



**HAL**  
open science

# Statistical Properties of Pairwise Distances between Leaves on a Random Yule Tree

M. Sheinman, F. Massip, P. F. Arndt

► **To cite this version:**

M. Sheinman, F. Massip, P. F. Arndt. Statistical Properties of Pairwise Distances between Leaves on a Random Yule Tree. PLoS ONE, 2015, 10 (3), pp.e0120206. 10.1371/journal.pone.0120206 . hal-02018975

**HAL Id: hal-02018975**

**<https://univ-lyon1.hal.science/hal-02018975>**

Submitted on 28 May 2020

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution 4.0 International License

RESEARCH ARTICLE

# Statistical Properties of Pairwise Distances between Leaves on a Random Yule Tree

Michael Sheinman<sup>1\*</sup>, Florian Massip<sup>1,2</sup>, Peter F. Arndt<sup>1</sup>

**1** Max Planck Institute for Molecular Genetics, Berlin, Germany, **2** INRA, UR1077 Unite Mathematique Informatique et Genome, Jouy-en-Josas, France

\* [mishashe@gmail.com](mailto:mishashe@gmail.com)

## Abstract

A Yule tree is the result of a branching process with constant birth and death rates. Such a process serves as an instructive null model of many empirical systems, for instance, the evolution of species leading to a phylogenetic tree. However, often in phylogeny the only available information is the pairwise distances between a small fraction of extant species representing the leaves of the tree. In this article we study statistical properties of the pairwise distances in a Yule tree. Using a method based on a recursion, we derive an exact, analytic and compact formula for the expected number of pairs separated by a certain time distance. This number turns out to follow an increasing exponential function. This property of a Yule tree can serve as a simple test for empirical data to be well described by a Yule process. We further use this recursive method to calculate the expected number of the  $n$ -most closely related pairs of leaves and the number of cherries separated by a certain time distance. To make our results more useful for realistic scenarios, we explicitly take into account that the leaves of a tree may be incompletely sampled and derive a criterion for poorly sampled phylogenies. We show that our result can account for empirical data, using two families of birds species.



## OPEN ACCESS

**Citation:** Sheinman M, Massip F, Arndt PF (2015) Statistical Properties of Pairwise Distances between Leaves on a Random Yule Tree. PLoS ONE 10(3): e0120206. doi:10.1371/journal.pone.0120206

**Academic Editor:** Arndt von Haeseler, Max F. Perutz Laboratories, AUSTRIA

**Received:** October 10, 2014

**Accepted:** January 20, 2015

**Published:** March 31, 2015

**Copyright:** © 2015 Sheinman et al. This is an open access article distributed under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

**Data Availability Statement:** All relevant data are within the paper.

**Funding:** The authors have no support or funding to report.

**Competing Interests:** The authors have declared that no competing interests exist.

## Introduction

The speciation process in evolution can be regarded as a branching process. One of the simplest stochastic models for a branching process is the so called Yule process [1, 2]. In this model branches are assumed to split with a constant rate and both resulting branches will evolve independently in time. Starting from one branch, a tree will grow, such that the number of leaves on average increases exponentially in time. In a more general version of the Yule tree each branch can also die and get extinct with a constant rate.

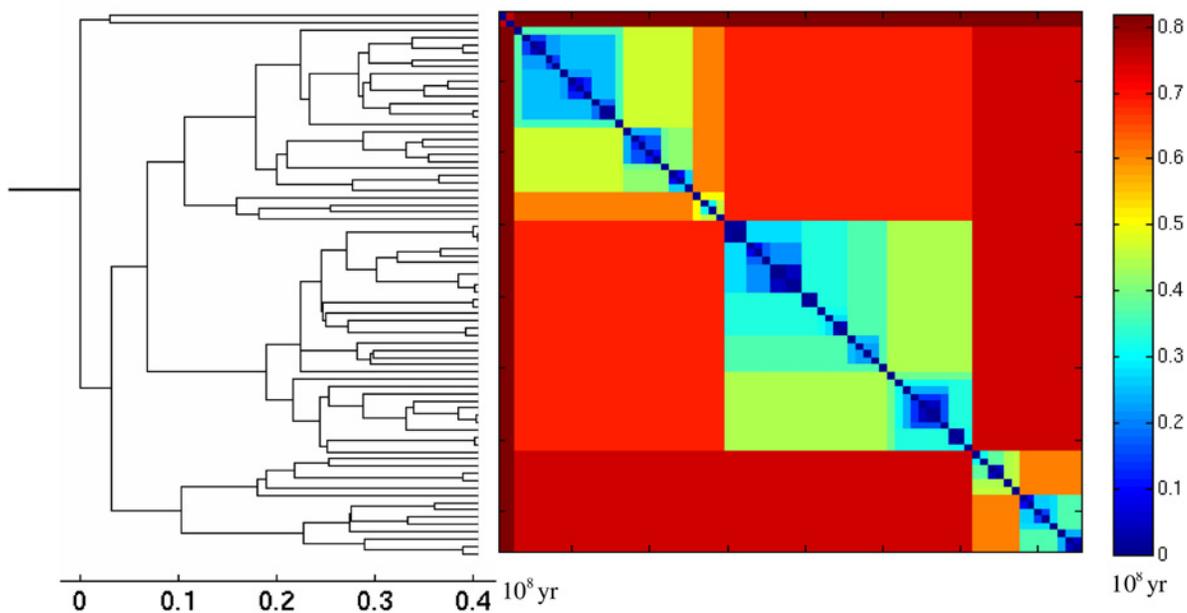
Despite its simplicity, many phenomena in different fields of science have been successfully modeled using the Yule process [3, 4]. Particular examples include statistical properties of the number of species in a genus [1], the number of members in protein and gene families [5, 6] and phoneme frequencies in languages [7]. In stochastic modelling of biological evolution, the Yule process is often useful as an instructive null hypothesis [8–11], even when its assumptions are clearly violated.

As an illustrative example of the branching process we present the reconstructed phylogenetic tree of species in the Siilvidae family of birds in the left panel of Fig. 1. The basis of such a reconstructed tree is pairwise distances between individual species. The color-coded matrix of such distances for the species is shown in the right panel of Fig. 1. The statistical properties of such a matrix for a Yule tree is the focus of our article.

Statistical properties of Yule trees have been intensively studied and much is already known. One of the most useful results is the distribution of the number of leaves on a Yule tree [12]. This exact analytical result is widely exploited, in particular, for reconstruction of phylogenetic trees and for estimation of rates of speciation and extinction [10, 11, 13]. Other discrete properties have been studied in Refs. [14–17] as well as properties of the distribution of branch lengths [18, 19].

Often the pairwise distances between all pairs of species in a group of species is the only available information useful for reconstruction of the evolutionary history of the group. For example, in phylogeny reconstruction, one can estimate the pairwise distance in time between two species (twice the time to their last common ancestor) using the molecular clock approach, together with morphological considerations and information about the fossil record [20]. Motivated by observations of mitochondrial DNA sequences with no recombination, the distribution of pairwise distances has been studied in Ref. [21] for a tree with discrete generations and a given number of leaves. In this study, the authors use a sort of mean-field approach, ignoring fluctuations in the number of leaves during the growth of the tree, to derive an approximate formula for the pairwise distances distribution on a tree.

Here we present a general method to derive the distribution of pairwise distances and other statistical properties on a continuous random Yule tree of a certain height with given birth and death rates. Using our method, we obtain exact, analytic, closed, non-recursive and compact formulas for the pairwise distance distribution, the distribution of distances to the closest neighbour, the distance distribution in so-called cherries, as well as a more general formula for the distribution distance to the  $n$ -th closest neighbour.



**Fig 1. One of the reconstructed trees for the Siilvidae family of species, taken from [28] (left) and its distance matrix (right).** The tree includes only the branches which lead to survived and observed leaves.

doi:10.1371/journal.pone.0120206.g001

Often, in biological context, one does not have an access to data about all existing species (i.e. leaves of a phylogenetic tree) [22]. Instead, species are incompletely sampled, or might have been subject to a recent massive extinction event [23]. As long as the extinction of species is random, both scenarios are equivalent on macroevolutionary timescales. In our study, we take the incomplete sampling explicitly into account, which allows us to make statements about the fraction of sampled species, using only the available data.

In the next section we will start with a formal definition of the Yule process and then derive the above mentioned distributions of pairwise distances. For illustrative purposes we also present numerical simulations perfectly matching our expectations. At the end of our article we apply our theoretical consideration to empirical data and analyze the speciation process in two families of birds for which data on speciation times and pairwise distances is available. One advantage of our approach is that we do not need to reconstruct a phylogenetic tree but can solely work with data on pairwise distances.

## A Yule tree with constant branching and extinction rates and incomplete sampling of leaves

### Definition of the Yule Tree

A Yule tree is defined as follows [1, 2]. At time  $t = 0$  there is one individual. As time progresses, this individual can branch and give birth to another individual. In an infinitesimally short time interval  $[t, t+dt]$ , all individuals can give birth to another one, each with the probability  $\lambda dt$ . The probability of an individual to die in the same time interval is  $\mu dt$ . We consider an ensemble of trees of age (height)  $T$ , referring to all existing individuals at this time as *leaves*. To make the model more realistic, we assume that due to incomplete sampling (or a short massive extinction event) just before the time  $T$ , each leaf is observed with a certain probability  $0 \leq \sigma \leq 1$ . The described process is illustrated in Fig. 2. We assume that the incompleteness of the sampling is random and ignore possible biases due to different sampling schemes [24].

### A Few Useful Results for Random Trees Generated by a Yule Process

Consider a Yule tree with birth rate  $\lambda$  and death rate  $\mu$ , that have been grown for total time (height)  $T$ . In the case where all leaves are sampled ( $\sigma = 1$ ), let  $P(M|T, \sigma = 1)$  be the probability that there are  $M$  leaves on a tree of age  $T$ . Following [25], we can then write the probability that no individual ( $M = 0$ ) survives through to time  $T$  as

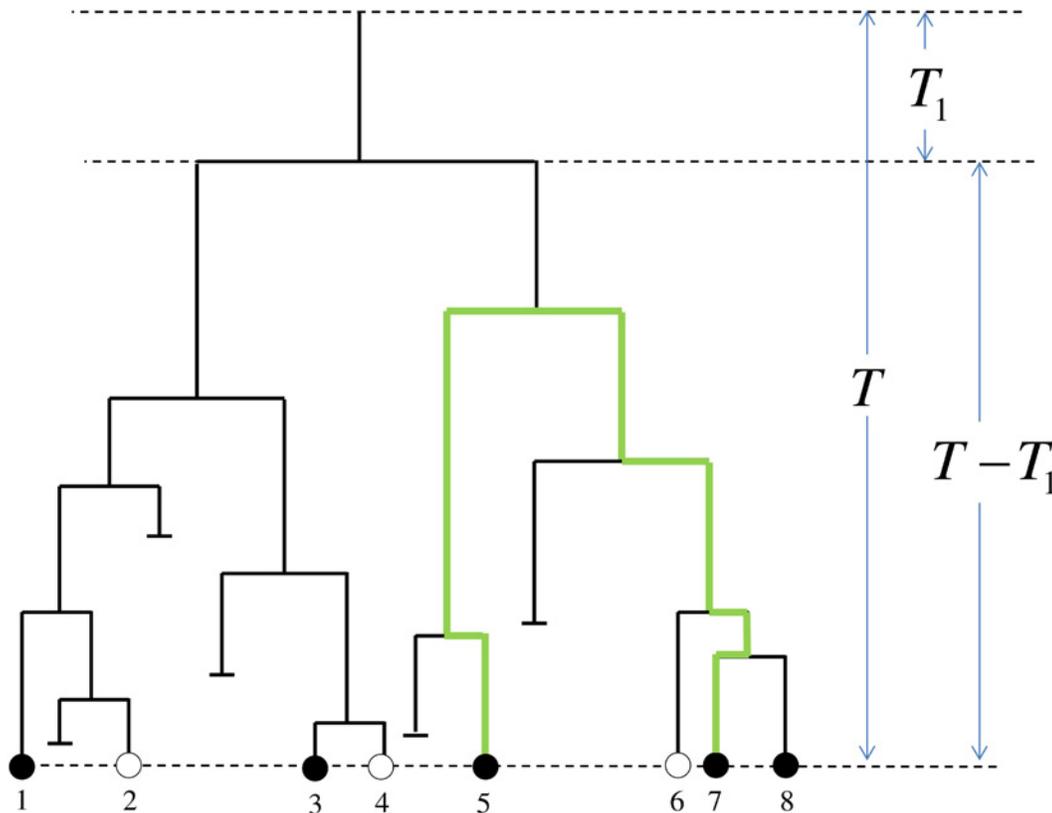
$$P(M = 0|T, \sigma = 1) = 1 - \frac{\lambda - \mu}{\lambda - \mu e^{-(\lambda-\mu)T}}. \tag{1}$$

For  $M > 0$  we have

$$P(M|T, \sigma = 1) = \frac{\lambda - \mu}{\lambda - \mu e^{-(\lambda-\mu)T}} \left[ 1 - \frac{1 - e^{-(\lambda-\mu)T}}{1 - \frac{\mu}{\lambda} e^{-(\lambda-\mu)T}} \right] \left[ \frac{1 - e^{-(\lambda-\mu)T}}{1 - \frac{\mu}{\lambda} e^{-(\lambda-\mu)T}} \right]^{M-1}. \tag{2}$$

We can derive corresponding equations also for the case where species are sampled incompletely. In this case, the probability that no species is observed is

$$P(M = 0|T) = P(0|T, \sigma = 1) + \sum_{m=1}^{\infty} \binom{m}{0} \sigma^0 (1 - \sigma)^{m-0} P(m|T, \sigma = 1) = \frac{e^{\mu T} (\mu - \lambda + \sigma \lambda) - e^{\lambda T} \mu \sigma}{e^{\mu T} (\mu - \lambda + \sigma \lambda) - e^{\lambda T} \lambda \sigma} \tag{3}$$



**Fig 2. An example of the rooted Yule tree of age  $T$ .** Filled circles (1, 3, 5, 7 and 8) denote observed leaves. Empty circles (2, 4 and 6) denote survived but not observed leaves. Short horizontal lines denotes an extinction event. Long, dashed horizontal lines denote the origin of the tree, the first branching event and the time of sampling the tree, from top to bottom. After the first branching at time  $T_1$  the two resulting subtrees both encompass  $M_1 = M_2 = 4$  leaves. However, the number of observed leaves is 2 (leaves 1 and 3) for the left subtree and 3 (leaves 5, 7 and 8) for the right one. The thick green line denotes the pairwise evolutionary distance between the two observed leaves 5 and 7. The horizontal dimension is meaningless. In this example for leaf 1 the first closest observed leaf is 3, the second (as well as the third and the fourth) is 5 (or 7 or 8). The tree has two observed cherry pairs: (1, 3) and (7, 8).

doi:10.1371/journal.pone.0120206.g002

and for  $M > 0$

$$P(M|T) = \sum_{m=M}^{\infty} \binom{m}{M} \sigma^M (1 - \sigma)^{m-M} P(M|T, \sigma = 1) = \frac{[e^{T(\mu-\lambda)} - 1]^{M-1} \lambda^{M-1} (\lambda - \mu)^2 \sigma^M e^{MT(\lambda-\mu)}}{[\lambda\sigma - \lambda + \mu - \lambda\sigma e^{T(\lambda-\mu)}]^{M+1}}. \quad (4)$$

Despite these complicated expressions, the average number of observed leaves in a tree of age  $T$  is simply given by

$$\langle M(T) \rangle = \sum_{m=0}^{\infty} m P(m|T) = \sigma e^{(\lambda-\mu)T} \quad (5)$$

and the average total number of pairs is

$$\sum_{m=0}^{\infty} \frac{m(m-1)}{2} P(m|T) = \frac{\sigma^2 \lambda}{\lambda - \mu} e^{(\lambda-\mu)T} [e^{(\lambda-\mu)T} - 1]. \quad (6)$$

The total length of all branches in a Yule tree is given by the integral:

$$\int_0^T \langle M(t) \rangle dt = \int_0^T e^{(\lambda-\mu)t} dt = \frac{1}{\lambda-\mu} [e^{(\lambda-\mu)T} - 1]. \tag{7}$$

To derive a corresponding expression for a tree reconstructed only from incompletely sampled leaves, we note that the average number of branches at time  $t$  with at least one observed descendant at time  $T$  is given by

$$\langle M(t, T) \rangle = e^{(\lambda-\mu)t} [1 - P(0|T-t, \sigma)]. \tag{8}$$

In the case where  $t = T$ , we have that  $\langle M(T, T) \rangle = \sigma \langle M(T) \rangle$ . The average total branch length on the tree of length  $T$  excluding the branches which do not lead to an observed leaf is then given by

$$\int_0^T \langle M(t, T) \rangle dt = \frac{\sigma e^{T(\lambda-\mu)}}{\mu - \lambda + \sigma\lambda} \ln \frac{\lambda\sigma + (\lambda - \sigma\lambda - \mu)e^{T(\mu-\lambda)}}{\lambda - \mu}. \tag{9}$$

In the limit of no extinction,  $\mu \rightarrow 0$ , and exhaustive sampling,  $\sigma \rightarrow 1$ , Equation (9) is identical to Equation (7). We turn now to calculations of the statistical properties of pairwise distances, using the above formulas.

### The Distribution of Pairwise Distances

In a biological context the available data often consist of the pairwise distances separating any pair in a group of species. Commonly these distances are used to reconstruct a phylogenetic tree representing the evolutionary history of a group of species. From such a tree one can then try to estimate rates of speciation and extinction [10, 11]. Here we propose another approach of analysing such data on pairwise distances circumventing the reconstruction of a phylogenetic tree, provided that the pairwise distances between the leaves are properly estimated.

Let  $N(t|T)dt$  be the average number of pairs of leaves on a tree of length (evolution time)  $T$ , separated by a time distance in the interval  $[t, t+dt]$ , i.e. their last common ancestor lived in the time interval  $[T-t/2-dt/2, T-t/2]$ . Now consider the branching process as illustrated in Fig. 2. The first branching happened at time  $T_1$  and the two resulting subtrees encompass, say,  $M_1$  and  $M_2$  leaves, respectively. In this situation one can derive the following recursion relation

$$N(t|T) = [2N(t|T - T_1) + \sigma^2 M_1 M_2 \delta(t - 2(T - T_1))I(0 \leq t \leq 2T)]e^{-\mu T_1} \tag{10}$$

where the first part in the summation on the right hand side counts the pairs inside each of the two subtrees and the second one counts the pairs between them. The common multiplicative factor,  $e^{-\mu T_1}$ , expresses the probability that the first branch survives to the time  $T_1$  (otherwise,  $N(t|T) = 0$ ). The function  $I$  is the indicator function, defined by:

$$I(\text{condition}) = \begin{cases} 1 & \text{if condition holds} \\ 0 & \text{otherwise} \end{cases} \tag{11}$$

and  $\delta(x)$  is the Dirac delta function. Averaging over  $M_1, M_2$  (using Equations (3, 4) with time  $T - T_1$ ) and then  $T_1$ , which follows an exponential distribution with mean  $1/\lambda$ , one obtains:

$$N(t|T) = 2\lambda \int_0^\infty N(t|T - T_1)e^{-(\lambda-\mu)T_1} dT_1 + \frac{\sigma^2 \lambda}{2} e^{\lambda t} e^{-(\lambda+\mu)(T-t/2)} I(0 \leq t \leq 2T). \tag{12}$$

In Laplace space one gets:

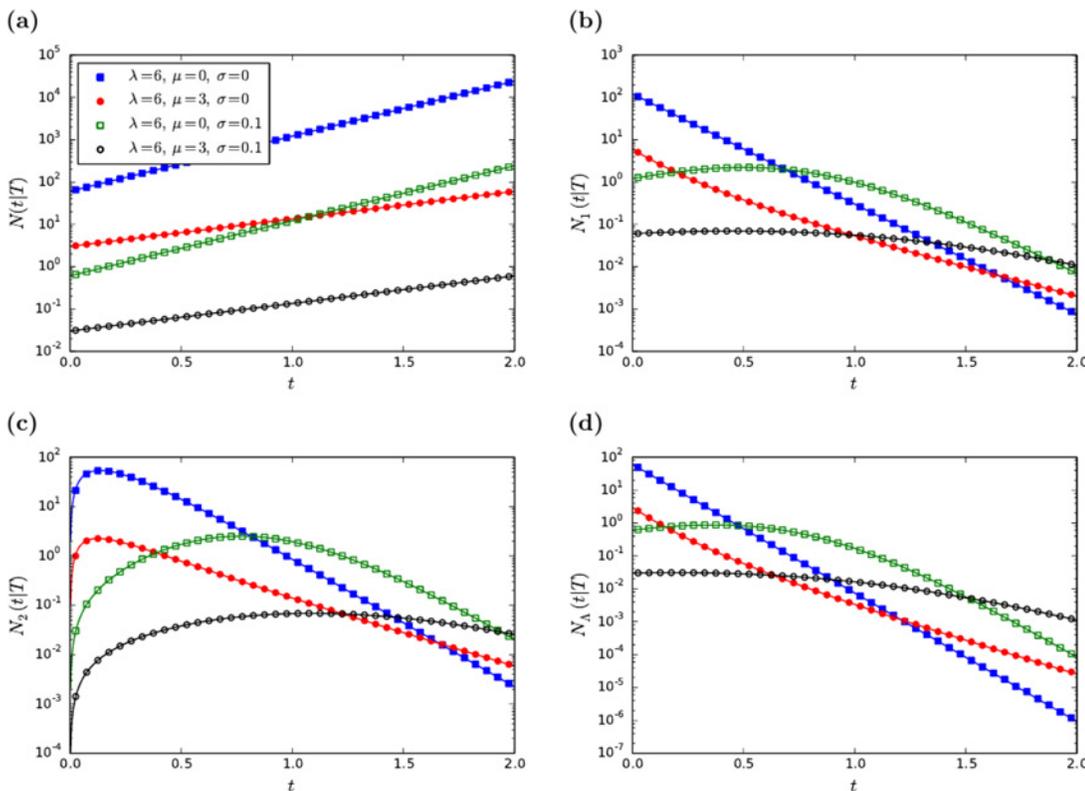
$$N(t|S) = 2\lambda \frac{N(t|S)}{S + \lambda + \mu} + \frac{\sigma^2 \lambda}{2} \frac{e^{\lambda t - St/2}}{S + \lambda + \mu}, \tag{13}$$

where  $S$  is the Laplace conjugate variable of  $T$ . Solving and inverting the Laplace transform one finally gets the solution:

$$N(t|T) = \frac{\sigma^2 \lambda}{2} e^{(\lambda - \mu)T} e^{(\lambda - \mu)t/2} \tag{14}$$

for  $0 \leq t \leq 2T$  and zero otherwise. Fascinatingly, this distribution is a simple exponential function in  $t$ . The distribution is cut off at  $t = 2T$  because in a tree of age  $T$  two leaves cannot be separated by a time larger than  $2T$ . In Fig. 3(a) we show this distribution of pairwise distances for several parameter values together with results of numerical simulations, which match perfectly our theoretical expectations. This result, applied for trees of DNA sequences can account for statistics of exact sequence matches in genomes of eukaryotes [26].

One can also derive the same result (14) using the following simple arguments. Pairs, separated by a time in the interval  $[t, t+dt]$ , branched at the time interval  $[T-t/2-dt/2, T-t/2]$ . The average number of branches in this interval is given by  $\lambda e^{(\lambda - \mu)(T-t/2)} dt/2$ . The average number



**Fig 3. Comparison of the analytic results with numerical simulations.** Markers indicate numerically obtained data using the following parameters set.  $T = 1$ ,  $\lambda = 6$ ,  $\mu = 0$  or  $3$  (circles or squares) and  $\sigma = 1$  or  $0.1$  (empty or filled symbols). Lines represent the analytic formulas. (a) Density of number of pairs separated by a certain time,  $t$ . Lines were obtained using Equation (14). (b) Density of number of leaves separated by a certain time,  $t$  with their closest leaf. Lines were obtained using Equation (17) or Equation (20) with  $n = 1$ . (c) Density of number of leaves separated by a certain time,  $t$  with their next-closest leaf. Lines were obtained using Equation (33) or Equation (20) with  $n = 2$ . (d) Density of number of cherries separated by a certain time,  $t$ . Lines were obtained using Equation (21).

doi:10.1371/journal.pone.0120206.g003

of observed pairs from a branch at this time is given by  $(\sigma e^{(\lambda-\mu)t/2})^2$ . Multiplying the two factors one gets Equation (14). However, for other quantities, derived below, the recursive equation approach is more effective.

### The Distribution of the Minimal-Distance to Other Leaves

Using the recursive method from the previous Section one can also compute other interesting quantities. For instance, in certain situations, the distance separating a leaf to its most closely relative may be estimated more precisely than its distance to other leaves in the tree. Thus, we might be interested in  $N_1(t|T)dt$ —the average number of leaves on the tree of age  $T$ , separated by the time distance between  $t$  and  $t+dt$  from their most closely related leaf. Interestingly, calculating this quantity lets us make certain statements on the value of the sampling rate  $\sigma$ .

To calculate this distribution, we can again write a recursion relation, assuming that the first branching occurred at time  $T_1$ . In this case one gets the distribution of the minimal distance time in the form

$$N_1(t|T) = \{2N_1(t|T - T_1) + 2P(1|T - T_1)[1 - P(0|T - T_1)]\delta(t - 2(T - T_1))I(0 \leq t \leq 2T)\}e^{-\mu T_1}, \quad (15)$$

where  $P(M|T)$  is the probability to observe  $M$  leaves after time  $T$ , as computed in Equations (3) and (4). In contrast to the recursion relation for the distribution of all pairwise distances, we count a branching point only if  $M_1 = 1$  and  $M_2 > 0$  or  $M_1 > 0$  and  $M_2 = 1$ , as expressed by the product  $2P(1|T-T_1)[1-P(0|T-T_1)]$  in Equation (15).

Averaging Equation (15) over  $T_1$ , one gets:

$$N_1(t|T) = 2\lambda \int_0^\infty N_1(t|T - T_1)e^{-(\lambda+\mu)T_1}dT_1 + \frac{e^{-(\lambda+\mu)T+(3\lambda/2+\mu)t}\lambda(\lambda - \mu)^3\sigma^2}{\left[e^{\frac{t}{2}}\lambda\sigma - e^{\frac{\mu}{2}}(\mu - \lambda + \sigma\lambda)\right]^3}I(0 \leq t \leq 2T). \quad (16)$$

The solution of this equation is given by

$$N_1(t|T) = \frac{e^{\frac{t}{2}+\lambda T+\mu t-\mu T}\lambda(\lambda - \mu)^3\sigma^2}{\left[e^{\frac{t}{2}}\lambda\sigma - e^{\frac{\mu}{2}}(\mu - \lambda + \sigma\lambda)\right]^3} \quad (17)$$

for  $0 \leq t \leq 2T$  and 0 otherwise. Results of numerical simulations perfectly match our theoretical expectations (see Fig. 3(b)). Interestingly, the function  $N_1(t|T)$  from Equation (17) possesses a maximum only if

$$\sigma < \frac{1}{3}\left(1 - \frac{\mu}{\lambda}\right) \leq \frac{1}{3} \quad (18)$$

and the position of the maximum

$$t_{\max}^1 \equiv \frac{2}{\lambda - \mu} \ln \frac{\lambda(1 - \sigma) - \mu}{2\lambda\sigma} \quad (19)$$

is in the range  $[0, 2T]$ . This result is useful for a quick estimation of the data completeness. In particular, a maximum in the distribution of the minimal distance implies that the sampling of the considered tree is not complete and  $\sigma < 1/3$ .

By similar arguments we can also derive expressions for the distributions of second minimal distances,  $N_2(t|T)$  (see Appendix) and of the  $n$ -th minimal distance  $N_n(t|T)$  (see Appendix) to

other leaves. The latter quantity is computed to be

$$N_n(t|T) = \frac{n(1+n)(\mu-\lambda)^3 \sigma(\lambda\sigma)^n}{2} \frac{[e^{\frac{1}{2}(\mu-\lambda)} - 1]^{n-1} e^{\frac{\mu\lambda}{2} + T\lambda + t\mu - T\mu}}{\left[ e^{\frac{t\mu}{2}}(\mu-\lambda+\sigma\lambda) - e^{\frac{t\lambda}{2}}\lambda\sigma \right]^{n+2}} \tag{20}$$

for  $0 \leq t \leq 2T$  and 0 otherwise. In Appendix we also calculate the distribution of distances in “cherries”. Cherries are adjacent pairs of leaves, such that they are reciprocal closest neighbors to each other (see Fig. 2 for illustration of cherries):

$$N_\Lambda(t|T) = \frac{\lambda(\lambda-\mu)^4 \sigma^2}{2} \frac{e^{\frac{t\lambda}{2} + T\lambda + \frac{3t\mu}{2} - T\mu}}{\left[ e^{\frac{t\mu}{2}}(\mu-\lambda+\sigma\lambda) - e^{\frac{t\lambda}{2}}\lambda\sigma \right]^4} \tag{21}$$

for  $0 \leq t \leq 2T$  and 0 otherwise. The function  $N_\Lambda(t|T)$  from Equation (21) possesses a maximum only if

$$\sigma < \frac{1}{4} \left( 1 - \frac{\mu}{\lambda} \right) \leq \frac{1}{4} \tag{22}$$

and the position of the maximum

$$t_{\max}^\Lambda \equiv \frac{2}{\lambda-\mu} \ln \frac{(1-\sigma)\lambda-\mu}{3\lambda\sigma} \tag{23}$$

is in the range  $[0, 2T]$ . This result is useful for a quick estimation of the data completeness. In particular, a maximum in the distribution of the distance between cherries implies that the sampling of the considered tree is not complete and  $\sigma < 1/4$ .

For illustration purposes we show the distributions for the second minimal distance in Fig. 3(c) and, for cherries, in Fig. 3(d).

### Beyond the Averages

Above results are average expectations. For instance, in The Distribution of Pairwise Distances Section we derive  $N(t|T)$ , defined as the average density number of pairs, separated by a certain time distance  $t$ , on a tree of length  $T$ . The average is over many realizations, say  $S$  many, of the Yule trees with a given set of parameters  $\lambda, \mu, \sigma$  and  $T$ . Namely,

$$N(t|T) = \langle N^s(t|T) \rangle_s = \lim_{S \rightarrow \infty} \frac{1}{S} \sum_{s=1}^S N^s(t|T), \tag{24}$$

where  $N^s(t|T)$  is the density number of pairs separated by a time distance in the interval  $[t, t+dt]$  in an individual sample tree number  $s$ . In reality one often possesses information only about one specific tree  $s = 1$ , i.e.  $N^1(t|T)$ . Therefore, we are interested not only in the derived averages of  $N(t|T), N_n(t|T), N_\Lambda(t|T)$  etc. but also their distributions in finite time intervals. The last becomes especially important in the maximum likelihood fitting and model testing. In the discussion below we refer to the distribution of the number of pairs separated by a certain time,  $N^1(t|T)$ . However, the same arguments can be applied to other quantities, like the  $n$ -th minimal distance or the distance in cherries, which we mention above.

Consider an infinitesimal (in practice very small) interval,  $[t, t+dt]$ , such that  $N(t|T)dt \ll 1$ . The number of pairs  $N^1(t|T)dt$  in this interval is distributed with the mean  $N(t|T)dt$ . However, in the considered small bin limit, the mean does not represent well the typical value because the distribution of  $N^1(t|T)dt$  is not well peaked but possesses a very small probability of having any positive value, while probability of having zero is almost one (see Appendix).

Pairs separated by the time in the interval  $[t, t+dt]$  branched at the time interval  $[T-t/2-dt/2, T-t/2]$ . The probability to have a branch in this interval is given by  $\lambda e^{(\lambda-\mu)(T-t/2)} dt/2$ . Given that there is a branching point in this interval it can lead to different number of leaves. The probability that no observed pairs survive from this branching is given by  $1-[1-P(0|t/2)]^2$ , where  $P(M|T)$  is the probability to observe  $M$  leaves on a tree of age  $T$  and is given in Equations (3, 4). Therefore, the probability that there are no observed pairs separated by the time in the interval  $[t, t+dt]$  is given by

$$\Pr(N^1(t|T)dt = 0) = 1 - \lambda e^{(\lambda-\mu)(T-t/2)} dt/2 \{1 - [1 - P(0|t/2)]^2\}. \tag{25}$$

In sum, in the small bin limit it is convenient to break the full distribution in two distributions: One comprising only the peak at zero and a second representing all samples with  $N^1(t|T) dt \neq 0$ . The total average can be broken as follow:

$$N(t|T)dt = 0 \times \Pr(N^1(t|T)dt = 0) + \tilde{N}(t|T)dt \times [1 - \Pr(N^1(t|T)dt = 0)]. \tag{26}$$

Here  $\tilde{N}(t | T)$  is the average of  $N^1(t|T)$  over the tree realizations with  $N^1(t|T) > 0$ . It can be computed to be:

$$\tilde{N}(t|T) = \lim_{S \rightarrow \infty} \frac{\sum_{s=1}^S N^s(t|T)}{\tilde{S}(t)} = \frac{N(t|T)}{1 - \Pr(N^1(t|T)dt = 0)} = \frac{1}{dt} \left( 1 + \sigma \lambda \frac{e^{\frac{\lambda-\mu}{2}t} - 1}{\lambda - \mu} \right)^2, \tag{27}$$

where  $\tilde{S}(t) = \sum_{s=1}^S [1 - \delta_{N^s(t|T),0}]$  is the number of samples with  $N^1(t|T) > 0$ . Since,  $1 - \Pr(N^1(t|T)dt = 0) \ll 1$ , the value of  $N(t|T)dt$  is not representative of the expected empirical average of  $N^1(t|T)dt$  for finite  $S$  and, in particular,  $S = 1$ . However, the value of  $\tilde{N}(t | T)$ , derived above (see Equation (27)), is representative of the expected empirical average of positive values of  $N^s(t|T)dt$ . We illustrate this in Fig. 4

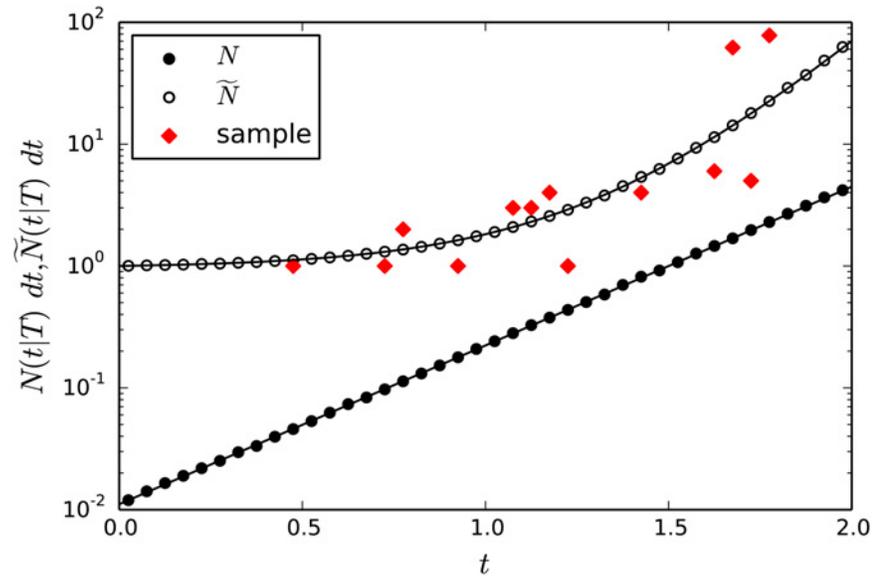
### Constrains on the sampling fraction

One can easily see that all the derived above results do not depend explicitly on the parameters  $\lambda, \mu$  and  $\sigma$ , but only on their combinations:  $\lambda - \mu$  and  $\sigma \lambda$ . Therefore, one cannot estimate the sampling fraction,  $\sigma$ , based on fitting the empirical data to the derived formulas (see examples in the next Section). The same loss of information in reconstructed trees was reported, based on an analysis of the density of bifurcation times in the reconstructed tree [27].

However, the information about the values of  $\lambda, \mu$  and, most intriguingly,  $\sigma$  is not lost completely. For instance, observing a maximum in the distribution of the minimal distances one can deduce that  $\sigma < 1/3$  (see Equation (18)). Observing a maximum in the distribution of the distances between cherries one can deduce that  $\sigma < 1/4$  (see Equation (23)). It is of an interest to construct other distributions which, possessing a maximum, provide information about the value of the sampling fraction,  $\sigma$ .

Consider an average density of pairs of leaves with the following property. Given that the first (second) leaf of the pair has a nearest neighbor at a distance (if a leaf is alone in the tree we define the distance to its nearest neighbor as twice the height of the tree)  $t_1$  ( $t_2$ ) the quantity  $\min(t_1, t_2)$  is given by  $t$ . We denote this density by  $N_{\min 2}(t|T)$ . The recursive equation for this quantity is given for a given time of first bifurcation,  $T_1$  by

$$N_{\min 2}(t|T) = \left\{ 2N_{\min 2}(t|T - T_1) + 2 \left[ \sigma e^{(\lambda-\mu)(T-T_1)} - \int_0^t N_1(t'|T - T_1) dt' \right] N_1(t|T - T_1) \right\} e^{-\mu T_1} \tag{28}$$



**Fig 4. The benefit to use  $\tilde{N}(t|T)$  instead of  $N(t|T)$  to estimate the parameters of the evolution process in a case of a small dataset.** In this plot  $T = 1, \lambda = 11, \mu = 5, \sigma = 0.01$  and  $dt = 0.005$ . After average over many samples ( $S \sim 10^6$  in this particular case) empirical averages of both  $N(t|T)$  (full circles) and  $\tilde{N}(t|T)$  (open circles) converge nicely to the analytic formulas. The last are given in Equations (14) and (27), respectively, and are denoted by the lines in the figure (see the legend). However, for a single random tree,  $S = 1$ , the values of  $N^1(t|T)$  (diamonds) are highly dispersed (most intervals show zero counts and do not show up in the semilogarithmic plot), such that their fit to the analytic formula of  $N(t|T)$  is not expected to lead to a good estimation of the model's parameters. In contrast, the values of  $N^1(t|T)$ , ignoring the bins where  $N^1(t|T) = 0$ , are well distributed around  $\tilde{N}(t|T)$ , although in this example the tree possesses only 19 observed leaves, such that the data is very poor (only 171 pairs in total).

doi:10.1371/journal.pone.0120206.g004

After average over  $T_1$  the solution is given by

$$N_{\min 2}(t|T) = \frac{2\lambda^2\sigma^3(\lambda - \mu)^4 e^{T(\lambda - \mu)} 2\lambda e^{t(\lambda - \mu)} + (\lambda - \mu)e^{-\frac{1}{2}t(\lambda + \mu)} + (\mu - 3\lambda)e^{\frac{1}{2}t(\lambda - \mu)} e^{T(\lambda - \mu)}}{3\lambda - \mu [\lambda\sigma - \lambda + \mu - \lambda\sigma e^{\frac{1}{2}t(\lambda - \mu)}]^5}. \quad (29)$$

This function possesses a maximum only if

$$\sigma < \frac{1}{5} \left(1 - \frac{\mu}{\lambda}\right) \leq \frac{1}{5} \quad (30)$$

Therefore, observing a maximum in the distribution of the minimal distance to the closest neighbors between two leaves one can deduce that  $\sigma < 1/5$ . Using our recursive method one can calculate different distributions (say, the minimal distance to the closest neighbor among three leaves etc.) which, exhibiting a maximum, provide direct information about an upper limit on the sampling fraction.

### Comparison of the derived results to empirical data

In this Section we demonstrate the relevance of the obtained analytic formulas to empirical data, studying the pairwise distances between species in families of the evolutionary tree. For comparison with the derived results we choose  $N(t|T), N_n(t|T)$  with  $n = 1, 2, 3, 4$  and  $N_\Lambda(t|T)$ . The results are presented in Fig. 5 for the Siilvidae family of birds (see one of the reconstructed trees for this family and its distance matrix in Fig. 1) and for the Tyrannidae family of birds in

**Fig. 6.** For every family we analyze Bayesian sampling of 1000 trees downloaded from the database [28]. Namely, we collect pairwise distances,  $n$ -minimal distances and distances between cherries of all 1000 trees and plot the histograms of these distances (with the  $y$ -axis divided by 1000) in Figs. 5 and 6. We fit all the points in a figure using the iterative reweighted least squares algorithm [29] in Matlab. Unfortunately, the explicit dependencies on  $\lambda$  and  $\mu$  in Equations (14, 20, 21) are insufficient to estimate all parameters. Instead one can estimate from the fit only the effective growth rate,  $\lambda - \mu$  and  $\lambda\sigma$ . The value of  $\sigma$  can be obtained assuming a certain ratio  $\mu/\lambda$ . In the captions of Figs. 5 and 6 we present the obtained estimates for  $\sigma$  for different assumptions about the ratio  $\mu/\lambda$ .

Over all, the fits to empirical data look satisfactory and result in a reasonable set of parameters, which roughly agree with the ones given in [28]. This indicates that certain statistical properties of speciation can be well captured by a simple Yule process. However, in some cases, deviations can be observed. For example, for the Sylviidae family the pairwise distances distribution deviates from the prediction for  $t > 30$  Myr, while for the Tyrannidae family we observe a clear deviation for distances around 55 Myr in all our estimates. This indicates a massive radiation event in the considered family of birds around 27.5 Myr ago, as already reported in [28], or other violation of the Yule process assumptions.

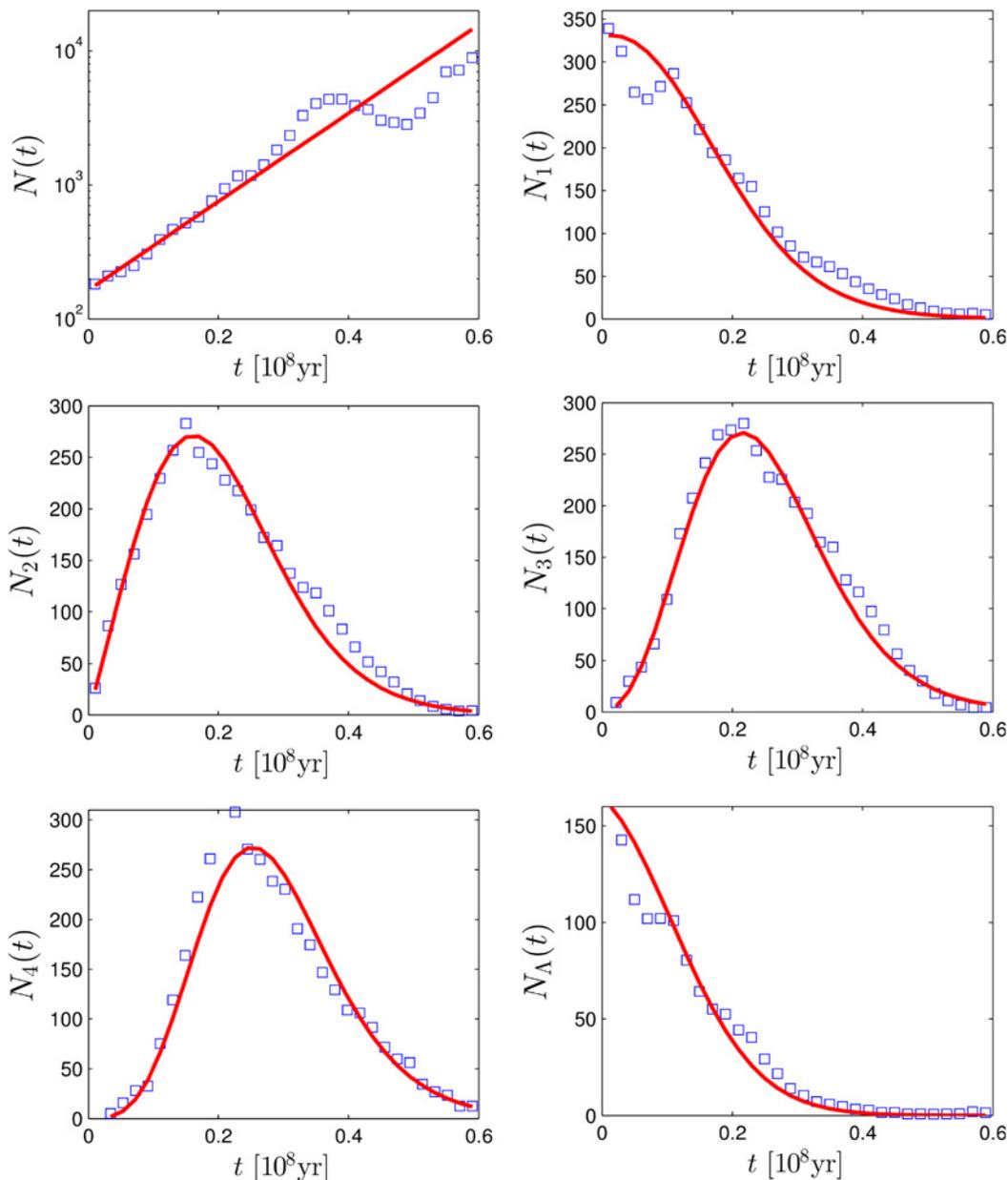
Interestingly, we can state that the Sylviidae family of birds is currently not well sampled. In fact, the estimator for the upper limit of the sampling fraction  $\sigma$  is 30% (see Fig. 5).

## Summary and concluding remarks

In this paper we present a novel method to calculate statistical properties of Yule trees. The method is based on a recursive equations which can be solved using the Laplace transform. We demonstrate the strength of our method deriving formulas for (i) average number of pairs separated by a certain time (Equation (14)), (ii) the number of most closely related pairs separated by a certain time (Equation (17)), (iii) the number of next-most closely related pairs separated by a certain time (Equation (33)), (iv) the number of  $n$ -most closely related pairs separated by a certain time (Equation (20)) and (v) the number of cherries separated by a certain time (Equation (21)).

Our results can be compared to empirical data using only the information about pairwise distances between leaves of a considered tree. We assume that the estimation of the pairwise distances is precise enough. If the distances are estimated using genetic divergence, this assume that the molecular clock reflect adequately the real time distance. If this holds the reconstruction of the tree structure is not required. This is a particular strength of our method because the reconstruction of such trees for a large number of leaves is sometimes problematic. In such cases one often considered a posterior distribution of trees which is generated by Bayesian sampling [30, 31]. Such a distribution of trees can still be easily analyzed using our method, based on recursive equations. Analyzing such ensembles of trees we use only their distance matrices.

We demonstrate the relevance of our results to statistical properties of pairwise evolutionary time distances between biological species. We find that in some cases the speciation process is well described by the Yule model. Significant deviations from the derived distributions are expected to be indicative for massive extinction or radiation events. In the case where the assumptions of the Yule process are justified, we expect our results to be useful for estimation of the incompleteness of the data sampling, i.e. the fraction of observed leaves out of all existing leaves,  $\sigma$ . However, similarly to the method developed in Ref. [11], all the derived results depend only on three parameters:  $\lambda - \mu$ ,  $\lambda\sigma$  and  $\sigma e^{(\lambda - \mu)T}$ . Therefore, even knowing those *three* parameters one cannot estimate the values of the *four* unknown parameters: the rates  $\lambda$ ,  $\mu$ , the height of the tree,  $T$  and the sampling fraction,  $\sigma$ , without an additional assumption about one



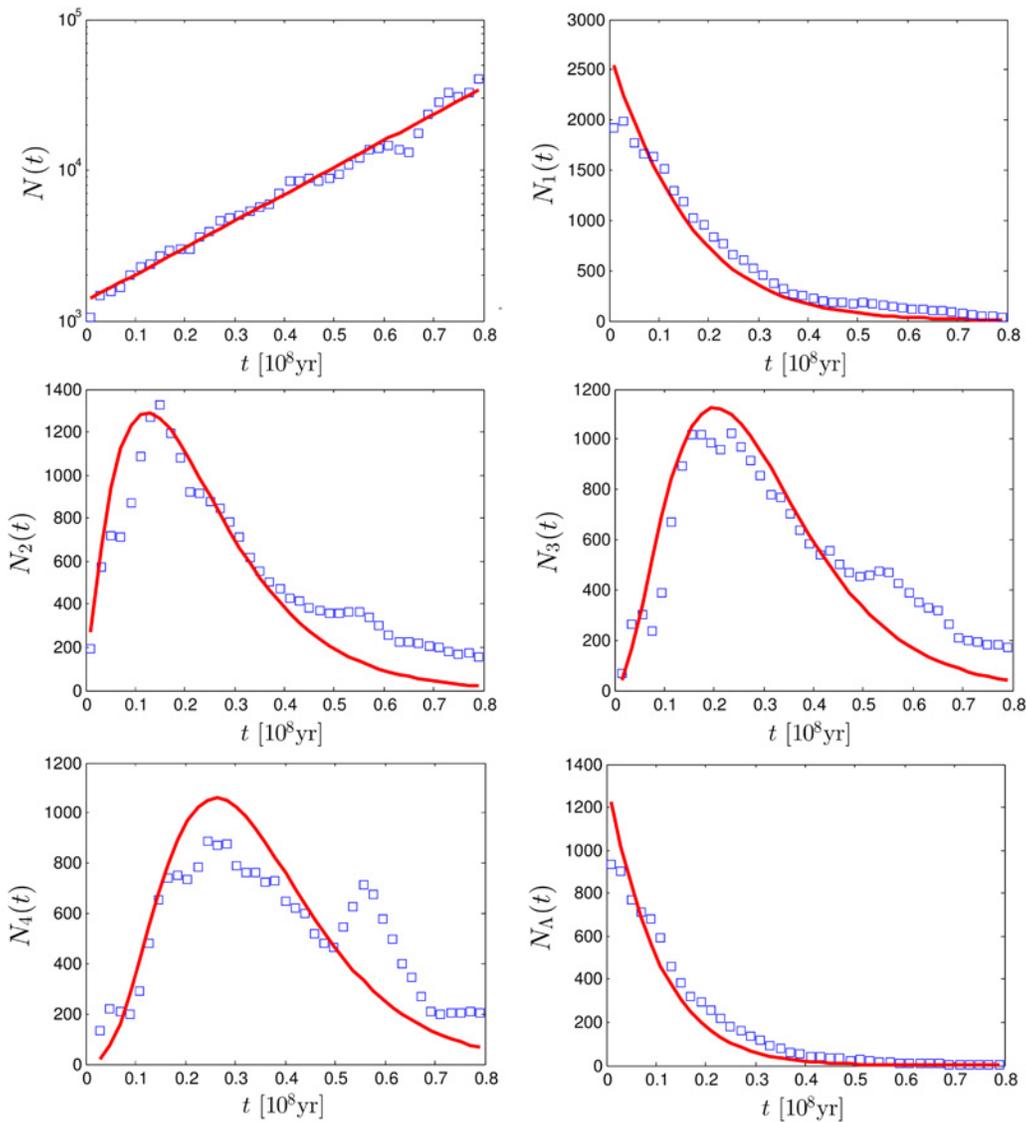
**Fig 5. Comparison of analytic predictions to the pairwise distances data of Sylviidae family with  $M = 75$  species taken from the database [28] with  $t \leq 0.6 \times 10^8$  Myr.** The markers represent the empirical data, while the lines represent the analytic formulas with fitted parameters. (a) Pairwise distance distribution. (b) Minimal distance distribution. (c-e)  $n$ -minimal distance distribution. (d) Cherries distance distribution. The lines are based on following set of parameters:  $\lambda - \mu = 15.2 \times 10^{-8} \text{yr}^{-1}$  and  $\lambda \sigma = 4.6 \times 10^{-8} \text{yr}^{-1}$ . For  $\mu = 0, 0.2, 0.4, 0.6, 0.8 \times \lambda$  this corresponds respectively to  $\sigma = 0.3, 0.24, 0.18, 0.12, 0.06$ .

doi:10.1371/journal.pone.0120206.g005

of these parameters, for instance the fraction  $\mu/\lambda$ . After estimation of  $(\lambda - \mu)$  and  $(\lambda \sigma)$  one can get an upper bound for the sampling fraction in the form (note that  $\mu \geq 0$ )

$$\sigma \leq \frac{(\sigma \lambda)}{(\lambda - \mu)}. \tag{31}$$

If the death rate is known to be much smaller than the birth rate,  $0 \leq \mu \ll \lambda$ , the upper bound is expected to be a good estimate for  $\sigma$ .



**Fig 6. Comparison of analytic predictions to the pairwise distances data of Tyrannidae family with  $M = 460$  species taken from the database [28] with  $t \leq 0.8 \times 10^8$  Myr.** The markers represent the empirical data, while the lines represent the analytic formulas with fitted parameters. (a) Pairwise distance distribution. (b) Minimal distance distribution. (c-e)  $n$ -minimal distance distribution. (d) Cherries distance distribution. The fit is performed for all points in the figure with  $t \leq 0.5$  to avoid clear break down of the Yule tree assumptions for larger distances (see text). The lines are based on following set of parameters:  $\lambda - \mu = 8 \times 10^{-8} \text{yr}^{-1}$  and  $\lambda\sigma = 6.4 \times 10^{-8} \text{yr}^{-1}$ . For  $\mu = 0, 0.2, 0.4, 0.6, 0.8 \times \lambda$  this corresponds respectively to  $\sigma = 0.8, 0.64, 0.48, 0.32, 0.16$ .

doi:10.1371/journal.pone.0120206.g006

If it is known that the sampling is perfect,  $\sigma = 1$ , one can estimate both the birth and the death rate. However, in contrast to Ref. [11], the method presented here does not require the reconstruction of the tree, but is solely based on statistical properties of pairwise distances between the leaves of the tree.

In the general case, one can get an upper limit for the sampling fraction and a lower limit for the birth rate by setting  $\mu/\lambda = 0$ . These bounds are expected to be useful for analysis of exponentially growing trees. Such trees can appear in phylogeny when analyzing the evolution of taxa, but also in population genetics, for instance, when considering an exponentially growing sub-population under the influence of a positive selection.

## Appendix

### Simulation details

To simulate Yule process for the generation of phylogenetic trees we use a Kinetic Monte Carlo algorithm. For a given birth rate  $\lambda$ , death rate  $\mu$ , and sampling fraction  $\sigma$ , the system is initiated with one “alive” lineage  $M = 1$  at time  $t = 0$ . The system is then iteratively propagated to the time  $t = T$ . In each iterative step one alive lineage is chosen at random and either either split into two alive lineages (with probability  $\lambda/(\lambda+\mu)$ ) or killed (with probability  $\mu/(\lambda+\mu)$ ). In each step the time is incremented by an amount  $\Delta t$  that is exponentially distributed with mean  $1/(M(\lambda+\mu))$ , where  $M$  is the number of alive lineages. After the time  $t = T$  has been reached, alive lineage are kept in the set of sampled leaves with probability  $\sigma$ .

During the whole simulation the complete tree—especially information about all branching points and branching times—are kept in memory. This way the distribution of pairwise distances or other quantities described in the text can easily be computed. To obtain the mean of such distributions we usually generated at least  $10^6$  trees and computed the averages.

### Second-minimal-distance distribution

Let  $N_2(t|T)dt$  be the average number of leaves on the tree of length  $T$ , separated by the time distance  $t$  from their second-most closely related leaf. Then, if the first branching occurs at time  $T_1$  and the two resulting subtrees possess  $M_1$  and  $M_2$  leaves, respectively, one gets the distribution of the minimal distance time in a form

$$N_2(t|T) = 2N_2(t|T - T_1)e^{-\mu T_1} + 2[2P(2|t/2)(1 - P(0|t/2)) + P(1|t/2)(1 - P(0|t/2) - P(1|t/2))] \times \delta(t - 2(T - T_1))I(0 \leq t \leq 2T)e^{-\mu T_1}. \tag{32}$$

After average over  $T_1$  and solving the resulting equation one obtains

$$N_2(t|T) = \frac{3\lambda^2(\lambda - \mu)^3\sigma^3\left(e^{\frac{t}{2}} - e^{\frac{\mu}{2}}\right)}{\left[e^{\frac{t}{2}}(\mu - \lambda + \sigma\lambda) - e^{\frac{\mu}{2}}\lambda\sigma\right]^4} e^{\frac{t}{2} + T\lambda + t\mu - T\mu} \tag{33}$$

for  $0 \leq t \leq 2T$ . Similarly, one can obtain any third-minimal distance distribution forth- etc. The general formula for the  $n$ -minimal-distance distribution is calculated in the following.

### $n$ -minimal-distance distribution

Let  $N_n(t|T)dt$  be the average number of leaves on the tree of length  $T$ , separated by the time distance  $t$  from their  $n$ -most closely related leaf. This notation means that 1-most closely related leaf is the closest one, 2-most closely related leaf is the second-most closest one etc. Then, if the first branching happens at time  $T_1$  and the two resulting subtrees possess  $M_1$  and  $M_2$  leaves, respectively, one gets the distribution of the minimal distance time in a form

$$N_n(t|T) = 2N_n(t|T - T_1)e^{-\mu T_1} + 2[nP(n|t/2)P_{>}(0|t/2) + (n - 1)P(n - 1|t/2)P_{>}(1|t/2) + \dots + P(1|t/2)P_{>}(n - 1|t/2)] \times \delta(t - 2(T - T_1))I(0 \leq t \leq 2T)e^{-\mu T_1} = \left[ 2N_n(t|T - T_1) + 2\delta(t - 2(T - T_1))I(0 \leq t \leq 2T) \sum_{k=1}^n kP(k|t/2)P_{>}(n - k|t/2) \right] e^{-\mu T_1}. \tag{34}$$

Here

$$P_{>}(k|T) = \frac{\sigma^{k+1}(\mu - \lambda)\lambda^k [e^{T(\mu-\lambda)} - 1]^k (e^{T\lambda} - e^{T\mu})^k e^{T\lambda}}{[e^{T\mu}(\mu - \lambda + \sigma\lambda) - e^{T\lambda}\lambda\sigma]^{k+1} [\lambda - e^{T(\mu-\lambda)}\mu]^k} \tag{35}$$

is the probability to observe more than  $k$  leaves on a tree of age  $T$  and  $P(n|T)$  is given in Equations (3, 4). After average over  $T_1$  and solving the resulting equation one obtains

$$N_n(t|T) = \frac{n(1+n)(\mu - \lambda)^3 \sigma(\lambda\sigma)^n [e^{\frac{t}{2}(\mu-\lambda)} - 1]^{n-1} e^{\frac{\mu\lambda}{2} + T\lambda + t\mu - T\mu}}{2 [e^{\frac{t\mu}{2}(\mu - \lambda + \sigma\lambda)} - e^{\frac{t\lambda}{2}\lambda\sigma}]^{n+2}} \tag{36}$$

for  $0 \leq t \leq 2T$  and 0 otherwise, resulting in Equation (20).

### Cherries-distance distribution

A cherry is a pair of adjacent tips on a tree (see Fig. 2). Let  $N_{\Lambda}(t|T)dt$  be the average number of cherry pairs on the tree of length  $T$ , separated by the time distance  $t$ . Then, if the first branch splits at time  $T_1$  and the two resulting subtrees possess  $M_1$  and  $M_2$  leaves, respectively, one gets the distribution in the form

$$N_{\Lambda}(t|T) = [2N_{\Lambda}(t|T - T_1) + P^2(1|T - T_1)\delta(t - 2(T - T_1))I(0 \leq t \leq 2T)]e^{-\mu T_1}. \tag{37}$$

After average over  $T_1$  and solving the resulting equation one obtains

$$N_{\Lambda}(t|T) = \frac{\lambda(\lambda - \mu)^4 \sigma^2 e^{\frac{t\lambda}{2} + T\lambda + \frac{3t\mu}{2} - T\mu}}{2 [e^{\frac{t\mu}{2}(\mu - \lambda + \sigma\lambda)} - e^{\frac{t\lambda}{2}\lambda\sigma}]^4} \tag{38}$$

for  $0 \leq t \leq 2T$  and 0 otherwise, resulting in Equation (21).

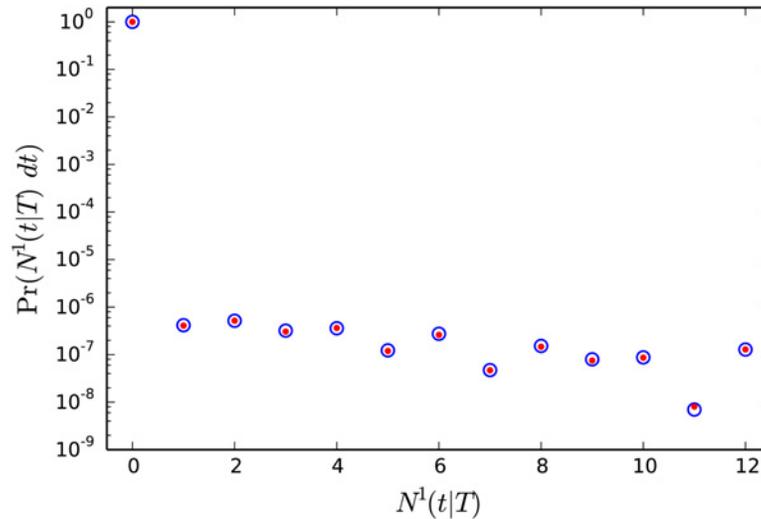
### The distribution of $N^1(t|T)dt$

In this Appendix we derive the distribution of  $N^1(t|T)dt$ . Consider an infinitesimal (in practice very small) interval,  $[t, t+dt]$ , such that  $N(t|T)dt \ll 1$ . The number of pairs  $N^1(t|T)dt$  in this interval is distributed with the mean  $N(t|T)dt$ . The full distribution can be derived using the following arguments.

Pairs, separated by the time in the interval  $[t, t+dt]$ , branched at the time interval  $[T-t/2-dt/2, T-t/2]$ . The probability to have a branch in this interval is given by  $\lambda e^{(\lambda-\mu)(T-t/2)} dt/2$ . Given that there is a branching point in this interval it can lead to different number of leaves and, therefore, pairs separated by the time in the interval  $[t, t+dt]$ . The probability that no observed pairs survive from this branching is given by  $1 - [1 - P(0|t/2)]^2$ , where  $P(n|T)$  is the probability to observe  $n$  leaves on a tree of age  $T$  and is given in Equations (3, 4). The probability that there are no observed pairs separated by the time in the interval  $[t, t+dt]$  is given by Equation (25). The probability that there are  $n > 0$  observed pairs separated by the time in the interval  $[t, t+dt]$  is given by

$$\begin{aligned} \Pr(N^1(t|T)dt = n) &= \lambda e^{(\lambda-\mu)(T-t/2)} dt/2 \sum_{n_1, n_2=1}^n P(n_1|t/2)P(n_2|t/2)\delta_{n_1 n_2, n} \\ &= \lambda e^{(\lambda-\mu)(T-t/2)} dt/2 \sum_{n_1|n} P(n_1|t/2)P(n/n_1|t/2). \end{aligned} \tag{39}$$

The last sum runs over all divisors of  $n$ , including 1 and  $n$ . One can see the comparison of Equations (25) and (39) to numerical results in Fig. 7.



**Fig 7. Probability to observe a certain number of pairs separated by the time in the interval  $[t, t+dt]$  on a tree of age  $T$ ,  $N^1(t|T)dt$ .** In this plot  $T = 1$ ,  $\lambda = 11$ ,  $\mu = 5$ ,  $\sigma = 0.01$ ,  $t = 1.5$  and  $dt = 0.00001$ . Circles denote the results of numerical simulation and dots were obtained using the analytic formulas (25) for zero value and (39) for non-zero values. Note the gap between zero and non-zero probabilities due to small bin size,  $dt$ .

doi:10.1371/journal.pone.0120206.g007

## Acknowledgments

The authors thank M. Mariadassou, P.W. Messer, and M. Vingron for helpful discussions.

## Author Contributions

Conceived and designed the experiments: MS FM PA. Performed the experiments: MS FM PA. Wrote the paper: MS FM PA.

## References

1. Yule G (1924) A mathematical theory of evolution, based on the conclusions of dr. jc willis. Philosophical Transactions of the Royal Society of London B B213: 21.
2. Karlin S, Taylor H (1975) A first course in stochastic processes. Academic Press, New York.
3. Newman ME (2005) Power laws, pareto distributions and zipf's law. Contemporary physics 46: 323. doi: [10.1080/00107510500052444](https://doi.org/10.1080/00107510500052444)
4. Novozhilov AS, Karev GP, Koonin EV (2006) Biological applications of the theory of birth-and-death processes. Briefings in bioinformatics 7: 70. doi: [10.1093/bib/bbk006](https://doi.org/10.1093/bib/bbk006) PMID: [16761366](https://pubmed.ncbi.nlm.nih.gov/16761366/)
5. Yanai I, Camacho CJ, DeLisi C (2000) Predictions of gene family distributions in microbial genomes: evolution by gene duplication and modification. Physical Review Letters 85: 2641. doi: [10.1103/PhysRevLett.85.2641](https://doi.org/10.1103/PhysRevLett.85.2641) PMID: [10978127](https://pubmed.ncbi.nlm.nih.gov/10978127/)
6. Reed WJ, Hughes BD (2004) A model explaining the size distribution of gene and protein families. Mathematical biosciences 189: 97. doi: [10.1016/j.mbs.2003.11.002](https://doi.org/10.1016/j.mbs.2003.11.002) PMID: [15051416](https://pubmed.ncbi.nlm.nih.gov/15051416/)
7. Tambovtsev Y, Martindale C (2007) Phoneme frequencies follow a yule distribution. SKASE Journal of Theoretical Linguistics 4: 1.
8. Raup DM (1985) Mathematical models of cladogenesis. Paleobiology 11: 42.
9. Aldous DJ (2001) Stochastic models and descriptive statistics for phylogenetic trees, from yule to today. Statistical Science: 23.
10. Nee S, May RM, Harvey PH (1994) The reconstructed evolutionary process. Philosophical Transactions of the Royal Society of London Series B: Biological Sciences 344: 305. doi: [10.1098/rstb.1994.0068](https://doi.org/10.1098/rstb.1994.0068) PMID: [7938201](https://pubmed.ncbi.nlm.nih.gov/7938201/)

11. Nee S, Holmes EC, May RM, Harvey PH (1994) Extinction rates can be estimated from molecular phylogenies. *Philosophical Transactions of the Royal Society of London Series B: Biological Sciences* 344: 77. doi: [10.1098/rstb.1994.0054](https://doi.org/10.1098/rstb.1994.0054) PMID: [8878259](https://pubmed.ncbi.nlm.nih.gov/8878259/)
12. Kendall DG (1949) Stochastic processes and population growth. *Journal of the Royal Statistical Society Series B (Methodological)* 11: 230.
13. Harvey PH, May RM, Nee S (1994) Phylogenies without fossils. *Evolution*: 523.
14. McKenzie A, Steel M (2000) Distributions of cherries for two models of trees. *Mathematical biosciences* 164: 81. doi: [10.1016/S0025-5564\(99\)00060-7](https://doi.org/10.1016/S0025-5564(99)00060-7) PMID: [10704639](https://pubmed.ncbi.nlm.nih.gov/10704639/)
15. Steel M, McKenzie A (2001) Properties of phylogenetic trees generated by yule-type speciation models. *Mathematical biosciences* 170: 91. doi: [10.1016/S0025-5564\(00\)00061-4](https://doi.org/10.1016/S0025-5564(00)00061-4) PMID: [11259805](https://pubmed.ncbi.nlm.nih.gov/11259805/)
16. Rosenberg NA (2006) The mean and variance of the numbers of r-pronged nodes and r-caterpillars in yule-generated genealogical trees. *Annals of Combinatorics* 10: 129. doi: [10.1007/s00026-006-0278-6](https://doi.org/10.1007/s00026-006-0278-6)
17. Mulder WH (2011) Probability distributions of ancestries and genealogical distances on stochastically generated rooted binary trees. *Journal of theoretical biology* 280: 139. doi: [10.1016/j.jtbi.2011.04.009](https://doi.org/10.1016/j.jtbi.2011.04.009) PMID: [21527261](https://pubmed.ncbi.nlm.nih.gov/21527261/)
18. Steel M, Mooers A (2010) The expected length of pendant and interior edges of a yule tree. *Applied Mathematics Letters* 23: 1315. doi: [10.1016/j.aml.2010.06.021](https://doi.org/10.1016/j.aml.2010.06.021)
19. Mooers A, Gascuel O, Stadler T, Li H, Steel M (2012) Branch lengths on birth–death trees and the expected loss of phylogenetic diversity. *Systematic biology* 61: 195. doi: [10.1093/sysbio/syr090](https://doi.org/10.1093/sysbio/syr090) PMID: [21865336](https://pubmed.ncbi.nlm.nih.gov/21865336/)
20. Kumar S (2005) Molecular clocks: four decades of evolution. *Nature Reviews Genetics* 6: 654. doi: [10.1038/nrg1659](https://doi.org/10.1038/nrg1659) PMID: [16136655](https://pubmed.ncbi.nlm.nih.gov/16136655/)
21. Slatkin M, Hudson RR (1991) Pairwise comparisons of mitochondrial dna sequences in stable and exponentially growing populations. *Genetics* 129: 555. PMID: [1743491](https://pubmed.ncbi.nlm.nih.gov/1743491/)
22. Mora C, Tittensor DP, Adl S, Simpson AG, Worm B (2011) How many species are there on earth and in the ocean? *PLoS biology* 9: e1001127. doi: [10.1371/journal.pbio.1001127](https://doi.org/10.1371/journal.pbio.1001127) PMID: [21886479](https://pubmed.ncbi.nlm.nih.gov/21886479/)
23. Pimm SL, Russell GJ, Gittleman JL, Brooks TM (1995) The future of biodiversity. *Science*: 347.
24. Hohna S, Stadler T, Ronquist F, Britton T (2011) Inferring speciation and extinction rates under different sampling schemes. *Molecular biology and evolution* 28: 2577. doi: [10.1093/molbev/msr095](https://doi.org/10.1093/molbev/msr095) PMID: [21482666](https://pubmed.ncbi.nlm.nih.gov/21482666/)
25. Kendall DG (1948) On some modes of population growth leading to ra fisher’s logarithmic series distribution. *Biometrika*: 6.
26. Massip F, Sheinman M, Schbath S and Arndt PF (2015) How Evolution of Genomes Is Reflected in Exact DNA Sequence Match Statistic. *Mol. Biol. Evol.* 32(2): 524. doi: [10.1093/molbev/msu313](https://doi.org/10.1093/molbev/msu313) PMID: [25398628](https://pubmed.ncbi.nlm.nih.gov/25398628/)
27. Stadler T (2009) On incomplete sampling under birth–death models and connections to the sampling-based coalescent. *Journal of Theoretical Biology* 261: 58. doi: [10.1016/j.jtbi.2009.07.018](https://doi.org/10.1016/j.jtbi.2009.07.018) PMID: [19631666](https://pubmed.ncbi.nlm.nih.gov/19631666/)
28. Jetz W, Thomas G, Joy J, Hartmann K, Mooers A (2012) The global diversity of birds in space and time. *Nature* 491: 444. doi: [10.1038/nature11631](https://doi.org/10.1038/nature11631) PMID: [23123857](https://pubmed.ncbi.nlm.nih.gov/23123857/)
29. Holland PW, Welsch RE (1977) Robust regression using iteratively reweighted least-squares. *Communications in Statistics-Theory and Methods* 6: 813. doi: [10.1080/03610927708827533](https://doi.org/10.1080/03610927708827533)
30. Bouckaert R, Heled J, Kuhnert D, Vaughan T, Wu CH, et al. (2014) Beast 2: a software platform for bayesian evolutionary analysis. *PLoS computational biology* 10: e1003537. doi: [10.1371/journal.pcbi.1003537](https://doi.org/10.1371/journal.pcbi.1003537) PMID: [24722319](https://pubmed.ncbi.nlm.nih.gov/24722319/)
31. Ronquist F, Huelsenbeck JP (2003) Mrbayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* 19: 1572. doi: [10.1093/bioinformatics/btg180](https://doi.org/10.1093/bioinformatics/btg180) PMID: [12912839](https://pubmed.ncbi.nlm.nih.gov/12912839/)