Chen-Yang Su 260729934
Simon Gravel Lab

Inferring the History of Admixed Populations

INTRODUCTION

Population genetics is a subdomain of the field of genetics where the focus is on the genetic composition of the population rather than the individual, and on how the genetic makeup such as allele frequencies changes over time. The change in frequency of the genotype and phenotype of populations is mainly affected by processes such as migration or gene flow, natural selection, genetic drift, and mutation. In this project, we consider how historical migration patterns can be traced and inferred from analyzing the genomic composition of an existing current population.

Over the course of human history, there have been many major migration events involving both voluntary and involuntary migrations of populations moving from one region to another. These large-scale migrations often change the genetic diversity and makeup of a population. The descendants from breeding between two or more ancestral populations are termed admixed, referring to their shared ancestry and genes from multiple source populations. For example, modern African Americans are an admixed population which share common ancestry from source populations involving individuals of European, West African, and Native American heritage. By exploring the genetic contributions to admixed populations, historical gene flow can be investigated and better understood to determine how it has shaped human history, demography, and genetic diversity.

Complex diseases such as diabetes, hypertension, obesity, and heart disease bear different burdens on an admixed population compared to its ancestral population[1]. Furthermore, admixed populations, such as African Americans and Latinx-Americans, are often under-represented in clinical and medical research due to socioeconomic disparities, and to a lesser extent difficulty in the modelling of heterogeneous genomic data. Therefore, with a better understanding of the contribution of genetic ancestry to admixed populations, better clinical care and outcomes can be achieved.

In this project, we attempt to model and simulate the migration between human populations through gene flow to infer admixture events. Continuous blocks or segments of the genome inherited from a population are termed “admixture tracts”. The existence of these contiguous segments is only possible through analysis of genomic data, also termed admixture deconvolution[2], since admixture tracts are unobservable. Segments of the genome which are passed down in non-admixed individuals can be compared to the same segments in admixed populations, and the variation in continuous ancestry tract lengths that exist due to recombination can provide information on the history of individuals. The tendency of earlier admixture events to give rise to shorter ancestry segments allows the inference of historical mass migration events, as well as patterns of historical gene flow between populations.

Currently, there are realistic models that study this. Here, we aim to improve the state-of-the-art with a novel and sensitive approach that combines high-replicate population genetic simulation[3] with performant composite likelihood inference on admixture tract distributions[4]. The software, *tracts*[4], allows the modeling of time-dependent gene-flow while *msprime*[3] is a coalescent simulator that allows the simulation of different evolutionary scenarios. After tracking the lineage to the most recent migration, we analyze the distribution of length and number of admixture tracts in comparison to expectation. We compare the exponential model from Gravel (2012)[4] to the *tracts* and *msprime* model to show the difference in prediction accuracy. Furthermore, we analyze our models on different parameters such as time since admixture, and proportions of ancestry from source populations, to elucidate the strengths and weaknesses of these models in predicting gene flow. In addition, we complete a speed analysis of *tracts* and *msprime* on a two-population pulse admixture model when chromosomes are varied in size to evaluate the effectiveness of the models. Finally, we use likelihood-based statistical inference to compute the predicted value of continuous migration rates in our *msprime* model.

The aim of this study is to use an efficient simulation approach instead of an analytical model. By replacing the analytical results with simulations, we gain better model generality. For example, we can model multiple source populations, or multiple pulses of migration. Nevertheless, the generality of simulations comes at a computational cost. To get sufficiently detailed distributions, a large replicate number is required. The objective is to determine if the computational improvements in *msprime* make this trade-off worthwhile. By developing our *msprime* model, which has better generality, to a wider range of evolutionary scenarios, we hope that events or processes such as natural selection, disease-association in genes, and sex-biased gene flow, can be further revealed and enable us to gain a better understanding of ancient human history.

METHODOLOGY

When analyzing recent ancestry in modern individuals, a substantial proportion of their genomes appear to come from at least two distinct populations. If these original source populations are statistically distinct, then local ancestry can be inferred from genome analysis with methods such as Lamp[5], PCAdmix[6], Saber[7], SupportMix[8], Hapmix[9], and many more.

In our modelling approach, we make a few assumptions. We assume that our population is homogeneous and identical with random mating occurring at every generation, and we assume that inbreeding is insignificant. We take an effective population size of 1,000,000 and sample size of 1,000 individuals and simulate their lineage to 5 generations or more in the past in order to identify the first migrant ancestor.

The Python code for all our simulations conducted in this report is available upon request. The software *msprime* is available through Python and documentation is available at https://msprime.readthedocs.io/en/stable/index.html. The software *tracts* was developed by Professor Simon Gravel and is available on GitHub at https://github.com/sgravel/tracts. The tree sequence toolkit *tskit* (a component of *msprime*) documentation is available at https://tskit.readthedocs.io/en/stable/.

msprime Model

msprime is a widely-used coalescent simulator that provides genome-wide simulations for millions of individuals[3]. Modern day simulations are often unrealistic, slow, require high memory usage, and produce results that are difficult to analyze and understand. The *msprime* model tackles these problems by providing not only realistic simulations but also ones that are scalable and have a low-memory footprint. The core innovation of *msprime* is the tree-sequence record, an efficient data structure that stores sets of correlated genealogical relationships. Tree sequence records allow the simulations to be significantly faster and use orders of magnitude less storage than other methods.

In our simulations in *msprime*, we use a pulse migration model to guarantee that every ancestor "moves out" by t generations into the past where t represents generation times of 5, 10, 15, or 20. Furthermore, we implement our simulations based on a backwards-in-time discrete Wright-Fisher model where there is no overlapping of generations and individuals are unable to mate with one another unless they are from the same existing predetermined generations (Nelson and Gravel, 2019)[10]. In this analysis, our simulations focus on the two-population model but simulations of k individuals with our simulator is possible.

After performing admixture simulations, we generate a migration table from the trees produced from *msprime* to determine how nodes at the end of the simulation relate to parent nodes. From this information, we can determine the tract distributions in our modern-day admixed population to infer which of the source population these genomic segments originated from.

tracts Model

The *tracts* model is a software developed by Simon Gravel in 2012 that matches the most accurate gene flow models to observed local ancestry patterns. It performs modeling through inferring the distribution of ancestry tract lengths. An example of tracts of segments originating from three identifiable source populations in an individual's genome can be seen in Figure 1. A distribution of tract length models the expected proportion of tract lengths in a set of genomes.

Modeling for admixture events in diploid individuals is computationally expensive since a full coalescent simulation would involve drift, migration, recombination, and finite chromosome length. Thus, the admixture model is simplified to that shown in Figure 2A. Here, the first migration event happens T generations ago where generations $s \in \{0, 1, 2, \ldots, T-1\}$ and each fraction of the entire population *m(s)* replaced by migrants at generation *s* are divided into fractions $m_p(s)$ from *M* migrant populations where $p \in \{1, 2, \ldots, M\}$. Random mating in a population occurs due to generations following a Wright-Fisher model. *tracts* uses a Markovian Wright-Fisher recombination model (Figure 2B, Model 2) which selects random parental chromosomes within a gene pool and follows a Markov path along any of these alleles to generate the diploid individual. The Markov assumption is satisfied since recombination events are assumed to be independent from one another. The states in the Hidden Markov Model are stored as a tuple in the form (generation, population) and the most relevant states are those that identify where migration has occurred.

The *tracts* model makes a few assumptions. First, it assumes that there is negligible occurrence of drift since the admixture began. Second, it assumes that recombination across the genome is uniform. Third, it assumes the absence of population structure within admixed populations and that mating occurs randomly without any restrictions (panmictic population).

The software *tracts* is a good approach since it provides a tractable model that predicts tract length distributions, which are sensitive summary statistics that depend on time of admixture and ancestry proportions. *tracts* models admixture effects in chromosomes of finite length and builds on top of the Pool and Nielsen model[11] with multiple improvements. Improvements such as allowing general time-dependent and strong migrations, modeling chromosomal end effects explicitly, and being able to incorporate tract assignment errors, make *tracts* a better predictor. Furthermore, it is computationally tractable to perform parameter inference which allows parameter tuning to be straightforward.

Setting up a demographic model in *tracts* involves defining functions that take parameters such as the migration rate *p*, and the time since migration has occurred *t*, as input and returning a migration matrix of dimension *t* x *p*. The migratory model takes an input vector containing the migration proportions over the last generations. Within this matrix, rows represent time, and each column defines a population. The current generation corresponds to row zero. At the last generation time, also called the “founding generation”, the migration rate should sum up to 1.

Since the model space and number of parameters is large, predefined parametrized models that introduce simplifying assumptions to describe the entire migration matrix based on a few parameters are used. An example of a function that implements single-pulse migration is given in the *tracts* github repository that formed the basis of our code for running admixture events through *tracts*.

When plotting *tracts* results, the expected number of tracts per bin for a diploid individual was simulated with distribution of chromosome lengths assumed to be of length 1 Morgan. To get the expected number of tracts per bin, the bin midpoint value was multiplied by the width.

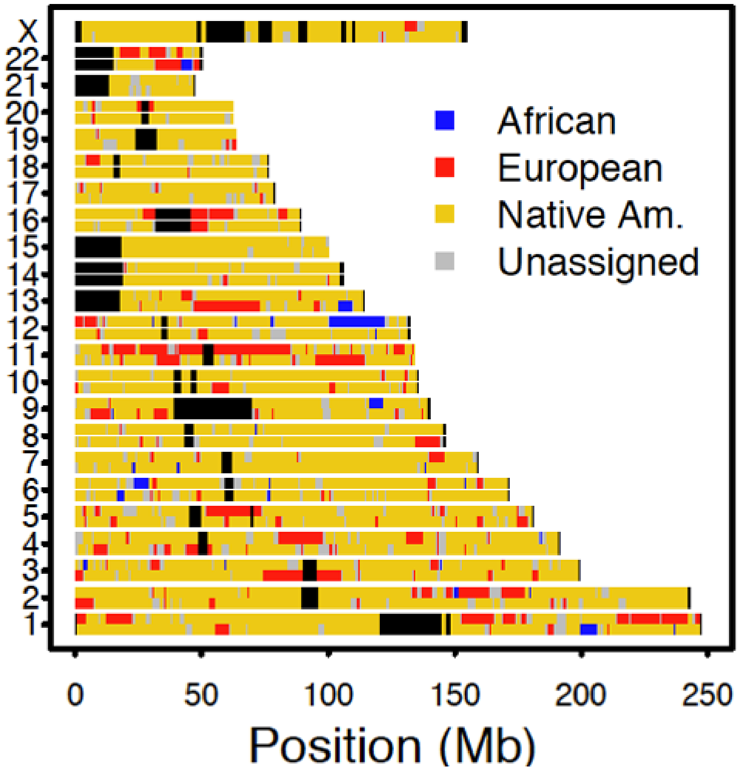


Figure 1. Tracts of ancestry for an individual with ancestry from three identifiable source populations. Retrieved from https://github.com/sgravel/tracts.

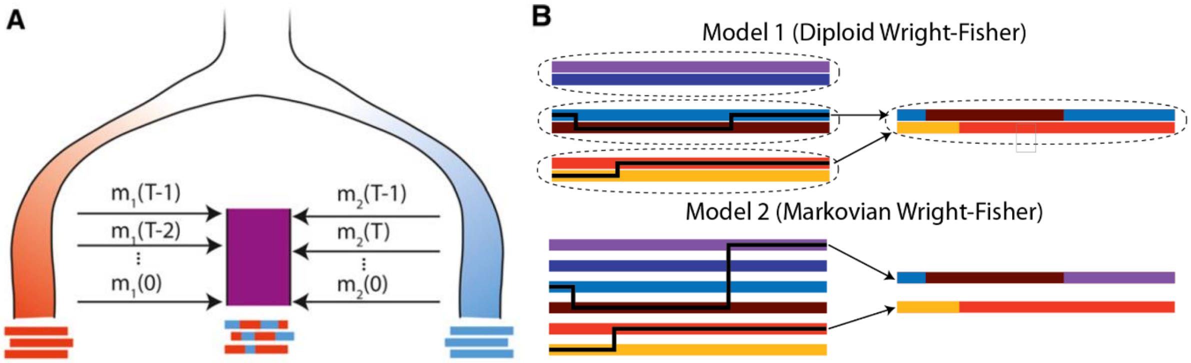

A
$m_1(T-1)$
$m_1(T-2)$
$m_1(0)$
$m_2(T-1)$
$m_2(T)$
$m_2(0)$
B
Model 1 (Diploid Wright-Fisher)
Model 2 (Markovian Wright-Fisher)

Figure 2. (A) An example of an admixture model beginning at generation $T-1$. The purple admixed population begins receiving m_it migrants at generation t from source populations coloured red (i = 1) and blue (i = 2). (B) Two different Wright-Fisher recombination models[4]. In the first model (top), two parents are randomly selected then a Markov path is followed along their chromosomes to generate the diploid individual. In the second model (bottom), any of the parental chromosomes within a pool can be selected and a Markov path is followed along these alleles to form the diploid individual.

RESULTS

In this section, we first reproduce the exponential model from equation (1) of Gravel (2012)[4], then we use it in a side by side comparison with the *tracts* model and *msprime* model to distinguish qualitative differences between the three models. Then, a speed analysis between *tracts* and *msprime* is performed by altering the chromosome length parameter for a quantitative analysis on simulation speed between the two models.

Exponential model compared to *msprime* simulation

In this study, the goal is to compare how *msprime*, a coalescent simulator matches up against the exponential model and *tracts* software. First, we compare the exponential model to our *msprime* simulator to confirm the fit of both models when common parameters such as *p* and *t* are simulated with.

Admixture tract lengths are independent and identically distributed and follow an exponential distribution. The exponential model below corresponds to equation (1) from Gravel (2012). Each equation represents different ancestry, red (R) or blue (B), representing arbitrary source populations. $\phi_i(x)$ is the distribution of tract lengths of these two ancestral populations. The parameter *p* represents the proportion of ancestry from a population or the fraction of migrants from each source population, and the time of split of the two populations going backwards in time is given by the parameter *t*.

$$\phi_R(x) = p(t-1)e^{-p(t-1)x}$$

$$\phi_B(x) = (1-p)(t-1)e^{-(1-p)(t-1)x}$$

Equation 1. The exponential distribution of two ancestral populations. *R* and *B* are arbitrary labels for two source populations.

In the *msprime* simulation, the mass migration happens at generation *t*, which is set as 20 generations by default in *msprime*. Thus, a starting population of 1000 individuals with an effective population size of 1 million was simulated for 20 generations into the past with a recombination rate (defined as the recombination rate per base per generation in *msprime*) of 1e-8 and a chromosome length of 1e8 base pairs. The final outputted segment lengths were converted to Morgans by using the conversion 1e8 base pairs = 1 Morgan and the values were plotted on a log scale to improve visualization. The resulting histogram was compared to the expectation from the exponential model shown in Equation 1 (Figure 3). The length distribution for both models were simulated with proportions of 0.5 from each ancestral population after 20 generations.

From initial observation, the exponential distribution appears to fit the data from the simulation well when 20 generations into the past were used. We then simulated with *t* = 19 and *t* = 21 in the exponential model while keeping *t* = 20 constant in the *msprime* model to check for an off-by-one error and found that by setting generation time to 19, the exponential model fit the data worse than before (Appendix 1 Left) as seen by segments of medium length not reaching the red exponential line. Similarly, when *t* = 21 in the exponential model (Appendix 1 Right), the fit between the two models is worsened as demonstrated by earlier overshooting of the *msprime* histogram y values. This suggested that an off-by-one generation time error was not present between the *msprime* simulation and the exponential model, and that both models match well in terms of their common parameter *t*.

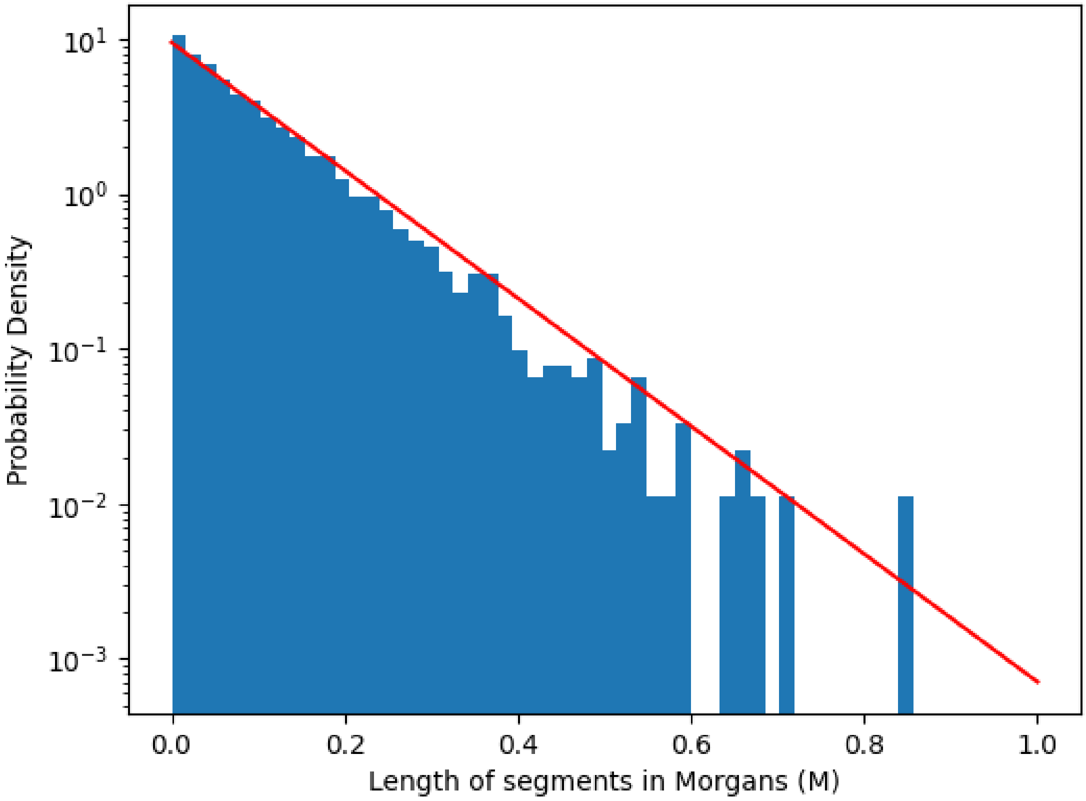


Fig 3. Exponential distribution versus *msprime* simulation model plotted on log scale. Filled histogram represents an *msprime* simulation with chromosome length of 1e8 base pairs (1 Morgan) with parameters *Ne* = 1e6, *n* = 1000, *t* = 20, *r* = 1e-8 (*Ne* = effective population size, *n* = sample size of the admixed population, *t* = number of generation the admixed population existed, *r* = recombination rate, length = chromosome length in base pairs). Each bar in the histogram is one of 50 bins. Red line represents the distribution of segment length under a simple exponential model modeled from Equation 1 with parameters *p* = 0.5, *t* = 20. (*p* = proportion of ancestry from a population; *t* = time of split).

Parameter search in *tracts* model

The *tracts* software allows the modeling of time-dependent gene-flow from multiple populations. Ancestry tracts of admixed individuals are used to model migration histories[4]. The *tracts* software contains bed files which carry the information from local ancestry calls and describes the local ancestry of segments along the genome. The bed files are generated from running a local ancestry inference method and in general, Rfmix is used[12]. Since *tracts* is object oriented, whole population information can be loaded in and used to calculate statistics. The generalizability of *tracts* allows it to take *k* populations as input and simulate arbitrary time-dependent migrations rates, m, for each population. A limitation of its extensibility that can be seen in most demographic models is the size of the model space. The number of parameters to optimize is much greater than information available for the model to learn from and in *tracts* the large model space can be seen in how it stores the migration rates as an array of dimension *t* x *k*. Thus, the model space must be parametrized and the risk of introducing biases must be kept to a minimum[4]. Fortunately, in *tracts*, many population models involving 2- or 3-populations are pre-defined and implemented already.

Instead of using the set of bed-style files in the software *tracts* as input to extract chromosome length information, we instead define a chromosome of length 1 Morgan to run in our simulation. By default, the software *tracts* supposes a single sample for a simulation and the results of a 2 population model simulated 20 generations into the past with ancestry proportions of 0.25 and 0.75 is shown in Figure 4 (Left). In general, the expected number of tract lengths decreases as a function of tract length. Tracts of shorter length tend to be present in a higher amount for both blue and orange curves in Figure 4 (left) and represent admixture events that happened further in the past. In contrast, tracts of greater length are present in greater portions in the genome due to admixture events happening closer to the present time (Figure 4 Right). When the time since split is much smaller and simulated back in time for only 5 generations (Figure 4 Right), a spike in the right-most bin representing continuous ancestry tracts of length 1 Morgan are visible. The orange curve representing p = 0.75 demonstrates a significantly larger uptick in the last bin than the blue curve that shows p = 0.25. The last bin contains chromosomes that have no ancestry switches and do not appear in the left graph of Figure 4 since enough time has passed for admixture events to happen on all chromosomes. In Gravel (2012)[4], any cases where the last bin contained chromosomes with no ancestry switches were excluded from the plotting as they do not correspond to a specific length value.

If proportions of ancestry of the two source populations is equal, for example by setting p = 0.5 for both source populations, then both populations are expected to have the same distribution (Appendix 2). Parameters that affect the steepness or slope of the distribution are the parameters p and t. A smaller value of p, representing proportions of ancestry from a source population, as well as a larger value of t, representing more time since the occurrence of admixture events, can both lead to a greater number of shorter segments along the genome and thus a steeper slope in the distribution.

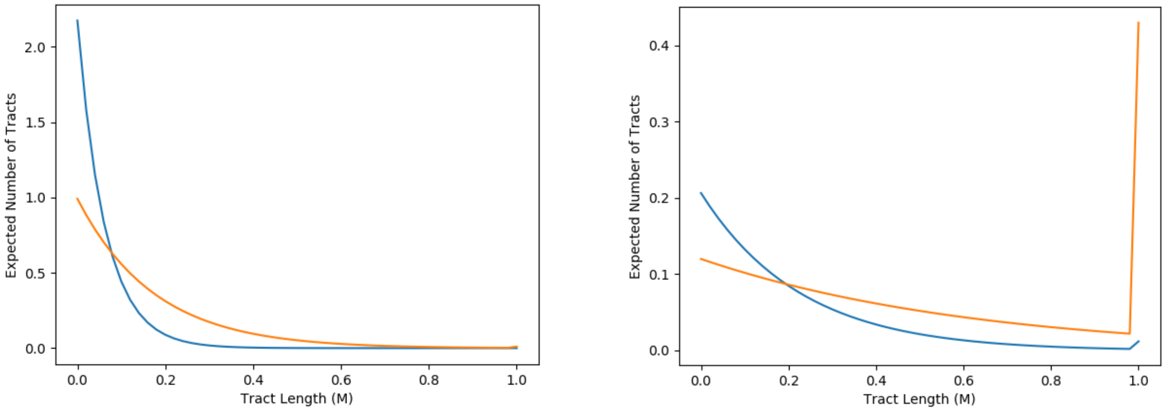


Figure 4. Expected number of tracts per bin of a single sample for 2 populations when simulated back in time, *t*, for 20 generations with proportions, *p*, of ancestry of 0.25 (blue) and 0.75 (orange) in the left figure and *t* = 5, *p* = 0.25, 0.75 in the right figure. (*p* = proportions of ancestry of the two source populations, *t* = time since the admixture in generations). Each data point represents a continuous segment along the genome whose length is contained in one of 50 bins.

A comparison of *msprime*, *tracts*, and the exponential model

In Figure 5, a comparison of all three models, *msprime*, *tracts*, and the exponential model are plotted when simulated back in time for 20 generations with initial migration proportions of 0.5 from each of the two source populations. For the exponential model the probability distribution is for a single observation. In *msprime,* we simulate with a single chromosome of length 1 Morgan and in *tracts* the default is a simulation with a single sample. For the *msprime* simulation, a sample size of 1000 admixed individuals, effective population size of 1e6, recombination rate of 1e-8, and chromosome length of 1 Morgan was used to obtain the results. The probability density is plotted on a log scale as a function of fragment lengths in Morgans. As expected, both graphs in Figure 5 appear almost identical with the *tracts* and exponential model showing similar results due to the deterministic nature of these models. The *msprime* model differs slightly between the two populations owing to its variability and stochasticity when simulating admixture events. From qualitative analysis, the three models appear to fit well with one another with *tracts* appearing to provide a lower bound and the exponential model bounding the *msprime* simulation above.

Using the same number of generations, *t* = 20, since the original admixed population split, and the same original parameters for *msprime*, we simulated all three models with initial migration proportions of *p* = 0.25 and *p* = 0.75 in the source populations (Figure 6). Once again, the data is plotted on a log scale and the *tracts* model appears to take on more conservative values for calculating tract length.

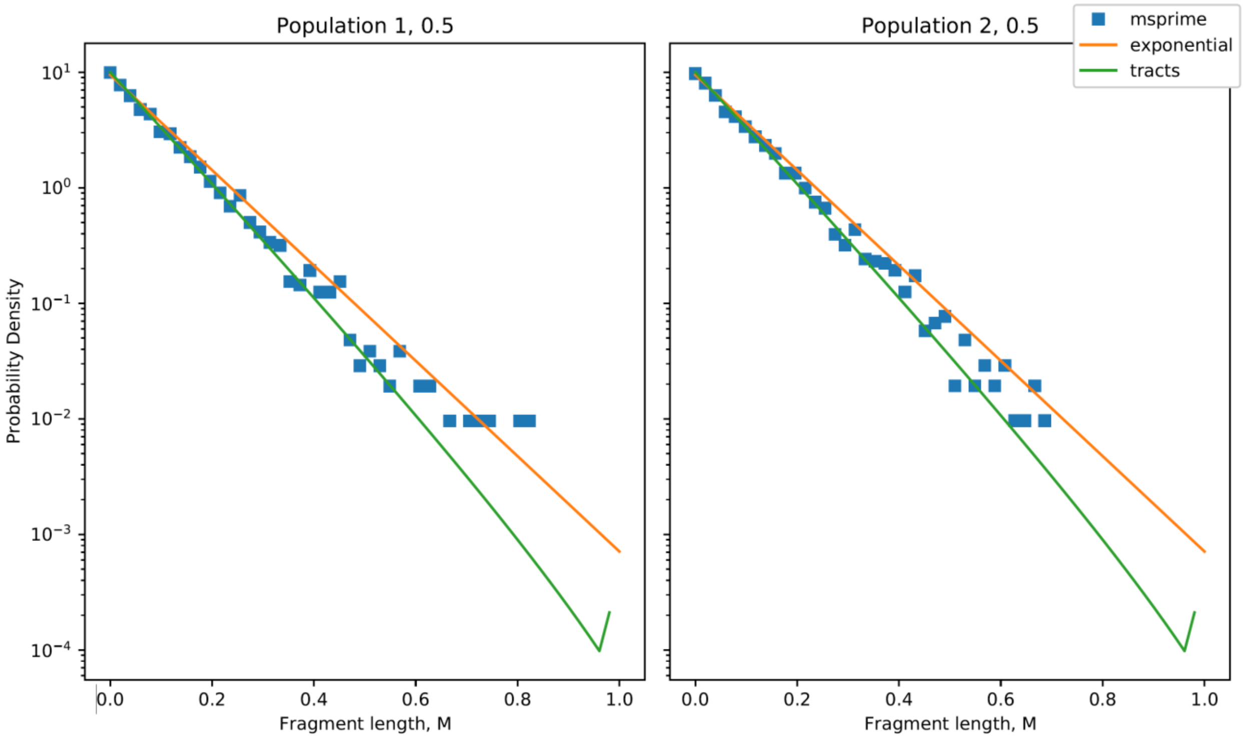


Figure 5. A comparison of the *msprime*, *tracts*, and exponential model when simulated back in time for 20 generations with initial ancestry migration proportions of 0.5 for each source population. Results are plotted on a log scale for probability density and each dot in the *msprime* model represents continuous fragments whose length is contained in one of 51 bins.

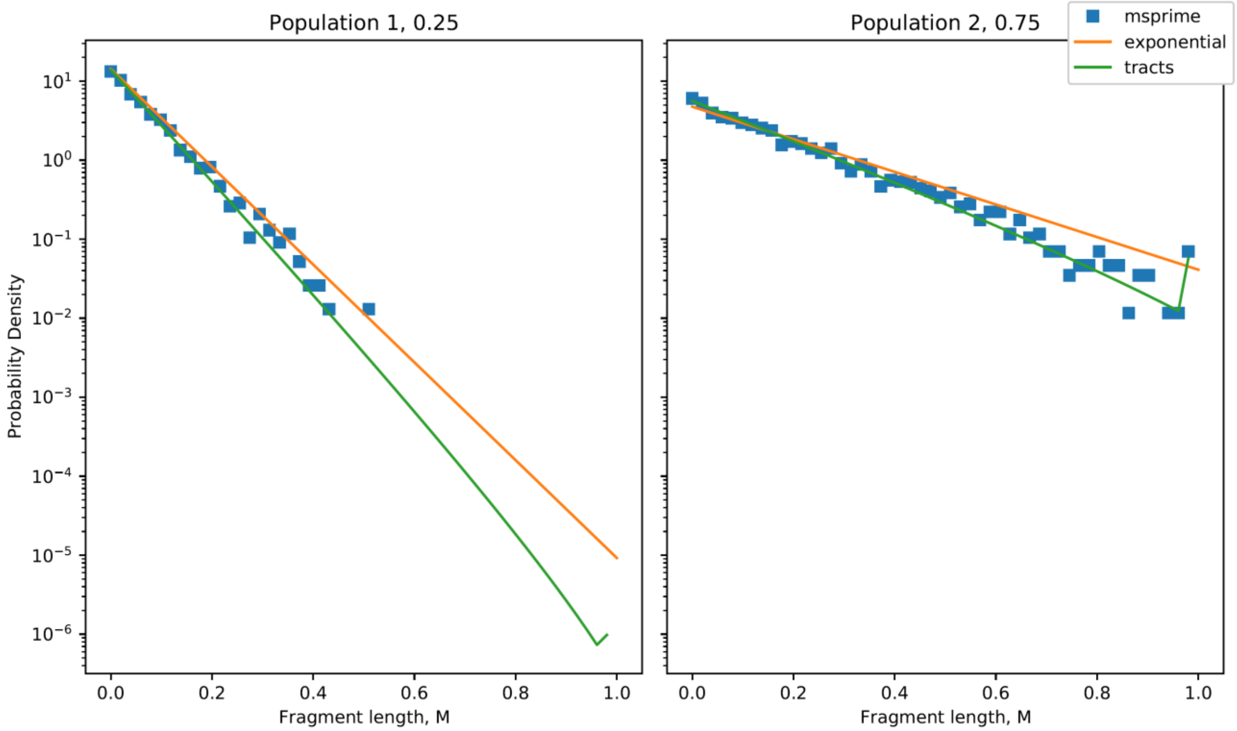


Figure 6. A comparison of the *msprime*, *tracts*, and exponential model when simulated back in time for 20 generations with initial ancestry migration proportions of 0.25 (left) and 0.75 (right) for each source population.

Results are plotted on a log scale for probability density and each dot in the *msprime* model represents continuous fragments whose length is contained in one of 51 bins.

Simulations with t = 10 and t = 15 were also done for initial migration proportions of p = 0.25 and p = 0.75 in the source populations and generally showed similar results to when t = 20. Figures for these simulations can be found in Appendix 3.

When simulations were done back in time for 5 generations instead of 20 generations, results become more interesting (Figure 7). Once again, initial migration proportions were set to 0.25 for one source population and 0.75 for the other source population. For both, the *msprime* simulation results in more spread-out data points. Furthermore, the exponential distribution tends to predict the lowest expected number of small fragments.

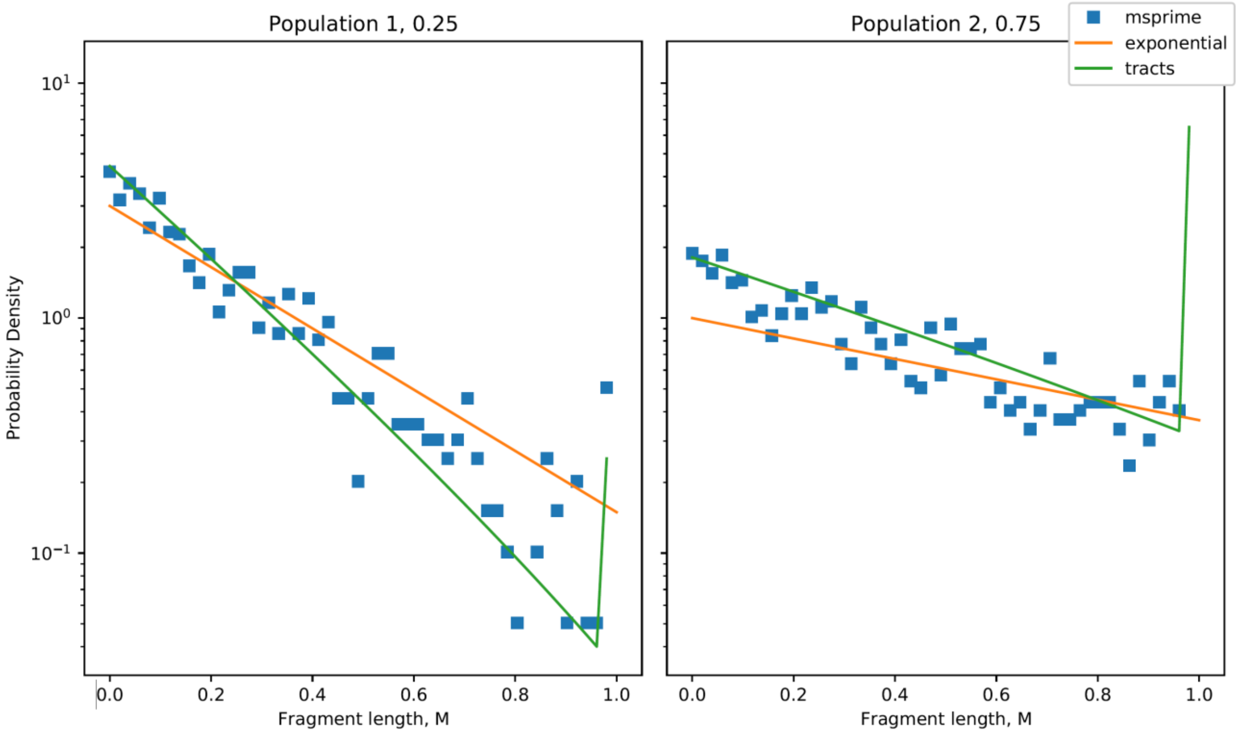


Figure 7. A comparison of the *msprime*, *tracts*, and exponential model when simulated back in time for 5 generations with initial ancestry migration proportions of 0.25 (left) and 0.75 (right) for each source population. Results are plotted on a log scale for probability density and each dot in the *msprime* model represents continuous fragments whose length is contained in one of 51 bins.

Speed analysis of *tracts* versus *msprime*

To test the efficiency of *msprime* compared to *tracts*, we run simulations on both models with different starting parameters that represent diverse evolutionary scenarios. In particular, we modify the chromosome length measured in Morgans between ranges of 0.1e8 to 3.5e9 to test out how run-time in our models change (Figure 8 and Table 1). Each run-time value in the table was averaged over 10 replicates and the simulation was done in a two-population model. The *msprime* simulation was ran with a sample size of 1000 admixed individuals, effective population size of 1e6, and using a two-source-population model with a recombination rate of 1e-8. Both models were simulated 20 generations into the past with proportions of initial migration of *p* = 0.5 for each of the two source populations.

From the *msprime* speed analysis results, the data suggests a linear increase in run-time as a function of chromosome length. In *tracts*, the simulations were instant and on the order of milliseconds and changing the magnitude of chromosome length did not have an effect on run-time which stayed relatively constant.

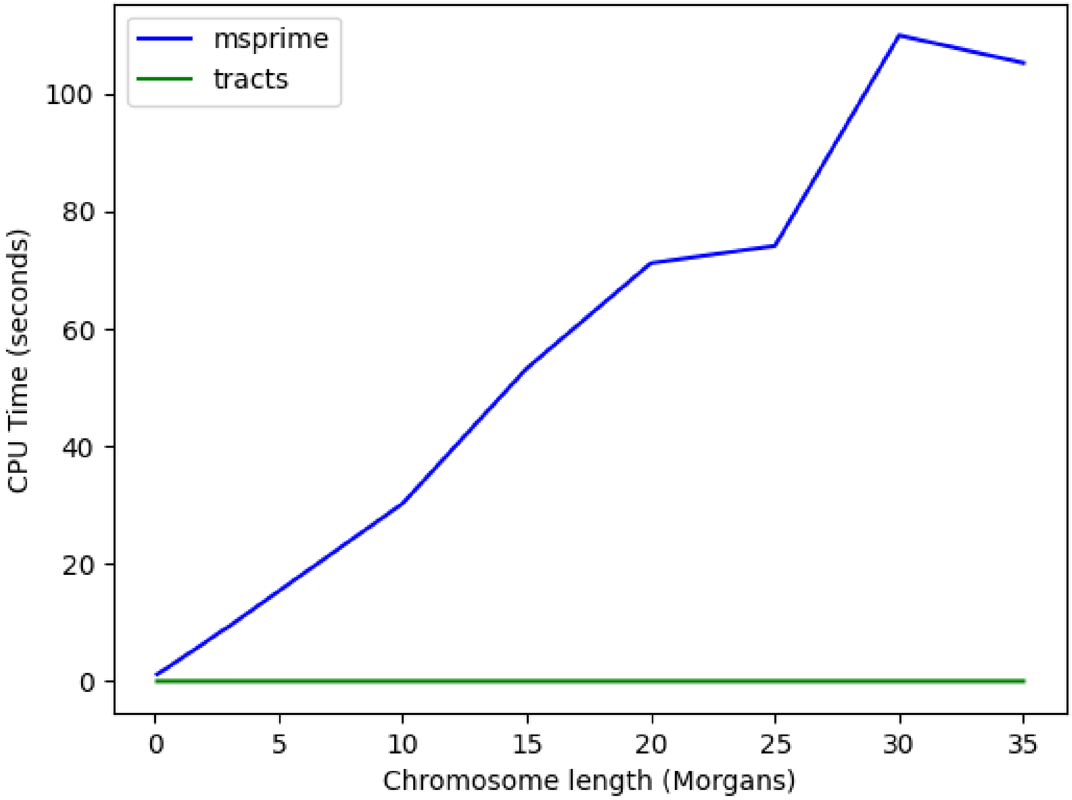


Figure 8. Speed analysis of *msprime* versus *tracts*. Run-time of the two models as a function of chromosome length in Morgans is compared in a 2-population model. The *msprime* model is simulated with parameters *n* = 1000, *Ne* = 1e6, *r* = 1e-8 (*n* = sample size of admixed individuals, *Ne* = effective population size, *r* = recombination rate). The time of split was set to 20 generations and initial migration proportions were set to 0.5 for each source population in the analysis and the actual chromosome lengths simulated as well as run-time can be found in Table 1. Each data point is averaged across 10 replicates.

Chromsome length (Morgans)	***tracts* run time (sec)**	***msprime* run time (sec)**
0.1	0.008304095268249512	1.202858909999486
0.5	0.00808389186859131	2.2368318900000306
1.0	0.00819101333618164	3.701139259999036
1.5	0.007835888862609863	5.075145690000499
2.0	0.008254504203796387	6.474344710001605
2.5	0.008240580558776855	7.9816020199999915
3.0	0.008127999305725098	9.339760659998865
5.0	0.008074212074279784	15.33858833000122
10.0	0.008188605308532715	30.281085049998364
15.0	0.007968688011169433	53.29472311999998
20.0	0.008150291442871094	71.20256120999997
25.0	0.008256101608276367	74.14945522000198
30.0	0.008202505111694337	109.97405449
35.0	0.008039283752441406	105.30333202000183

Table 1. Speed analysis of *msprime* versus *tracts*. Each run-time value is averaged across 10 replicates.

DISCUSSION

Recombination between the chromosomes of individuals results in different alleles appearing within an individual genome. In our study, we simulate a population of individuals backwards in time for 20 generations until ancestors have moved out to the original two source populations. Being able to infer the admixture tract lengths of individuals allows us to infer the history of migration, identify disease loci, and look for regions experiencing selection[4].

We develop our own simulation model in msprime and compare it to a pre-existing one, *tracts,* to evaluate three major results involving the accuracy, run-time, and generality of our model.

In terms of accuracy, our goal was to simulate with enough starting individuals such that the distribution of tract lengths resulted in a small confidence interval with smooth, less noisy data. In Figure 3, when comparing the exponential distribution with the *msprime* simulation model, the fit between the two models is mostly by qualitative observation. To decrease subjectivity and bias, possible ways to analyze the actual fit include calculating the difference or relative error as well as root mean square deviation. We compare the distributions of the *msprime*, *tracts*, and exponential models (Figures 5, 6, 7) when run with different initial migration proportions and different number of generations into the past and show that our models generally fit well with one another. In the *msprime* model, the chromosome length (l) was set to 1e8 base pairs, and the recombination rate (r) to 1e-8, so the equation $r*l = 1$ is satisfied. The *tracts* and exponential models both assume lengths are in Morgans, while *msprime* computes recombination distances in base pairs. We performed a conversion of 1e8 base pairs = 1 Morgan for the *msprime* model to avoid distorting the distribution because of hotspots with short segments and recombination deserts with long segments which can also cause more variance in tract length than in the constant recombination model. In doing so, we were able to reproduce the distributions from Figure 4 of the *tracts* paper[4], also shown in Appendix 4 as a modified version. In the original paper, the simulations were carried out under continuous gene flow or continuous migration while in our simulations, a pulse admixture migration approach was used. The difference is not too crucial since the main goal of our study is not to compare pulses or continuous migration, but to compare the same models in both simulation frameworks.

Speed and scalability are one the main reasons for using *msprime* over alternate simulators[10]. *Msprime* makes storage more efficient and increases simulation speed allowing simulations with genome-scale data of hundreds of thousands to millions of samples to be finished in minutes and data processing to be completed in seconds[3]. Before analysis of *msprime* versus *tracts*, we discovered that population size in our *msprime* simulation was a major contributing factor to the run-time. By simulating with a smaller starting population, the run time of our model became much quicker. For instance, with 1e3 starting individuals, our model was able to finish the calculations in just over 3 seconds (Table 1). However, simulations with upwards of 1e6 individuals would increase the run time by almost one order of magnitude, 10 seconds (data not shown). This is consistent with Figure 3 of Kelleher et al. (2016)[3]. There is a trade-off between speed and accuracy as a faster simulation using a smaller sample size of individuals would result in a noisier distribution while a larger sample size would take longer to simulate but result in better, less noisy results. In this study, we focused on how chromosome length would affect the speed in *tracts* compared to *msprime*. Our results (Figure 8 and Table 1) suggested that *tracts* was not only much more efficient than *msprime*, but also remained relatively constant in terms of run-time when increasing chromosome lengths by an order of magnitude, while a linear increase in run-time was seen in *msprime*. In this analysis of run-time speed as a function of chromosome length, our demographic model in *msprime* was applied with constant parameters for the number of generations the admixed population existed (t = 20), and for the proportion of initial migration from the source populations (p = 0.5 for each source population). The parameter p should not affect the run-time, and as a check when running our *msprime* model with ancestry proportions of 0.25 and 0.75 for a 1 Morgan length chromosome the time took 3.57 seconds (data not shown) instead of 3.7 seconds (Table 1). Different parameters of t were not tested since it does not matter as much in pulse admixture events. If simulating continuous migration however, the parameter t would start to make a difference and would need to be tested. Moreover, past 20 generations into the past, the distribution of tract lengths becomes uninformative (Gravel, 2012). Therefore, in this application, t = 20 acts as an upper limit to the relevant time window we want to investigate.

Despite having much slower run-time than *tracts*, the strength of our *msprime* model lies in its ability to generalize to different evolutionary scenarios. This generality allows a more realistic representation of true human migration patterns. The simulation of a starting cohort of individuals back in time until two source populations are identified was the main feature of this study as our main goal was to compare our novel simulator with the pre-existing *tracts* software which is deterministic in nature. The robustness of our model lies in its ability to generalize and simulate *k* source populations where *k* > 2, or model multiple pulses of migration. In addition, our model is able to take different recombination rates for different populations as input which is a feature not possible in *tracts*.

We also performed likelihood-based statistical inference with our *msprime* demographic model by using constant parameters *r* = 1e-8 and chromosome length = 1e8 base pairs and testing on a range of continuous migration rates of 50 evenly spaced values between 1e-6 and 0.3. A grid search was carried on these different continuous gene flow rates to determine the most optimal continuous migration rate. By optimization with maximum likelihood inference, the best continuous migration rate was found to be 0.1 (Figure 9). The optimization of likelihood-based statistical inference is the next step and our current results are preliminary. Due to long run-times of *msprime*, this inference framework is not particularly efficient and requires further development.

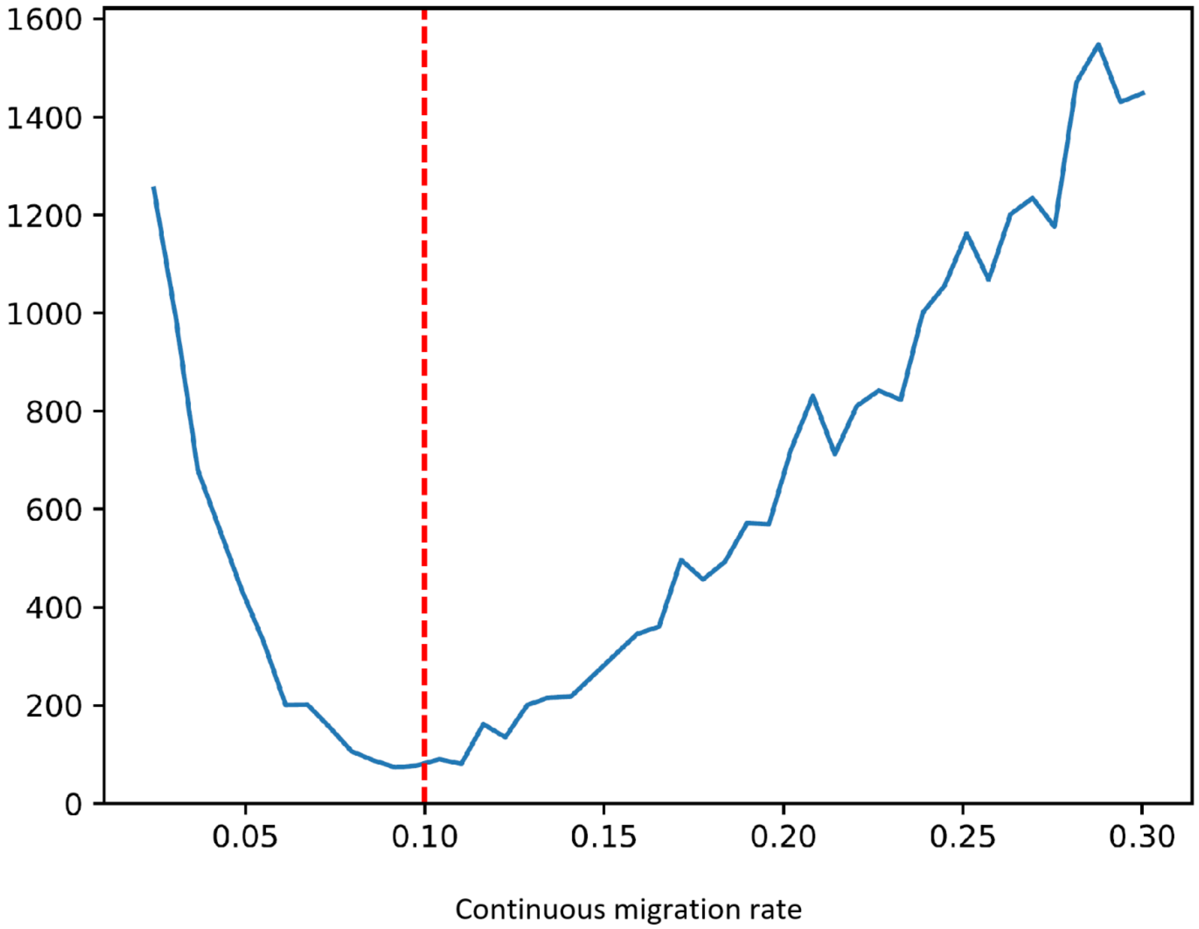


Figure 9. Likelihood-based statistical inference to determine most optimal continuous migration rate in *msprime* model. Red line shows the most optimal migration rate which is close to 0.1.

Limitations of the *tracts* model

Limitations of the *tracts* model have been reported by Liang and Nielsen (2014)[2] where it was discovered that admixture tract lengths are neither independent and identically distributed nor exponentially distributed. The Pool and Nielsen (2009)[11] model assumed the insignificance of inbreeding and the rarity of tracts such that recombination of tracts with one another is unlikely. *Tracts* built on top of this model by relaxing the second assumption and used a simplified model of reproduction based on a Markovian Wright-Fisher Model which follows a Markov process along the genome. However, Liang and Nielsen (2014) pointed out that for recent or for ancient admixture events, the assumption of iid exponential tract lengths does not hold. Their results suggested that two major factors result in deviation from this assumption: having a finite number of ancestors for small generation times and having inbreeding for large generation times[2] caused by inherited fragments being highly correlated with one another. With a small value of *t*, the high correlation between tracts is due to inheritance from a pedigree that is small and fixed, and with a large value of *t*, inbreeding may cause similar correlation between fragments. In addition, the number of migrant ancestors an individual has can contain variance that contributes to this correlation in segments also. Thus, false-positives may be present in the data if this iid assumption is relied on when determining the number of admixture events[2]. The assumption of tract lengths being exponentially distributed is also incorrect, which is shown by Theorem 3 in Liang and Nielsen (2014) which explains how the ancestry-copying process (N_x) in the Markovian Wright-Fisher model is not Markov because of individuals having finite number of ancestors and recombination fragments inherited from the same ancestor being brought together again.

FUTURE DIRECTIONS

One of the advantages of our *msprime* approach is its extensibility. For example, we may attempt to assign sexes to individuals within the population to allow simulation of sex chromosomes, in particular the X chromosome, and infer sex-based migration patterns. Another possible future direction for this project would involve incorporating assortative mating within the model to replace the oversimplified random mating assumption.

Often it is difficult to account for realistic scenarios with simulations. In this study, our *msprime* model was simulated under constant recombination rate and it would be interesting to compare results given by our model when realistic recombination rates from HapMap data were used. We could generate similar histograms to begin fitting demographic models to our HapMap population data. By testing our simulations on real data, our models can potentially become more applicable and relevant to real life situations and increase contribution to the field.

To model realistic scenarios we can use chromosome-wide recombination rates from the HapMap genetic map data (https://github.com/adimitromanolakis/geneticMap-GRCh37). In Figure 10, a HapMap of chromosome 22 is shown. An even better method would be to build a dictionary for every chromosome from the HapMap file, storing the variable recombination rates along the chromosome or just storing the average recombination rate of each autosome. It is important to note that the derivation of the exponential distribution only works in the constant recombination model so if comparing the exponential model to the *msprime* model using a realistic recombination map with variable recombination rate, the only possible way to compare the models would be by expressing segment lengths in *msprime* in units of Morgans.

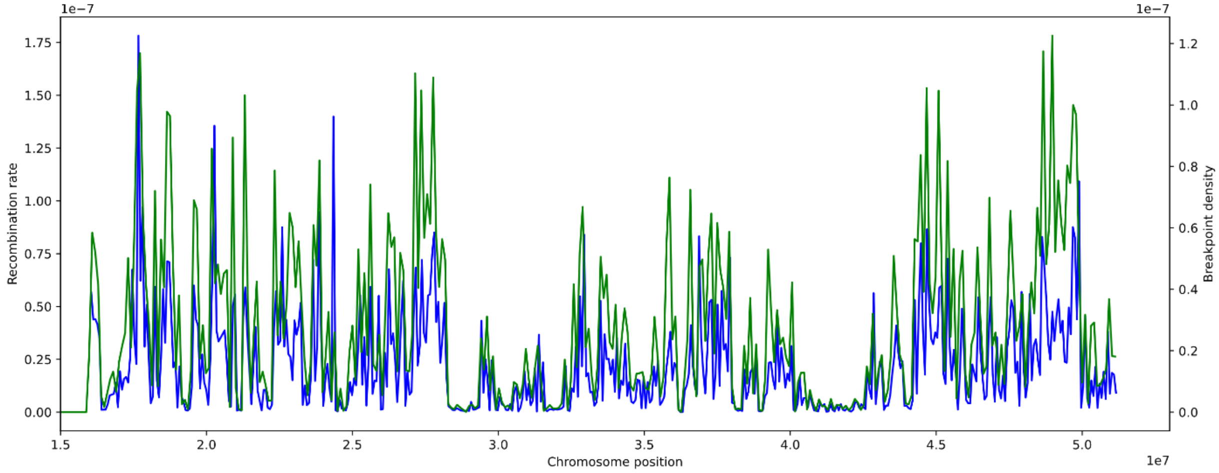


Figure 10. HapMap of chromosome 22 showing variable recombination rates (blue) and density of breakpoints (green) along the chromosome. Reference: https://msprime.readthedocs.io/en/stable/tutorial.html#recombination-maps

In the study, we performed analysis with discrete migration such as single pulse migration events and our data demonstrated that *tracts* runs extremely fast when run on a single pulse admixture model compared to *msprime*. A particularly interesting direction would be to perform continuous migration on our models to see how speed differs between the *msprime* and *tracts* model as different starting parameters are chosen. In theory, *tracts* should be much slower than *msprime* in a continuous model. In the continuous migration approach, the parameter *t* (number of generations the admixed population existed) would begin to significantly affect run-time and in the speed analysis between both models, this parameter would need to be assessed and probed for different values.

In our study, we focused on determining how chromosome length affected the run-time of *tracts* and our *msprime* model. As mentioned before, sample size is also a parameter that would affect the run-time of our models and a future direction could be to determine the smallest sample size and effective population size such that a balance between the accuracy and speed of our simulation could be obtained and maximized. If in doing so, results yield too much noise, we can instead defer to approximate Bayesian computation (ABC) to estimate the distribution. In our case throughout the study, ABC was not needed as simulating with enough individuals (1000), and a large enough effective population size (1e6) allowed less noisy data distributions.

CONCLUSION

Overall, we showed that the *msprime* model was able to accurately simulate pulse admixture events and can output similar distributions as that of *tracts*. Despite the slower speed of our *msprime* simulation model, its strength lies in its ability to generalize to many diverse evolutionary scenarios and we briefly showed the inference capabilities of our model in a continuous migration scenario as well. Although inferring demographic parameters from real data was not shown in this study, our model is capable of doing so and as a consequence, our model can potentially improve the ability to infer the complex histories that underlie the demographics in current populations. In doing so, populations that are under-represented in medical genetic studies due to complexities in modeling heterogenic data can hopefully begin becoming better represented, and genetic diversity can be minimized as a factor in causing disparities in testing and medical care.

ACKNOWLEDGEMENTS

I would like to thank Professor Simon Gravel for allowing me to conduct this research project under his supervision. As he is one of the top emerging population geneticists in the world, I am deeply humbled to be able to collaborate with him and work in his lab. I am extremely grateful for the help of Dr. Ivan Kryukov for his supervision and support throughout this entire project. The success of this project would not have been possible without his experience and expertise. I would also like to thank the Biology and Computer Science program at McGill for the opportunity to conduct this independent research project for credit, as well as Professor Mathieu Blanchette for coordinating this course and allowing flexibility during the ongoing COVID-19 pandemic.

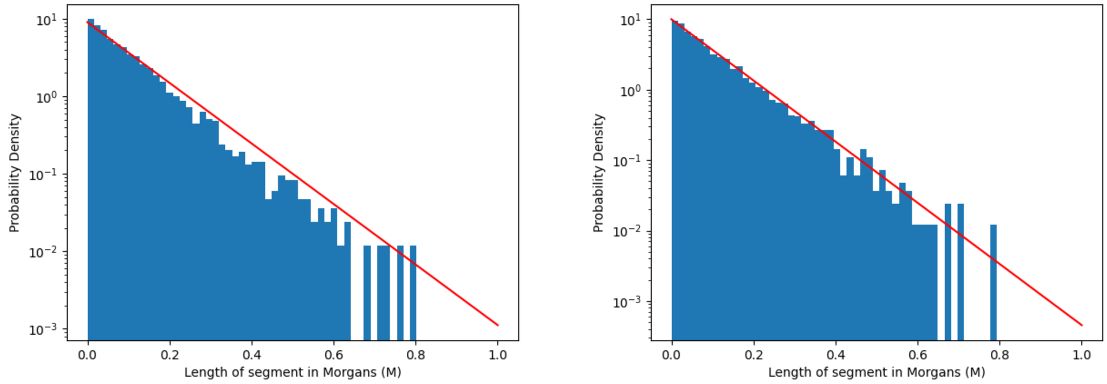
WORKS CITED

1. Qin, H., Zhao, J. & Zhu, X. Identifying Rare Variant Associations in Admixed Populations. *Sci. Rep.* **9**, 1–11 (2019).
2. Liang, M. & Nielsen, R. The lengths of admixture tracts. *Genetics* **197**, 953–967 (2014).
3. Kelleher, J., Etheridge, A. M. & McVean, G. Efficient Coalescent Simulation and Genealogical Analysis for Large Sample Sizes. *PLoS Comput. Biol.* **12**, 1–22 (2016).
4. Gravel, S. Population genetics models of local ancestry. *Genetics* **191**, 607–619 (2012).
5. Sankararaman, S., Sridhar, S., Kimmel, G. & Halperin, E. Estimating Local Ancestry in Admixed Populations. *Am. J. Hum. Genet.* **82**, 290–303 (2008).
6. Brisbin, A. *et al.* PCAdmix: Principal Components-Based Assignment of Ancestry along Each Chromosome in Individuals with Admixed Ancestry from Two or More Populations. *Hum. Biol.* **84**, 343–364 (2013).
7. Tang, H. *et al.* Recent genetic selection in the ancestral admixture of Puerto Ricans. *Am. J. Hum. Genet.* **81**, 626–633 (2007).
8. Omberg, L. *et al.* Inferring genome-wide patterns of admixture in Qataris using fifty-five ancestral populations. *BMC Genet.* **13**, (2012).
9. Price, A. L. *et al.* Sensitive detection of chromosomal segments of distinct ancestry in admixed populations. *PLoS Genet.* **5**, (2009).
10. Nelson, D., Kelleher, J., Ragsdale, A. P., McVean, G. & Gravel, S. Coupling Wright-Fisher and coalescent dynamics for realistic simulation of population-scale datasets. *bioRxiv* 674440 (2019). doi:10.1101/674440
11. Pool, J. E. & Nielsen, R. Inference of historical changes in migration rate from the lengths of migrant tracts. *Genetics* **181**, 711–719 (2009).
12. Maples, B. K., Gravel, S., Kenny, E. E. & Bustamante, C. D. RFMix: A discriminative modeling approach for rapid and robust local-ancestry inference. *Am. J. Hum. Genet.* **93**, 278–288 (2013).

APPENDIX

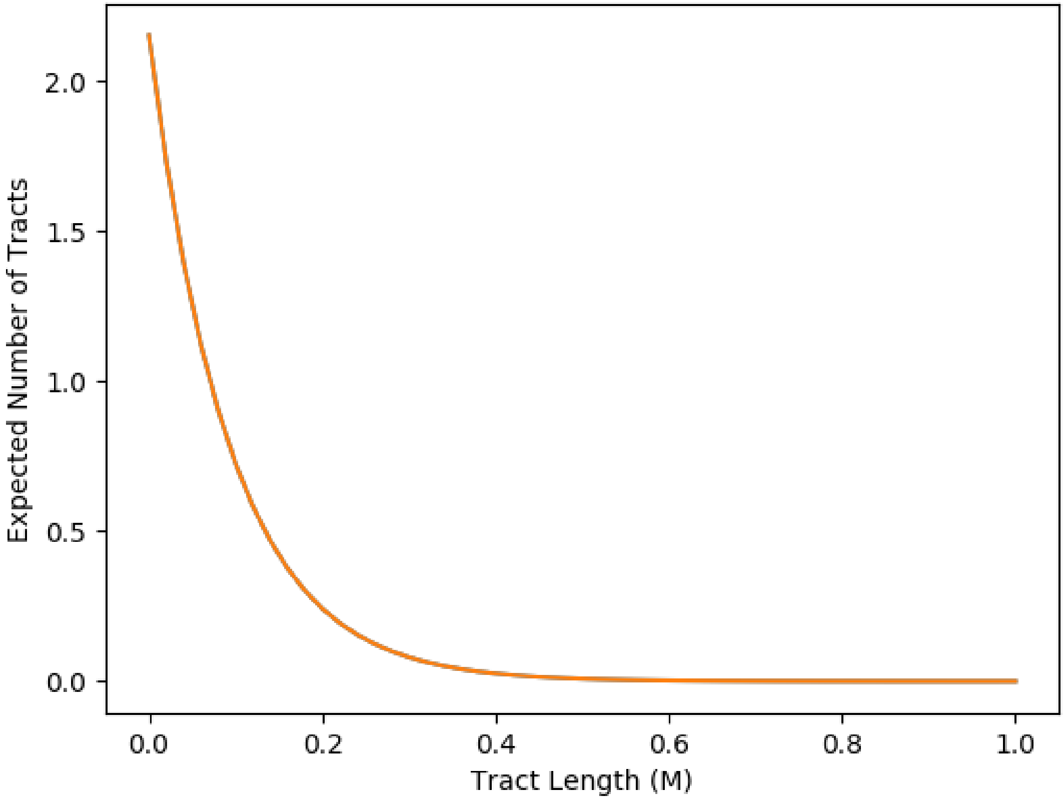
Appendix 1.

Exponential distribution versus *msprime* simulation model plotted on a log scale. Filled histogram represents an *msprime* simulation with chromosome length of 1e8 base pairs (1 Morgan) with parameters *Ne* = 1e6, *n* = 1000, *t* = 20, *r* = 1e-8, length = 1e8 (*Ne* = effective population size, *n* = sample size of the admixed population, *t* = number of generation the admixed population existed, *r* = recombination rate, length = length in base pairs). Each bar in the histogram is one of 50 bins. Red line represents the distribution of segment length under a simple exponential model modeled from Equation 1 with parameters *p* = 0.5, *t* = 19 (left) and *t* = 21 (right). (*p* = proportion of ancestry from a population; *t* = time of split). Setting generation time to 20 in the exponential model while keeping time as 20 in the *msprime* simulation appears to give the best fit suggesting the two models match well (Figure 3).



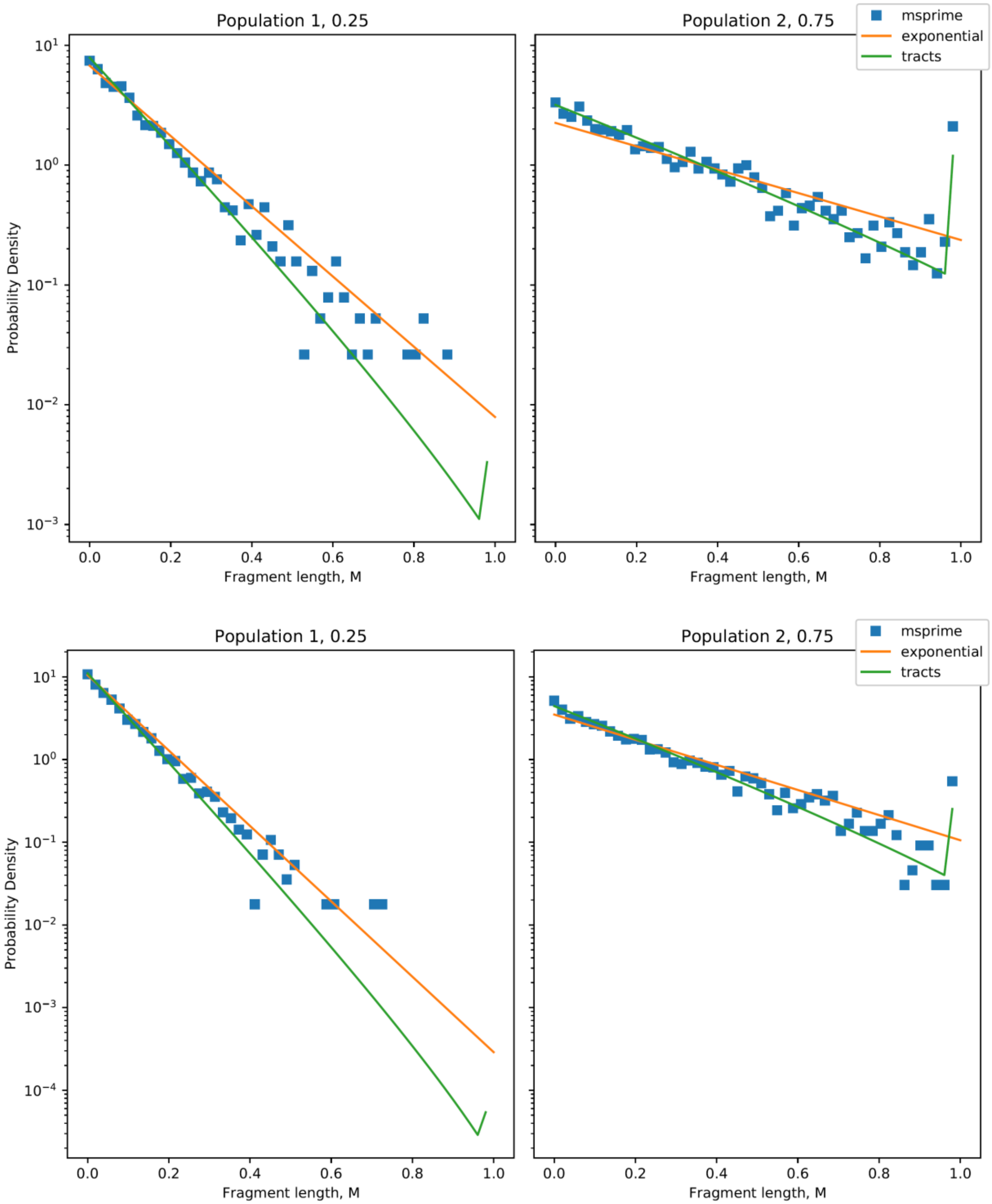
Appendix 2.

Expected number of tracts per bin of a single sample for 2 populations when simulated back in time *t*, for 20 generations with proportions *p*, of ancestry of 0.5 for both populations. (*p* = proportions of ancestry of the two source populations, *t* = time since the admixture in generations). Each data point represents a continuous segment along the genome whose length is contained in one of 50 bins. As expected, both populations have the same curve, so the orange curve masks the blue curve.



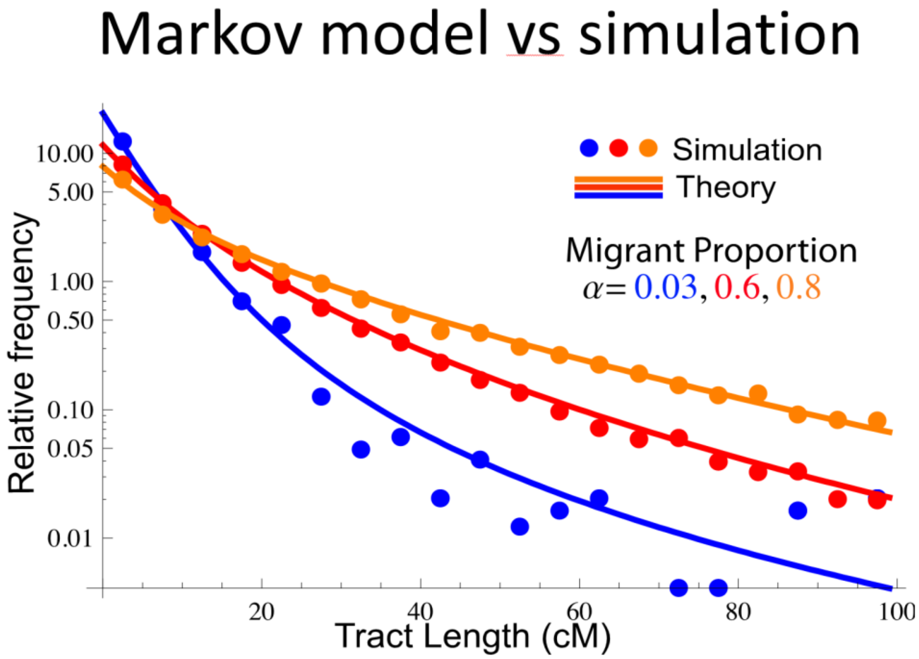
Appendix 3.

A comparison of the *msprime*, *tracts*, and exponential model when simulated back in time for 10 generations (Top) and 15 generations (Bottom) with initial ancestry migration proportions of 0.25 and 0.75 for each source population. Results are plotted on a log scale for probability density and each dot in the *msprime* model represents continuous fragments whose length is contained in one of 51 bins.



Appendix 4.

A modified version of Figure 4 from Gravel (2012)[4] comparing the *tracts* model, and Wright-Fisher simulation for distributions of tract length across different migrant proportions. Figure retrieved from https://github.com/sgravel/tracts/tree/master/docs.


Markov model vs simulation
● ● ● Simulation
Theory
Migrant Proportion
α= 0.03, 0.6, 0.8
Relative frequency
Tract Length (cM)