded continuous function $f : \mathbb{R}_+^2 \to \mathbb{R}$. By definition,

$$\mathbb{E}\left[\left(\int f d\eta_{t,n}\right)^2\right] = n^{-2}\mathbb{E}\left[\sum_i f^2(H_t^i) + \sum_{i\neq j} f(H_t^i)f(H_t^j)\right]$$
$$= n^{-2}\left(n\mathbb{E}[f^2(H_t^1)] + n(n-1)\mathbb{E}[f(H_t^1)f(H_t^2)]\right).$$

Hence, $\mathbb{E}[(\int f d\eta_{t,n})^2]$ is asymptotically equivalent to

$$\mathbb{E}[f(H_t^1)f(H_t^2)]. \tag{6}$$

By definition, (6) is equal to

$$\mathbb{E}[f(2^{-t}P_t^1, 2^{-t}G_t^1)f(2^{-t}P_t^2, 2^{-t}G_t^2)]. \tag{7}$$

Lemma 9 establishes a coupling such that $P_t^i = G_t^i = B_t^i$ with probability tending to one in the limit under our hypotheses. Thus, under our conditions on $t = t_n$ the expectation (7), and hence $\mathbb{E}[(\int f d\eta_{t,n})^2]$, converges to

$$\lim_{t\to\infty} \mathbb{E}[f(W_t^1)f(W_t^2)] = \lim_{t\to\infty} \mathbb{E}[f(W_t^1, W_t^1)]\mathbb{E}[f(W_t^2, W_t^2)]$$
$$= \left(\int_{\mathbb{R}_+} f(x,x)R(dx)\right)^2$$
$$= \left(\int_{\mathbb{R}_+^2} f d\eta\right)^2.$$

A similar but simpler argument shows that $\mathbb{E}[\int f d\eta_{t,n}]$ converges to $\int f d\eta$. Combining these two facts shows that $\mathrm{Var}[\int f d\eta_{t,n}]$ converges to zero.

Therefore, $\int f d\eta_{t,n}$ converges in probability to $\int f d\eta$ for all bounded continuous functions $f$, as required. ■

**Proof of Proposition 2.** It suffices by Lemma 8 to show that

$$\frac{1}{n(b-a)} \cdot \#\left\{an \leq k \leq bn : 2^{-t}P_t^{F(t,k)} \in [2\beta(b), 2\beta(a)]\right\}$$

converges to 1 in probability as $t = t_n$ and $n$ go to infinity in such a way that $2^{2t_n}/n \to 0$.

For $\gamma > 0$ and an integer $1 \leq k \leq n$,

$$\eta_{t,n}(\mathbb{R}_+ \times [\gamma, \infty)) \geq \frac{k}{n}$$
$$\Leftrightarrow \#\{1 \leq i \leq n : 2^{-t}G_t^i \geq \gamma\} \geq k$$
$$\Leftrightarrow 2^{-t}G_t^{F(t,k)} \geq \gamma,$$

by definition of the empirical distribution $\eta_{t,n}$ and the indices $F(t,k)$. Because the limit measure $\eta$ assigns zero mass to the boundary $\mathbb{R}_+ \times \{\gamma\}$ of the set $\mathbb{R}_+ \times [\gamma, \infty)$, it follows from Lemma 10 that

$$\frac{1}{n} \cdot \#\{1 \leq i \leq n : 2^{-t}G_t^i \geq \gamma\}$$

converges to $\eta(\mathbb{R}_+ \times [\gamma, \infty)) = R([\gamma, \infty))$ in probability. In particular,

$$\frac{1}{n} \cdot \#\{1 \leq i \leq n : 2^{-t}G_t^i \geq 2\beta(c)\}$$

converges to $c$ in probability for $0 < c < 1 - R(\{0\})$. Thus, $2^{-t}G_t^{F(t,\lfloor cn \rfloor)}$ converges in probability to $2\beta(c)$ for such a $c$.

With $0 < a < b < 1 - R(\{0\})$ as in the statement of the proposition, it follows that

$$\frac{1}{n} \cdot \#\left\{an \leq k \leq bn : 2^{-t}G_t^{F(t,k)} \in [2\beta(b-\epsilon), 2\beta(a+\epsilon)]\right\}$$

converges in probability to $(b - a - 2\epsilon)$ for $0 < \epsilon < (b-a)/2$.

Note by Lemma 10 that

$$\frac{1}{n} \cdot \#\left\{1 \leq k \leq n : \left|2^{-t}P_t^{F(t,k)} - 2^{-t}G_t^{F(t,k)}\right| > \delta\right\}$$
$$= \eta_{t,n}(\{(x,y) \in \mathbb{R}_+^2 : |x-y| > \delta\})$$

converges in probability to 0 for any $\delta > 0$, because the probability measure $\eta$ assigns all of its mass to the diagonal $\{(x,y) \in \mathbb{R}_+^2 : x = y\}$.

Taking $\delta < 2\min\{\beta(a) - \beta(a+\epsilon), \beta(b-\epsilon) - \beta(b)\}$ so that

$$[2\beta(b-\varepsilon) - \delta, 2\beta(a+\varepsilon) + \delta] \subseteq [2\beta(b), 2\beta(a)],$$

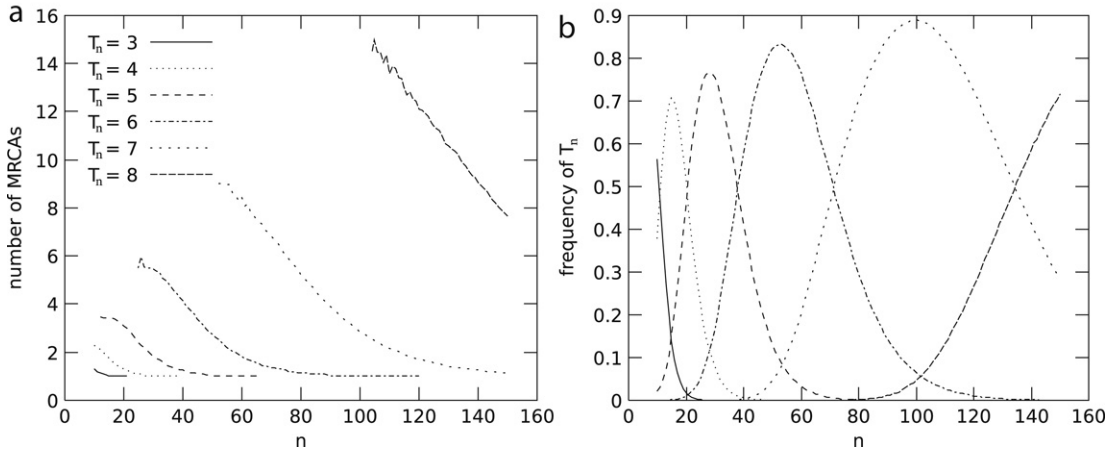


**Fig. 5.** The number of MRCAs where the dependence on $T_n$ (the time to MRCA) is taken into account. Average of 10 000 simulations. Figure (a) shows the number of MRCAs at $T_n$ conditioned on $T_n$. Figure (b) shows the dependence of the distribution of times to MRCA on population size. As described in the text, it is the combination of these two distributions using the law of total expectation that produces the "bumps" of Fig. 4. Note that several simulations with "extreme" values of $T$ have been eliminated from (a) for clarity; these combinations of $T_n$ and population size are rare and thus we would not get an accurate estimate of the expectation.

letting $n$ tend to infinity, and then sending $\epsilon$ to zero completes the proof. ■

As an application of this proposition, one might wonder about the number of descendant alleles of those individuals with many genealogical descendants. It is imaginable that the number of descendant alleles of each individual would stay bounded; however, this is not the case.

**Corollary 1.** *Fix $y > 0$, and suppose $t = t_n$ satisfies $\lim_{n\to\infty} 2^{2t_n}/n \to 0$. With probability tending to one as n goes to infinity, there will be an individual i in the population at time 0 such that $\mathbb{E}\left[Q_t^i|\mathfrak{G}_t\right] > y$.*

**Proof.** Because the support of the probability distribution $R$ is all of $\mathbb{R}_+$, the function $\beta$ is unbounded. The result is then immediate from Proposition 2. ■

## 4. The number of MRCAs and the number of descendant alleles per MRCA

There are a number of other interesting phenomena that seem more difficult to investigate analytically but are interesting enough to deserve mention. For the simulations of this section (and the one mentioned in the introduction) we wrote a series of simple `ocaml` programs which are available upon request.

As mentioned in the introduction, it is not uncommon to get several genealogical MRCAs simultaneously. We denote the (random) time to achieve a genealogical MRCA for a population of size $n$ by $T_n$. We denote the (random) number of genealogical MRCAs for a population of size $n$ by $M_n$. The surprising dependence of $\mathbb{E}[M_n]$ on $n$ is shown in Fig. 4.

However, the situation becomes clear by investigating the conditional expectation $\mathbb{E}[M_n|T_n]$ as shown in Fig. 5. According to the law of total expectation, one can gain the expectation by taking the sum of conditional expectations weighted by their probability. In this setting,

$$\mathbb{E}[M_n] = \sum_k \mathbb{E}[M_n \mid T_n = k]\mathbb{P}\{T_n = k\}. \tag{8}$$

First note in Fig. 5(a) that $\mathbb{E}[M_n|T_n = k]$ appears to be a decreasing function of $n$ when $k$ is fixed. This is not too surprising: imagine that we are doing simulations with $n$ individuals, but only looking at the results of simulations such that $T_n = k$. When $n$ gets large, simulations such that $T_n = k$ are ones which take an unusually short time to reach $T$. It's not surprising to find that the number of MRCAs would be small in this case. Conversely, simulations such that $T_n$ is significantly bigger than $\log_2 n$ are ones that take an unusually long time; it is not surprising that such simulations have a larger number of MRCAs as they have more individuals "ready" to become MRCAs just before $T_n$. This argument is bolstered by Fig. 6 which shows that simulations resulting in different $T_n$'s have remarkably similar behavior. Specifically, the distribution of the number of genealogical descendants sorted by rank does not show a very strong dependence on the time to most recent common ancestor $T_n$. Therefore simulations for a given population size that have a smaller $T_n$ have fewer individuals who are close to being MRCAs while individuals with larger $T_n$ have more.

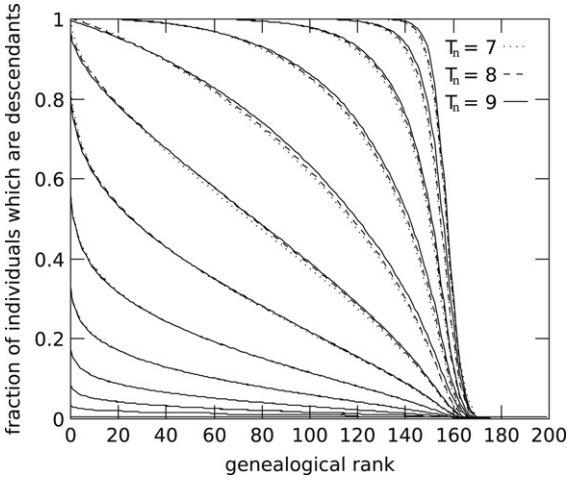


**Fig. 6.** A plot of $\mathbb{E}\left[G_t^{F(t,k)}|T_n\right]$ through time conditioning on $T_n$. Each curve for a given choice of $T_n$ represents the expected state of the process at a given time. That is, each curve represents the image of the map $k \mapsto \mathbb{E}\left[G_t^{F(t,k)}\middle| T_n\right]$ for some choice of $t$ and $T_n$. As described in the text, the curves show surprisingly little dependence on $T_n$, rather depending almost exclusively on $t$. Average of 10 000 simulations with $n = 200$.

Second, note in Fig. 5(b) that the distribution of $T_n$ has bumps such that (at least for integers $k > 3$), there is an interval of $n$ such that $\mathbb{P}\{T_n = k\}$ is large in that interval. In such an interval we are approximately on a single line of Fig. 5(a), that is, $\mathbb{E}[M_n]$ is approximately $\mathbb{E}[M_n \mid T_n = \kappa_n]$ where $\kappa_n$ is the most likely