Time-conditional properties of branches in coalescent gene trees

Alexander Platt[1,*]

[1]Temple University, Department of Biology, Center for Computational Genetics and Genomics, Philadelphia, 19123, USA
[*]alex@alexanderplatt.org

ABSTRACT

Coalescent gene trees have proven to be a powerful framework for formulating and solving problems in population genetics both in theory and practice. Using them, geneticists have been able to generate expectations for many attributes of a random sample of genotypes from a population given a model of the history of the population. This paper derives three new properties of coalescent gene trees that will help characterize the present-day impacts of historical events. Considering a single branch sampled at a given time t_s in the past, it presents distributions describing 1) the length of time a branch sampled t_s generations in the past had existed at the time of sampling, 2) the length of time that branch continues from time t_s towards the present, and 3) the the probability that the branch is ancestral to x individuals in a modern sample.

Introduction

The introduction of coalescent theory to the field of population genetics brought with it two key insights: that modeling the genealogy of a sample backwards in time to the point of common ancestry would eliminate the need to keep track of an entire population's worth of lineages and that properties of the shape of the resulting gene tree would be robust to a wide range of generative evolutionary models[1]. Where most previous results focus on properties of distributions of times of coalescent events across the entire tree, the properties considered here are those of individual lineages as a function of their positions within a tree. This is accomplished by augmenting the traditional backwards-looking approach to coalescent analysis with forward-looking death models to describe four new properties of coalescent tree branches conditional on their age within the tree. These properties are illustrated in Figure 1, and all are concerned with properties of an arbitrary branch $\mathscr{B}$ identified in the genealogy of the sample at time t_s before the sample was taken. The first result is the probability density function of t_a, the time of the most oldest end of branch $\mathscr{B}$. The second result is a derivation of the probability density function of t_d, the time of the most recent end of branch $\mathscr{B}$. The third result is a derivation of the probability mass function of x, the number of individuals in the sample that are descendants of branch $\mathscr{B}$, with particular attention paid to the probability that $x = 1$, a special case where branch $\mathscr{B}$ is considered to be external on the tree. Together, these describe properties of the coalescent tree that are critical to our ability to answer questions about how different parts of the tree relate to each other, how different individuals and groups of individuals in a sample relate to each other, and what kinds of signals historic events may have left in contemporary or archaic samples.

Results

Length of $\mathscr{B}$ from t_s to t_a

Caliebe *et al.*[2] designate the length of a randomly chosen external branch as Z_n in time measured in coalescent units (equal to $2N$ generations for a population of constant size). They do not present a density function for Z_n directly but rather the limiting case of the product of the length of a random branch and n, the number of branches sampled:

$$\lim_{n\to\infty} p(nZ_n = x) = \frac{8}{(2+x)^3}. \tag{1}$$

For sufficiently large values of n, a change of variable transformation letting $y = \frac{x}{n}$ yields

$$p(Z_n = y) = \frac{8}{(2+ny)^3}\frac{dx}{dy}$$

=

$$8n(2+ny)^3. \tag{2}$$

The general distribution of time until the next coalescent event involving a random branch after time τ is then $p(Z_{n(\tau)} = y) = \frac{8n(\tau)}{(2+n(\tau)y)^3}$ where $n(\tau)$ is the number of lineages extant at time τ (measured in coalescent units). The full probability distribution for the time-conditional number of lineages is an infinite sum of terms of alternating sign[3] and is cumbersome to work with. For even fairly small values of n, however, it behaves nearly deterministically[4] and for which numerous approximations exist[5–8]. Most straightforwardly, Slatkin & Rannala[6] propose

$$n(\tau) = \frac{n}{1+n\tau/2}. \tag{3}$$

This approximation produces the result

$$p(Z_{n(\tau)} = y) \approx \frac{2n(n\tau+2)^2}{(n\tau+ny+2)^3}. \tag{4}$$

To get an answer in terms of generations requires another transformation of variables where $\frac{t_s}{2N} = y$. This gives

$$p(Z_n = t_s) = \frac{d\frac{t_s}{2N}}{dt_s}\frac{8n}{(2+\frac{nt_s}{2N})^3} = \frac{4n}{N(2+\frac{nt_s}{2N})^3}. \tag{5}$$

For a variable population size, however, the rescaling of time is not a constant multiple of 2N, but rather the coalescent intensity $\tau = \int_0^t \frac{1}{2N(\hat{t})}d\hat{t}$[9]. The change of variable now is $\int_0^{t_s} \frac{1}{2N(t)}dt = y$ giving

$$p(Z_n = t_s) = \frac{d(\int_0^{t_s} \frac{1}{2N(t)}dt)}{dt_s}\frac{8n}{(2+n\int_0^{t_s} \frac{1}{2N(t)}dt)^3}$$

This evaluates to

$$\frac{8n}{2N(t_s)(2+n\int_0^{t_s} \frac{1}{2N(t)}dt)^3}$$

=

$$\frac{4n}{N(t_s)(2+n\tau_{t_s})^3}, \tag{6}$$

where $N(t_s)$ is the population size t_s generations ago and $\tau_{t_s} = \int_0^{t_s} \frac{1}{2N(t)}dt$.

Finally, the time of the ancestral node t_a is simply t_s plus the length of the preceding branch and is distributed as:

$$p(t_a|t_s)t_s + \frac{4n}{N(t_s)(2+n\tau_{t_s})^3} \tag{7}$$

Length of $\mathscr{B}$ from t_s to t_d

The distribution of the time ϕ to the more recent end of a branch identified at time t_s is derived using a hazard model parameterized forwards in time.

The first quantity necessary to derive is the instantaneous rate of coalescence in a population as a function of ϕ. Substituting $t = (t_s - \phi)$ in equation 3 gives an approximation of the number of lineages extant as a function ϕ, and the first derivative of this equation, $n'(\phi)$ gives the instantaneous rate of change of the number of lineages through time. As the number of lineages changes by exactly one for every coalescent event, this is equivalent to the instantaneous rate of coalescence:

$$n'(\phi) = \left(\frac{n}{2N(t_s-\phi)\tau(t_s-\phi)+4}\right)^2. \tag{8}$$

This is the rate of coalescence in the entire population. The only coalescent events of interest, however, are those that would fall along $\mathscr{B}$. The probability that any given coalescent event involves $\mathscr{B}$ is $1/n(\phi)$ as $\mathscr{B}$ always one of $n(\phi)$ lineages. This gives the instantaneous rate of coalescent events on $\mathscr{B}$ as

$$\frac{n'(\phi)}{n(\phi)}, \tag{9}$$

which is referred to as the hazard function $\lambda(\phi)$[10]. When considering k lineages instead of just one, the rate is k times faster and

$$\lambda(\phi|k) = \frac{kn'(\phi)}{n(\phi)}. \tag{10}$$

The probability that an event that occurs with an instantaneous rate $\lambda(x)$ has not happened over the interval $(0,\phi)$ is $e^{-\int_0^\phi \lambda(x)dx}$, a quantity often referred to as the reliability function or $R(\phi)$. The probability that there have been no events on the interval $(0,\phi)$ followed by an event at ϕ is simply the product $f(\phi) = R(\phi)\lambda(\phi)$. Substituting the previous expression for the hazard function gives

$$f(\phi|k,t_s) = \frac{kn'(\phi)e^{-\int_0^\phi (kn'(x)/n(x))dx}}{n(\phi)}. \tag{11}$$

This is the distribution of time until the first coalescent event more recent than t_s within a specified k lineages and is true for $\phi < t_s$. The full distribution gets truncated such that $f(\phi|k,t_s) = 0$ when $\phi > t_s$ and has a point mass at $\phi = t_s$ equal to the $k = 1$ probability from equation (12).

Distribution of descendants of $\mathscr{B}$

A branch is described as external if it the proximal (youngest) end of the branch is a sampled individual, not a coalescent event. The probability that a single branch identified at time t_s is an external branch can be treated as a special case of a more general problem: what is the probability that a set of k branches extant at time t_s are all external? Using equation 11 this can be further generalized to the full probability mass function of the number of individuals in the sample descended from $\mathscr{B}$.

Consider first the case of $k = 2$ branches identified at time t_s, when there are $n(t_s)$ lineages. The probability that the next coalescent event (forward in time) is *not* among the two selected lineages is

$$\frac{n(t_s)-2}{n(t_s)}.$$

The probability that neither of the first two coalescent events involve the selected lineages is

$$\frac{n(t_s)-2}{n(t_s)} \times \frac{n(t_s)-1}{n(t_s)+1}.$$

The probability that none of the first three coalescent events involve the selected lineages is then:

$$\frac{n(t_s)-2}{n(t_s)} \times \frac{n(t_s)-1}{n(t_s)+1} \times \frac{n(t_s)}{n(t_s)+2}.$$

And the probability that none of the first four coalescent events involve the selected lineages is:

$$\frac{n(t_s)-2}{n(t_s)} \times \frac{n(t_s)-1}{n(t_s)+1} \times \frac{n(t_s)}{n(t_s)+2} \times \frac{n(t_s)+1}{n(t_s)+3}.$$

There are in total $n(t_s) - n - 1$ coalescent events between times 0 and t_s. The probability that *none* of those events have involved the selected lineages is

$$\frac{n(t_s)-2}{n(t_s)} \times \frac{n(t_s)-1}{n(t_s)+1} \times \frac{n(t_s)}{n(t_s)+2} \times \frac{n(t_s)+1}{n(t_s)+3} \times \cdots \times \frac{n-5}{n-3} \times \frac{n-4}{n-2} \times \frac{n-3}{n-1}.$$

Many of these terms cancel.

$$\frac{n(t_s)-2}{\cancel{n(t_s)}} \times \frac{n(t_s)-1}{\cancel{n(t_s)+1}} \times \frac{\cancel{n(t_s)}}{\cancel{n(t_s)+2}} \times \frac{\cancel{n(t_s)+1}}{\cancel{n(t_s)+3}} \times \cdots \times \frac{\cancel{n-5}}{\cancel{n-3}} \times \frac{\cancel{n-4}}{n-2} \times \frac{\cancel{n-3}}{n-1}.$$

The only remaining terms are the first k numerators and the last k denominators, leaving the compact probability

$$p(k|t_s) = \prod_{i=1}^{k} \frac{n(t_s)-i}{n-i}. \tag{12}$$

A particular insight from this derivation is the clear observation that $p(k|t_s)$ depends only on the value of $n(t_s)$ and not the entire function $n(t)$. Once we know the number of lineages extant at time t_s it doesn't matter how or when they got that way. Similarly, the demographic history of the population, $N(t)$ matters only to the extent that it influences $n(t_s)$ and is otherwise immaterial.

The distribution of the number of sampled descendants of a branch identified at time t_s, $g(x = X|t_s)$, is derived by integrating over the possible timing of the intervening coalescent events.

The probability that a branch identified at time t_s leaves exactly one sampled descendant is by definition equivalent to the probability that it is an external branch. The probability that it leaves exactly two sampled descendants is the probability that the lineage splits once at some time t_{d1}, and the two resulting lineages are both external branches. From equations 11 and 12 we get

$$g(x = 2|t_s) = \int_0^{t_s} f(t_s - t_{d1}|1, t_s) p(2|t_{d1}) dt_{d1}.$$

The probability that the branch leaves exactly three sampled descendants is the probability that there have been two coalescent events on lineages descended from $\mathscr{B}$ at times t_{d1} and t_{d2}, after which all three resulting lineages are external:

$$g(x = 3|t_s) = \int_0^{t_s} \int_0^{t_{d1}} f(t_s - t_{d1}|1, t_s) f(t_{d1} - t_{d2}|1, t_{d1}) p(3|t_{d2}) dt_{d2} dt_{d1}.$$

Continuing in this fashion, each successive extra descendant requires integrating over an additional coalescence time and incrementing the number of terminal lineages determined to be external. As a generic function, this gives

$$g(x = X|t_s) = \int_0^{t_s} \int_0^{t_{d1}} \dots \int_0^{t_{dX-2}} f(t_s - t_{d1}|1, t_s) f(t_{d1} - t_{d2}|1, t_{d1}) \dots f(t_{dX-2} - t_{dX-1}|t_{dX-2}) p(X|t_{dX-1}) dt_{dX-1} \dots dt_{d2} dt_{d1}. \quad (13)$$

This probability is directly applicable as the number of copies of a variant allele created by a mutation at time t_s (conditional on the mutation being present in a sample). In general, this expression will be of most use for problems involving branches with relatively few descendants, where the multiplicity of integrals (there are $X - 1$ of them) won't be burdensome.

Discussion

By using standard coalescence theory to describe the overall shape of a gene tree and forward-in-time death process analysis it is possible to get simple closed form expressions related to arbitrary branches within the tree as functions of the time at which those branches existed.

References

1. Kingman, J. F. C. The coalescent. *Stoch. processes their applications* **13**, 235–248 (1982).
2. Caliebe, A., Neininger, R., Krawczak, M. & Rösler, U. On the length distribution of external branches in coalescence trees: genetic diversity within species. *Theor. population biology* **72**, 245–252 (2007).
3. Griffiths, R. C. Lines of descent in the diffusion approximation of neutral wright-fisher models. *Theor. population biology* **17**, 37–50 (1980).
4. Maruvka, Y. E., Shnerb, N. M., Bar-Yam, Y. & Wakeley, J. Recovering population parameters from a single gene genealogy: an unbiased estimator of the growth rate. *Mol. biology evolution* **28**, 1617–1631 (2010).
5. Griffiths, R. C. Coalescent lineage distributions. *Adv. applied probability* **38**, 405–429 (2006).
6. Slatkin, M. & Rannala, B. Estimating the age of alleles by use of intraallelic variability. *Am. journal human genetics* **60**, 447 (1997).
7. Chen, H. & Chen, K. Asymptotic distributions of coalescence times and ancestral lineage numbers for populations with temporally varying size. *Genet.* **194**, 721–736 (2013).
8. Jewett, E. M. & Rosenberg, N. A. Theory and applications of a deterministic approximation to the coalescent model. *Theor. population biology* **93**, 14–29 (2014).
9. Griffiths, R. C. & Tavare, S. Sampling theory for neutral alleles in a varying environment. *Philos. transactions: biological sciences* 403–410 (1994).
10. Watson, G. S. & Leadbetter, M. R. Hazard analysis. i. *Biom.* **51**, 175–184 (1964). URL http://www.jstor.org/stable/2334205.

Acknowledgements

With thanks to Jody Hey for helpful comments and discussion, and funding from NIH Grant RO1GM078204 to Dr. Hey.

Additional information

Competing financial interests The author declares no competing financial interests.

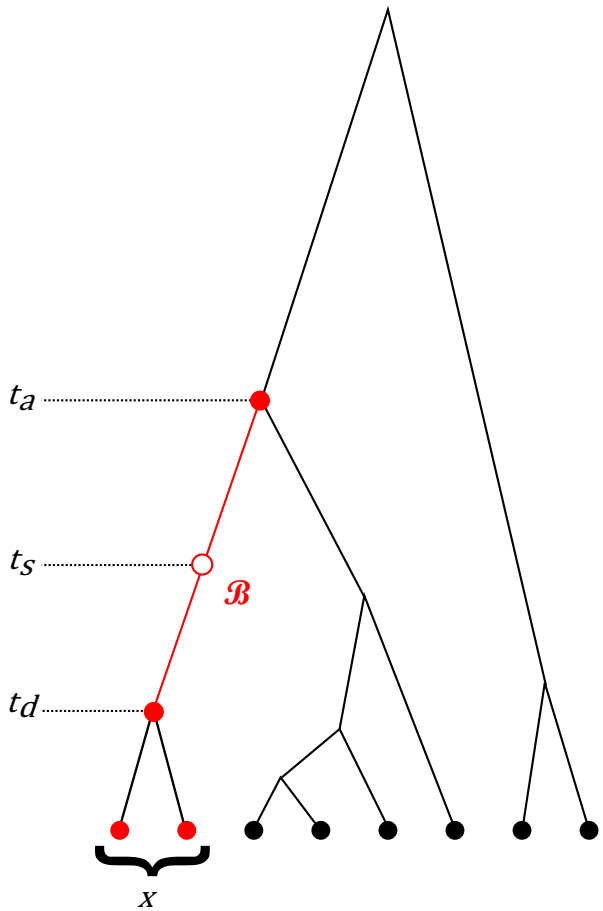

Figure 1. Diagram of important parameters. Branch $\mathscr{B}$ is identified at time t_s. It has an ancestral node at time t_a, a descendant node at time t_d and leaves x descendants in the current sample.