RESEARCH ARTICLE

Coalescence computations for large samples drawn from populations of time-varying sizes

Andrzej Polanski[1]*, **Agnieszka Szczesna**[1], **Mateusz Garbulowski**[1,2], **Marek Kimmel**[3,4]

1 Institute of Informatics, Silesian University of Technology, ul. Akademicka 16, 44-100 Gliwice, Poland, **2** The Linnaeus Centre for Bioinformatics, Uppsala University, BMC, Uppsala, Sweden, **3** Systems Engineering Group, Silesian University of Technology, ul. Akademicka 16, 44-100 Gliwice, Poland, **4** Department of Statistics, Rice University, M.S. 138, 6100 Main Street, Houston, TX 77005, United States of America

* Andrzej.Polanski@polsl.pl

Citation: Polanski A, Szczesna A, Garbulowski M, Kimmel M (2017) Coalescence computations for large samples drawn from populations of time-varying sizes. PLoS ONE 12(2): e0170701. doi:10.1371/journal.pone.0170701

Editor: Zaid Abdo, Colorado State University, UNITED STATES

Received: July 12, 2016

Accepted: January 9, 2017

Published: February 7, 2017

Data Availability Statement: Data in table 1 are from the study: Chen H., Chen K., (2013), Asymptotic Distributions of Coalescence Times and Ancestral Lineage Numbers for Populations with Temporally Varying Size, Genetics, vol. 194, pp. 721. Data in table 2 are from the mtDB database (http://www.mtdb.igp.uu.se/). The related paper is: Ingman, M., Gyllensten, U., (2006), mtDB: Human Mitochondrial Genome Database, a resource for population genetics and medical sciences. Nucleic Acids Res 34, D749-D751. Our Matlab code is avaiable as a supplement to the paper at the publisher's page and as the post at the

Abstract

We present new results concerning probability distributions of times in the coalescence tree and expected allele frequencies for coalescent with large sample size. The obtained results are based on computational methodologies, which involve combining coalescence time scale changes with techniques of integral transformations and using analytical formulae for infinite products. We show applications of the proposed methodologies for computing probability distributions of times in the coalescence tree and their limits, for evaluation of accuracy of approximate expressions for times in the coalescence tree and expected allele frequencies, and for analysis of large human mitochondrial DNA dataset.

Introduction

Coalescent theory [1, 2], widely used for statistical inference on genetic parameters and structures of evolving populations is a thoroughly studied area with many results published over decades. The classical coalescent model concerns a sample drawn from a population which has evolved with constant size over many generations in the past. For such a model many results concerning e.g., probability distributions of times in the coalescence tree [3, 4], expected ages [5, 6] and frequencies of mutations and recombinations [3, 4] were developed. Since majority of populations undergo changes in their size in the course of their evolution several authors developed coalescence computations for the case of time dependent population sizes, either by deriving analytical approaches [5, 7–9] or by using stochastic coalescence simulations [5, 10]. Other directions of developing coalescent modeling involve different scenarios of populations evolution, constant or undergoing expansions or bottlenecks, combined with possible inhomogeneity of their structures [11, 12], as well as different models of mutation, infinite size, infinite alleles, recurrent, stepwise. There are also several studies concerning coalescence modeling for populations under selection [13–15].

Emergence of large datasets resulting from contemporary sequencing technologies has drawn attention of researchers to problems in the coalescent theory arising in the situation where the coalescent sample size becomes large. There are several areas of analysis of such problems. Below we describe those of them, which are in the scope of our interest in the

GitHub repository (https://github.com/agnieszkaszczesna/Coalescence-Computations-for-Large-Samples).

Funding: This work was supported by Polish National Science Centre, OPUS grant (2016/21/B/ST6/02153) (Prof. Andrzej Polanski [AP]); Polish National Science Centre, Maestro grant (DEC-012/04/A/ST7/00353) (Prof. Marek Kimmel [MK]); Silesian University of Technology, Institute of Informatics, Statute project, (BK/RAU-2/2016) (PhD Agnieszka Szczesna [AS]), and Polish National Science Centre, SYMFONIA 3 grant (UMO 2015/16/W/NZ2/00314) (MSc Mateusz Garbulowski [MG]). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

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

present paper. The first area is pursuing computations for existing algorithms for the case of large sizes of the coalescence tree. Many authors have pointed out (e.g., [7, 9, 16, 17] that some of the computational algorithms for coalescence modeling, published in the literature, are applicable only for relatively small sizes of samples, below 50. In view of availability of data of much larger sizes it is interesting and important to study and develop analogous or corresponding methodologies suitable for larger sample sizes. The second area includes computing limits and/or growth/shrinkage patterns of distributions (expectations) of coalescence times and allele frequencies in the coalescence process. When the sample size tends to large values it becomes a natural and important question whether limits of distributions/values exist, how they can be computed and used in population and statistical genetics research. The third area includes developing large sample approximations. Large sample approximations usually have the form of simple analytical expressions and therefore may be very useful for (approximate) statistical inference. Large sample approximations also give a valuable support for intuitive understanding of mechanisms of evolution of large samples. Two types of approximations have been applied for large samples, deterministic and normal. Deterministic approximation is based on the fact that the coalescence process, which in principle has a stochastic nature, converges partly to a deterministic scenario when sample size goes to infinity [15, 18]. In the deterministic approach times in the coalescence process are represented by deterministic values. In the normal approximation approach times in the coalescence process are modeled (approximated) by normal distributions [16] on the basis of the fact that, under the constant population scenario, they are sums of many independent components [19].

The aim of the present paper is to show new results and conclusions in the three areas listed above. We derive our results by using new methods for computing exact probability distributions and expectations of times to coalescences for trees of arbitrary large sizes and for arbitrary scenarios of population time change. In previous publications [16, 20] such distributions and values were computed by using approximations or estimated by stochastic coalescence simulations. The proposed approach is based on deriving the inverse of the integral transform introduced in [7]. Further we derive the limit distribution of the time to most recent common ancestor, under different scenarios of population size change, which uses the gamma quotient formula for infinite products [21]. We show the following applications of the proposed approaches:

- Computing probability distributions and expectations of coalescence times for genealogies of large samples of DNA sequences, with high accuracy. In previous articles such distributions and expectations were estimated on the basis of coalescence simulations or approximate methods.
- Computing limit distributions of times to most recent common ancestor in the coalescene tree under different rates of population growth.
- Evaluation of accuracy of published large sample approximations [16] for times in the coalescence tree and expected allele frequencies.
- Estimation of rates of convergence of distributions of times in the coalescence tree to their limits.
- Fitting the exponential growth model to DNA polymorphisms data from the whole database of mitochondrial DNA for over 2000 individuals [22].

In the "Discussion" section we show some other possible applications of the derived results and some possible further directions of the research.

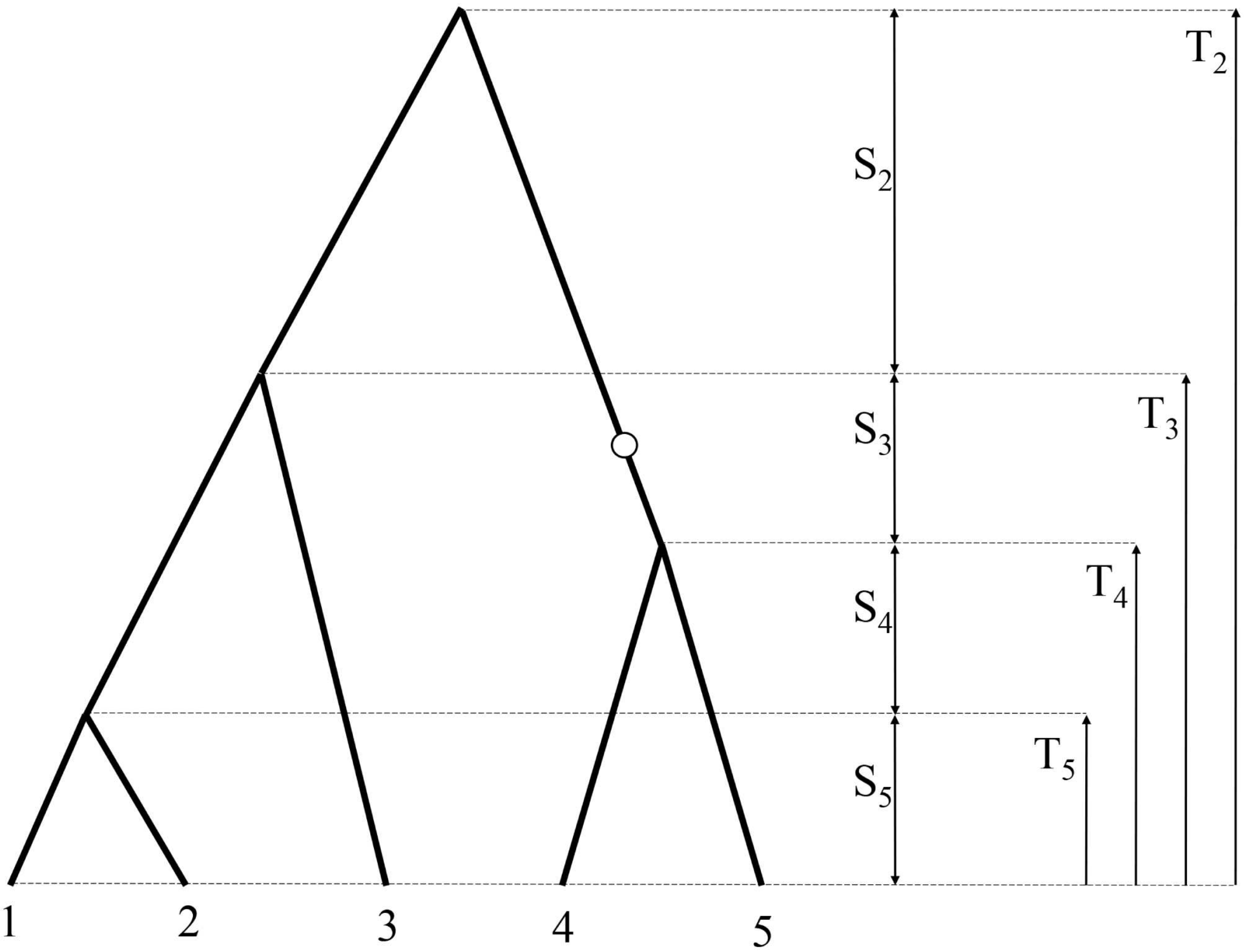

Fig 1. Coalescence tree and notation for ancestral history of a sample of *n* = 5 DNA sequences. Times to coalescence events are denoted by capital letters *T*, times between coalescences are denoted by capital letters *S*. A mutation event is marked by an open circle.

doi:10.1371/journal.pone.0170701.g001

Results

Results, which we show in this paper concern the past history of an n-sample (of DNA sequences) taken at present, as illustrated in Fig 1 where samples are numbered from 1 to $n = 5$. Time t is measured from the present to the past with the units defined by number of generations. We assume validity of the diffusion approximation [23], so t is a continuous variable. Coalescences are events of merging (joining) branches of the phylogenic tree of samples. Random coalescence times from sample of size n to sample of size $k - 1$ are denoted by T_k, $k = 2$, $3\ldots n$, and their realizations by corresponding lower case letters $t_n, t_{n-1}, \ldots, t_2$, $0 < t_n < t_{n-1}\ldots<t_2$. Times between coalescence events are denoted by the capital and lower case letters S, s; in Fig 1 these times are denoted by $S_5, S_4, \ldots, S_2$. Apart from coalescnce times $T_2, \ldots, T_{n-1}$, T_n and times between coalescence events $S_2, \ldots, S_{n-1}, S_n$ of special interest (e.g., [5, 7, 8, 16]) are also the time to the most recent common ancestor (*TMRCA*) and total length of branches

in the coalescence tree (*TLBT*), defined as follows

$$TMRCA = T_2, \tag{1}$$

and

$$TLBT = \sum_{k=2}^{n} kS_k = T_2 + \sum_{k=2}^{n} T_k. \tag{2}$$

During the DNA replication process mutation events occur along branches of the coalescence tree. In Fig 1 an exemplary mutation event is marked by an open circle. Consequently, sequences 4 and 5 have mutant alleles (bases), while sequences 1, 2 and 3—ancestral ones. We assume that the mutation process is described by a Poisson point process and that assumptions of the infinite sites model are satisfied (e.g., [5, 16, 24]). Allele frequencies corresponding to mutations depend on times in the coalescence tree S_k and on mutation intensity μ. Expected allele frequency, f_{nb}, of mutation of type b, i.e., having b mutant bases versus $n - b$ ancestral bases in the leaves of the coalescence tree (in Fig 1 $b = 2$, $n - b = 3$), is given by the following expression (e.g., [5, 16, 24])

$$f_{nb} = \mu \frac{(n-b-1)!(b-1)!}{(n-1)!} \sum_{k=2}^{n} \binom{n-k}{b-1} k(k-1)E(S_k), \ 1 \le b \le n-1. \tag{3}$$

Under the additional hypothesis that μ is close to zero, probability, p_{nb}, that a randomly chosen mutation is of type b is given by the following expression [5]

$$p_{nb} = \frac{f_{nb}}{\sum_b f_{nb}} = \frac{f_{nb}}{\mu E(TLBT)}. \tag{4}$$

The effective size of the underlying population is assumed to be given by a deterministic function $N(t)$, $t \in [0, \infty)$. Two special cases of population size change (growth) scenarios are often researched, constant and exponential. The constant population size scenario is denoted as

$$N(t) = N^C(t) = N_0. \tag{5}$$

The exponential growth scenario is given by

$$N(t) = N^E(t) = N_0 \exp(-rt) \tag{6}$$

where r is the exponent parameter. For exponential growth we also denote

$$\rho = rN_0, \tag{7}$$

and we call ρ the product parameter of the population exponential growth. With $r = 0$ the exponential scenario Eq (6) becomes the constant scenario Eq (5).

Probability distributions of coalescence times

In this subsection we present results concerning probability distributions of times in the coalescence tree, which according to our best knowledge were not published before. We obtain them by using the methods described in subsections "Inversion of the integral transform" and "Limit distributions" of the "Methods" section. In Fig 2 we show probability distributions of *TMRCA*, $\pi_{TMRCA}\left(\frac{t}{N_0}\right)$, for genealogy sizes $n = 10$, $n = 100$ and $n = \infty$, for different scenarios of

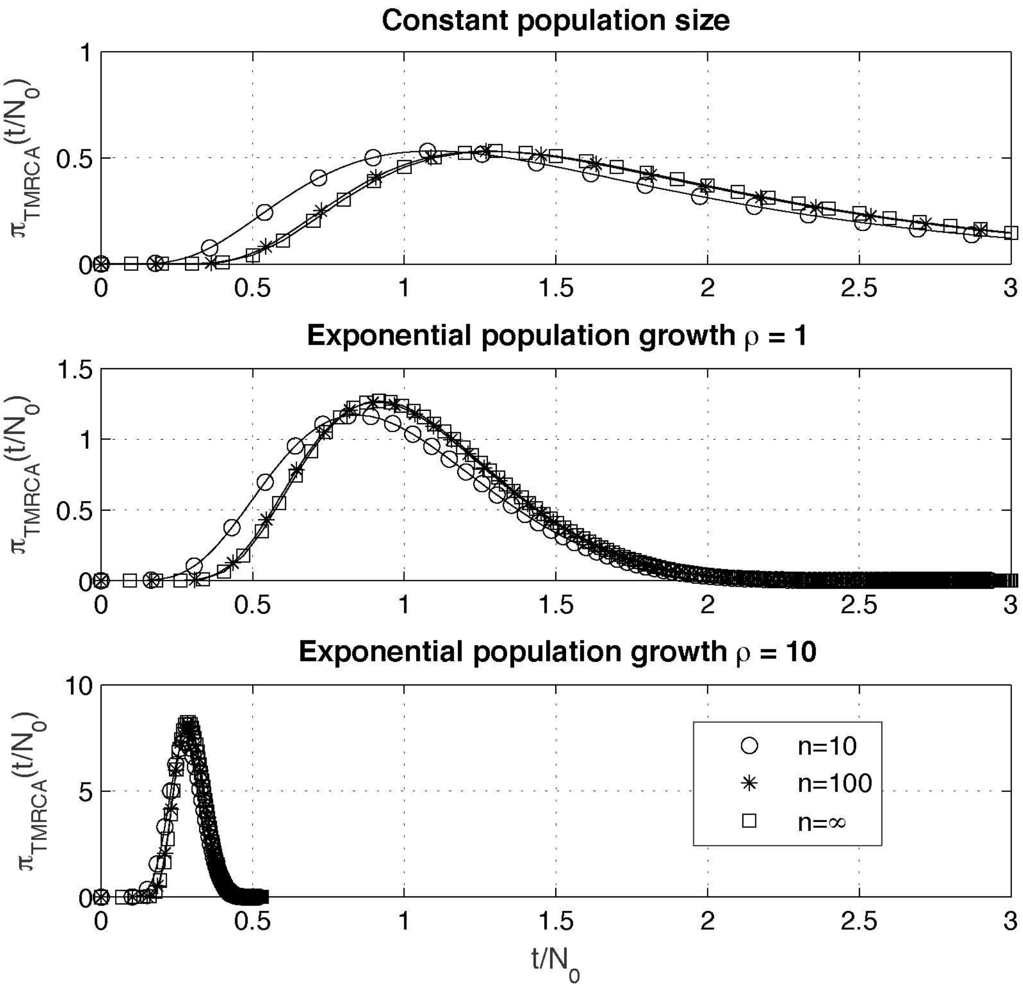

Fig 2. Probability density functions, $\pi_{TMRCA}\left(\frac{t}{N_0}\right)$, for different scenarios of populations size change, constant (upper plot) and exponentially growing with $\rho = 1$ (middle plot) and $\rho = 10$ (lower plot). Probability distributions are shown for different genealogy sizes $n = 10$, $n = 100$ and $n = \infty$ (limit distribution).

doi:10.1371/journal.pone.0170701.g002

populations size change, constant (upper plot) and exponentially growing with $\rho = 1$ (middle plot) and $\rho = 10$ (lower plot). Convergence of probability density functions of *TMRCA* to the limit distribution, derived in subsection “Limit distributions”, is rather fast. One can observe that time scale change related to exponential scenario of population growth with increasing ρ results in probability distribution of *TMRCA* with increasing similarity to normal distribution.

The result published by [19] states that probability distributions of times T_k in the middle of the coalescence tree converge to normal distribution when $n \to \infty$. Using our expressions from subsection “Inversion of the integral transform” of the “Methods” section, we have numerically studied the rate of this convergence by computing skewness coefficients $\gamma(T_k)$,

$$\gamma(T_k) = E\left[\left(\frac{T_k - E(T_k)}{Std(T_k)}\right)^3\right] \tag{8}$$

for distributions of times T_k when the index k changed from top to the bottom of the

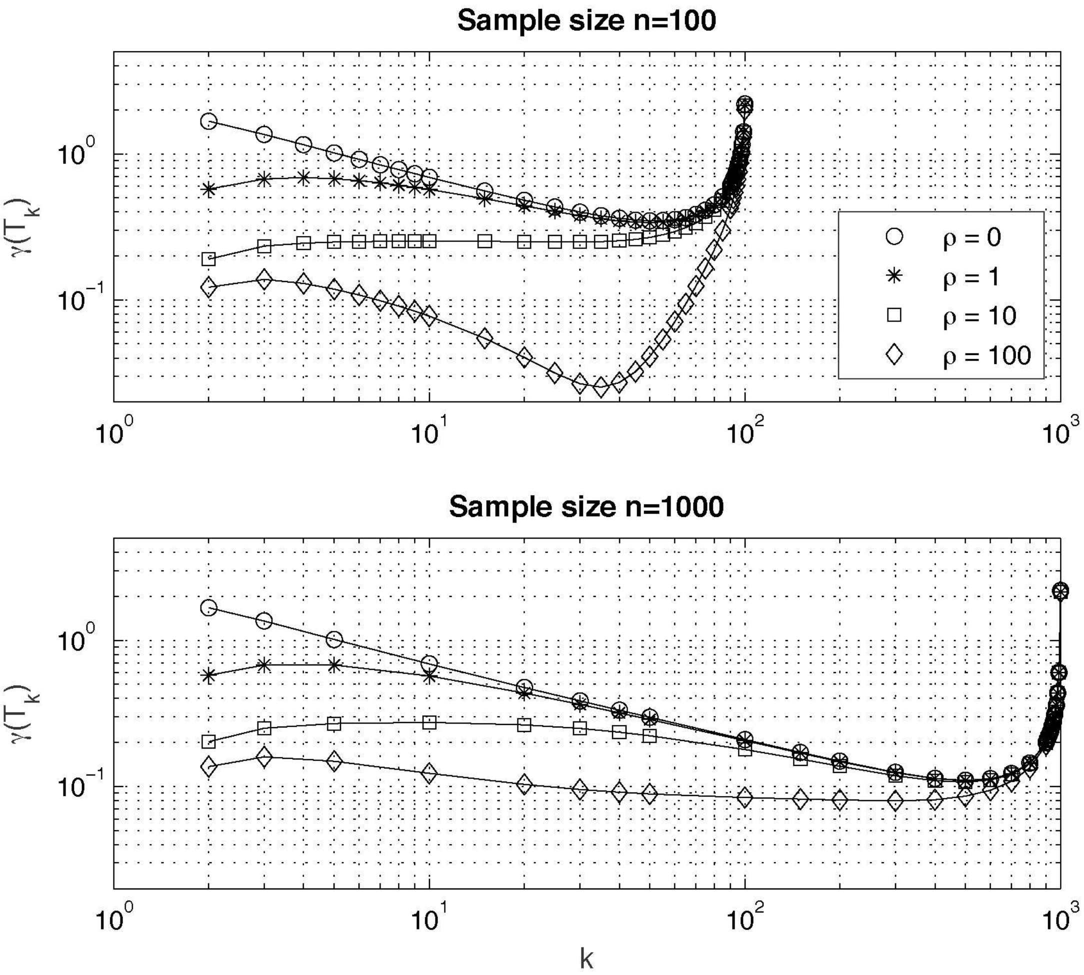

Fig 3. Values of skewness coefficient $\gamma(T_k)$ of probability distributions of times in the coalescence tree computed for different genealogy sizes, $n = 100$ (upper plot) and $n = 1000$ (lower plot) and for different scenarios of population size change constant ($\rho = 0$) and exponentially growing with $\rho = 1$, $\rho = 10$ and $\rho = 100$.

doi:10.1371/journal.pone.0170701.g003

coalescence tree. Values of skewness coefficient allow for estimating departure of the distribution of interest from normality.

Plots of skewness coefficient $\gamma(T_k)$ for different genealogy sizes, $n = 100$ (upper plot) and $n = 1000$ (lower plot) and for different scenarios of population size change constant ($\rho = 0$) and exponentially growing with $\rho = 1$, $\rho = 10$ and $\rho = 100$ are shown in Fig 3.

From the plots in Fig 3 one can see that skewness of probability distributions of times T_k decrease for increasing values of the product parameter ρ. For all scenarios, constant and exponential with different ρ, one observes a sharp increase of values of skewness coefficient $\gamma(T_k)$ in the fourth quartile of the range of values of the coalescence tree index k.

Accuracy of approximate formulae for expectations of coalescence times

Large sample approximations for probability distributions and expectations of coalescence times are very useful due to both their simple forms, and applicability to samples of arbitrary large size. Chen and Chen, (2013) [16], derived large sample approximations for expected

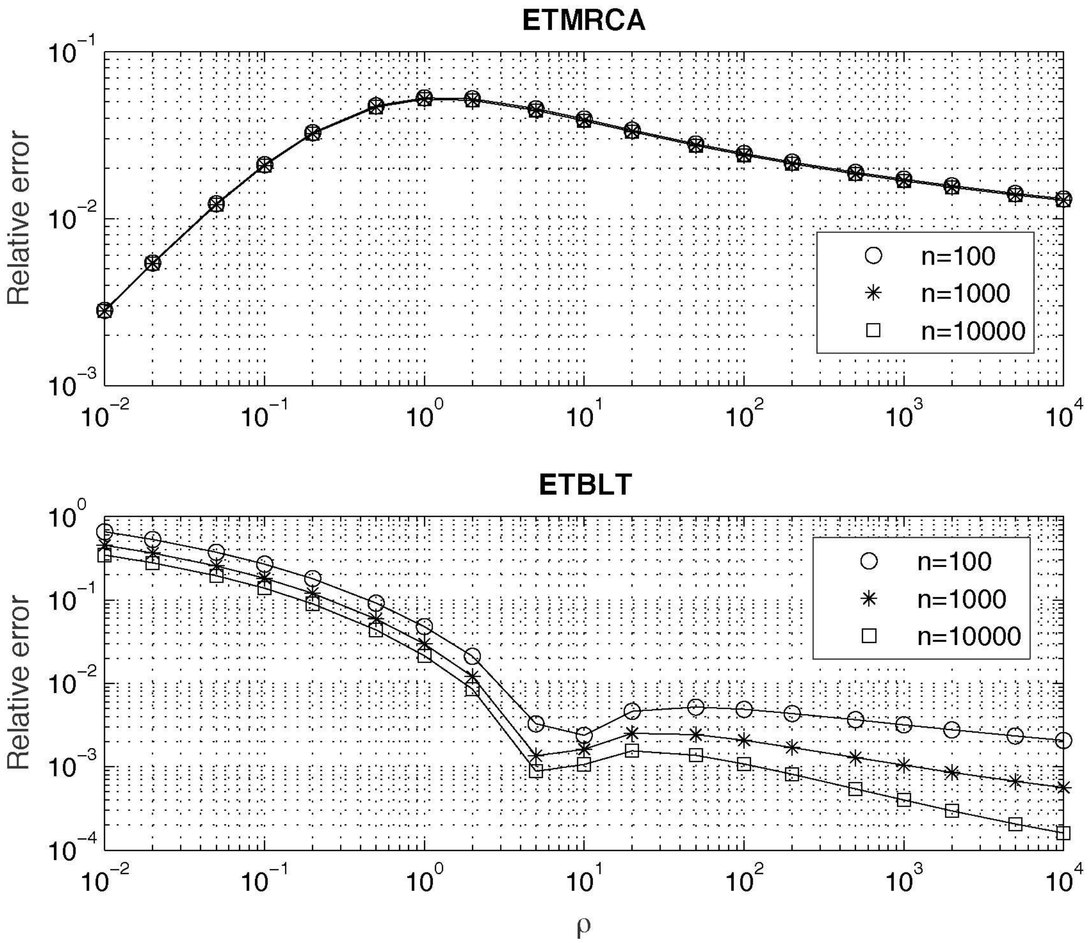

Fig 4. Relative errors of approximations for *ETMRCA* (upper plot) and *ETBLT* (lower plot) proposed by Chen and Chen (2013) [16].

doi:10.1371/journal.pone.0170701.g004

coalescence times, $E(T_k^E)$, *ETMRCA*, *ETBLT* for the case of exponential growth of population underlying the coalescence process. An important issue for applications of Chen and Chen's approximate formulae ("Methods" section, Eqs (26)–(28), is how accurately they approximate exact values. Chen and Chen (2013) [16] in their Figure 5 show plots, which demonstrate accuracy of their approximations of *ETMRCA* Eq (27) and *ETBLT* Eq (28). However, estimation of accuracy of approximation of *ETMRCA* (upper plot "A" in Figure 5 in Chen and Chen, (2013) [16] is based only on simulations, which reduces precision of estimation. The lower plot "B" in Figure 5 in Chen and Chen, (2013) [16] shows good accuracy of approximation Eq (28). However, one can examine accuracy only in qualitative terms.

Here we precisely evaluate accuarcy of Chen and Chen's approximations by using the approach presented in the "Methods" section. This approach is justified by results of the numerical study which proves that for sample sizes of orders of hundreds of thousands relative accuracy is better than 10^{-6} (see the Methods section).

In Fig 4 we show plots of relative errors of Chen and Chen's, (2013) [16] approximations ("Methods" section, Eqs (27) and (28) for *ETMRCA* and *ETBLT* computed by using our expressions given in "Methods" section, Eqs (22)–(25). It is easily seen, that relative error for both *ETMRCA* and *ETBLT*, for a given sample size *n* depends only on the value of the product

Table 1. Comparison of exact expectations and standard deviations of times to coalescence T_k to their asymptotic approximations proposed by Chen and Chen (2013) [16], for $n = 800$.

r	k	Mean T_k			Standard deviation of T_k		
		Exact	Asymptotic	Bias%	Exact	Asymptotic	Bias%
0.001	6	6647.928	6679.599	0.476	250.612	258.485	3.141
0.001	11	5964.931	5981.414	0.276	181.193	184.235	1.678
0.001	51	4327.067	4330.733	0.085	85.654	85.933	0.326
0.001	201	2771.310	2772.589	0.046	50.608	50.631	0.045
0.001	401	1790.748	1791.759	0.057	45.001	45.005	0.009
0.001	796	30.870	30.962	0.299	13.556	13.635	0.581
0.005	6	1651.262	1657.606	0.385	50.175	51.749	3.136
0.005	11	1514.458	1517.766	0.218	36.314	36.922	1.674
0.005	51	1185.169	1185.918	0.063	17.314	17.369	0.317
0.005	201	865.864	866.147	0.033	10.656	10.659	0.030
0.005	401	651.350	651.619	0.041	10.389	10.386	0.030
0.005	796	28.849	29.206	1.243	11.891	12.153	2.204
0.01	6	894.933	898.105	0.354	25.091	25.878	3.137
0.01	11	826.518	828.172	0.200	18.167	18.465	1.670
0.01	51	661.765	662.141	0.057	8.669	8.696	0.315
0.01	201	501.585	501.728	0.029	5.363	5.365	0.032
0.01	401	393.043	393.183	0.036	5.297	5.295	0.034
0.01	796	26.796	27.343	2.043	10.361	10.699	3.262

doi:10.1371/journal.pone.0170701.t001

parameter of the population growth ρ Eq (7). When computing approximate *ETBLT* one needs to replace the value on the right hand side of Eq (28) by its limit, $2N_0$, in the case when $n = 2rN_0 = 2\rho$. As seen from upper and lower plots in Fig 4 relative approximation errors of *ETMRCA* and *ETBLT* show quite complicated, nonlinear dependence on ρ. Relative error committed when using Chen and Chen's approximation approximation for *ETMRCA* ("Methods" section, Eq (27)) is of order of percents. For $n > 100$ this error practically does not depend on n, which is consistent with fast convergence of distribution of *TMRCA* (shown in Fig 2). Relative error committed when using Chen and Chen's approximation for *ETBLT* ("Methods" section, Eq (28)) increases for small values of ρ and decreases for large ρ and for large sample sizes n. For values of $\rho > 10$ and for sample sizes $n > 100$ accuracy of approximation Eq (28) is very good, of the order of 10^{-3} or better, consistently to results already shown in [16]).

In their Table 1 Chen and Chen (2013) [16] show values of asymptotic approximations of expectations and standard deviations of coalescence times. In order to evaluate accuracy of these approximations they compare them to averages computed over stochastic simulations, which results in committing error resulting from random variation in simulations. Below in our Table 1 we have reproduced one part of Chen and Chen's (2013) [16] Table 1, corresponding to sample size $n = 800$. In our Table 1 we have replaced estimates of expectations and standard deviations of coalescence times obtained by simulations by their exact values computed with the use of our algorithm described in subsection "Inversion of the integral transform" of the "Methods" section. Chen and Chen (2013) [16] use index named m to number coalescence times. Our index used for numbering coalescence times is k. Due to different notation conventions these indexes must be shifted by 1 in order to obtain corresponding results.

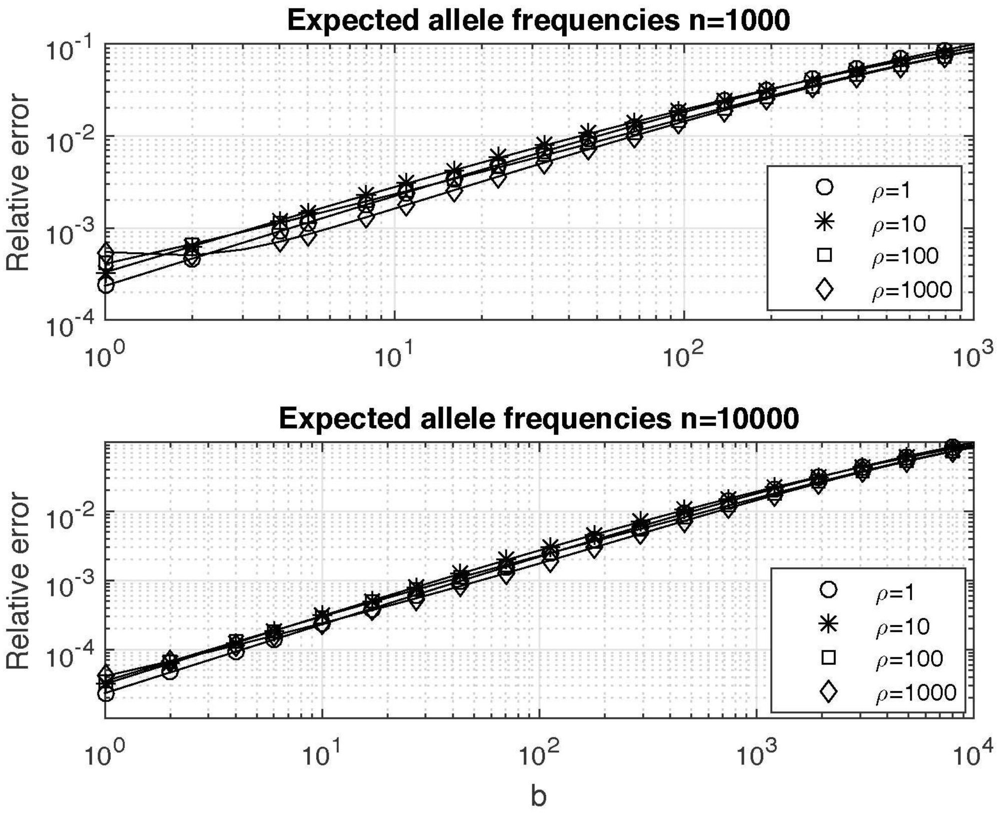

Fig 5. Relative errors of expected allele frequencies q_{nb} versus allele type b for two values of genealogy size n = 1000 (upper plot) and n = 10000 (lower plot) for different values of the product parameter of the population growth ρ = 1, ρ = 10, ρ = 100 and ρ = 1000.

doi:10.1371/journal.pone.0170701.g005

Comparing values of biases in our Table 1 to their counterparts in Chen and Chen's (2013) [16] Table 1 one can see that Chen and Chen's (2013) [16] were able to estimate magnitudes of biases, however it was not possible for them to compute exact values.

Accuracy of approximate formulae for expected allele frequencies

We evaluate accuracy of Chen and Chen's (2013) [16] method for approximate computation of expected allele frequencies based on the idea of replacing expected times in coalescence process with underlying exponentially growing population, by their approximations ("Methods" section, Eq (26)). In Chen and Chen's (2013) [16] Figure 6 one can observe that approximate method proposed by Chen and Chen (2013) [16] leads to estimates, which properly reflect patterns of change of expected allele frequencies. However, this obsevation can be done only qualitatively.

In Fig 5 we show values of relative errors of expected allele frequencies q_{nb} versus allele type b for two values of genealogy size n = 1000 (upper plot) and n = 10000 (lower plot) for different values of the product parameter of the population growth ρ = 1, ρ = 10, ρ = 100 and ρ = 1000. Relative error shows nonlinear behavior with respect to changes in ρ. For the range of values of ρ depicted in Fig 5 values of the relative error are small (of the order of 10^{-4} – 10^{-3}) for low values of b and grow to about 10% for high values of b. Since accurate computation of expected

Table 2. Statistics of segregating sites in mtDNA data from Human mtDNA database [22]. Elements in *b* are possible numbers of copies of the rare allele, and elements in c_k are numbers of segregating sites in the sample that have the number of copies of the rare allele equal *b*.

b	c_k	*b*	c_k	*b*	c_k	*b*	c_k	*b*	c_k
1	1231	28	10	55	5	87	1	156	1
2	542	29	5	56	2	88	1	174	1
3	298	30	5	57	1	89	1	176	1
4	170	31	8	58	4	90	1	204	1
5	149	32	10	59	1	91	1	213	1
6	95	33	7	60	3	94	1	218	1
7	66	34	5	61	4	95	2	234	1
8	67	35	8	62	3	96	3	235	1
9	35	36	9	63	2	98	2	244	1
10	33	37	8	64	1	104	1	264	1
11	28	38	2	65	3	110	1	272	1
12	21	39	8	66	1	111	1	299	1
13	17	40	1	67	1	127	1	347	2
14	24	41	3	68	1	128	1	390	1
15	16	42	5	69	2	129	2	444	1
16	22	43	5	70	1	131	2	505	1
17	15	44	7	72	1	132	2	550	1
18	19	45	5	74	1	133	1	604	1
19	13	46	1	76	1	134	1	610	1
20	17	47	6	77	2	135	1	720	1
21	10	48	4	78	1	138	2	724	1
22	13	49	1	79	1	139	1	777	1
23	14	50	1	81	1	144	1	867	1
24	8	51	3	83	3	147	3	933	1
25	5	52	2	84	4	149	3	943	1
26	15	53	2	85	3	150	1	944	1
27	13	54	1	86	6	152	1		

doi:10.1371/journal.pone.0170701.t002

frequencies for alleles corresponding to low values of *b* is more important than for those corresponding to high values of *b*, Chen and Chen's (2013) [16] approximation seems useful for many applications.

Analysis of mitochondrial DNA dataset

The last result, which we show is the analysis of human mitochondrial DNA polymorphisms from the Human Mitochondrial Genome mtDB database [25]. The mtDB database contains in total 3857 polymorphic sites quantified in 2704 individuals. For our analysis we have chosen only a subset of 3857 polymorphic sites in mtDB, namely those sites whose status was determined for all 2704 individuals and which were diallelic. There were 3213 such segregating sites. Table 2 shows their allelic frequencies.

We have fitted the model of exponential growth for the data given in Table 2. Analogously to [8] we treated each segregating site as a separate SNP. The model fit is based on maximizing the likelihood function defined in [8] (Eq (24)). In Fig 6 we present plots of log likelihood functions versus exponential growth product parameter ρ, obtained with the use of exact

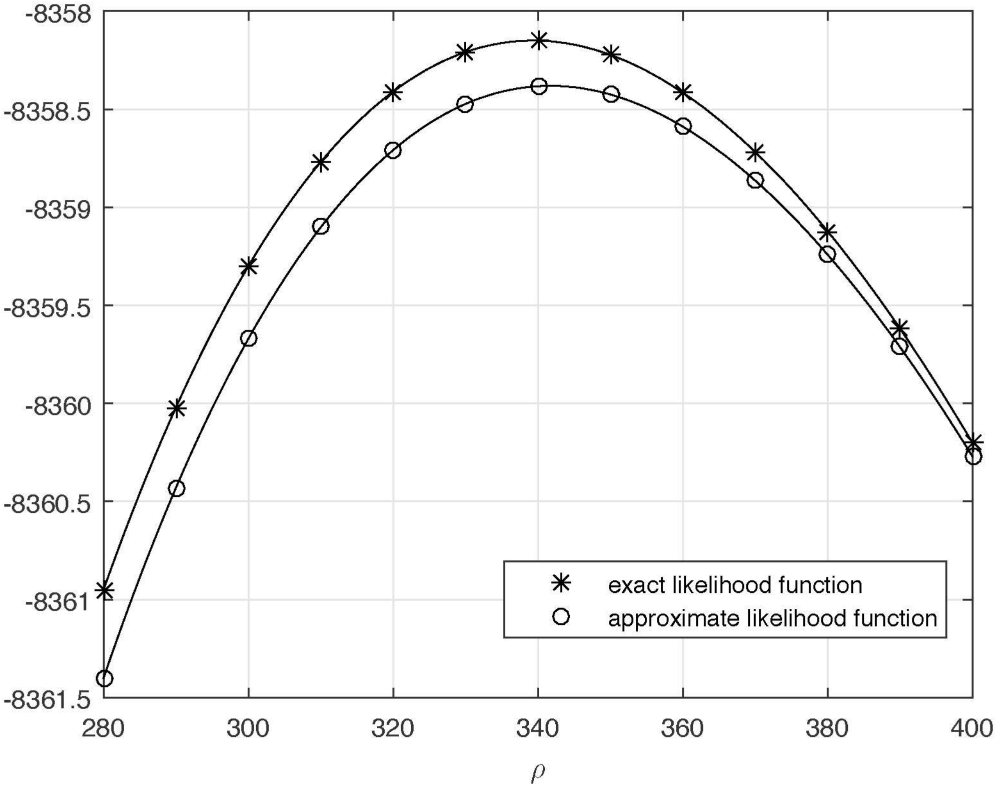

Fig 6. Log-likelihood curves for the exponential model of population growth for data on segregating sites from the mtDB database [25]. Each segregating site from Table 2 was treated as a separate SNP. The curve marked with asterisks shows the exact log likelihood function and the one marked with open circles is the approximate log likelihood function. The maximum of the exact log likelihood function is attained at $\hat{\rho}_{exact} = 339.3$ and the maximum of the approximate log likelihood function is attained at $\hat{\rho}_{approx} = 341.7$.

doi:10.1371/journal.pone.0170701.g006

method (marked with asterisks) and with the use of approximate expectations of coalescence times proposed by Chen and Chen (2013) [16] (marked with open circles). One can see that plots are quite close one to another, consistently to results presented in subsection "Accuracy of approximate formulae for expected allele frequencies". Maximum likelihood estimate obtained by using the exact method is $\hat{\rho}_{exact} = 339.3$ and the estimate obtained by using the approximate method is $\hat{\rho}_{approx} = 341.7$. Additionally, we used Hudson's program "ms" [26] to perform 1000 coalescence simulations, with appropriate parameters, which allowed us to estimate 95% confidence interval as $285 < \rho < 403$. Similar estimates can be obtained on the basis of the likelihood ratio statistics [8]. The obtained values and bounds of confidence intervals, fit into the range of values (50–500) of exponential growth product parameter of human population, which we estimated in [8] on the basis of different datasets.

Discussion

In this paper we evaluate the accuracy of the approximations for times in the coalescence tree and expected allele frequencies as proposed by [16] and we compute the probability distributions of times in the coalescence tree and their limits. We also use Human Mitochondrial Genome mtDB database to present a comparison of exact versus approximate log likelihood function for solving the inverse problem of estimating population size history from observed allele frequencies [18, 20, 24].

The presented resuts are based on new approaches described in the "Methods" section. We propose new methods for coalescence computations for large sample sizes, based on inverting the integral transform defined in [7] ("Methods" section, Eq (12)) and on using analytical expressions for infinite products for computing limit distributions. Both the integral transform Eq (12) and its inverse Eq (29) use techniques well known in statistical genetics—the Laplace (Fourier) transformation and coalescence time scale change. However, combining them together allows for deriving new results, unavailable in the previous literature.

Methodologies for efficient computation of probability distributions and expectations of coalescence times for large sample sizes, presented in the "Methods" section of this paper, can lead to many further applications. The inverse transform Eq (29) can be generalized to higher dimensions. Two-dimensional generalization of Eq (29) can be used for computing second-order moments of coalescence times [27] for large sample sizes. Methodology for computing second-order moments of coalescence times can be useful e.g., for analyzing statistics of triallelic DNA loci [28].

In this paper, by large sample sizes n we understand numbers comparable to the present throughput capabilities of experimental techniques for DNA sequencing, i.e. thousands of people, with data publically available in databases in 1000 Genomes Project [29] or mtDB [25]. The sample size is going to increase to hundreds of thousands or even millions in short order, with ongoing projects including UK10K (about 10 thousand human genomes) and the Million Veteran Program (about 1 million human genomes). The computational methods proposed in this paper are likely to be relevant to many aspects of statistical analyses of these datasets.

Sequencing data for even larger number of cell samples are already available in the cancer genomics TCGA database [30]. Cancer tissue is an evolving population of cancer cells with diversity increasing as the tumour advances in development, and using coalescence modeling for the analysis of cancer genomics data [17, 31] is of great interest and possibly of significant practical importance. Cell count in cancer tissues exceeds bilions, and biopsies include upwards of millions of cells [32]. However, developing algorithms for coalescence analyses of cancer genomics sequencing data requires addressing not only the problem of large sample size but also numerous additional issues specific to that type of data. Typical sequencing cancer genomics data include reads obtained from a mixture rather than from separate cancer cells, which calls for the development of integrative approaches combining large sample coalescence modeling with ascertainment models (e.g., [33]). Another concern would be the presence of mutational events such as chromosomal duplications, the loss of heterozygosity and rearangements [31], which interfere with the point mutation processes. Additionally, point mutations seen in the cancer sequencing samples are classified as either driver or passenger, which is related with their roles in the selection mechanisms in the carcinogenesis process. Driver mutations are defining new clones creating cancer cell population subdivisions, leading to the need for further model refinement. Some of the problems listed above are possible topics of present studies on coalescence modeling methods for applications in cancer genomics.

Methods

In the case of the evolutionary scenario with the constant population size times between coalescence events, $S_n^C, S_{n-1}^C, \ldots, S_2^C$, are mutually independent random variables, each distributed exponentially, with expectations (e.g., [5])

$$E(S_k^C) = \frac{N_0}{\binom{k}{2}}, \; k = 2, 3, \ldots, n. \tag{9}$$

For the case of constant population size one can obtain analytical expressions for expected allele frequencies f_{nb}^{C} and probabilities p_{nb}^{C} [5, 34].

In the general case of the population size history given by a function $N(t)$, times between coalescences, $S_2, \ldots, S_{n-1}, S_n$, are not independent. Joint probability density function of the distribution of times $T_2, \ldots T_{n-1}, T_n$ can be computed by using the following expression [35].

$$p(t_2, \ldots, t_{n-1}, t_n) = \prod_{j=2}^{n} \frac{\binom{j}{2}}{N(t_j)} \exp\left(-\int_{t_{j+1}}^{t_j} \frac{\binom{j}{2} d\sigma}{N(\sigma)}\right) \tag{10}$$

where $0 = t_{n+1} < t_n < t_{n-1} \ldots < t_2$, and $\binom{j}{2}$ is the binomial symbol. Marginal distributions of times $T_2, \ldots, T_{n-1}, T_n$, denoted by $\pi_{T2}(t), \ldots, \pi_{Tn-1}(t), \pi_{Tn}(t)$ follow from multiple integrations of the above formula Eq (10).

A method for computing marginal distributions, $\pi_{T2}(t), \ldots, \pi_{Tn-1}(t), \pi_{Tn}(t)$, based on combining the time scale change

$$\tau = g(t) = \int_0^t \frac{d\sigma}{N(\sigma)} \tag{11}$$

with the technique of integral transformations was proposed in [7]. The proposed integral transform $\Upsilon\{.\}$ with the underlying function $N(t)$ was defined as follows

$$P(s) = \Upsilon\{\pi(t)\} = E\left[\exp\left(-s\int_0^t \frac{d\sigma}{N(\sigma)}\right)\right] = \int_0^\infty \pi(t) \exp\left(-s\int_0^t \frac{d\sigma}{N(\sigma)}\right) dt. \tag{12}$$

Application of Υ transformation led to analytical expressions for Υ transforms of marginal distributions

$$\Upsilon\{\pi_{Tk}(t)\} = \prod_{j=k}^{n} \frac{\binom{j}{2}}{s + \binom{j}{2}}, \tag{13}$$

and to expressions for marginal distributions

$$\pi_{Tk}(t) = \sum_{j=k}^{n} A_{jk}^{n} q_j(t) \; k = 2, 3, \ldots, n, \tag{14}$$

where coefficients A_{jk}^{n} followed from partial fraction expansion of the product in Eq (13) (see Eq (7) in [7]), and

$$q_j(t) = \frac{\binom{j}{2}}{N(t)} \exp\left(-\int_0^t \frac{\binom{j}{2} d\sigma}{N(\sigma)}\right). \tag{15}$$

The probability distribution, $q_j(t)$, in the above formula Eq (15) is the distribution of the time to the first coalescence event in the sample of size j. The formula Eq (14) was also independently derived by Zivkovic and Wiehe (2008) [27] by repeated integration and using mathematical induction.

By using Eqs (14) and (15) one can compute expectations of times to coalescences $E(T_n)$, $E(T_{n-1}), \ldots, E(T_2)$

$$E(T_k) = \sum_{j=k}^{n} A_{jk}^{n} e_j \; k = 2, 3, \ldots, n, \tag{16}$$

where

$$e_j = \int_0^\infty t q_j(t) dt, \tag{17}$$

are expected times to the first coalescence event in a sample of size j. Expressions for expectations of times T_n, T_{n-1}, . . ., T_2 Eq (16) can be used for computing expectations $E(S_2)$, . . ., $E(S_{n-1})$, $E(S_n)$ and $ETLBT$. Consequently, formula (Eq (3)) can be applied for computing allele frequencies.

For the case where $N(t)$ follows the exponential scenario, $N(t) = N^E(t)$ Eq (6) time scale change in Eq (11) becomes

$$\tau = g_E(t) = \frac{1}{rN_0}(\exp(rt) - 1) \tag{18}$$

and $q_j(t)$ in Eq (15) becomes

$$q_j^E(t) = \frac{\binom{j}{2}}{N_0} \exp\left[rt + \frac{\binom{j}{2}}{rN_0}(1 - \exp(rt))\right]. \tag{19}$$

Analogously to Eqs (16) and (17), by using Eqs (14) and (19) one can compute expectations for the exponential scenario

$$E(T_k^E) = \sum_{j=k}^{n} A_{jk}^n e_j^E \tag{20}$$

where e_j^E are expectations of times with probability distributions given in Eq (19), equal to [8, 10]

$$e_j^E = e_j^E(N_0, r) = -\frac{\exp[\binom{j}{2}(rN_0)^{-1}]}{r} Ei[-\binom{j}{2}(rN_0)^{-1}]. \tag{21}$$

In the above Ei denotes the exponential integral, $\mathrm{Ei}(-\mu) = -\int_1^\infty [\exp(-\mu x)/x]dx$, $\mathrm{Re}(\mu) > 0$, ([36], 3.351.5). Eq (20) can be used for computing $E(S_2^E)$, ..., $E(S_{n-1}^E)$, $E(S_n^E)$, $ETMRCA^E$, $ETLBT^E$ and, substituted in Eq (3), for computing allele frequencies.

As we have already mentioned in the introduction section, many authors [7–9, 16, 17] have reported that expressions for probability distributions and expectations of times, and for allele frequencies of mutations for the general case of population size history $N(t)$ are applicable only for small sample sizes $n < 50$, due to the fact that coefficients A_{jk}^n very quickly diverge to very large numbers with alternating signs when n increases.

The approach based on application of combinatorial identities and methods of summing hypergeometric series given in [7, 8], which allows for obtaining numerically stable expressions for $ETMRCA$, $ETLBT$ and expected allele frequencies f_{nb}, applicable for large values of n. These expressions have the following forms

$$ETMRCA = \sum_{j=2}^{n}(2j-1)\frac{n!(n-1)!}{(n+j-1)!(n-j)!}(-1)^j e_j, \tag{22}$$

$$ETLBT = \sum_{j=2}^{n}(2j-1)\frac{n!(n-1)!}{(n+j-1)!(n-j)!}[1+(-1)^j]e_j \tag{23}$$

and

$$f_{nb} = \mu \sum_{j=2}^{n} W_{bj}^{n} e_j, \ \ b = 1, 2, ..., n-1. \tag{24}$$

Coefficients W_{bj}^{n} in Eq (24) are given by the recursion below

$$W_{b2}^{n} = \frac{6}{(n+1)}, \ \ W_{b3}^{n} = 30 \frac{(n-2b)}{(n+1)(n+2)},$$

$$W_{b,j+2}^{n} = -\frac{(1+j)(3+2j)(n-j)}{j(2j-1)(n+j+1)} W_{bj}^{n} + \frac{(3+2j)(n-2b)}{j(n+j+1)} W_{b,j+1}^{n}, \tag{25}$$

$j = 2, 3, \ldots, n-2$. In Eqs (22)–(24) e_j denote expected times to the first coalescence event in a sample of size j given in Eq (17). By using expressions Eqs (9), (15)–(17) and (21) one can compute expectations e_j and further *ETMRCA*, *ETLBT* and f_{nb} for any scenario of population size change, constant (e_j^C), given by generally defined function $N(t)$ (e_j) and exponential (e_j^E).

Expressions Eqs (22)–(25) were used by several authors for computing exact values of expected times *ETMRCA*, *ETLBT* and expected allele frequencies, e.g., or for studying properties of coalescence process [37], for studies on pupulation size histories [38] and for comparisons between exact and approximate methods [16].

By applying time scale change $g^{-1}(\tau)$, given in Eq (36) Chen and Chen, (2013) [16] have obtained the following approximation of expected coalescence times for the exponential growth scenario

$$E(T_k^E) \simeq \frac{1}{r} \ln \left[2rN_0 \left(\frac{1}{k-1} - \frac{1}{n} \right) + 1 \right], \tag{26}$$

and

$$ETMRCA^E = E(T_2^E) \simeq \frac{1}{r} \ln \left[2rN_0 \left(1 - \frac{1}{n} \right) + 1 \right]. \tag{27}$$

By integrating over time expectation of the pure death process describing merging of branches of the coalescence tree Chen and Chen, (2013) [16] have also derived the following approximation for expected total length of branches in the coalescence tree under exponential scenario $ETBLT^E$

$$ETBLT^E \simeq \frac{2nN_0 \ln \frac{2rN_0}{n}}{2rN_0 - n}. \tag{28}$$

Finally Chen and Chen, (2013) [16] have proposed to substitute approximate expectations of coalescence times Eq (26) in expression Eq (3) to obtain approximate expected allele frequencies in the coalescence process with the underlying exponential population growth.

Inversion of the integral transform

The limitation of applicability of computations based on combinatorial identities and summing hypergeometric series is that they can only be used for expectations *ETMRCA*, *ETLBT* and for expected allele frequencies f_{nb}. Computing probability distributions of times T_k is not possible.

In this subsection we present a new approach, which allows for computing distributions and expectations of times to coalescence events, $T_2, \ldots, T_{n-1}, T_n$, with (theoretically) arbitrary accuracy, applicable for large genealogies. The approach is based on construction of the transformation inverse to Eq (12). The inverse of Eq (12), denoted by $\Upsilon^{-1}\{.\}$, has the following form

$$\pi(t) = \Upsilon^{-1}\{P(s)\} = \frac{1}{N(t)}\frac{1}{2\pi i}\int_{c-i\infty}^{c+i\infty} P(s)\exp\left(s\int_0^t \frac{d\sigma}{N(\sigma)}\right)ds. \tag{29}$$

In Eq (29) c is a suitably chosen constant and $i = \sqrt{-1}$. The above formula is constructed by analogy to the inverse of the Laplace transform, Mellin—Fourier integral [39]. Since $\pi(t)$ is a density function of the probability distribution we can set $c = 0$, $s = i\omega$ and replace Eq (29) by

$$\pi(t) = \Upsilon^{-1}\{P(i\omega)\} = \frac{1}{N(t)}\frac{1}{2\pi}\int_{-\infty}^{\infty} P(i\omega)\exp\left(i\omega\int_0^t \frac{d\sigma}{N(\sigma)}\right)d\omega. \tag{30}$$

Verification that $\Upsilon^{-1}[\Upsilon(\pi(t))] = \Upsilon^{-1}[P(s)] = \pi(t)$ is straightforward, since either Eq (29) or Eq (30) can be understood as a two-step procedure. The first step is the inverse Laplace

$$\pi^{C_0}(\tau) = \frac{1}{2\pi i}\int_{c-i\infty}^{c+i\infty} P(s)\exp(s\tau)ds \tag{31}$$

or inverse Fourier transform

$$\pi^{C_0}(\tau) = \frac{1}{2\pi}\int_{-\infty}^{\infty} P(i\omega)\exp(i\omega\tau)d\omega \tag{32}$$

of $P(s) = \Upsilon(\pi(t))$ or of $P(i\omega) = P(s)|_{s = i\omega}$, with τ given by Eq (11). The second step is the time scale change $g^{-1}(\tau)$ inverse to Eq (11). Since $[g^{-1}(\tau)]^{-1} = \tau = g(t)$, then the second step is

$$\pi(t) = \frac{d}{dt}(g(t))\pi^{C_0}(g(t)) = \frac{1}{N(t)}\pi^{C_0}(g(t)). \tag{33}$$

It is obvious that in the first step we obtain probability distribution $\pi^{C_0}(\tau)$ which is the original of the Laplace transform $P(s)$ or Fourier transform $P(i\omega)$ and corresponds to the constant population size scenario with $N_0 = 1$, while in the second step, by the time scale change $t = g^{-1}(\tau)$ we obtain probability distribution $\pi(t)$ under the scenario of the population size change given by $N(t)$.

Using Eq (30) we can write expression for probability distribution of time T_k in the following integral form

$$\pi_{Tk}(t) = \frac{1}{N(t)}\frac{1}{2\pi}\int_{-\infty}^{\infty}\prod_{j=k}^{n}\frac{\binom{j}{2}}{i\omega + \binom{j}{2}}\exp\left(i\omega\int_0^t\frac{d\sigma}{N(\sigma)}\right)d\omega, \tag{34}$$

valid for the general case of the population size history $N(t)$.

For the case of the exponential scenario of time change of the population size, the two steps mentioned above would assume the following forms. In the first step we compute probability distribution $\pi_{Tk}^{C_0}(\tau)$, of time to coalescence T_k under constant population size scenario with $N_0 = 1$

$$\pi_{Tk}^{C_0}(\tau) = \frac{1}{2\pi}\int_{-\infty}^{\infty}\prod_{j=k}^{n}\frac{\binom{j}{2}}{i\omega + \binom{j}{2}}\exp(i\omega\tau)d\omega. \tag{35}$$

In the second step we transform $\pi_{Tk}^{C_0}(\tau)$ using Eq (11), which leads to

$$t = g^{-1}(\tau) = \frac{1}{r}\ln(1 + N_0 r\tau), \tag{36}$$

and

$$\pi^{E}_{Tk}(t) = \frac{\exp(rt)}{N_0}\pi^{C_0}_{Tk}(\tau(t)) = \frac{1 + N_0 r\tau(t)}{N_0}\pi^{C_0}_{Tk}(\tau(t)). \tag{37}$$

Distributions $\pi^{C_0}_{Tk}(\tau)$, Eq (35) and $\pi^{E}_{Tk}(t)$, Eq (37) are computed numerically. Numerical computations can be done in two ways, by numerical integration procedures, according to Eq (30), separately for each time point, or by using the inverse fast Fourier transform algorithm. In our computational examples, presented further in this paper, we have used both these approaches. For numerical integration we have used adaptive Gauss-Kronrod quadrature procedure [40] implemented as the Matlab function "quadgk". Advantage of using numerical integration is that the time points can be freely located according to needs, which leads to lower errors. The disadvantage of the method of numerical integration is that it is slower compared to the fast Fourier transform algorithm.

The advantage of the fast Fourier transform algorithm is that it its much faster. However, estimating and controlling accuracy is more difficult. Despite problems with controlling accuracy, in the majority of computational examples we have computed Fourier integrals in Eq (35) by using Matlab inverse fast Fourier transform function "ifft", taking advantage of its speed. In more detail, at first we have defined the time axis range and grid for $\pi^{C_0}_{Tk}(\tau)$ by using information on the first and second moments of $T^{C_0}_{k}$ [16] and assuming some additional margin related to skewness of the distributions. Time axis range and grid allows for defining the corresponding frequency axis ranges and grid and for computing $\pi^{C_0}_{Tk}(\tau)$ by the inverse fast Fourier transform procedure. We have estimated accuracy of computations by comparing known values of moments of times to values computed on the basis of numerically obtained distributions. In this way we have estimated that a grid with 500 equidistant time points was sufficient for obtaining relative error $\leq 10^{-4}$ for the case of computations for constant population size for $n \leq 10^4$. Nonlinear transformations of the time scale, necessary for computations for population exponential growth scenarios, result in nonuniformity of the time axis grid resolution, which leads to increase of the error. For the case of exponential scenarios of population growth we have experimentally verified that a grid with 1000 equidistant time points for τ was sufficient for obtaining relative error $\leq 10^{-3}$ for moments of distribution with the transformed time, with $n \leq 10^4$ and $\rho \leq 10^4$.

As supporting files (S1 File) to this paper we provide Matlab functions and scripts for computing probability distributions of times in the coalescence tree for exponential scenario of population growth, based on the direct method of numerical integration. We have tested these functions for the range of values of the product parameter $0 \leq \rho \leq 10^6$ and genealogy sizes $n < = 10^4$. In the provided programs all parameters are set automatically and the relative errors are 10^{-5} or better.

Limit distributions

According to our best knowledge no results concerning limit distributions of times close to the root of the coalescence tree, in particular *TMRCA*, were published in the literature. In this subsection we compute limit distributions for *TMRCA* for both constant and time varying population size scenarios. Denote the limit distribution of *TMRCA* under the population size scenario $N(t)$ by $\pi_{TMRCA,\infty}(t)$. On the basis of results from the previous subsection we have

$$\pi_{TMRCA,\infty}(t) = \frac{1}{N(t)}\frac{1}{2\pi}\int_{-\infty}^{\infty}\prod_{j=2}^{\infty}\frac{\binom{j}{2}}{i\omega + \binom{j}{2}}\exp\left(i\omega\int_0^t\frac{d\sigma}{N(\sigma)}\right)d\omega. \tag{38}$$

Infinite product, which appears on the right hand side of the above formula can be analytically computed by using the following well known identity involving quotients of gamma functions (e.g., [21])

$$\prod_{k=0}^{\infty}\frac{(k+a_1)(k+a_2)}{(k+b_1)(k+b_2)}=\frac{\Gamma(b_1)\Gamma(b_2)}{\Gamma(a_1)\Gamma(a_2)}, \tag{39}$$

where $\Gamma(.)$ denotes Euler's gamma function and a_1, a_2, b_1, b_2 are any complex numbers satisfying $a_1 + a_2 = b_1 + b_2$. Using Eq (39) with $a_1 = 1$, $a_2 = 2$, $b_{1,2} = 1.5 \pm \sqrt{\frac{1}{4} - 2i\omega}$, $i = \sqrt{-1}$, allows for deriving the following expression for the infinite product in Eq (38)

$$\prod_{j=2}^{\infty}\frac{j(j-1)}{2i\omega + j(j-1)}=\frac{2\pi i\omega}{\cos\left[\pi\sqrt{\frac{1}{4}-2i\omega}\right]}. \tag{40}$$

The above identity is listed in A. Dieckmann's internet collection of infinite products [41]. Subsituting the above identity in Eq (38) one can compute the limit distribution of *TMRCA*, $\pi_{TMRCA,\infty}(t)$ as follows

$$\pi_{TMRCA,\infty}(t)=\frac{1}{N(t)}\frac{1}{2\pi}\int_{-\infty}^{\infty}\frac{2\pi i\omega}{\cos\left[\pi\sqrt{\frac{1}{4}-2i\omega}\right]}\exp\left(i\omega\int_0^t\frac{d\sigma}{N(\sigma)}\right)d\omega. \tag{41}$$

We have numerically computed limit distribution $\pi_{TMRCA,\infty}(t)$ given by the above formula Eq (41) using the direct method of numerical integration. When trying to apply fast Fourier transform algorithm we have encountered problems with the proper control of the accuracy of computations.

Round-off errors in computing allelic frequencies

We have conducted a computational study on effects of round-off errors on accuracy of computation of expected allele frequencies by using expression Eq (24). Results in this subsection can be useful for such researches as those reported in [37], [38] and [16].

We denote the computed and the true expected allele frequencies by f_{nb}^{comp} and f_{nb}^{true} respectively, and we define the maximum relative error commited when computing allele frequencies, *MxRelErr*, as follows

$$MxRelErr = \max_{1\le b\le n-1}\left|\frac{f_{nb}^{comp}-f_{nb}^{true}}{f_{nb}^{true}}\right|. \tag{42}$$

By "computed allele frequencies" we mean values obtained by using expressions Eqs (24) and (25). One can estimate upper bound for *MxRelErr* by using error analysis technique. We define $f_{nb}^{comp}(\sigma)$ as representing values computed by using Eqs (24) and (25) in the case where expected times e_j in Eq (24) are additionally corrupted by Gaussian, relative error with standard deviation σ. By assuming the value of σ of one or two orders of magnitude higher than true relative round-off errors in computing e_j we can obtain the following, conservative, upper bound on *MxRelErr*

$$MxRelErr < \max_{1\le b\le n-1}\left|\frac{f_{nb}^{comp}(\sigma)-f_{nb}^{comp}}{f_{nb}^{comp}}\right|. \tag{43}$$

In Fig 7 we show upper bounds of *MxRelErr* for the scenario of exponential growth of population with different values of product parameter ρ, obtained by assuming $\sigma = 10^{-13}$. The assumed value of σ is approximately of one-two orders of magnitude higher than accuracy of

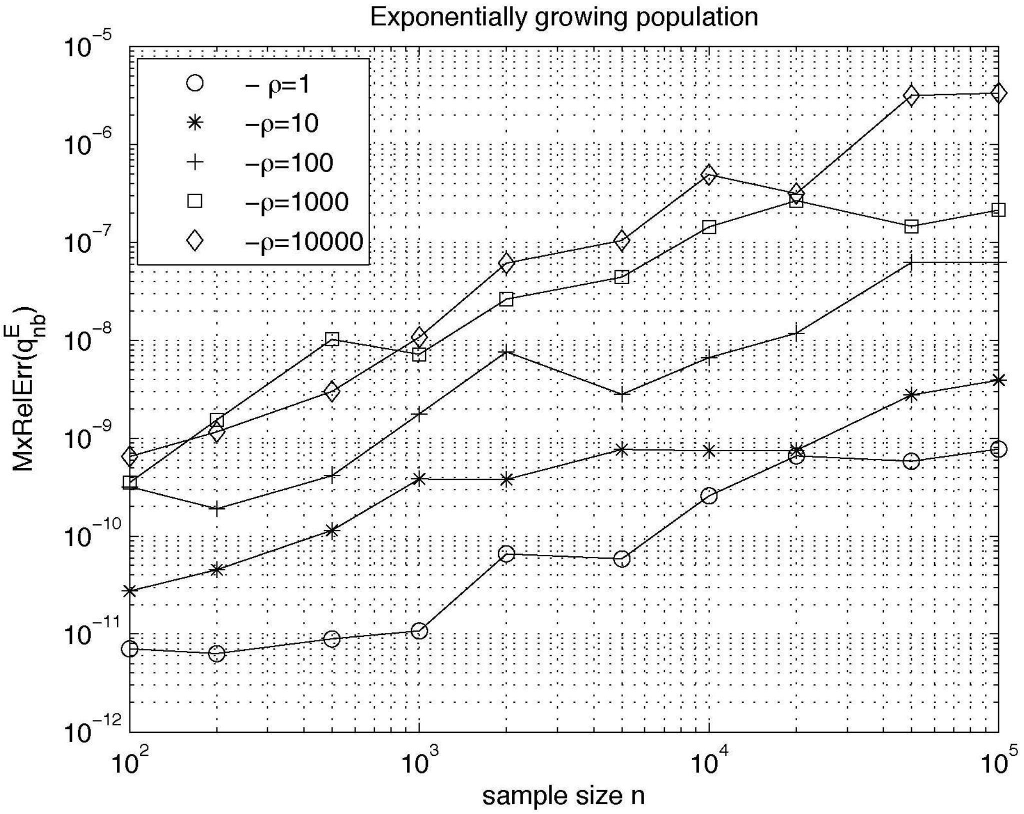

Fig 7. Influence of round-off errors on accuracy of computation of expected allele frequencies by using expressions Eqs (22)–(25). The plot shows upper bounds of maximum relative error for the scenario of exponential growth of population with different values of product parameter ρ, obtained by corrupting values of expected times e_j by Gaussian, relative error with standard deviation $\sigma = 10^{-13}$.

doi:10.1371/journal.pone.0170701.g007

computing e_j^E in Eq (21), which we estimate to be in the range $10^{-14} - 10^{-15}$. Values of e_j^E were computed by using Matlab function "expint" with the modification described in [8], which allows for obtaining exact function values for wide ranges of argument values.

Contemplating values of upper bounds for *MxRelErr* computed by using Eq (43), shown in Fig 7, we come to the conclusion that formulae Eqs (24) and (25) for computing allele frequencies f_{nb} can be safely used for sample sizes n up to the range of hundreds of thousands and expect relative errors not higher than 10^{-6}.

Formulas for distributions of times in the coalescence tree, Eqs (31) and (32) can be used for computing expectations of coalescence times by numerical integration, which can be then substituted in Eq (3). This provides alternative method for computing allelic frequencies. Due to errors in numerical integration, higher by approximately two orders of magnitude than errors in computing values of special functions $Ei(x)$, maximal relative round-off errors of allelic frequencies obtained by using Eq (3) and numerically computed expectations of coalescence times are in the range $10^{-4} - 10^{-3}$. They are significantly higher than those in Fig 7 but still acceptable in many applications.

Supporting information

S1 File. Software for computing probability distributions of times in the coalescence process. Archive with Matlab functions and scripts for computing probability distributions of times in the coalescence tree for exponential scenario of population growth, based on the direct method of numerical integration. Our Matlab functions and scripts are also available as a GitHub repository (https://github.com/agnieszkaszczesna/Coalescence-Computations-for-Large-Samples).
(ZIP)

Author contributions

Conceptualization: AP MK.

Data curation: AP AS.

Formal analysis: AP.

Funding acquisition: AP MK AS.

Investigation: AP AS MG.

Methodology: AP AS.

Project administration: AS MG.

Resources: AS.

Software: AP AS.

Supervision: AP.

Validation: AP AS.

Visualization: AP AS.

Writing – original draft: AP.

Writing – review & editing: AP AS.

References

1. Kingman JFC, (1982a), The Coalescent, Stoch. Proc. Appl., vol. 13, pp. 235–248, 1982. doi: 10.1016/0304-4149(82)90011-4
2. Kimura M, (1983), The neutral theory of molecular evolution, Cambridge (UK): Cambridge University Press.
3. Griffiths RC, (1999), The Time to the Ancestor along Sequences with Recombination, Theor. Pop. Biol., 55:137–144. doi: 10.1006/tpbi.1998.1390 PMID: 10329513
4. Griffiths RC, Marjoram P, (1996), Ancestral Inference from Samples of DNA Sequences with Recombination, J. Comput. Biol., 3: 479–502. doi: 10.1089/cmb.1996.3.479 PMID: 9018600
5. Griffiths RC, Tavare S, (1998), The Age of a Mutation in at General Coalescent Tree, Stochastic Models, 14: 273–295. doi: 10.1080/15326349808807471
6. Stephens M, (2000), Times on Trees and the Age of an Allele, Theor. Pop. Biol., 57: 109–119. doi: 10.1006/tpbi.1999.1442 PMID: 10792976
7. Polanski A, Bobrowski A, Kimmel M, (2003), A note on distributions of times to coalescence, under time dependent population size, Theoretical Population Biology, vol. 63, pp. 33–40, 2003. doi: 10.1016/S0040-5809(02)00010-2 PMID: 12464493

8. Polanski A, Kimmel M, (2003), New Explicit Expressions for Relative Frequencies of SNPs with Application to Statistical Inference on Population Growth, Genetics, vol. 165, pp. 427–436, 2003. PMID: 14504247

9. Wooding S, Rogers A, (2002), The matrix coalescent and an application to human single—nucleotide polymorphisms, Genetics, 161:1641–1650. PMID: 12196407

10. Slatkin M, Hudson RR, (1991), Pairwise comparisons of mitochondrial DNA in stable and exponentialy growing populations, Genetics, 129: 555–562. PMID: 1743491

11. Wakeley J, (2001), The coalescent in an island model of population subdivision with variation among demes, Theor. Popul. Biol., 59: 133–144. doi: 10.1006/tpbi.2000.1495 PMID: 11302758

12. Chen H, (2012), The joint allele frequency spectrum of multiple populations: A coalescent theory approach, Theoretical Population Biology 81, pp. 179–195. doi: 10.1016/j.tpb.2011.11.004 PMID: 22155588

13. Krone SM, Neuhauser C. (1997), Ancestral processes with selection. Theoretical Population Biology 51, 210–237. doi: 10.1006/tpbi.1997.1299 PMID: 9245777

14. Neuhauser C, Krone SM. (1997) The genealogy of samples in models with selection, Genetics 145 519–534. PMID: 9071604

15. Campbell R, (2007) Coalescent size vs. coalescent time with strong selection, Bull. Math. Biol., 69: 2249–2259. doi: 10.1007/s11538-007-9218-9 PMID: 17546476

16. Chen H, Chen K, (2013), Asymptotic Distributions of Coalescence Times and Ancestral Lineage Numbers for Populations with Temporally Varying Size, Genetics, vol. 194, pp. 721–736. doi: 10.1534/genetics.113.151522 PMID: 23666939

17. Durrett R, (2013), Population genetics of neutral mutations in exponentially growing cancer cell populations, The Annals of Applied Probability, Vol. 23, pp. 230–250. doi: 10.1214/11-AAP824 PMID: 23471293

18. Maruvka YE, Shnerb NM, Bam-Yam Y, Wakeley J, (2011), Recovering Population Parameters from a Single Gene Genealogy: An Unbiased Estimator of the Growth Rate Mol. Biol. Evol. 28(5):1617–1631. doi: 10.1093/molbev/msq331 PMID: 21172828

19. Griffiths RC, (1984), Asymptotic line-of-descent distributions. J. Math. Biol., 21: 67–75. doi: 10.1007/BF00275223

20. Chen H, Hey J, Chen K, (2015), Inferring Very Recent Population Growth Rate from Population-Scale Sequencing Data: Using a Large-Sample Coalescent Estimator, Mol Biol Evol., 32, pp: 2996–3011. doi: 10.1093/molbev/msv158 PMID: 26187437

21. Chamberland M, Straub A, (2013), On gamma quotients and infinite products. Adv. in Appl. Math. 51 (5), 546–562. doi: 10.1016/j.aam.2013.07.003

22. Ingman M, Gyllensten U, (2006), mtDB: Human Mitochondrial Genome Database, a resource for population genetics and medical sciences. Nucleic Acids Res 34, D749–D751. doi: 10.1093/nar/gkj010 PMID: 16381973

23. Kingman JFC, (1982b) On the genealogy of large populations, Journal of Applied Probability, vol. 19, pp. 27–43. doi: 10.1017/S0021900200034446

24. Liu X, Fu YX, (2015), Exploring Population Size Changes Using SNP Frequency Spectra, Nat Genet, 47(5): 555–559. doi: 10.1038/ng.3254 PMID: 25848749

25. Ingman M, Gyllensten U, (2006), mtDB: Human Mitochondrial Genome Database, a resource for population genetics and medical sciences. Nucleic Acids Res 34, D749–D751. doi: 10.1093/nar/gkj010 PMID: 16381973

26. Hudson RR. (2002) Generating samples under a Wright-Fisher neutral model. Bioinformatics, vol. 18, pp. 337–338. doi: 10.1093/bioinformatics/18.2.337 PMID: 11847089

27. Zivkovic D, Wiehe T, (2008), Second-order moments of seg- regating sites under variable population size, Genetics, vol. 180, pp. 341–357. doi: 10.1534/genetics.108.091231 PMID: 18716326

28. Jenkins PA, Mueller JW, Song YS, (2014), General Triallelic Frequency Spectrum Under Demographic Models with Variable Population Size, Genetics, Vol. 196, pp. 295–311. doi: 10.1534/genetics.113.158584 PMID: 24214345

29. The 1000 Genomes Project Consortium, (2015), A global reference for human genetic variation, Nature 526 pp. 68–74. doi: 10.1038/nature15393 PMID: 26432245

30. The Cancer Genome Atlas Research Network, (2008), Comprehensive genomic characterization defines human glioblastoma genes and core pathways. Nature 455, 1061–1068. doi: 10.1038/nature07385 PMID: 18772890

31. Beerenwinkel N, Schwarz RF, Gerstung M, Markowetz F, (2015), Cancer evolution: mathematical models and computational inference. Syst Biol., 64(1):e1–e25. doi: 10.1093/sysbio/syu081 PMID: 25293804

32. Sidow A, Spies N, (2015), Concepts in solid tumor evolution, Trends in Genetics, vol. 31, no. 4, pp. 208–201. doi: 10.1016/j.tig.2015.02.001 PMID: 25733351

33. Clark AG, Hubisz MJ, Bustamante CD, Williamson SH, Nielsen R, (2005), Ascertainment bias in studies of human genome-wide polymorphism, Genome Res, 15 pp:1496–1502. doi: 10.1101/gr.4107905 PMID: 16251459

34. Fu YX, (1995), Statistical properties of segregating sites. Theor. Popul. Biol. 48: 172–197. doi: 10.1006/tpbi.1995.1025 PMID: 7482370

35. Griffiths RC, Tavare S, (1994), Sampling theory for neutral alleles in a varying environment. Philos. Trans. R. Soc. Lond. B 344: 403–410. doi: 10.1098/rstb.1994.0079 PMID: 7800710

36. Gradshteyn IS, Ryzhik IM, Table of integrals, series and products, fifth ed., Academic Press, 1980.

37. Eldon B, Birkner M, Blath J, Freund F, (2015), Can the Site-Frequency Spectrum Distinguish Exponential Population Growth from Multiple-Merger Coalescents?, Genetics, vol. 199, pp. 841–856. doi: 10.1534/genetics.114.173807 PMID: 25575536

38. Bhaskar A, Wang R, Song YS, (2015), Efficient inference of population size histories and locus-specific mutation rates from large-sample genomic variation data, Genome Res. 25:268–279. doi: 10.1101/gr.178756.114 PMID: 25564017

39. Davies BJ. (2002), Integral transforms and their applications (3rd ed.), Berlin, New York. doi: 10.1007/978-1-4684-9283-5

40. Shampine LF, (2008), Vectorized Adaptive Quadrature in MATLAB, Journal of Computational and Applied Mathematics, 211, pp.131–140. doi: 10.1016/j.cam.2006.11.021

41. Dieckmann A, Collection of Infinite Products and Series, http://www-elsa.physik.uni-bonn.de/~dieckman/InfProd/InfProd.html, accessed 10.02.2016.