Predicting the time distribution of the MRCA between two haploid hermaphrodites with recombination

Zhao, Zehui[1]

[1] *Department of Physics, Emory University, Atlanta, Georgia, 30324, USA*

(Dated: December 22, 2021)

Considering recombination, a prediction on the time distribution of the most recent common ancestor of two haploid hermaphrodites is presented. As approximations, coalescences between lineages that do not overlap are only considered when they represent the same individual, and the stochastic process of recombination is treated as deterministic. The prediction is then compared to simulated results using msprime.

I. INTRODUCTION

The time distribution of the most recent common ancestor (MRCA) of a sample is well studied using the Wright Fisher model, and results have been applied to practical problems [1, 2]. However, the process of recombination is mostly ignored in previous analyses, and though this may be a good approximation when the recombination rate is low, there is the probability of it significantly affecting the time distribution in reality. Considering the recombination process exactly in the theoretical prediction is difficult, and it is previously proposed that one may simplify the analysis by ignoring coalescences between lineages that do not overlap [3]. In this report, an analytic prediction of the time distribution of the MRCA of two haploid hermaphrodites is made with recombination in mind. The rest of the report will first derive the prediction from scratch based on the Wright Fisher model, and then the prediction will be compared to several simulations to test its accuracy and verify that nothing is apparently wrong.

II. THEORY

A. Coalescence without recombination

Using the Wright Fisher model, suppose there are N single chromosome haploid hermaphrodites, and in each generation, a new generation with the same population size is generated by randomly picking individuals from the previous generation and copying them to the next. Then, for two individuals in some generation, the probability of their lineages coalescing, i.e. they having the same parent by chance, is

$$P = \frac{1}{N}; \tag{1}$$

one of the two can choose any of the N individuals in the previous generation to be its parent, and the other has to pick the same one out of N. This implies that the probability of two present lineages coalescing exactly at the t-th previous generation is

$$P(t) = (1 - \frac{1}{N})^{t-1}\frac{1}{N}, \tag{2}$$

and this is by definition the time distribution of their MRCA. Note here that t is dimensionless.

For the ease of manipulating the probabilities, turn Equation 2 into a probability distribution, where t now is any positive real number. One way of doing this is to discretize each generation into M subgenerations and then take the limit, where the probability of two lineages coalescing in one generation is

$$P = \frac{1}{MN} = \frac{1}{N}\,\Delta t. \tag{3}$$

Writing this out explicitly,

$$\begin{aligned} P(t) &= (1 - \frac{1}{MN})^{Mt-1}\frac{1}{MN} \\ \lim_{M\to\infty} P(t) &= \lim_{M\to\infty}[\frac{1}{N}\frac{(1-\frac{1}{MN})^{Mt}}{1-\frac{1}{MN}}\frac{1}{M}] \\ P(t)\,dt &= \frac{1}{N}\exp(-\frac{t}{N})\,dt. \end{aligned} \tag{4}$$

The dt on the left side in the last line comes from how the right side corresponds to the probability of coalescing over a small interval in continuous time, and the measure of a singleton in the reals is zero. This may be the simplest method for this problem, but it will be hard to generalize when recombination is added.

So consider instead a large number of trials and the quantity $\bar{F}(t)$ being the fraction of trials that have not coalesced yet at the t-th previous generation. The name of the quantity $\bar{F}(t)$ is the complementary cumulative distribution function (CCDF). One may then write

$$\bar{F}(t + \Delta t) = \bar{F}(t) - \frac{\Delta t}{N}\bar{F}(t), \tag{5}$$

because the fraction of trials left at the next time step is the fraction of trials currently remaining minus the amount that will coalesce away out of the currently remaining ones. Taking the limit of Equation 5 gives a differential equation

$$\begin{aligned} \lim_{\Delta t\to 0}\frac{\bar{F}(t+\Delta t) - \bar{F}(t)}{\Delta t} &= \lim_{\Delta t\to 0}[-\frac{1}{N}\bar{F}(t)] \\ \bar{F}'(t) &= -\frac{1}{N}\bar{F}(t), \end{aligned} \tag{6}$$

whose solution is

$$\bar{F}(t) = \exp(-\frac{t}{N}) \tag{7}$$

after normalization. The probability density function (PDF) is then the negative derivative of the CCDF, so

$$P(t) = \frac{1}{N}\exp(-\frac{t}{N}). \tag{8}$$

The advantage of this method is that Equation 5 and therefore Equation 6 can be generalized to arbitrary single-generation coalescing probabilities. Assuming the general single-generation coalescing probability P does not depend on the fineness of time discretization, which should always be the case, one can apply the reasoning for Equation 5 to write

$$\begin{aligned} \bar{F}(t+\Delta t) &= \bar{F}(t) - P\,\Delta t\,\bar{F}(t) \\ \bar{F}'(t) &= -P\bar{F}(t), \end{aligned} \tag{9}$$

which recovers Equation 6 using Equation 1. Note here that P may be time dependent.

Then, to find the time distribution of the MRCA of two individuals under recombination, it suffices to find the single-generation coalescing probability under recombination. The next subsection will show that the lineage number $n(t)$ is an important quantity when recombination is present, and the relationship between n and P will be found. The third subsection will then show that the interpretation of $n(t)$ and its dynamics are more complicated than they seem, and an approximation to the dynamics of n will be made. Finally, the fourth subsection will obtain two predictions for the time distribution, one ignoring non-overlapping coalescences and one partly considering them.

B. Recombination and the single-generation coalescing probability

In reality, one can measure a recombination rate π for every population, and π is the number of recombinations per generation per nucleotide. To add in recombination to the Wright Fisher model, when finding the parent of every individual in the next generation, also generate a random number from 0 to 1 and see if it is less than πL, where L is the chromosome length counted in nucleotides. It is assumed here that $\pi L \ll 1$. If the randomly generated number is less than πL, then generate two parents for the individual. Additionally, generate an integer a from 1 to L such that the individual's first to a-th nucleotides are inherited from the first parent, and the rest are inherited from the second.

The effect of recombination on coalescence can be seen in a simple example. Suppose $L = 5000$, and both individuals are represented by/offsprings of two lineages in the t-th previous generation. The first individual inherited its 1-st to 2397-th bases (write $(1, 2397)$ as a shorthand) from its first lineage 11 and $(2397, 5000)$ from its second lineage 12. The second individual inherited $(1, 4627)$ from its first lineage 21 and $(4627, 5000)$ from its second lineage 22. Then tracing back to the $(t-1)$-th generation, there is a coalescence if either

- lineage 11 coalesces with lineage 21 (since $(1, 2397)$ overlaps with $(1, 4627)$), or
- lineage 12 coalesces with lineage 21 (since $(2397, 5000)$ overlaps with $(1, 4627)$), or
- lineage 12 coalesces with lineage 22 (since $(2397, 5000)$ overlaps with $(4627, 5000)$).

Note here that lineage 11 coalescing with lineage 22 does not find the two current individuals a common ancestor, because $(1, 2397)$ does not overlap with $(4627, 5000)$, so the parent found here does not contribute a same piece of genetic information to both current individuals. Then the probability of coalescing in 1 generation is

$$P = \frac{1}{N} + \frac{1}{N} + \frac{1}{N} = \frac{3}{N} \tag{10}$$

for this example.

For now, ignore lineages that give separated genetic information to a current individual, such as one from which an individual inherits its $(237, 3722)$ *and* $(3927, 4327)$. Such lineages may form through coalescences within an individual's lineages, for example 2 of an individual's 5 lineages had a common ancestor at a previous generation and became 1 in earlier generations. The next subsection will discuss the effect of those lineages. Additionally, assume that no two splitting points are at the same position. This assumption is made because it makes the following analysis simpler, and the likelihood of both chromosomes splitting at a same position is small given the recombination rate and the chromosome length.

The lineage number in the previous example is $n = 4$, and after coming up with more examples, one may guess that the number of overlapping pairs is always $n-1$, and the single-generation coalescing probability is given by

$$P = \frac{n-1}{N}. \tag{11}$$

This is true, and to see this, first abstract the setup a little bit. Consider representing each current individual's chromosome by a unit interval. Then, every number in the unit interval becomes a relative position on the chromosome, and since it is assumed that lineages of the same individual do not coalesce, the individual's lineages can be represented by an ordered list of real numbers starting from 0 and ending at 1. As an example, an individual having 6 lineages contributing to its

$$\begin{matrix} (1, 171) & (171, 1774) & (1774, 2342) \\ (2342, 3141) & (3141, 4916) & (4916, 5000) \end{matrix}$$

at a previous generation will be represented by the list

$$\begin{aligned} l &= (0, \frac{171}{5000}, \frac{1774}{5000}, \frac{2342}{5000}, \frac{3141}{5000}, \frac{4916}{5000}, 1) \\ &= (0, 0.0342, 0.3548, 0.4684, 0.6282, 0.9832, 1). \end{aligned} \tag{12}$$

Call a pair of adjacent numbers in a list an adjacent pair. Then with this abstraction, Equation 11 becomes equivalent to the claim that for any two such ordered lists, the number of overlapping adjacent pairs o is equal to the total number n of adjacent pairs in both lists minus 1.

For the proof, let the number of adjacent pairs in one list be n_1, the number of adjacent pairs in the other list be n_2, and the total number of adjacent pairs be $n = n_1+n_2$. Then the number of overlapping adjacent pairs o equals

$$o = \sum_{i=1}^{n_1}(1 + a_i), \tag{13}$$

where i ranges over the adjacent pairs in the first list, and a_i is the count of numbers in the second list that are surrounded by the i-th adjacent pair, excluding 0 and 1. This is because for every adjacent pair indexed by i in the first list, every adjacent pair in the second list overlapping i has its left end surrounded by i except the one most to the left. This is true because the two lists cannot contain a common number except 0 and 1, because the two chromosomes cannot have a common splitting point. So

$$\begin{aligned} o &= \sum_{i=1}^{n_1} 1 + \sum_{i=1}^{n_1} a_i \\ &= n_1 + n_2 - 1 \\ &= n - 1, \end{aligned} \tag{14}$$

where the second sum is equal to $n_2 - 1$ because the total count of numbers in the second list that are not 0 or 1 is the number of adjacent pairs there minus 1.

As a conclusion, when all lineages of both individuals at a previous generation do not contribute separated genetic information to them, Equation 11 holds.

C. Recombination and the effective lineage number

Now one only has to find the dynamics of the lineage number. To account for recombination, consider letting the lineage number grow deterministically at a rate of $2\pi L$ per generation. This means the stochasticity of recombination is ignored, and every recombination event is occurring according to the previously measured average rate.

What is left is much more complicated to deal with. When there are multiple lineages, lineages that do not overlap may coalesce, and each of these may or may not reduce the number of overlapping pairs by 1. As shown in the next section, ignoring non-overlapping coalescences may be a good enough approximation, but there is a significant/noticeable difference between the resulting prediction and simulated results. To see how non-overlapping lineages may or may not reduce the number of overlapping pairs, consider the example where the first individual has lineages represented by

$$l_1 = (0, 0.6142, 0.8944, 1), \tag{15}$$

and the second individual has lineages represented by

$$l_2 = (0, 0.4398, 0.8382, 1). \tag{16}$$

1. First suppose lineages $(0, 0.6142)$ and $(0.6142, 0.8944)$ of the first individual coalesce. Then counting the overlapping pairs before and after the coalescence gives 5 vs 4, where there is 1 reduced because $(0.4398, 0.8382)$ could match to both $(0, 0.6142)$ and $(0.6142, 0.8944)$ but can only match to $(0, 0.8944)$ after coalescence. The probability of this coalescence happening is $1/N$ in one generation.

2. Now suppose instead that lineages $(0, 0.6142)$ and $(0.8944, 1)$ of the first individuals coalesce. Then the number of overlapping pairs does not change, because no lineages of the second individual could match to both $(0, 0.6142)$ and $(0.8944, 1)$. The probability of this coalescence is also $1/N$ in one generation.

3. Alternatively, suppose instead that lineage $(0, 0.6142)$ of the first individual and lineage $(0.8382, 1)$ of the second coalesce. The number of overlapping pairs also does not change, because when a lineage of one individual matches to an overlapping lineage of the other individual, it is irrelevant whether it is matching at the same time to a lineage of its own. The probability of this coalescence is also $1/N$ in one generation.

4. Now suppose that in addition to the coalescence in case 2, lineages $(0, 0.4398)$ and $(0.8382, 1)$ of the second individual also coalesce. Then the number of overlapping pairs is reduced by 1, because after both coalescences, the pair of $(0, 0.4398)$ and $(0, 0.6142)$ and the pair of $(0.8382, 1)$ and $(0.8944, 1)$ becomes one. The probability of this coalescence is $1/N^2$ in one generation.

5. Alternatively, suppose the in addition to the coalescence in case 3, lineage $(0, 0.4398)$ of the second individual also coalesces with $(0.8944, 1)$ of the first. Then as in case 4, the number of overlapping pairs is reduced by 1 for similar reasons. The probability of this coalescence is also $1/N^2$ in one generation.

In case 1, both the lineage and the number of overlapping pairs are reduced by 1. In cases 2 and 3, the lineage number is reduced by 1, but the number of overlapping pairs is unaffected. In cases 4 and 5, the lineage number is reduced by 2, but the number of overlapping pairs is

only reduced by 1. Since Equation 11 is true and useful only because $n-1$ is the number of overlapping pairs o, redefine $n := o+1$ to be the effective lineage number. Then, in each generation, case 1 represents a process that occurs at a probability of $1/N$ and reduces n by 1, cases 2 and 3 represent processes that occur at a probability of $1/N$ but do not affect n, and cases 4 and 5 represent processes that occur at a probability of $1/N^2$ and reduce n by 1. Therefore, to account for non-overlapping coalescences only up to processes with probabilities of order $1/N$ in each generation, one only has to consider process 1, that is coalescences between lineages of a same individual that have a common overlapping lineage of the other individual.

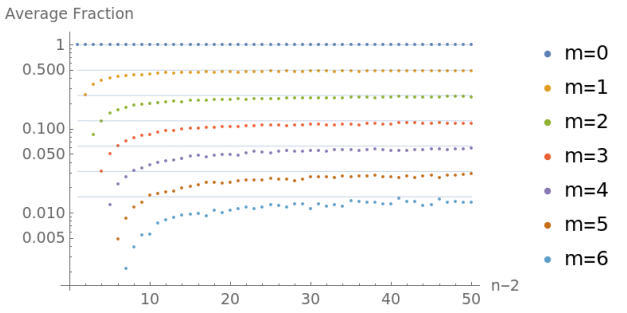

FIG. 1. The estimated average fractions of separated pairs that have a common overlapping lineage using the monte carlo method are shown here. The vertical axis is on a log scale. The thin horizontal lines are the predicted limits of $1/2^m$'s, and each dotted curve represents a fixed number of separating segments. The lineage number increases along the horizontal axis, and one can see that each curve approaches its claimed limit as the lineage number increases.

To count those, consider counting by the number of separating lineages of two lineages. When the separation number is 0, every two adjacent lineages will have a common overlapping lineages of the other individual. When the separation number is 1, on average, only a certain fraction of separation 1 pairs will have common overlapping lineages, and so on. By running a monte carlo to estimate the average, it turns out that the average fraction of separation m pairs having a common overlapping lineage is $1/2^m$. To be more specific or rigorous, given a pair of unit intervals randomly split into n pieces, when randomly picking one of the $n-2$ splitting points on the two unit intervals, the probability that the piece to the left has a $(m-1)-th$ piece to the right, and the two has a common overlapping piece on the other chromosome, tends to $1/2^m$ as n tends to infinity. A simulation with conditions ranging from $m=0$ and $n=3$ to $m=6$ and $n=52$, with $50,000$ trials per condition is performed, and its result is shown in Figure 1. When the lineage number is small relative to the separation number, the $1/2^m$ rule is quite accurate visually. The error between the simulated sum over m and the predicted sum over m for every n is shown in Figure 2. In the regime of interest, that is when n is large due to the high recombination rate, the error is below 0.1. Using this result, the effect of recombination and the effect of non-overlapping coalescences on the effective lineage number can be summarized as

$$\begin{aligned} n' &= 2\pi L - \left(\sum_{m=0}^{\infty} \frac{n-2}{2^m}\right)\frac{1}{N} \\ &= 2\pi L - \frac{2n-4}{N}. \end{aligned} \tag{17}$$

Here, n is the effective lineage number, $n-2$ is the number of splitting points on both chromosomes, $(n-2)/2^m$ is the average number of separation m pairs that have a common overlapping lineage, and $1/N$ is the probability of each pair coalescing. It is worth noting here that Equation 17 made some additional approximations. One is that once a non-overlapping coalescence occurs, the further dynamics of the lineages with the separated lineage formed (the genetic information the new lineage contributes to one of the two current individuals is separated on its chromosome) will be the same as the dynamics when all the lineages are connected. This is certainly false; a lineage coalesced from 2 pieces will have 3 adjacent lineages when going to previous generations, while Equation 17 will still assume that it has 2, like a connected lineage. Another similar approximation is that non-overlapping coalescences that do not have an immediate effect on n will not have long term effects. This is again not true; cases 4 and 5 can both be broken down into two steps, where each step can take place in one generation with a probability of $1/N$, and each individual step does not have an immediate effect. These assumptions, however, are hard to remove. And to slightly settle these concerns, the next section will compare the resulting prediction with simulated results.

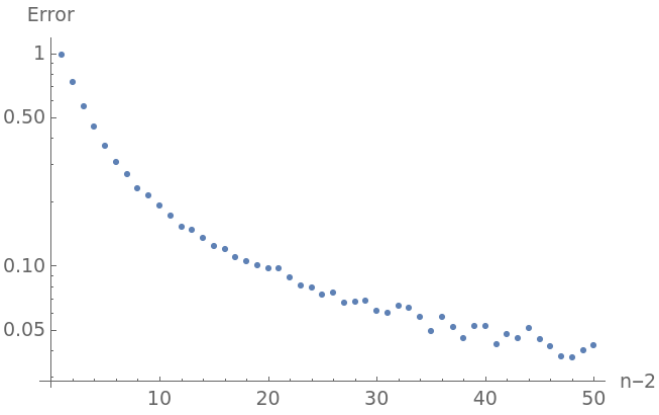

FIG. 2. For each lineage number on the horizontal axis, the sum of the first six simulated averages is compared to the sum of the first six $1/2^m$'s. The vertical axis is again on a log scale. Though the total overestimation is large when the lineage number is small, it soon decreases as the lineage number increases.

D. Two predictions for the time distribution

Using Equation 9, Equation 11, and Equation 17, one can obtain a prediction for the time distribution of the MRCA of two haploid hermaphrodites. One can also obtain a prediction by ignoring non-overlapping coalescences altogether. To do that, all one has to do is to replace Equation 17 with

$$n' = 2\pi L. \tag{18}$$

By Mathematica, The solutions to Equation 17 and Equation 18 specified by the boundary condition

$$n(0) = 2 \tag{19}$$

are (when non-overlapping coalescences are partly considered, call it the complicated case)

$$n(t) = -\pi LN \exp(-\frac{2t}{N}) + \pi LN + 2 \tag{20}$$

and (when non-overlapping coalescences are ignored, call it the simple case)

$$n(t) = 2\pi Lt + 2 \tag{21}$$

respectively. Substituting Equation 20 and Equation 21 into Equation 11 and then into Equation 9 gives

$$\bar{F}'(t) = -(-\pi L \exp(-\frac{2t}{N}) + \pi L + \frac{1}{N})\bar{F}(t) \tag{22}$$

for the complicated case and

$$\bar{F}'(t) = -(\frac{2\pi Lt}{N} + \frac{1}{N})\bar{F}(t) \tag{23}$$

for the simple case. Again by Mathematica, the general solutions to Equation 22 and Equation 23 are both normalizable, and the normalized CCDF for the complicated case is

$$\bar{F}(t) = \exp(-\frac{\pi LN}{2}\exp(-\frac{2t}{N}) - \frac{(\pi LN+1)t}{N} + \frac{\pi LN}{2}), \tag{24}$$

and the normalized CCDF for the simple case is

$$\bar{F}(t) = \exp(-\frac{\pi Lt^2}{N} - \frac{t}{N}). \tag{25}$$

Before taking the negative derivatives of Equation 24 and Equation 25 to obtain the time distributions, first notice that both equations can be simplified by defining the combined parameter

$$\rho := \pi LN \tag{26}$$

and the rescaled time

$$\tau := \frac{t}{N}. \tag{27}$$

The combined parameter is common to many other problems in coalescent theory, and it signifies the rate at which recombinations occur in the total population. With these two definitions, Equation 24 can be rewritten as

$$\bar{F}(\tau) = \exp(-\frac{\rho}{2}e^{-2\tau} - (\rho+1)\tau + \frac{\rho}{2}), \tag{28}$$

and Equation 29 can be rewritten as

$$\bar{F}(\tau) = \exp(-\rho\tau^2 - \tau). \tag{29}$$

The new equations now have only one parameter being the combined parameter ρ. The time distributions, or strictly speaking the rescaled time distributions, are then

$$P(\tau) = (-\rho e^{-2\tau} + \rho + 1)\exp(-\frac{\rho}{2}e^{-2\tau} - (\rho+1)\tau + \frac{\rho}{2}) \tag{30}$$

for the complicated case and

$$P(\tau) = (2\rho\tau + 1)\exp(-\rho\tau^2 - \tau) \tag{31}$$

for the simple case.

III. SIMULATIONS

A. Verify parameter reduction

The first thing to verify about Equation 30 and Equation 31 is the validity of parameter reduction. If it really can be assumed that different combinations of π, L, and N's that produce the same ρ give the same rescaled time distribution, then the histograms generated by the simulation under those combinations should all look the same after rescaling time. To verify this, 12 combined parameters being

$$\rho = 0, 0.25, 0.5, 1, 2, 4, 8, 16, 20, 24, 28, 32$$

are chosen for testing. For each combined parameter, sequence length ranges over

$$L = 10^2, 10^3, 10^4,$$

population size ranges over

$$N = 10^3, 10^4, 10^5,$$

and the recombination rates are then determined according to Equation 26. For each of the $12 \times 3 \times 3$ conditions/choices of parameters, $90,000$ simulations are performed using msprime. For each trial, the time of 2 random individuals' MRCA is recorded after rescaled according to Equation 27. Then, for each condition, a density histogram plot with 200 bins is generated for the $90,000$ recorded rescaled times. Finally, for each of the 12 combined parameters, the 9 histograms that all have this combined parameter are plotted together to see if any is deviated from the other. The results are shown in

Figure 3, and one can see that for all 12 combined parameters, the 9 conditions that have different combinations of π, L, and N have the same rescaled time distribution, or to put it more carefully, the between group (grouped according to their combined parameters) differences are larger than the within group differences.

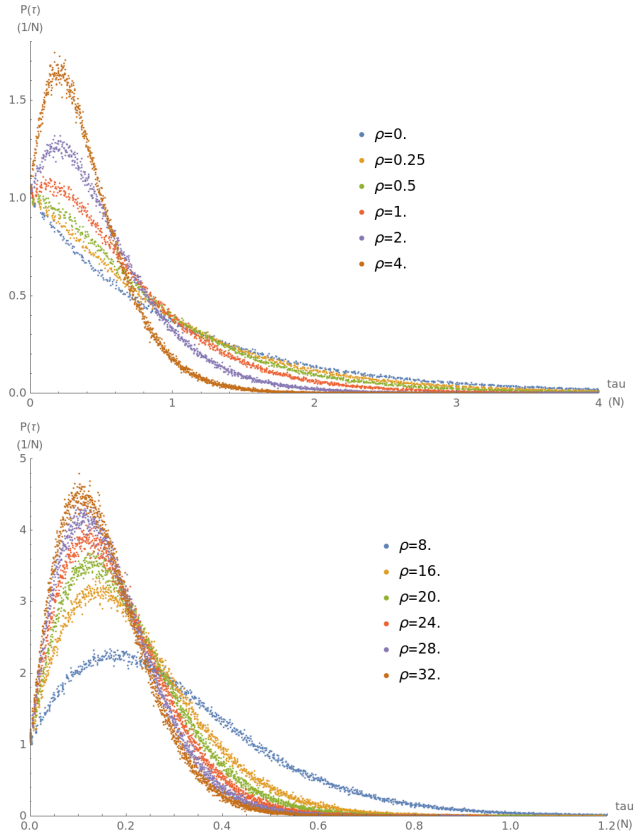

FIG. 3. The 9 histograms for all 12 combined parameters are plotted together. The horizontal axis is rescaled time, and the vertical axis is the probability density. Histograms with the same ρ are given the same color, so that one can distinguish which data point belongs to which group. The data are plotted on two graphs because their ranges vary by a lot when ρ varies from 0 to 24. Adjacent ρ's are plotted on the same graph, so that one can compare the between group and within group differences. A comparison between $\rho = 4$ and $\rho = 8$ may be difficult, but the difference between the two is apparently large given their maxima. Though some data points with different ρ are close to each other when ρ is large, the point of this graph is that curves with the same ρ but different π, L, and N are close to each other, which is still true at large ρ's.

B. Verify predictions

Once parameter reduction is verified, one can then test the accuracy of the predictions in Equation 30 and Equation 31. To do this, simply add the two prediction curves for each of the 12 combined parameters to Figure 3, and the scattered data points can inform one about the uncertainties in the simulation. Figure 4 includes the prediction curves, and one can see that the complicated predictions, i.e. the ones taking some non-overlapping coalescences into account, are within the scattered data points for all combined parameters simulated. The simple predictions, i.e. the ones ignoring non-overlapping coalescences, may also be considered as accurate when the combined parameters are relatively small, but their deviations from the simulated data are quite significant when the combined parameters are relatively large. One may infer from this that the effect of non-overlapping coalescences is not negligible when the recombination rate is high relative to the population size.

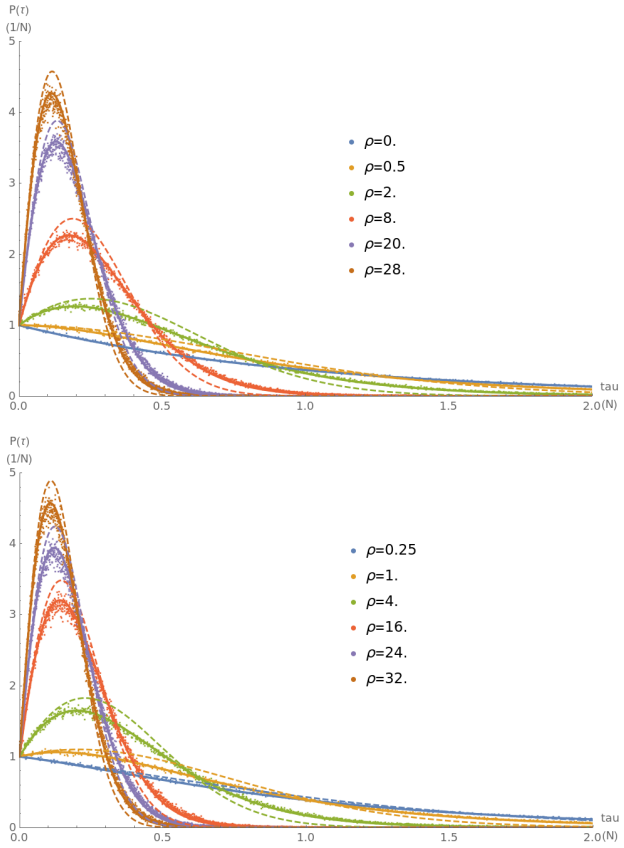

FIG. 4. Simulated data and their predictions are plotted together for comparison. Predictions and data of the same ρ are given the same color, and adjacent ρ's are plotted on different graphs for clarity. The complicated predictions are always surround by the simulated data of the same ρ, and the simple predictions are though with the right shape, always overestimating the coalescent time. The overestimation grows as ρ increases.

IV. CONCLUSION

In this report, an analytic prediction is made for the time distribution of 2 haploid hermaphrodites' MRCA. The recombination rate may or may not be small, and the final prediction compares nicely to simulated data. When analyzing recombination, non-overlapping coalescences may be negligible when the combined parameter is small, but may not when it is large.

ACKNOWLEDGMENTS

I thank msprime's developers for making msprime available. I also thank Daniel Weissman for his advice during this rotation.

[1] J. Wakely, *Coalescent Theory: An Introduction* (Macmillan Learning, 2016).
[2] Magnus Nordborg, "On the probability of neanderthal ancestry," The American Journal of Human Genetics **63**, 1237–1240 (1998).
[3] Gilean A.T McVean and Niall J Cardin, "Approximating the coalescent with recombination," Philosophical Transactions of the Royal Society B: Biological Sciences **360**, 1387–1393 (2005).