Rare variant method for identity-by-descent detection in sequence data, and the time to most recent common ancestor given identity-by-descent segment length

Steven Mason Foltz

A thesis submitted in partial fulfillment of the
requirements for the degree of

Master of Science

University of Washington

2013

Reading Committee:

Sharon Browning, Chair

Brian Browning, Chair

Program Authorized to Offer Degree:
Biostatistics

University of Washington

Abstract

Rare variant method for identity-by-descent detection in sequence data, and the time to most recent common ancestor given identity-by-descent segment length

Steven Mason Foltz

Co-Chairs of the Supervisory Committee:
Research Associate Professor Sharon Browning
Biostatistics

Associate Professor of Medicine Brian Browning
Medical Genetics

Identity-by-descent (IBD) is defined as two individuals sharing a haplotype they have both inherited from a common ancestor, where a haplotype is a segment of genetic material found on a single homologue. Existing methods for IBD detection were written to analyze low-density SNP arrays. These methods may not be well-suited to effectively report shared haplotype segments in sequence data because sequence data sets have an abundance of rare variants and an overall higher density of markers. We present a new method, RV-IBD, which uses the sharing of rare variants in sequence data to report IBD segments. Compared to existing methods for IBD detection, RV-IBD by itself did not show improved power and accuracy, but merging reported tracts from RV-IBD and the best-performing existing method did lead to improved power and accuracy. The RV-IBD method complements and is competitive with existing IBD detection methods using sequence data, showing that rare variants can be informative for IBD detection.

We investigate the distribution of time to most recent common ancestor given the length of a shared IBD segment. In populations of constant size, we find the distribution to be Gamma with a scale parameter dependent on the length variable. In exponentially growing populations, we discuss trends in the conditional distributions.

TABLE OF CONTENTS

Page

LIST OF FIGURES

LIST OF TABLES

Chapter 1

RARE VARIANT METHOD FOR IDENTITY-BY-DESCENT DETECTION IN SEQUENCE DATA

1.1 Background

Identity-by-descent (IBD) is defined as two individuals sharing a haplotype they have both inherited from a common ancestor, where a haplotype is a segment of genetic material found on a single homologue. More closely related pairs of individuals are expected to share a higher proportion of their genomes identical by descent. In general, for two individuals separated genealogically by m meioses, they share a haplotype identical by descent at a proportion 2^{1-m} of the genome, and the length of a shared haplotype has an approximately exponential distribution with mean $100 \cdot m^{-1}$ centiMorgans (cM). Since IBD segments are broken up by recombination events, the greater the number of meioses separating two individuals, the shorter the expected length of a shared segment is.

Several methods exist for the detection of IBD segments from population-level genetics data sets. There are three main approaches to IBD detection: length-based methods, probabilistic models that do not model linkage disequilibrium (LD), and probabilistic models that do model LD [8]. Input data consists of the genotype data of each sample at markers across the genome.

GERMLINE [12] utilizes the length-based approach. The program uses a hashing algorithm and a sliding window of markers to group samples in each window that have identical haplotypes. An IBD segment is reported if a pair of samples is grouped together in multiple adjacent windows such that the total length exceeds some minimum value. GERMLINE works best when presented with phased data and relies on other programs for this.

PLINK [22] uses a probabilistic model that does not model LD by implementing a hidden Markov model (HMM) based on the overall relatedness of samples and the local genotype identity-by-state (IBS) sharing of each pair. Genotypes are not phased before being given to PLINK, so only the number of shared alleles between samples is counted (0, 1, or 2). Two segments with a shared allele at each position are identical by state, but may not have come from a common ancestor. The longer an IBS segment is, the more likely that segment is actually an IBD segment. Short IBS segments are not indicative of IBD – only long (*e.g.* 1-2 cM) stretches of haplotype IBS should be used to infer IBD. PLINK requires markers to be approximately independent (*i.e.* in linkage equilibrium). Thus, only a thinned selection of markers is kept for use in the analysis, although a related method IBD_Haplo [3] was used without thinning markers with higher false-positive rates but also with increased power.

Unlike PLINK and IBD_Haplo, RELATE [2] and IBDLD [13] both incorporate LD by conditioning genotype probabilities on the IBD state and genotypes of neighboring SNPs. RELATE conditions only on the neighboring SNP in highest LD with the current marker, while IBDLD conditions on the neighboring twenty SNPs.

Like PLINK, Beagle [7] implements an HMM but utilizes LD to improve phasing accuracy. Beagle fastIBD [5] uses an approach similar to GERMLINE's hashing algorithm; instead of shared length it uses haplotype frequency to determine the statistical significance of reported segments. fastIBD is potentially more robust to haplotype phasing errors than GERMLINE because it samples haplotypes from multiple iterations of the phasing step. Refined IBD [6] is the latest program in the Beagle family. Like fastIBD, it first implements the GERMLINE algorithm to identify shared haplotypes of a minimum length and then evaluates the IBD probability of each candidate segment. One run of Refined IBD was found to be more powerful than ten runs of fastIBD.

Methods that only use pairs of individuals to report IBD may miss short IBD segments, while information gained from multiple pairwise comparisons may improve power. Methods for IBD detection between multiple individuals include MCMC_IBDfinder [17] (without LD

modeling, small number of individuals only) and ALADIN [1] (with limited LD modeling, given a pedigree).

SNP arrays and sequence data differ in the proportion of reported rare variants and the overall density of markers. In a study of 14,002 individuals, Nelson, *et al.* [18] found an average of 1 variant every 17 bases in sequence data, or 5.9×10^{-2} variants per base pair. Over 95% of variants reported were rare ($< 0.5\%$). Another study by Tennessen, *et al.* [24] with 2440 individuals reported 86% of variants in protein-coding genes to be rare using sequence data. With a larger sample size, one expects a higher rate of rare variants since there is a greater opportunity for novel mutations to be found. In contrast, SNP array data from the Wellcome Trust Case Control Consortium (WTCCC2) [11] had 2.5×10^{-5} variants per base pair, only 6.74% of which were rare (using the Affymetrix GeneChip 500K Mapping Array Set).

A previous study [23] examined the performance of GERMLINE and fastIBD on SNP array data and higher density sequence data. Such methods were written for use on SNP arrays with density up to 1 marker per 3 kb, but applying them to sequence data presents challenges. For instance, Beagle documentation [4] suggests removing markers with low minor allele frequency (MAF) because these markers are difficult to phase accurately. However, in sequence data, a majority of markers have low MAF. These markers would be ignored even though rare variants can be highly informative for IBD detection.

Rare variants can be highly informative for IBD detection because the same rare variant showing up at the same position in two individuals who are not identical by descent at that position is highly unlikely. This is not true for common variants because a common variant is more likely to be shared by chance than by common ancestry. Thus, the presence of the same rare variant at the same marker can be used as a clue that two individuals share a haplotype including that marker from a recent common ancestor.

1.2 Methods

The rare variant IBD detection method (RV-IBD) is a new, independent algorithm that uses the sharing of markers with low MAF in sequence data to report IBD segments. The algorithm is a simple, rule-based approach that mirrors the scientific logic of the underlying biology: when two individuals share a rare variant, they have most likely inherited it from a recent common ancestor. RV-IBD operates irrespective of phase and LD but could also be used in concert with existing IBD detection methods written for SNP array data.

1.2.1 Rare variant algorithm

The rare variant algorithm runs a comparison of all pairs of samples in search of IBD segments. The following description is generalized for multi-allelic data. For a given pair of samples ($s1$ and $s2$):

1. List the markers at which $s1$ and $s2$ share a rare allele. Suppose there are L such markers.

2. For the $L-1$ pairs of sequential markers in the list, report a segment between that pair of markers if

   a.) The distance between the markers is less than a specified length, and

   b.) The samples share an allele at all intervening markers.

3. Merge contiguous segments.

The first user-specified parameter is the maximum number of copies of a minor allele a marker may have to be defined as having a rare allele. The second user-specified parameter is the maximum length between two shared rare alleles for a segment to be defined between them.

1.2.2 Defining rare alleles

Excluding singletons, markers that have fewer than a user-specified number of copies of the minor allele are defined to be rare. Dividing the maximum user-specified number of copies by the number of haplotypes in the study gives the maximum MAF threshold. In practice, the lower bound of the threshold could be raised to help control for genotyping error. The maximum MAF for inclusion as a rare variant is up to the user and is a decision made by balancing the yield and accuracy of the parameter choice.

1.2.3 Maximum length between shared rare alleles

Two markers must be less than a user-specified distance apart in order for a candidate segment to be defined between them. The farther two sets of shared rare variants are apart, the more likely an intervening recombination has split the shared ancestral haplotype into separate segments. Distance here may be genetic distance (measured in centiMorgans) or physical distance (measured in base pairs).

1.2.4 Sharing alleles at all intervening markers

A candidate segment may be eliminated if at some marker between the endpoints of that segment the two samples do not share an allele. In dialleic data, this is known as discordant homozygosity and means that one sample is a major allele homozygote while the other sample is a minor allele homozygote. Every marker in the data set between the two endpoints is considered, whether it has a rare allele or not. Although a segment could represent true IBD and also have a homozygote discordant position within it due to mutations since the common ancestor (or due to genotype error), this step improves accuracy. (A current median estimate of the mutation rate in humans is 1.38×10^{-8} mutations per base pair, per generation [18].)

1.2.5 Defining segment endpoints

Two segments may be merged together if their endpoints coincide. For instance, a segment from marker $m1$ to marker $m2$ could be merged with a segment from marker $m2$ to marker

$m3$ to form a new segment from marker $m1$ to marker $m3$ (*i.e.* the union of the two segments). Even after merging, the true segment likely extends beyond what is reported because segment endpoints are defined only by the positions of the shared rare variants. However, ruled-based methods for extending segments may introduce unnecessary false positives. One anti-conservative approach is to extend segments in either direction until they no longer share an allele in common.

1.3 Results and discussion

1.3.1 Data sets

Five sequence data sets were simulated using the Markovian Coalescent Simulator (MaCS) [10] with 2000 individuals (4000 haplotypes) from a single population. The data was modeled to match the modern United Kingdom population using a variable-rate exponential growth model. True IBD segments were defined as two individuals sharing an identical by state haplotype with length 0.2 cM or greater. Markers with MAF $< .0025$ were not included when determining IBS because those rare alleles could be the result of recent mutations occurring after the most recent common ancestor. After simulation, phase information was ignored and genotype error was introduced by changing an allele at a rate of 0.1% in markers with MAF $\geq .0005$, and rate $2 \times$ MAF in markers with MAF $< .0005$.

1.3.2 IBD detection methods

In addition to RV-IBD, we applied the existing methods Beagle Refined IBD, GERMLINE, and PLINK for a comparison of their performance on sequence data. We selected an array of parameters for RV-IBD to demonstrate the relative effects of each setting, and we chose the parameters for the other three methods to balance yield and accuracy.

Excluding markers with MAF $< .02$, we used Beagle Refined IBD [6] version r1041 with default settings to phase each data set and report IBD segments. We ran GERMLINE [12] version 1.5.1 on default settings (except with the -haploid option and -min_m 2). GERMLINE received phased data from the prior run of Beagle Refined IBD and thus

was also run excluding markers with MAF $< .02$. We ran PLINK [22] version 1.07 excluding markers with MAF $< .0005$ and window parameters 776 SNPs per window, 194 SNPs overlapping between windows, and LD threshold $r^2 = 0.2$. (In these data sets there were about 776 markers with MAF $\geq .0005$ per 100 kb.)

After running each method, we ignored the phase indicators of the reported IBD and re-merged each set of segments. For methods that handle phased data, phase indicators denote which haplotype the IBD segments come from in each individual. We ignored the phased indicators to allow for a more direct comparison between the methods that report haplotype indicators (Beagle Refined IBD and GERMLINE) and those that do not (PLINK and RV-IBD). For another comparison, we also merged the segments reported by RV-IBD and Beagle Refined IBD to see the extent to which the reported segments were the same. To merge segments, we concatenated the two files, and if two samples shared more than one IBD segment and those segments overlapped, we reported the union of the overlapping segments. For all methods we removed any IBD segment of length less than 0.4 cM. We chose 0.4 cM instead of the minimum true IBD length 0.2 cM in order to avoid situations where overestimation led to a completely wrong reported segment because the true segment had length close to but less than 0.2 cM.

1.3.3 Comparison of methods

Figure 1.1 shows the relative performance of each method in true versus false discovery. A method demonstrates better performance when it appears above and to the left of another method. In Figure 1.1, Beagle Refined IBD outperforms each method, and RV-IBD performs slightly better than GERMLINE.

Figure 1.2 shows the relative performance of each method for power, accuracy, the amounts missed, and the amounts overestimated. For power, Beagle Refined IBD shows the best performance up to lengths of 3 cM, after which GERMLINE and PLINK do slightly better. For accuracy, RV-IBD and Beagle Refined IBD show top performance across all

sizes of reported segments, though all methods have similar measures on reported segments of length greater than 3 cM. RV-IBD misses the most true IBD and underestimates the least, while GERMLINE misses the least true IBD and overestimating the most. Combining RV-IBD and Beagle Refined IBD gives the best power and accuracy, while also reducing the amount missed for true IBD lengths above 3 cM. The merged set of segments shows increased power to detect true segments of length greater than 1.5 cM.

Figure 1.3 demonstrates how using different parameters with RV-IBD may give different levels of accuracy and yield. Holding the MAF parameter constant, setting the length threshold at 1 Mb or 2 Mb produces equally good power and false discovery rate (FDR). A higher MAF parameter gives higher power but higher FDR, too; setting MAF to be highest value of 0.02 gave an unreasonably high FDR. The settings that produced results most similar to Beagle Refined IBD were a length of 1 Mb or 2 Mb with a MAF threshold of 0.01. Merging Beagle Refined IBD with RV-IBD with length 1 Mb and MAF threshold 0.005 gave superior results to Beagle Refined IBD alone, indicating that a large portion of the segments reported by RV-IBD were distinct from those reported by Beagle Refined IBD.

Figure 1.4 shows a story similar to Figure 1.3 but without having added error to the data. Comparing the "no error added" case to the previously reported "error added" case, RV-IBD showed an increase in Power and decrease in FDR, while the Power and FDR of the other methods both increased. "RV plus Refined IBD" refers to the merged set of segments reported by RV 1Mb .005 and Beagle Refined IBD. Tables 1.1 and 1.2 report the performance of each method on a variety of accuracy and yield statistics including FDR and Power which are shown graphically in Figures 1.3 and 1.4.

Other accuracy measures include the false discovery rate due to completely wrong reported segments (FDR (cw)) and the proportion of reported segments that are completely wrong (CW). Other yield statistics are the mean proportion of markers included in one or more reported segments (Coverage) and the mean number of segments covering a marker (Depth). We also report the median length of reported segments (Median).

Moving from Table 1.1 (with error added) to Table 1.2 (no error added), all methods improved according to FDR (cw), CW, Power, Coverage, and Depth. Furthermore, all methods had longer median segment lengths. Finally, Beagle Refined IBD, GERMLINE, and PLINK had the same or increased FDR, while RV-IBD had lower FDR, which may suggest that as sequencing accuracy improves RV-IBD may have increased utility.

1.4 Conclusions

The RV-IBD method is a simple first step in IBD detection using sequence data. We have shown that it is comparable to existing methods in power and accuracy, and we have shown its value as a complementary algorithm in concert with existing or future methods.

One limitation of the RV-IBD method and the comparisons made in the previous section is that RV-IBD is phase-agnostic. It would be preferable to have the algorithm report what haplotype each segment exists on in each sample, but since the algorithm receives unphased data and has no phasing mechanism, it cannot report segments with phase indicators. However, this issue may resolve itself because sequencing technologies will soon be able to reliably determine phase, even that of rare variants [15]. The same RV-IBD framework would be applicable with the phase known data, with four comparisons (one for each pair of haplotypes) to make for each pair of samples instead of one.

A further limitation of the RV-IBD method is that no attempt has been made to accurately determine the true ends of reported segments. A more sophisticated method would model the appropriate length to extend each segment while minimizing false discovery. More accurately determined segment lengths would be useful from a population genetics viewpoint, as segment length can be used to indicate the number of generations back to a common ancestor. Also, in this study we have looked only at shared rare variants in a single population. It is not yet known whether RV-IBD as currently described could be applied to IBD analyses involving multiple distinct or admixed populations.

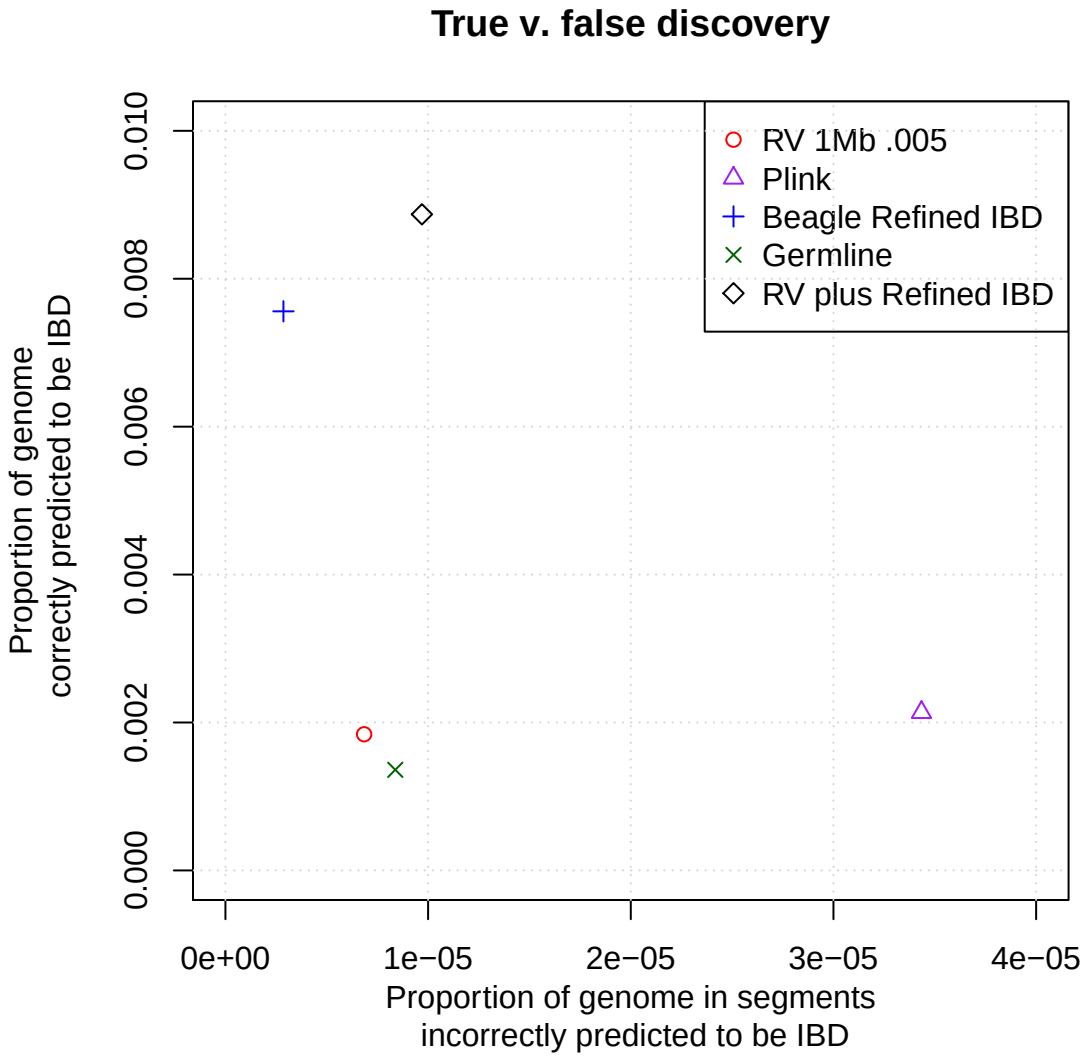

Figure 1.1: True versus false discovery of several IBD detection methods. "RV 1Mb .005" refers to RV-IBD using a length threshold of 1 Mb (converted to an average of 1.84 cM to fit this data) and MAF threshold of 0.005, meaning that only those markers with MAF less than or equal to 0.005 were used to report candidate segments. "RV plus Refined IBD" refers to the merged set of segments reported by RV 1 MB .005 and Beagle Refined IBD. On the x-axis, the proportion of the genome in segments incorrectly predicted to be IBD is the total length of all reported segments that have less than 25% overlap with true IBD divided by the maximum possible value of the numerator (the number of pairs of individuals times the length of the region). On the y-axis, the proportion is the total length of all reported IBD that is also true IBD, divided by the number of pairs of individuals times the length of the region.

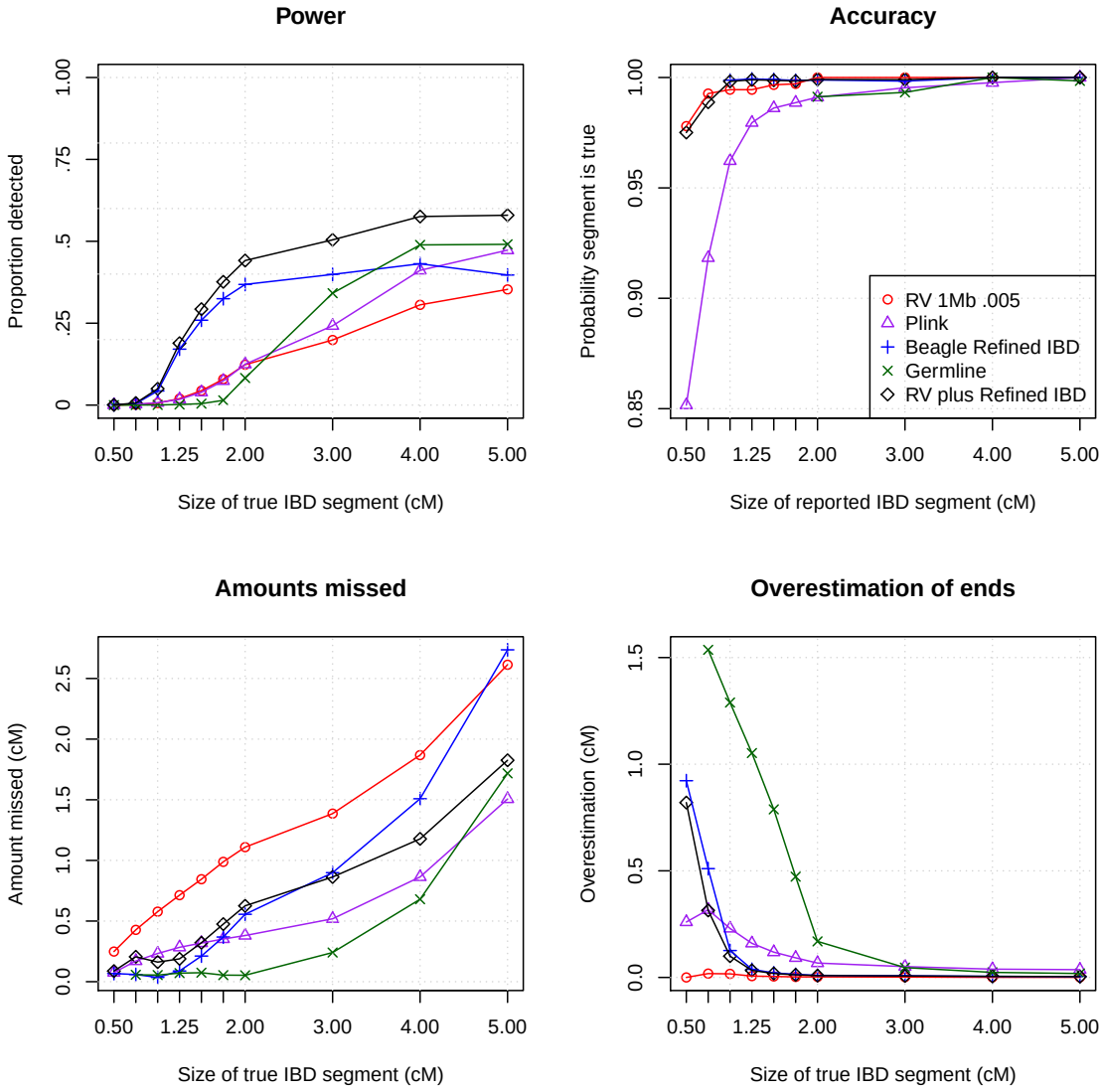

Figure 1.2: Power, accuracy, amount missed, and amount overestimated of several IBD detection methods. The common x-axis consists of bins of segments based on their length. The bins have cut points 0.625, 0.875, 1.125, 1.375, 1.625, 1.875, 2.125, 3.125, and 4.125 cM. The average measure for each bin is plotted at the points 0.5, 0.75, 1.0, 1.25, 1.5, 1.75, 2.0, 3.0, 4.0, and 5.0 cM. For the plots of power, amount missed, and overestimation, the x-axis lengths refer to true segments. For the plot of accuracy, the segment lengths refer to reported segments. Power is reported as the average proportion of a true segment that is detected, including all the true segments that were not reported. Accuracy is given as the probability that at least 50% of a reported segment is true. Amounts missed is the average amount of the true segment that was not reported, given that some part of a true segment was reported. On the other hand, overestimation of the ends reports the average length that reported segments extend beyond the end of the true segment, given that some part of a true segment was reported. "RV plus Refined IBD" refers to the merged set of segments reported by RV 1 MB .005 and Beagle Refined IBD.

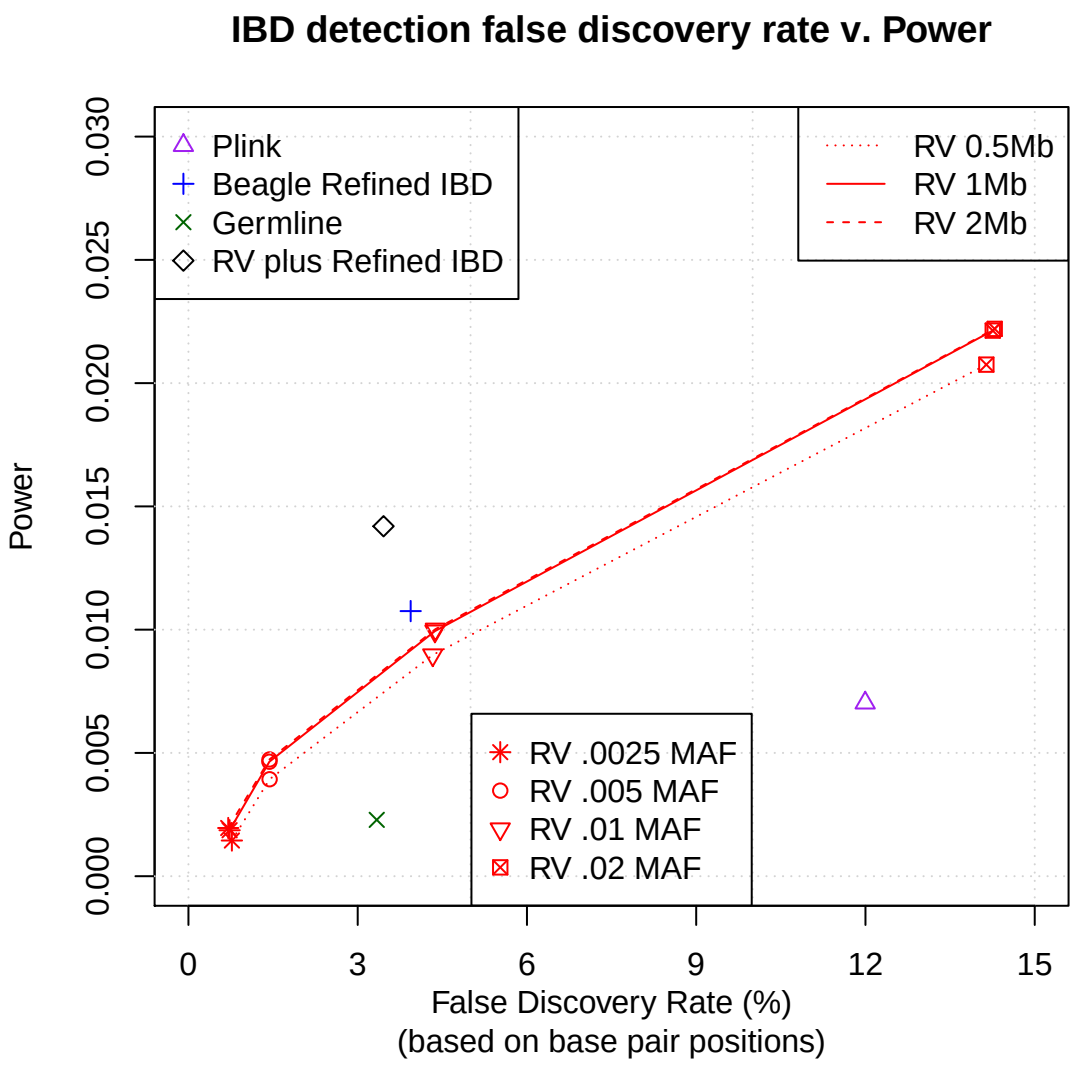

Figure 1.3: Power to detect IBD versus false discovery rate using different RV-IBD parameter settings. In this figure and Figure 1.4, the x-axis False Discovery Rate (%) (FDR) is based on base pairs, not genetic distance. FDR is computed for each method by merging the reported and true segments, then taking the difference in the total length of the merged and true data sets divided by the total length of the reported segments. Here power is the total length of reported segments multiplied by (1-FDR), divided by the total length of true segments. For RV-IBD, the different line types refer to different length parameters, and the different icons refer to different MAF levels. "RV plus Refined IBD" refers to the merged set of segments reported by RV 1 MB .005 and Beagle Refined IBD.

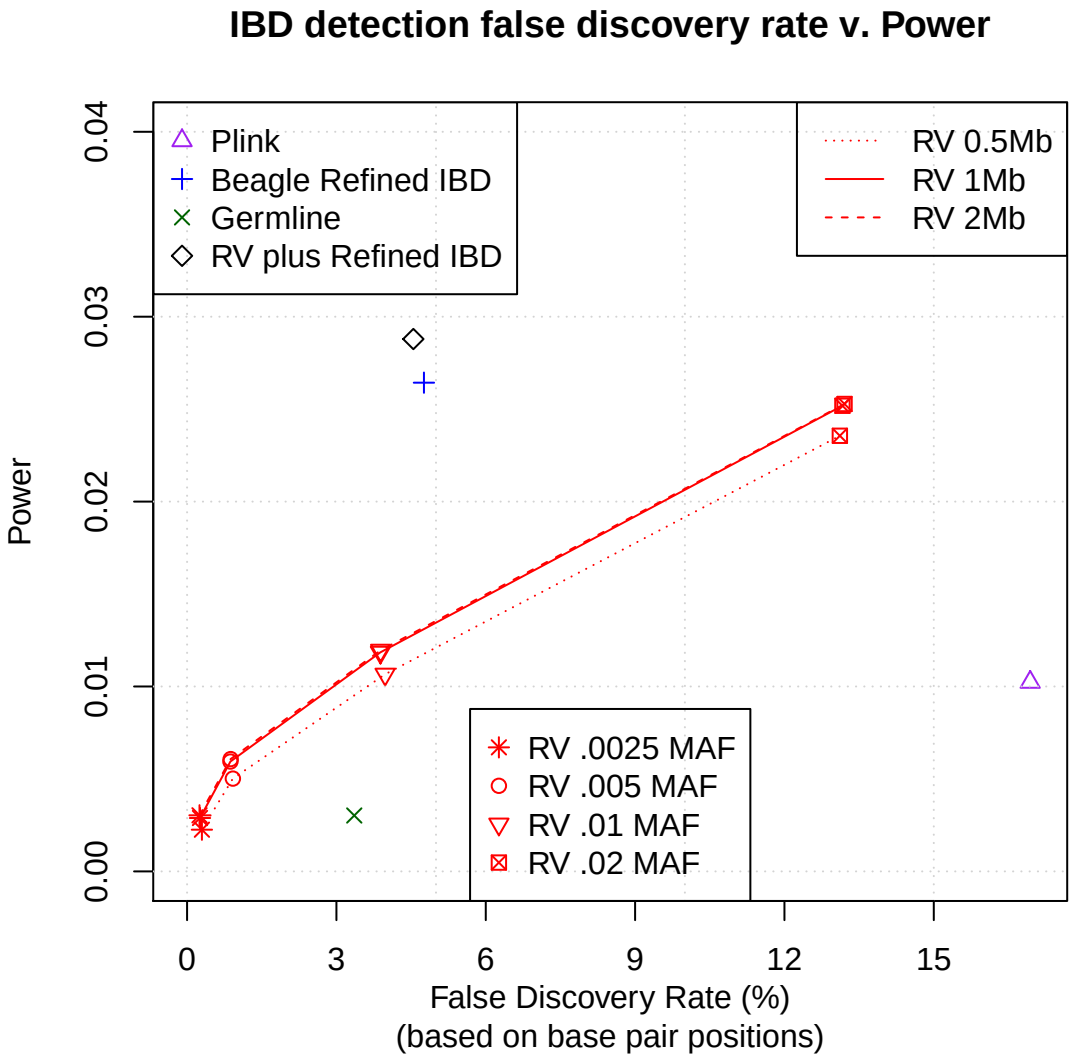

Figure 1.4: Power to detect IBD versus false discovery rate using different RV-IBD parameter settings, no error added. Refer to Figure 1.3 for an explanation of the measures reported here. Comparing the "no error added" case to the previously reported "error added" case, RV-IBD showed an increase in Power and decrease in FDR, while the Power and FDR of the other methods both increased. "RV plus Refined IBD" refers to the merged set of segments reported by RV 1 MB .005 and Beagle Refined IBD.

RV Length (Mb)	Max MAF (%)	FDR (%)	FDR (cw) (%)	CW (%)	Power	Coverage	Depth	Median (Mb)
0.5	0.25	0.77	0.47	0.68	0.001	0.21	0.41	0.41
	0.5	1.44	0.78	1.17	0.004	0.35	1.12	0.40
	1	4.33	2.08	2.99	0.009	0.50	2.64	0.39
	2	14.14	7.25	9.34	0.021	0.69	6.80	0.37
1.0	0.25	0.73	0.36	0.61	0.002	0.28	0.53	0.46
	0.5	1.43	0.67	1.09	0.005	0.45	1.33	0.43
	1	4.37	1.88	2.87	0.010	0.60	2.92	0.40
	2	14.26	6.80	9.06	0.022	0.78	7.26	0.38
2.0	0.25	0.71	0.35	0.61	0.002	0.30	0.55	0.47
	0.5	1.44	0.66	1.09	0.005	0.46	1.35	0.44
	1	4.38	1.87	2.87	0.010	0.61	2.95	0.40
	2	14.29	6.77	9.05	0.022	0.78	7.28	0.38
Beagle Refined IBD		3.94	0.00	0.00	0.011	0.62	3.11	0.56
RV + Refined IBD		3.46	0.21	0.38	0.014	0.77	4.10	0.53
GERMLINE		3.34	0.07	0.08	0.002	0.35	0.67	1.29
PLINK		12.00	1.23	1.47	0.007	0.51	2.28	1.23

-For RV-IBD results, RV Length (Mb) is the maximum distance between shared rare variants for a candidate segment to be reported. Max MAF (%) is the MAF that defines a rare variant.

-FDR (%) is the false discovery rate, the proportion of each reported segment that is false positive.

-FDR (cw) (%) is the false discovery rate due to reported segments that are completely wrong.

-CW (%) is the proportion of reported segments that are completely wrong, *i.e.* do not overlap with any true segment.

-Power is the ratio of the total length of reported segments to the total length of true segments, multiplied by 1-FDR.

-Coverage is the mean proportion of markers included in one or more reported segments.

-Depth is the mean number of segments covering a marker.

-Median (Mb) is the median length of reported segments.

Table 1.1: Yield and accuracy statistics of several IBD detection methods, including different parameter settings of RV-IBD

RV Length (Mb)	Max MAF (%)	FDR (%)	FDR (cw) (%)	CW (%)	Power	Coverage	Depth	Median (Mb)
0.5	0.25	0.30	0.25	0.45	0.002	0.26	0.64	0.43
	0.5	0.92	0.58	0.96	0.005	0.38	1.43	0.42
	1	3.98	2.01	3.12	0.011	0.52	3.13	0.40
	2	13.11	6.64	8.94	0.024	0.69	7.65	0.38
1.0	0.25	0.26	0.20	0.41	0.003	0.36	0.82	0.48
	0.5	0.87	0.49	0.91	0.006	0.50	1.69	0.45
	1	3.88	1.81	3.00	0.012	0.63	3.48	0.42
	2	13.16	6.20	8.68	0.025	0.79	8.18	0.39
2.0	0.25	0.25	0.19	0.40	0.003	0.39	0.86	0.49
	0.5	0.88	0.48	0.90	0.006	0.52	1.73	0.46
	1	3.89	1.80	3.00	0.012	0.64	3.51	0.42
	2	13.21	6.18	8.67	0.025	0.80	8.22	0.39
Beagle Refined IBD		4.76	0.00	0.00	0.026	0.88	7.80	0.68
RV + Refined IBD		4.55	0.10	0.23	0.029	0.91	8.48	0.65
GERMLINE		3.36	0.06	0.06	0.003	0.43	0.88	1.33
PLINK		16.94	1.95	2.32	0.010	0.60	3.51	1.23
-See Table 1.1 for column heading definitions.								

Table 1.2: Yield and accuracy statistics of several IBD detection method, including different parameter settings of RV-IBD, no error added. Compared to the "with error added" case: a.) All methods improved according to FDR (cw), CW, Power, Coverage, and Depth. b.) All methods had longer median segment lengths. c.) Beagle Refined IBD, GERMLINE, and PLINK had the same or increased FDR, while RV-IBD had lower FDR, which may suggest that as sequencing accuracy improves RV-IBD may have increased utility.

Chapter 2

TIME TO MOST RECENT COMMON ANCESTOR GIVEN IDENTITY-BY-DESCENT SEGMENT LENGTH

In this chapter we derive the distribution of the time to most recent common ancestor given identity-by-descent segment length with a constant population size. We first describe the Wright-Fisher model which defines our population and then combine three known distributions using Bayes' Rule. We also describe results when the population size grows exponentially.

2.1 Wright-Fisher model

Writing in the 1920s and 30s, Sewall Wright, Ronald Fisher, and J.B.S. Haldane, followed later by Motoo Kimura, drove many of the early developments in population genetics. Their work focused on what would happen in a population under certain constraints, such as migration, selection, and drift. Without large data sets from actual genomes, these prospective models were the necessary object of study. The more recent availability of real and simulated population-wide genomic data has presented an opportunity to use information from the present day to study past demography. Demographic history shapes the genetic relationship between members of a population, so studying current genetic relationships in a population such as IBD sharing could give information about past demography. [14]

The Wright-Fisher model provides a framework for making probabilistic statements about a population's genealogical processes and can be formulated for a diploid or haploid population. From [14], common assumptions defining the model for a haploid population include:

1. Discrete, non-overlapping generations: each haplotype is the descendent of a haplotype

in the previous generation

2. Constant population size: the number of haplotypes in each generation is the same over time

3. Equal fitness and fecundity among individuals: each haplotype has the same ability to live and reproduce

4. No population structure: a haplotype has an equal chance of being the descendent of any haplotype in the previous generation

5. No recombination: reenforces the notion that the model consists of discrete units of genetic material (such as alleles) that remain fully intact over time, and we need to extend each unit according to a well-defined distribution in order to answer questions about lengths of IBD that are affected by recombination over time.

Such a population is highly idealized, but even so there are some interesting facts to examine. For a given haplotype in the current generation, the probability that a certain haplotype in the next generation is a descendent of the haplotype in the current generation is $1/N$, where N is the constant number of haplotypes in the population. Thus, repeating this Bernoulli trial N times for all the haplotypes in the next generation, the number of descendants of a given haplotype in the current generation is distributed Binomial(N,$1/N$), with expectation 1 (maintaining a constant population size) and variance $(N-1)/N$. The probability of a haplotype having zero descendants in the next generation is $(1-1/N)^N$, which for large N has the approximate value of $e^{-1} \approx .37$.

For two haplotypes in the current generation, the probability they are both descendants of the same haplotype in the previous generation is $1/N$, as

$$P(\text{Hap 1 and Hap 2 share same ancestor}) =$$

$$\sum_{i=1}^{N} P(\text{Hap 2 has ancestor } i | \text{Hap 1 has ancestor } i) P(\text{Hap 1 has ancestor } i) =$$

$$\sum_{i=1}^{N} P(\text{Hap 2 has ancestor } i)P(\text{Hap 1 has ancestor } i) =$$

$$\sum_{i=1}^{N} \frac{1}{N^2} = \frac{1}{N},$$

assuming that Hap 1 and Hap 2 are independent. The probability they are descendants of different haplotypes is thus $1-1/N$. In fact, the same probabilities of success and failure apply independently to any past number of generations until the most recent common ancestor.

Given a set of N haplotypes, we can trace back a possible lineage any number of generations. For each haplotype, randomly draw from the pool of haplotypes in the previous generation, with replacement. The haplotype drawn from the previous generation is the direct ancestor of the haplotype in the current generation. Repeating this process over several generations and sorting the lines of descent may result in a picture such as Figure 2.1.

Such a process has the memoryless property that the likelihood of an event occurring in the present state is independent of events occurring in previous states. That is, for a random variable X, $P(X > a+b|X \geq a) = P(X > b)$. For continuous processes, the exponential distribution has this memoryless property, and for discrete processes, the geometric distribution has this property. [9]

2.2 Time to most recent common ancestor given length of IBD with constant population size

An emerging question in population genetics is, given the length of an IBD segment shared between two individuals (L), what information does that give about the age of their most recent common ancestor (T)? That is, we are interested in the distribution of $T|L$. Bayes' Rule allows

$$f_{T|L}(t|l) = \frac{f_{L|T}(l|t)f_T(t)}{f_L(l)}.$$

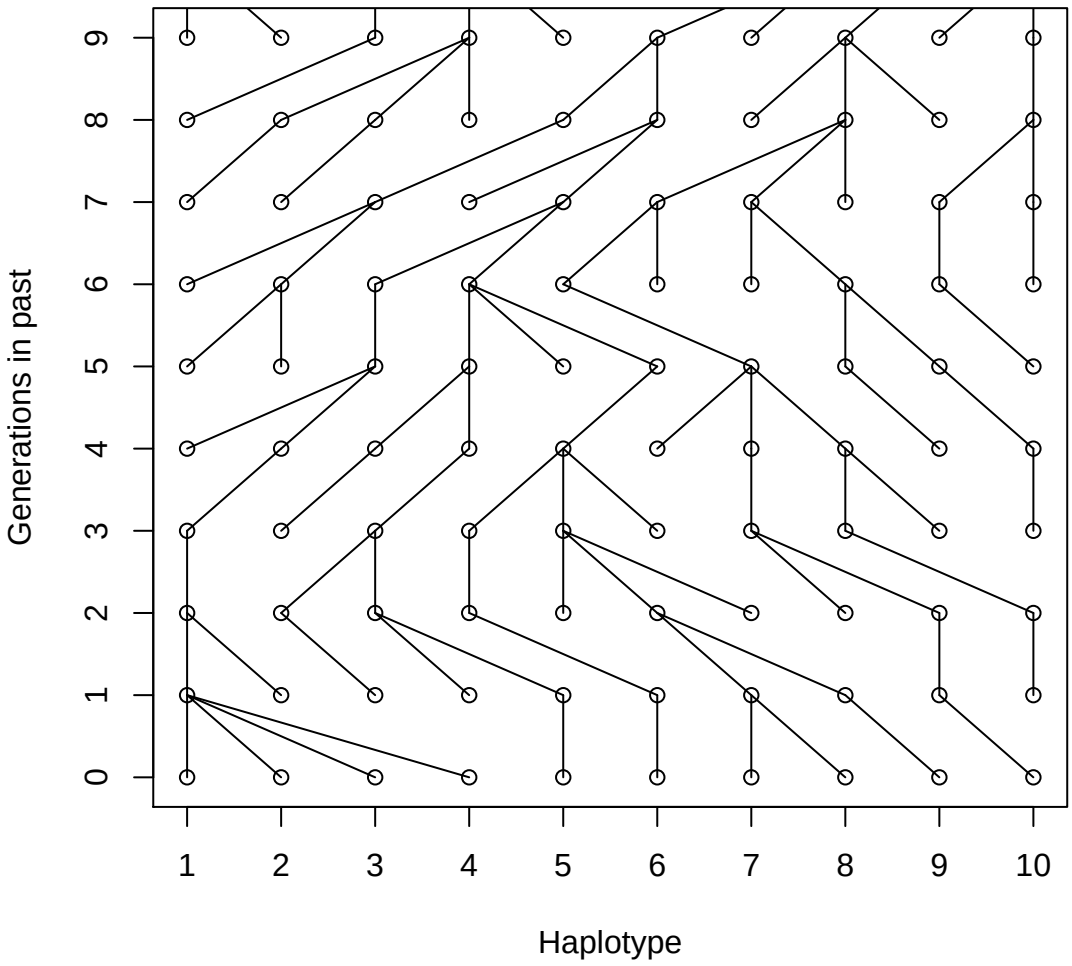

Figure 2.1: A sample genealogy from a Wright-Fisher model population with ten haplotypes. This sample genealogy from a Wright-Fisher population with ten haplotypes illustrates several assumptions of the model. For instance, each generation consists only of descendants of the previous generation, and the population size remains constant over time. In the construction of this plot, there was no fitness or reproductive advantage given to any haplotype, and each haplotype had an equal chance of being the descendent of any haplotype in the previous generation.

Finding an analytic expression for $f_{L|T}(l|t)$, $f_T(t)$, and $f_L(l)$ will allow the computation of $f_{T|L}(t|l)$.

2.2.1 Time to most recent common ancestor with constant population size

The number of generations, T, back to the most recent common ancestor of two haplotypes in a Wright-Fisher model with N haplotypes has a Geometric$(1/N)$ distribution. Random variables from a Geometric(p) distribution represent the number of trials necessary to achieve a success, where the probability of a success in each trial is p, and each trial is independent. Thus, probability mass function of T is

$$p(t) = (1 - 1/N)^{t-1}(1/N)$$

where the common ancestor is not found in generations 1 through $t-1$, but is found in generation t. The expected value of T is N, and the variance of T is $N(N-1)$, which for large N is approximately N^2.

As the time between trials becomes negligible (*i.e.* as we move from discrete to continuous time), the geometric distribution becomes the exponential distribution with the same probability of success. The moment generating function of a Geometric(p) random variable $X \geq 1$ is

$$\begin{aligned}
M_X(u) &= E(\exp(ux)) \\
&= \sum_{x=1}^{\infty} \exp(ux)(1-p)^{x-1}p \\
&= p\exp(u)\sum_{x=1}^{\infty} \exp(u(x-1))(1-p)^{x-1} \\
&= p\exp(u)\sum_{x=0}^{\infty} \exp(ux)(1-p)^{x} \\
&= p\exp(u)\sum_{x=0}^{\infty} (\exp(u)(1-p))^{x} \\
&= \frac{p\exp(u)}{1-\exp(u)(1-p)}
\end{aligned}$$

where $\sum_{x=0}^{\infty} (\exp(u)(1-p))^x = \frac{1}{1-\exp(u)(1-p)}$ when $\exp(u)(1-p) < 1$, or $u < -\log(1-p)$.

Now let $Y \sim$ Geometric(p) where $p = \lambda/n$ and $u = v/n$. The limit of the MGF is

$$\begin{aligned}
\lim_{n\to\infty} M_Y(u) &= \lim_{n\to\infty} \frac{(\lambda/n)\exp(v/n)}{1-\exp(v/n)(1-(\lambda/n))} \\
&= \lim_{n\to\infty} \frac{\lambda\exp(v/n)}{n(1-\exp(v/n)) + \lambda\exp(v/n)} \\
&= \frac{\lambda}{\lambda - v} \\
&= \frac{p}{p-u}
\end{aligned}$$

which is the MGF of an Exponential(p) random variable. Hence, we may think of the continuous distribution of T as an Exponential(rate=$1/N$) random variable. Like discrete T, the mean of continuous T is N with variance N^2. The probability density function of continuous T is

$$f_T(t) = \frac{1}{N}\exp\left(-\frac{t}{N}\right).$$

2.2.2 Length of IBD given time to most recent common ancestor

For two individuals from a population with any generalized demographic history, knowledge of the time to the most recent common ancestor at a certain point in the genome determines the distribution of the length of IBD segments shared between the two individuals. Recombination events down each path from the most recent common ancestor have likely reduced the length of the shared haplotype. Suppose two individuals share a haplotype identical by descent at a certain point from a common ancestor t generations ago. Then the distance either up or downstream from that point to the end of the currently shared haplotype has the same distribution as the minimum of $R_1, \ldots, R_{2t}$, where t is the number of generations back to the most recent common ancestor, and $R_i \sim$ Exponential(1/100).

Thus, the total length of the currently shared haplotype is distributed as the sum of two random variables each defined as $\min(R_1, \ldots, R_{2t})$. The R random variables are indepen-

dent since recombination events in different generations are independent, and each R_i has an Exponential(1/100) distribution because recombinations happen at a rate of 1 per 100 centiMorgans (cM). We are interested in the minimum of the R_i because it is the minimum that determines how much of the original haplotype remains intact. There are $2t$ random variables to consider because recombinations that would shorten the length of a shared segment may happen down either side of the genealogy connecting the two individuals with their most recent common ancestor t generations ago. We take the sum of two such minima to get the total distance up and downstream from the original point in the currently shared haplotype.

Thus, we have that the distribution of the length of an IBD segment given the time to the most recent common ancestor has distribution

$$L|T \sim \min(R_1, \ldots, R_{2t}) + \min(R_1, \ldots, R_{2t})$$

where each $R_i \sim$ Exponential(1/100). In general, the minimum of multiple exponential random variables has an exponential distribution with rate equal to the sum of the individual rates. [9] Therefore,

$$\min(R_1, \ldots, R_{2t}) \sim \text{Exponential}\left(\frac{2t}{100} = \frac{t}{50}\right).$$

The sum of identically distributed exponential random variables has a gamma distribution with shape parameter equal the number of terms and rate parameter equal to the rate of the exponentials. Hence,

$$L|T \sim \text{Gamma}\left(2, \text{rate} = \frac{t}{50}\right) \text{ or } \text{Gamma}\left(2, \text{scale} = \frac{50}{t}\right) \text{ and}$$

$$f_{L|T}(l|t) = \frac{1}{\Gamma(2)\left(\frac{50}{t}\right)^2} l \exp\left(-l\frac{t}{50}\right).$$

See [20] for an alternate explanation.

2.2.3 Length of IBD with constant population size

The joint probability density function of the time to most recent common ancestor and the length of IBD with constant population size is a product of the conditional distribution

of $L|T$ and the marginal distribution of T. If we assume without proof that $f(t,l) = f_{L|T}(l|t)f_T(t)$, then the marginal distribution of the length of IBD, conditional on being IBD at a certain point in the genome, with constant haploid population size N is

$$\begin{aligned} f_L(l) &= \int_0^\infty f_{L|T}(l|t)f_T(t)dt \\ &= \int_0^\infty \frac{1}{\Gamma(2)\left(\frac{50}{t}\right)^2} l \exp\left(-l\frac{t}{50}\right) \cdot \frac{1}{N}\exp\left(-\frac{t}{N}\right) dt \\ &= \frac{l}{50^2 N}\Gamma(3)\left(\frac{50N}{Nl+50}\right)^3 \int_0^\infty \frac{1}{\Gamma(3)\left(\frac{50N}{Nl+50}\right)^3} t^2 \exp\left(-t\left(\frac{Nl+50}{50N}\right)\right) dt \ (*) \\ &= \frac{l}{50^2 N}\Gamma(3)\left(\frac{50N}{Nl+50}\right)^3 \end{aligned}$$

The integrand in $(*)$ above is a gamma probability density function and integrates to one. $f_L(l)$ integrates to one over its range 0 to ∞.

To motivate our assumption that $f(t,l) = f_{L|T}(l|t)f_T(t)$, we can show that the relative difference between $f_L(l)$ and

$$\sum_{t=1}^{t_u} f_{L|T}(l|t)p_T(t) = \sum_{t=1}^{t_u} \frac{1}{\Gamma(2)\left(\frac{50}{t}\right)^2} l \exp\left(-l\frac{t}{50}\right) \cdot \frac{1}{N}\left(1-\frac{1}{N}\right)^{t-1}$$

goes to zero as t_u goes to infinity. Table 2.1 gives the relative difference for several different haploid effective population sizes N and an array of values for the length, l, and upper bound on the time to most recent common ancestor, t_u. Since the relative difference between the discrete and continuous form of $f_L(l)$ appear to be going to zero as t_u goes to infinity, we assume without further proof that using the continuous form is justified.

N=1000	t_u=10	t_u=100	t_u=1000	t_u=10000	t_u=1e+05	t_u=1e+06
L=0.01	1.00E+00	1.00E+00	8.79E-01	7.70E-04	2.50E-04	2.50E-04
L=0.1	1.00E+00	9.96E-01	4.22E-01	5.00E-04	5.00E-04	5.00E-04
L=1	9.98E-01	6.46E-01	9.29E-04	9.29E-04	9.29E-04	9.29E-04
L=10	6.46E-01	9.86E-04	9.87E-04	9.87E-04	9.87E-04	9.87E-04
N=10000	t_u=10	t_u=100	t_u=1000	t_u=10000	t_u=1e+05	t_u=1e+06
L=0.01	1.00E+00	1.00E+00	9.96E-01	4.23E-01	5.00E-05	5.00E-05
L=0.1	1.00E+00	9.99E-01	6.49E-01	9.27E-05	9.29E-05	9.29E-05
L=1	9.99E-01	6.71E-01	9.89E-05	9.93E-05	9.93E-05	9.93E-05
L=10	6.49E-01	9.29E-05	9.33E-05	9.33E-05	9.33E-05	9.33E-05
N=1e+05	t_u=10	t_u=100	t_u=1000	t_u=10000	t_u=1e+05	t_u=1e+06
L=0.01	1.00E+00	1.00E+00	9.99E-01	5.61E+00	9.10E-06	9.29E-06
L=0.1	1.00E+00	9.99E-01	6.74E-01	9.37E-07	9.93E-06	9.93E-06
L=1	9.99E-01	6.74E-01	9.54E-06	9.98E-09	9.99E-06	9.99E-06
L=10	6.50E-01	2.94E-06	3.35E-06	3.35E-11	3.35E-06	3.35E-06

Table 2.1: Numerical examination of the relative difference between the discrete and continuous forms of $f_L(l)$ for several values of population size.

2.2.4 Distribution of time to most recent common ancestor given length of IBD with constant population size

With the analytic expressions of $f_{L|T}(l|t)$, $f_T(t)$, and $f_L(l)$, we can solve for the probability distribution function of $T|L$ with constant haploid population size N using Bayes' Rule.

$$\begin{aligned} f_{T|L}(t|l) &= \frac{f_{L|T}(l|t)f_T(t)}{f_L(l)} \\ &= \frac{\frac{1}{\Gamma(2)\left(\frac{50}{t}\right)^2} l \exp\left(-l\frac{t}{50}\right) \cdot \frac{1}{N}\exp\left(-\frac{t}{N}\right)}{\frac{l}{50^2 N}\Gamma(3)\left(\frac{50N}{Nl+50}\right)^3} \\ &= \frac{1}{\Gamma(3)\left(\frac{50N}{Nl+50}\right)^3} t^2 \exp\left(-t\left(\frac{Nl+50}{50N}\right)\right) \end{aligned}$$

Therefore,

$$T|L \sim \text{ Gamma}\left(3, scale = \frac{50N}{Nl+50}\right)$$

with

$$E(T|L) = 3\left(\frac{50N}{Nl+50}\right) \text{ and } Var(T|L) = 3\left(\frac{50N}{Nl+50}\right)^2.$$

As the given length of an IBD segment increases, both the expected value and variance of the time to the most recent common ancestor decreases.

2.2.5 Evaluation of theoretical distribution function using simulated data

We simulated 10 cM regions of sequence data from 1000 populations of constant size (10,000 diploids, 20,000 haploids) using MaCS. [10] After simulation, we sampled 100 haplotypes from each population and found the time to most recent common ancestor along the whole sequence for each pair. We defined IBD segments to be where a pair of sampled haplotypes had the same time to most recent common ancestor over a stretch of 0.1 cM or greater. Segment endpoints were reported along with the age of the most recent common ancestor for that segment. Using this information, we plotted the corresponding density function of the time to most recent common ancestor given the length of an IBD segment to see how well the theoretical distribution derived in Section 2.2.4 matches the simulated data.

In Figures 2.2 and 2.3, the different colored lines show the empirical densities of the actual data, stratified by segment length. The empirical densities were estimated using the density() function in R. The dotted lines each show a Gamma probability density function using an approximation of the maximum likelihood estimators for the shape and scale parameters. The solid black lines show the probability density function of the Gamma distribution derived in Section 2.2.4 using the bounds of each segment length bin for the length. Both figures refer to the same data, but Figure 2.2 focuses on the the range of generations from zero to two-thousand while Figure 2.3 focuses on generations zero to five-hundred.

2.2.6 Approximation of Gamma MLEs

The maximum likelihood estimators of the Gamma(α,scale=β) distribution do not have a closed form solution when both parameters are unknown. [9] We can, however, find an approximation for $\hat{\alpha}$ accurate on average to within 1.1% of actual α values in our range of interest and use that to find $\hat{\beta}$.

Determine the log-likelihood function.

$$\begin{aligned} L(\alpha,\beta|\vec{x}) &= \prod_{i=1}^{n} \frac{1}{\Gamma(\alpha)\beta^{\alpha}} x_i^{\alpha-1} \exp\left(-\frac{x_i}{\beta}\right) \\ &= \left(\frac{1}{\Gamma(\alpha)\beta^{\alpha}}\right)^{n} \left[\prod_{i=1}^{n} x_i\right]^{\alpha-1} \exp\left(-\frac{1}{\beta}\sum_{i=1}^{n} x_i\right) \\ \log L(\alpha,\beta|\vec{x}) &= -n\log(\Gamma(\alpha)) - n\alpha\log(\beta) + (\alpha-1)\sum_{i=1}^{n}\log(x_i) - \frac{1}{\beta}\sum_{i=1}^{n} x_i \end{aligned}$$

Maximize the log-likelihood first with respect to β, treating $\hat{\alpha}$ as known.

$$\begin{aligned} \frac{\delta \log L(\alpha,\beta|\vec{x})}{\delta\beta} = 0 &= \frac{-n\alpha}{\beta} + \frac{1}{\beta^2}\sum_{i=1}^{n} x_i \\ \frac{n\alpha}{\beta} &= \frac{1}{\beta^2}\sum_{i=1}^{n} x_i \\ \hat{\beta} &= \frac{1}{\hat{\alpha}}\overline{x} \end{aligned}$$

Next, maximize the log-likelihood with respect to α.

$$\frac{\delta \log L(\alpha, \beta|\vec{x})}{\delta\alpha} = 0 = -n\frac{\delta}{\delta\alpha}\log(\Gamma(\alpha)) - n\frac{\delta}{\delta\alpha}\alpha\log(\beta) + n\overline{\log(x)} + \frac{\delta}{\delta\alpha}\frac{1}{\beta}\sum_{i=1}^{n} x_i$$

$\frac{\delta}{\delta\alpha}\log(\Gamma(\alpha))$ is the digamma function $\Psi(\alpha) = \log(\alpha) - \frac{1}{2\alpha} - \frac{1}{12\alpha^2} + \mathcal{O}(\alpha^{-4})$ [19], and β may be replaced by $\frac{\overline{x}}{\alpha}$.

Now solve for α:

$$\begin{aligned}
0 &= -n\Psi(\alpha) - n\frac{\delta}{\delta\alpha}\alpha\log\left(\frac{\overline{x}}{\alpha}\right) + n\overline{\log(x)} + \frac{\delta}{\delta\alpha}\frac{\alpha}{\overline{x}}\sum_{i=1}^{n} x_i \\
0 &= -n\Psi(\alpha) - n\frac{\delta}{\delta\alpha}[\alpha\log(\overline{x}) - \alpha\log(\alpha)] + n\overline{\log(x)} - n \\
0 &= -n\left(\Psi(\alpha) + \log(\overline{x}) - \frac{\alpha}{\alpha} - \log(\alpha) - \overline{\log(x)} + 1\right) \\
0 &= \Psi(\alpha) - \log(\alpha) + s \text{ where } s = \log(\overline{x}) - \overline{\log(x)} \\
0 &= \log(\alpha) - \frac{1}{2\alpha} - \frac{1}{12\alpha^2} + \mathcal{O}(\alpha^{-4}) - \log(\alpha) + s \\
0 &\approx -\left(\frac{6\alpha+1}{12\alpha^2}\right) + s \\
0 &\approx (12s)\alpha^2 - 6\alpha - 1 \\
\hat{\alpha} &\approx \frac{3+\sqrt{9+12s}}{12s}
\end{aligned}$$

We need $s = \log(\overline{x}) - \overline{\log(x)} > 0$ so that the estimate of $\hat{\alpha}$ makes sense (must be greater than zero) and so that the argument of the square root in the approximation can be evaluated nicely. The log function is convex downward, so Jensen's Inequality applies in the following form:

$$\phi(E(X)) \geq E(\phi(X))$$

where $\phi(X) = \log(X)$. So we have

$$\log(E(X)) \geq E(\log(X))$$

$$\log(E(X)) - E(\log(X)) > 0.$$

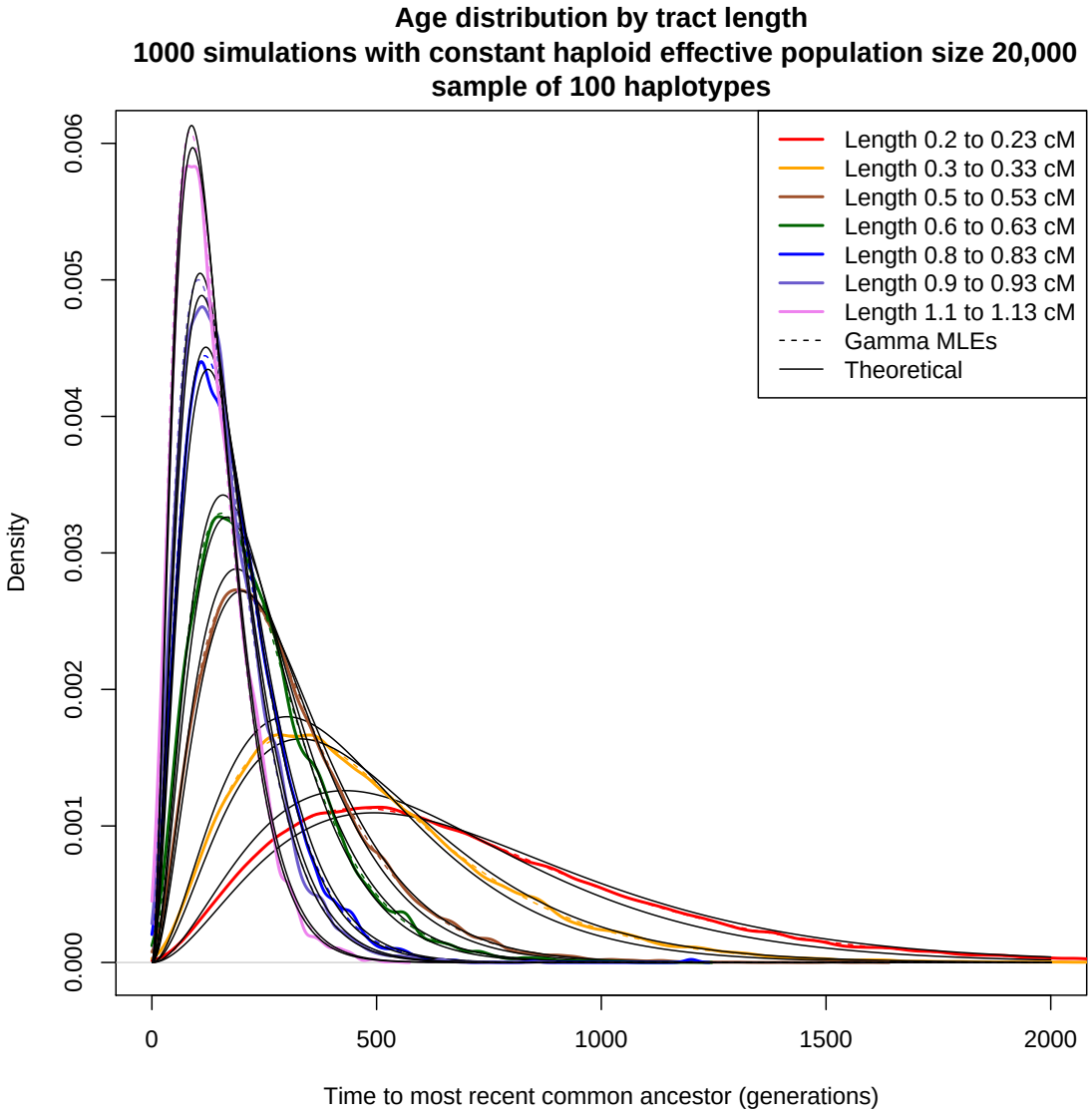

Figure 2.2: Time to most recent common ancestor given length of shared IBD segment, 0-2000 generations. The shorter length bins exhibit greater variance and expected time to most recent common ancestor. As the length of a given IBD segment increases, we see lower variance and expected time to most recent common ancestor. This makes sense because long segments tend to be broken up more (become shorter) as the number of meioses separating two individuals increases.

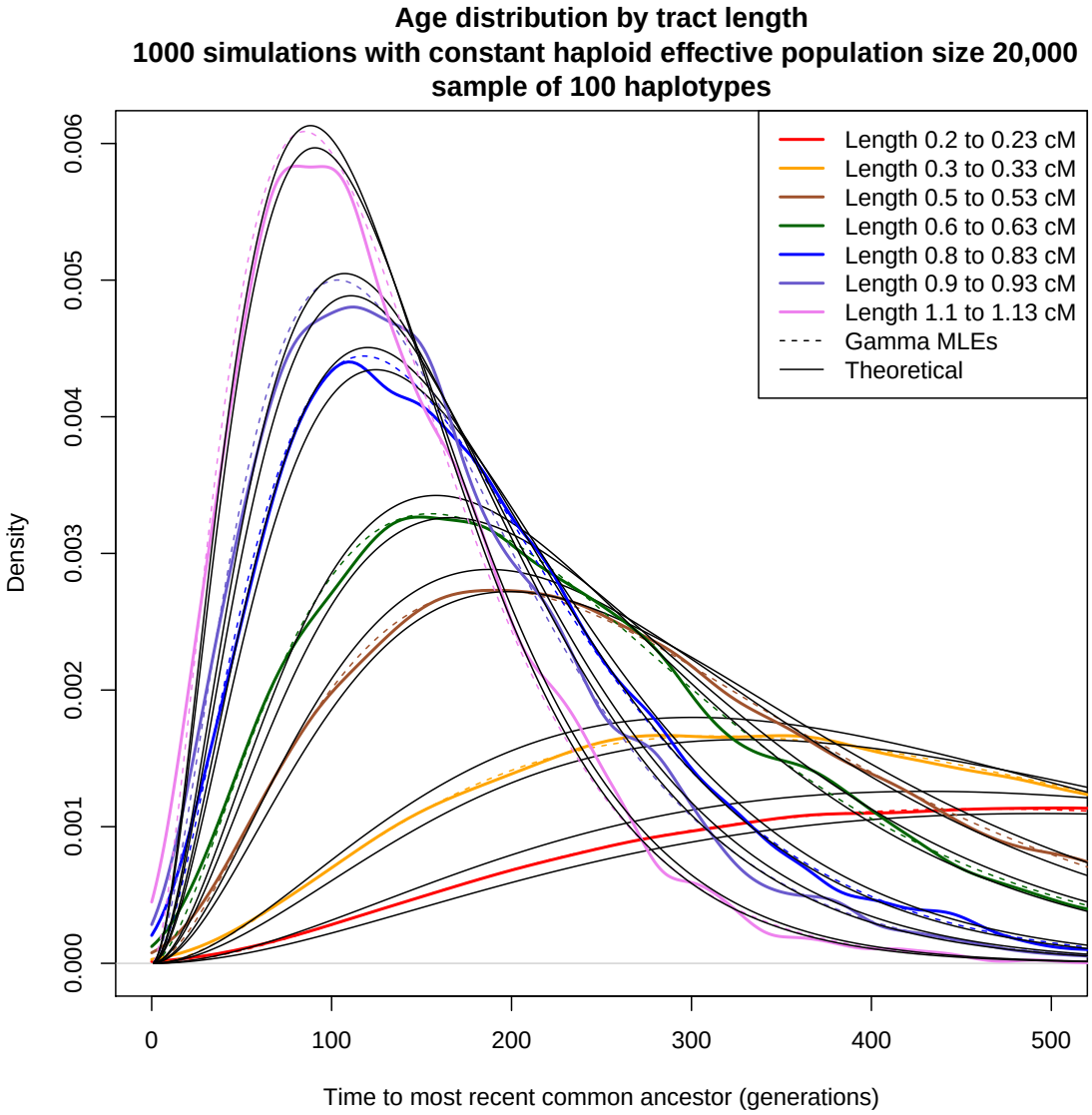

Figure 2.3: Time to most recent common ancestor given length of shared IBD segment, 0-500 generations. Focusing on the first five-hundred generations back, we can see where the theoretical and simulated results differ. Although the black theoretical lines defining each length bin do cross (they must cross if the area underneath each curve is one), the observed densities are bounded for the most part. Imperfect smoothness in the observed densities is due to random variation still present with 1000 simulations.

2.3 Time to most recent common ancestor given length of IBD with exponentially growing population size

To create exponentially growing populations with different growth rates, we used MaCS [10] to simulate 1000 populations each with the same historical and current population sizes, but with a different number of generations of growth separating the historical and current populations. We ran 1000 simulations for each number of generations of growth. The historical populations each had a constant diploid effective size of 5,000 individuals until their growth started, and the current populations each had a diploid effective size of 25,000 individuals. The numbers of generations of growth were 200, 500, 1000, and 2000. The growth rates in the four situations were approximately .008, .0032, .0016, and .0008, respectively. The rate may be calculated by solving the following equation for r:

$$N_{current} = N_{historical} \exp(rt_0)$$

where t_0 is the number of generations of growth. Solving for r gives

$$r = \frac{1}{t_0} \log\left(\frac{N_{current}}{N_{historical}}\right).$$

With an exponentially increasing population size, we no longer have the the distributions for T and L which we derived previously assuming a constant population size. Although the exponential growth scenario is easily parameterized and can be modeled knowing the exact population sizes each generation, a distribution function for the time to most recent common ancestor given length of shared IBD segment is does not appear to have a closed form solution. Figures 2.4-2.7 show the simulated empirical densities of T given L, with a separate graph for each length bin and a separate line on each graph for each different number of generations of growth.

In the following descriptions of Figures 2.4-2.7, length bins are referred to by their lower bound, and different curves are referred to by the corresponding times of growth. Separating curves in this way helps to illustrate some trends.

Figure 2.4: For lengths 0.2-0.3 cM, each density has very high variance and there is little differentiation between curves. For lengths 0.4-0.6 cM, a pattern between times 500, 1000, and 2000 becomes apparent. Time 200 is shifted up and to the right (higher mean, lower variance) of where it would be if it followed the pattern of the other curves. In general, for times 500, 1000, and 2000, the greater the number of generations of growth, the peak of the curve is higher and to the left (lower mean and variance).

Figure 2.5: For lengths 0.7-1.1 cM, the pattern between times 500, 1000, and 2000 continues, and the time 200 curve approaches its place in order below the time 500 curve as the length increases.

Figure 2.6: For lengths 1.2-1.6 cM, each curve tends toward a lower mean and variance as the length increases. Visually, the curves shift left and become more peaked.

Figure 2.7: For lengths 1.7-2.1 cM, the curves reach their most peaked state and look increasingly similar for larger lengths. This illustrates the main observable trend: the curves become more alike as the segment length increases. Since the populations are more similar the fewer generations we look into the past, and longer segments come from less-distantly-past most recent common ancestors, the longer the given length of shared segments is, the more alike the curves are likely to be since they are coming from more similar populations.

Another trend is that for a single length bin, populations with a greater number of generations of growth (*i.e.* slower growth rate) have curves that are more peaked. This means that in populations with slower growth rates, the usefulness of the information offered by segment length is greater as it leads to a distribution with lower variance.

Figures 2.8-2.11 show the distribution of time to most recent common ancestor given shared segment length, one plot for each growth rate. As expected, longer shared segments tend to come from less-distantly-past most recent common ancestors and the associated

densities have lower variance. The figures also illustrate further the trend that slower growth rates have density curves with smaller variance and means.

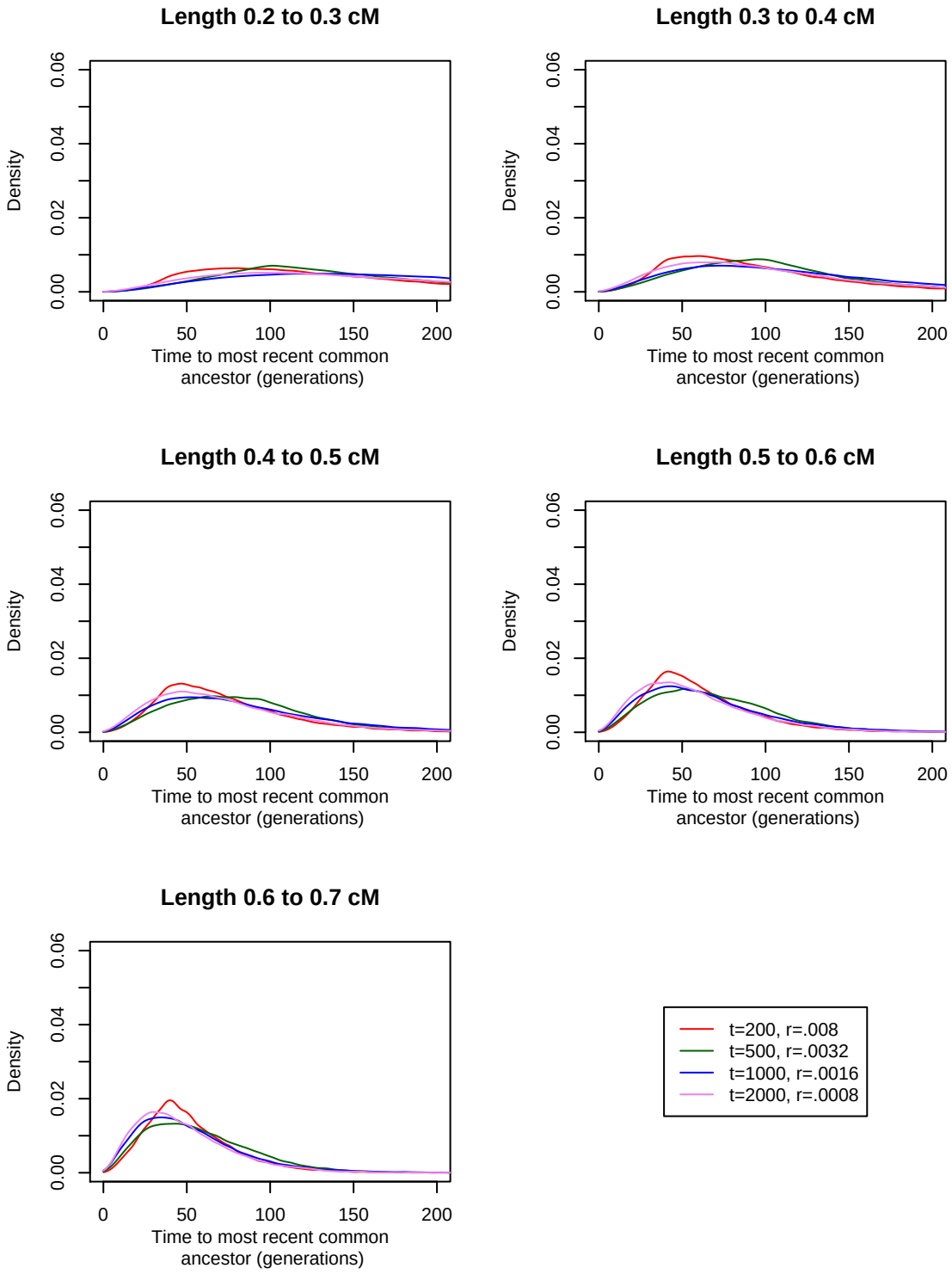

Figure 2.4: Time to most recent common ancestor given length of shared IBD segment, length 0.2-0.6 cM. t is the number of generations of growth from the ancestral population size (5,000) to the current population size (25,000). r is the growth rate over the generations of growth.

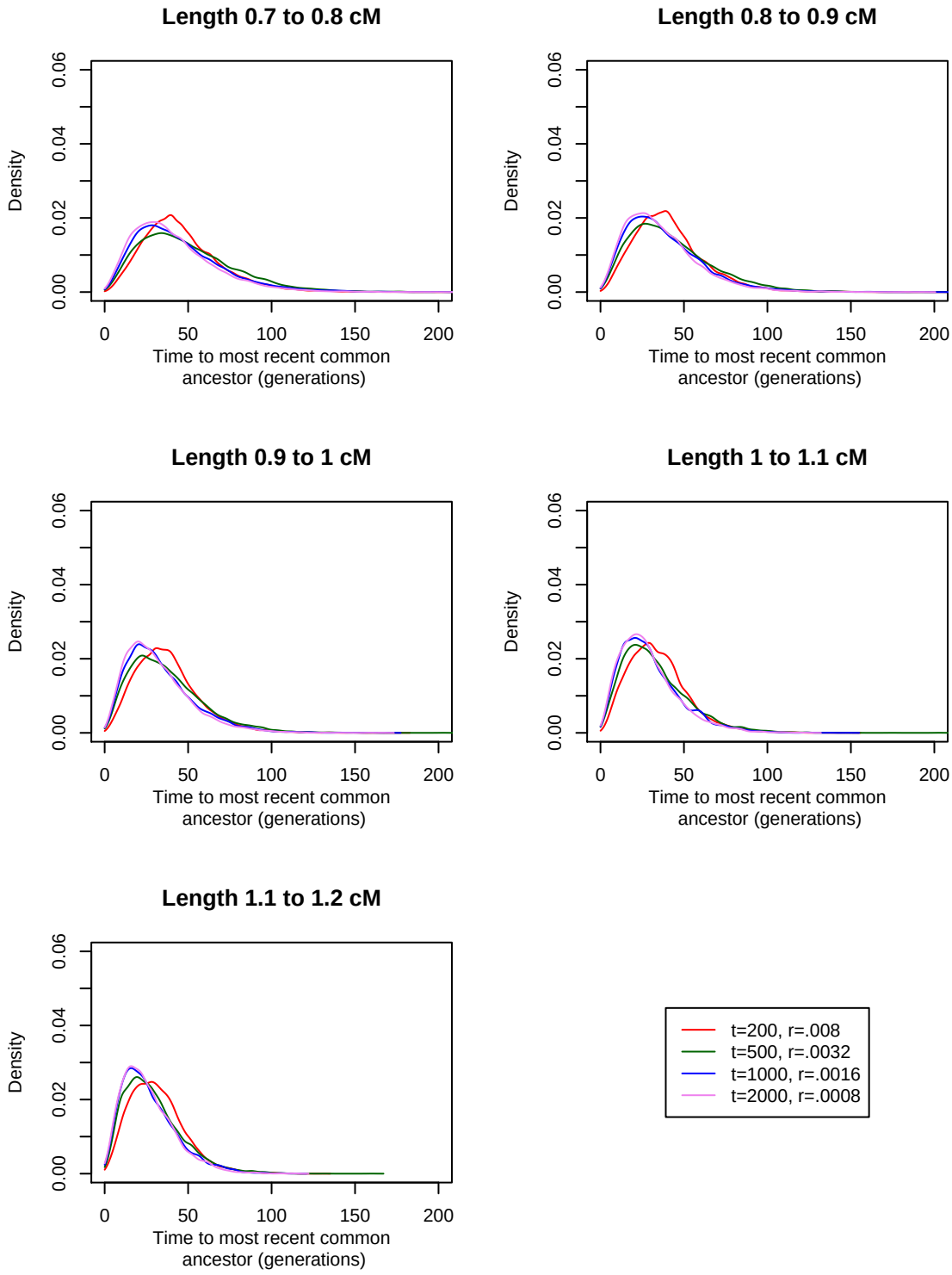

Figure 2.5: Time to most recent common ancestor given length of shared IBD segment, length 0.7-1.1 cM. t is the number of generations of growth from the ancestral population size (5,000) to the current population size (25,000). r is the growth rate over the generations of growth.

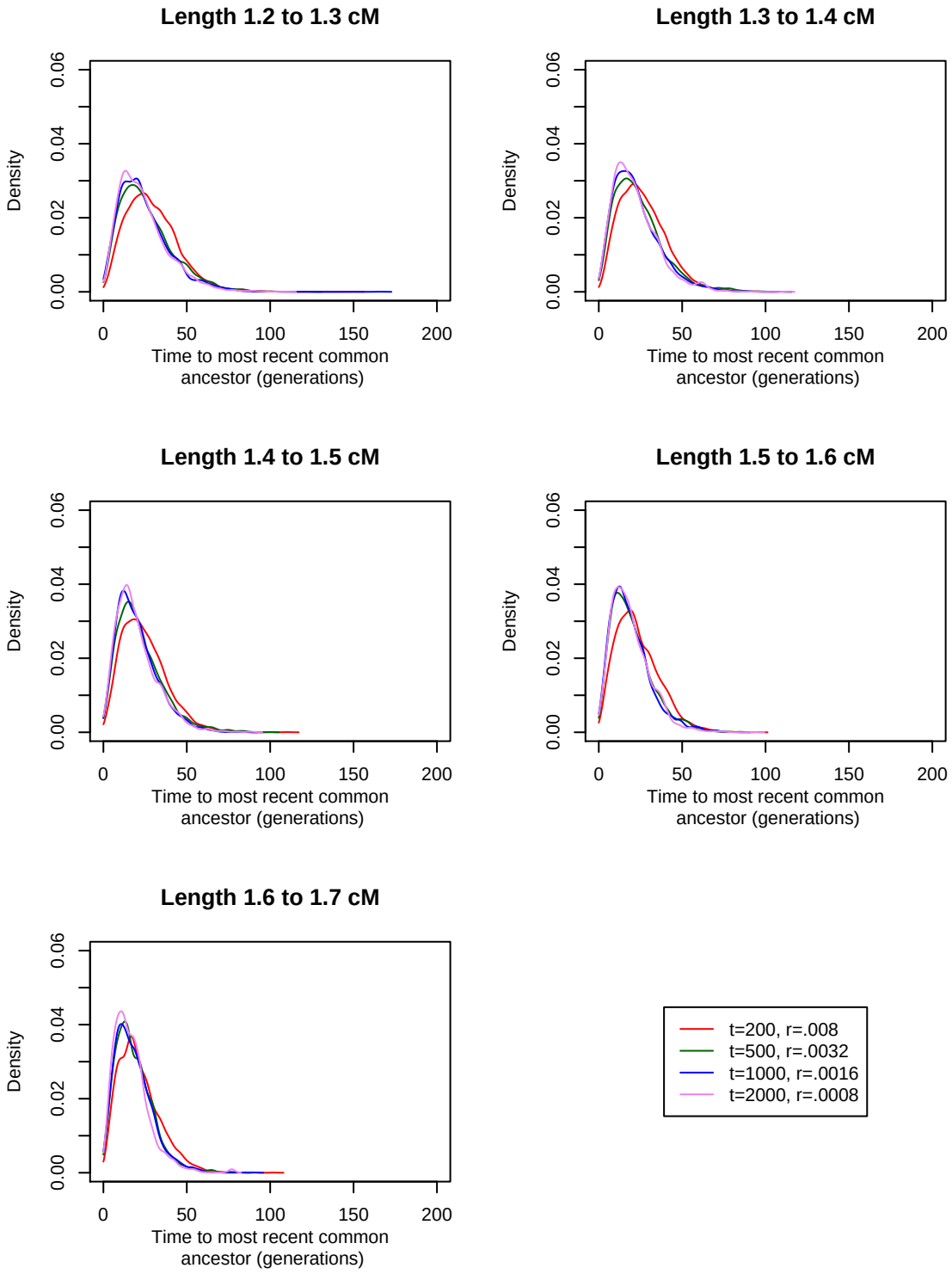

Figure 2.6: Time to most recent common ancestor given length of shared IBD segment, length 1.2-1.6 cM. t is the number of generations of growth from the ancestral population size (5,000) to the current population size (25,000). r is the growth rate over the generations of growth.

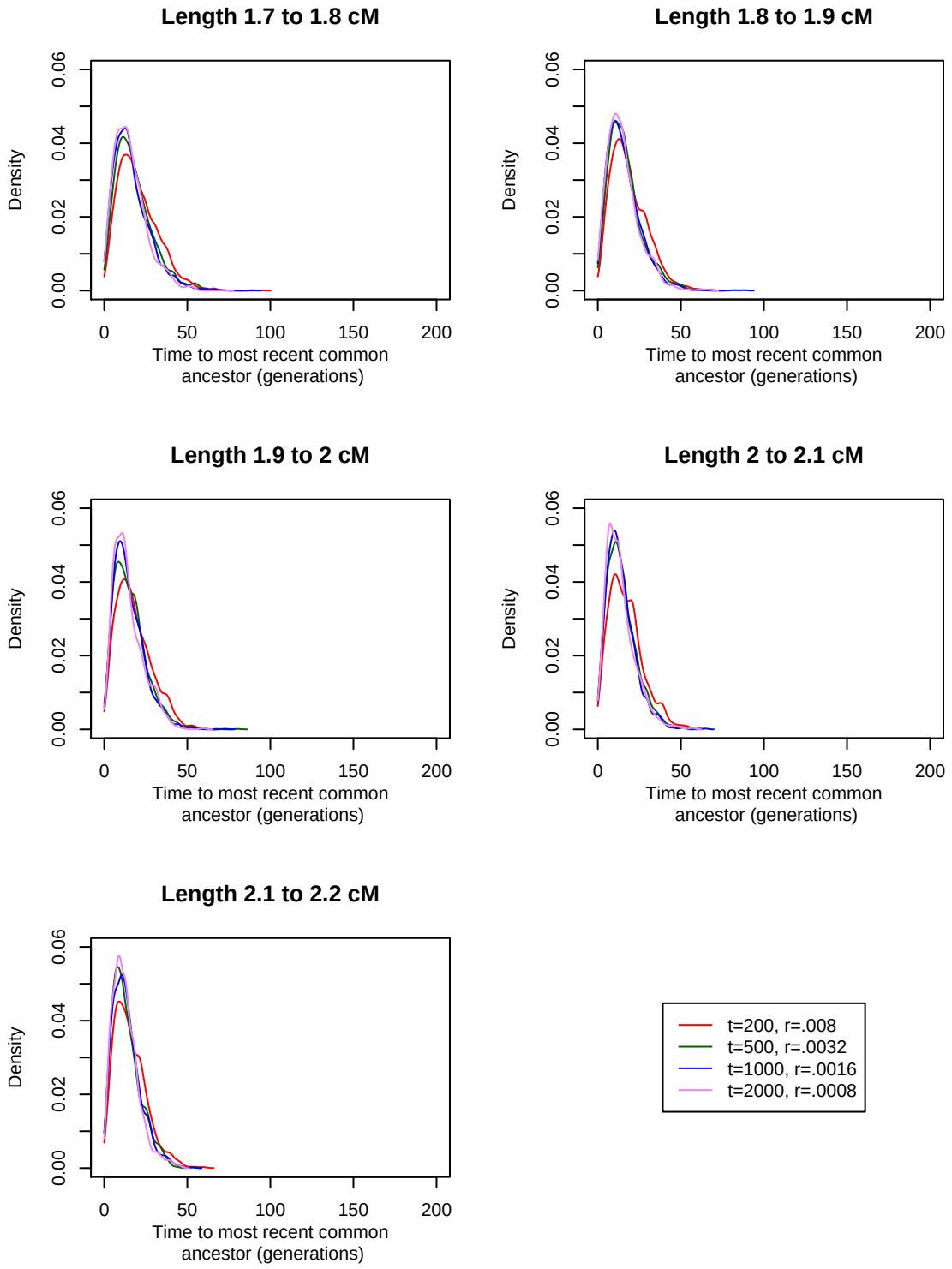

Figure 2.7: Time to most recent common ancestor given length of shared IBD segment, length 1.7-2.1 cM. t is the number of generations of growth from the ancestral population size (5,000) to the current population size (25,000). r is the growth rate over the generations of growth.

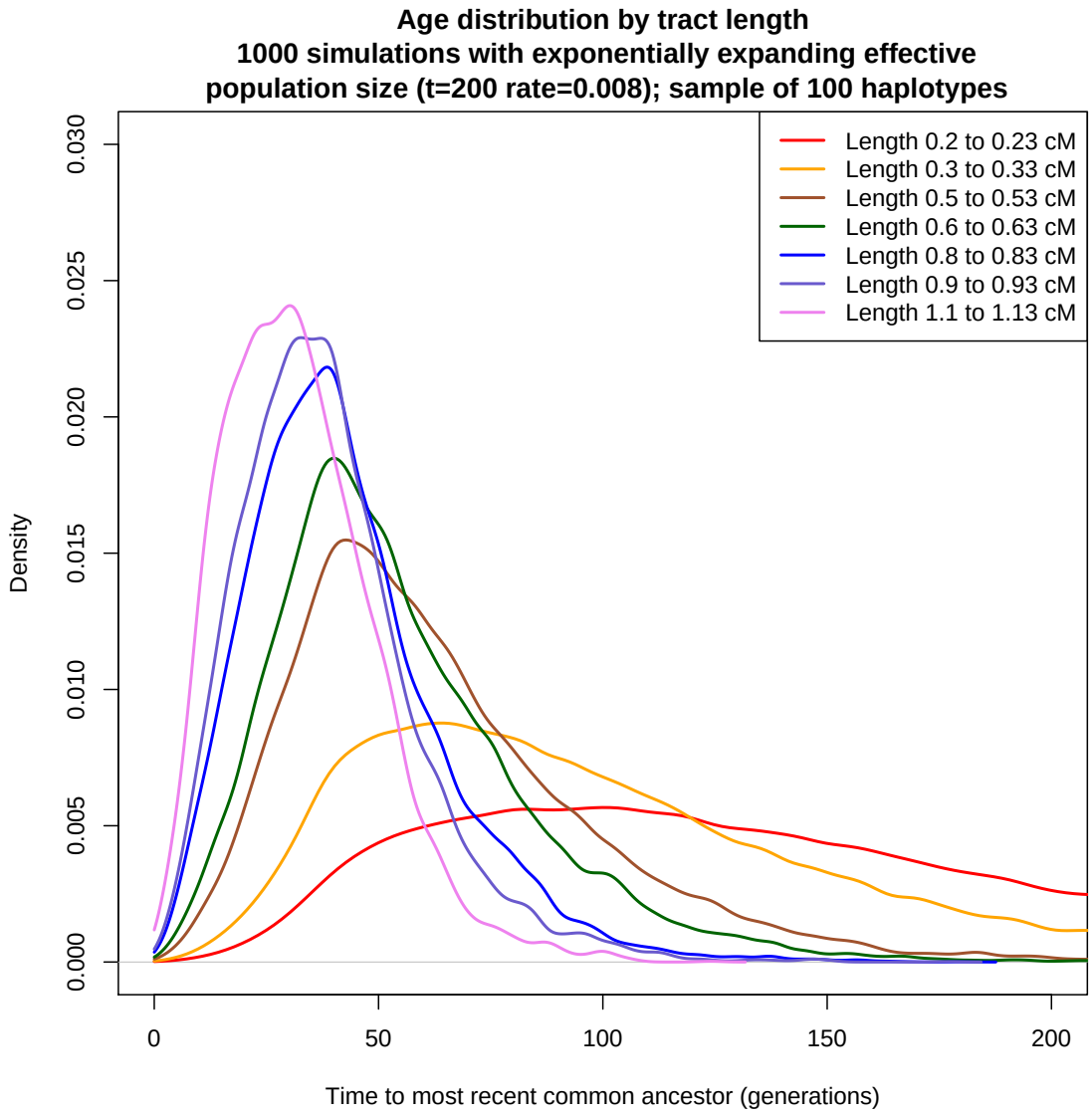

Figure 2.8: Time to most recent common ancestor given length of shared IBD segment, exponentially growing population (rate = 0.008)

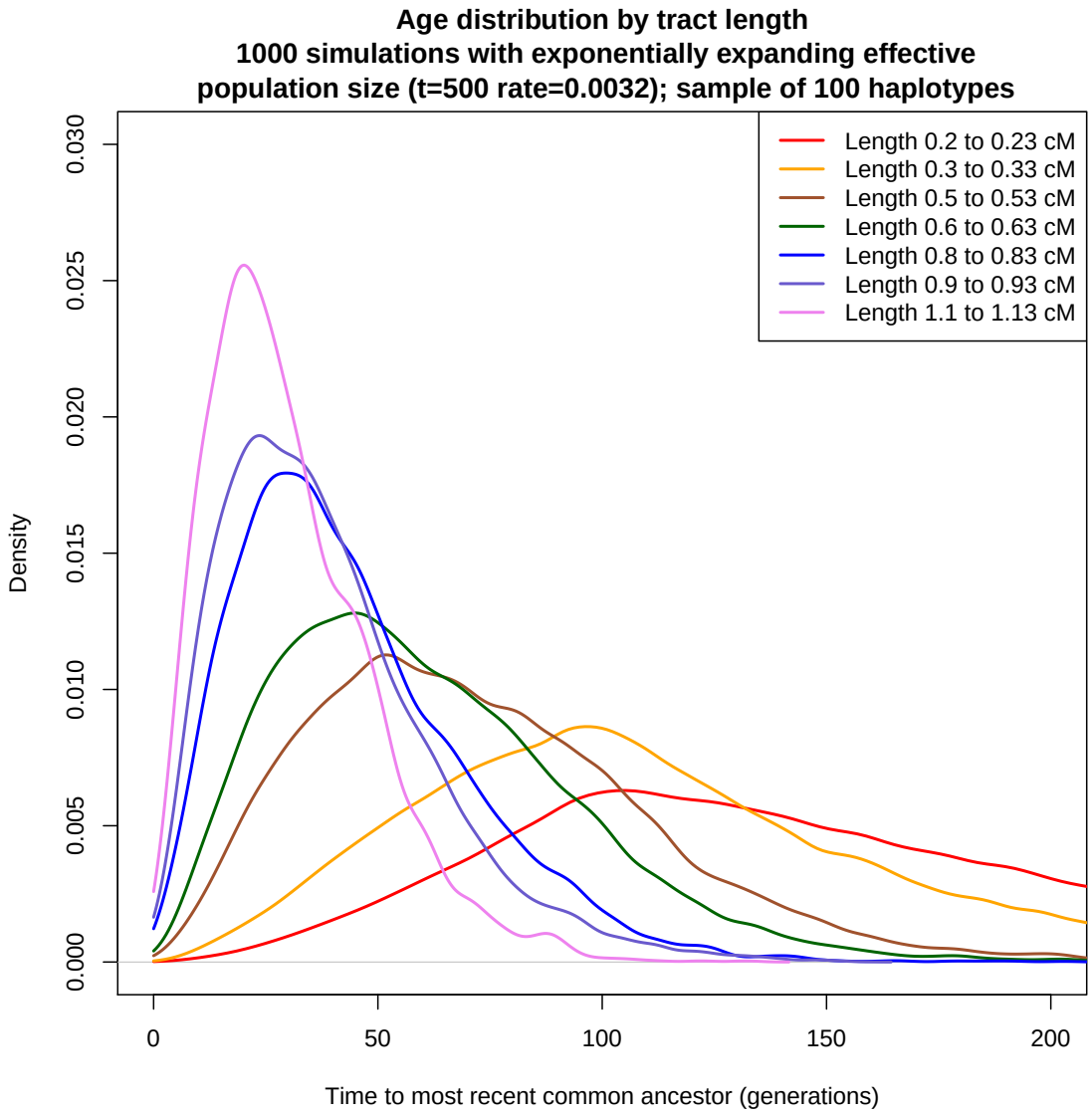

Figure 2.9: Time to most recent common ancestor given length of shared IBD segment, exponentially growing population (rate = 0.0032)

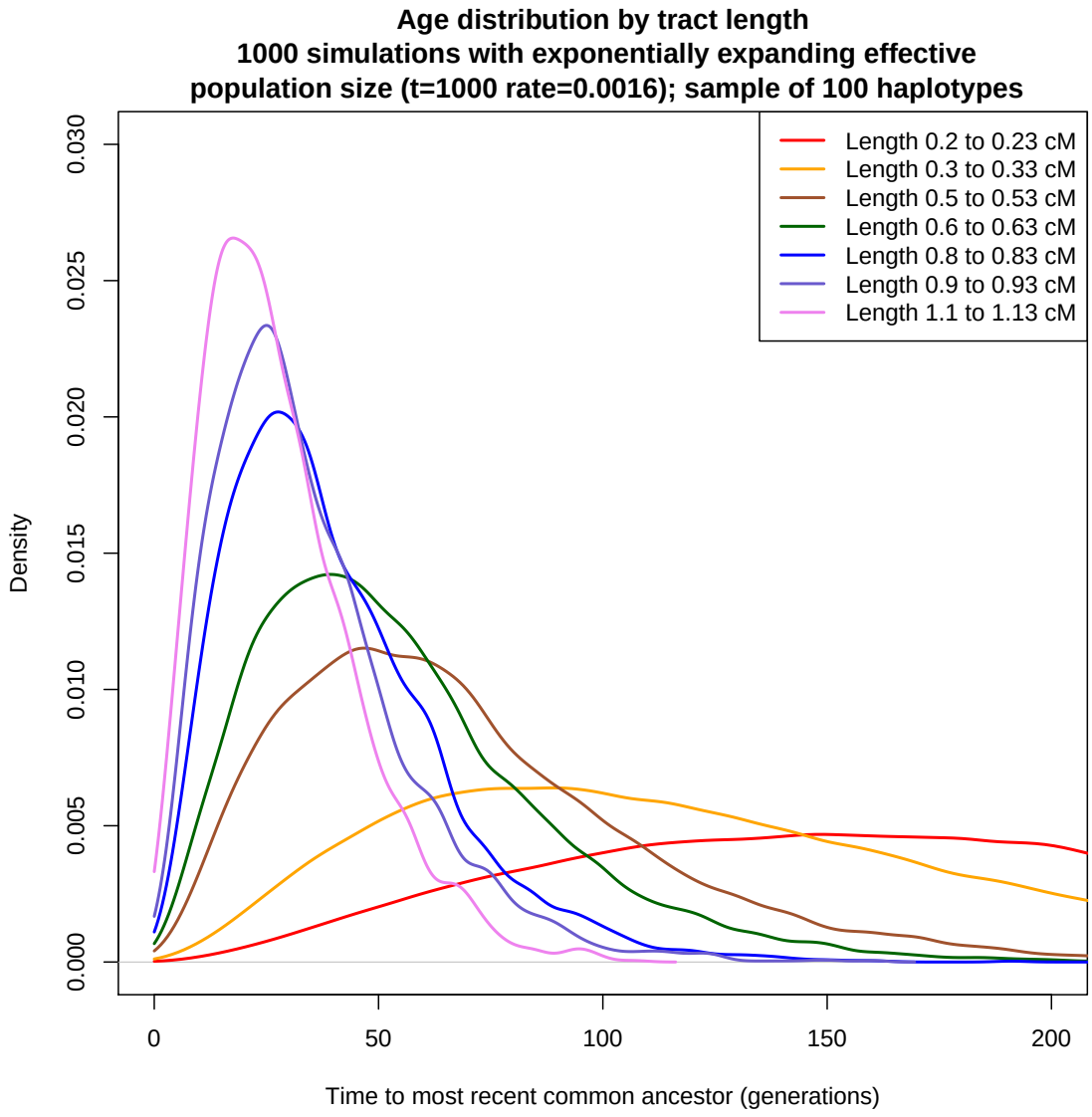

Figure 2.10: Time to most recent common ancestor given length of shared IBD segment, exponentially growing population (rate = 0.0016)

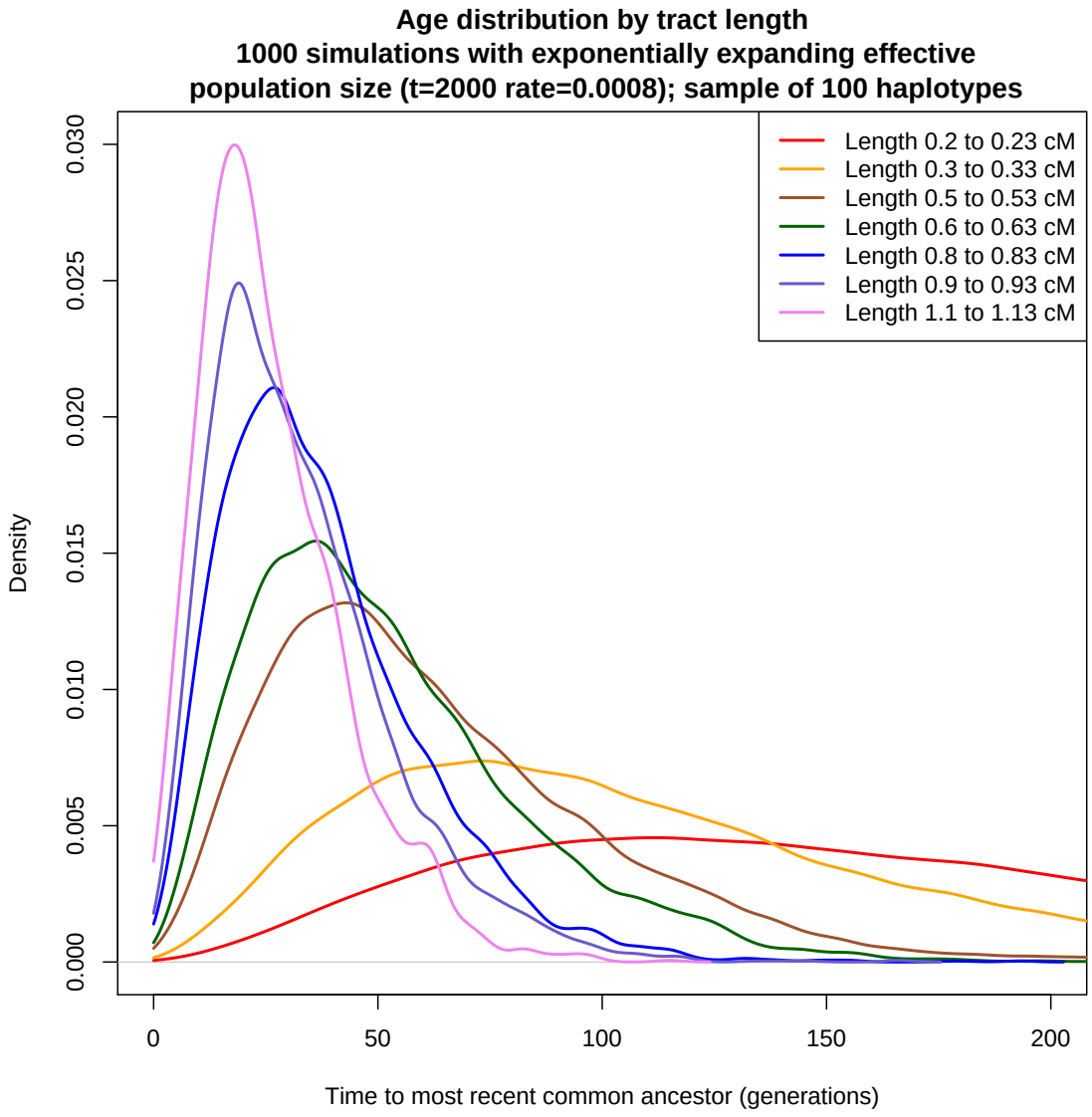

Figure 2.11: Time to most recent common ancestor given length of shared IBD segment, exponentially growing population (rate = 0.0008)

BIBLIOGRAPHY

[1] CA Albers, J Stankovich, R Thomson, M Bahlo, and HJ Kappen. Multipoint approximations of identity-by-descent probabilities for accurate linkage analysis of distantly related individuals. *Am J Hum Genet*, 82:607–622, 2008.

[2] A Albrechtsen, TS Korneliussen, I Moltke, TV Hansen, FC Nielsen, and R Nielsen. Relatedness mapping and tracts of relatedness for genome-wide data in the presence of linkage disequilibrium. *Genet Epi*, 33:266–274, 2009.

[3] MD Brown, CG Glazner, C Zheng, and EA Thompson. Inferring coancestry in population samples in the presence of linkage disequilibrium. *Genetics*, 190:1447–1460, 2012.

[4] B Browning. Beagle genetic analysis software package. `http://faculty.washington.edu/browning/beagle/beagle.html`, 2013. [Online; accessed 18-July-2013].

[5] BL Browning and SR Browning. A fast, powerful method for detecting identity by descent. *Am J Hum Genet*, 88:173–182, 2011.

[6] BL Browning and SR Browning. Improving the accuracy and efficiency of identity by descent detection in population data. *Genetics*, 194:459–471, 2013.

[7] SR Browning and BL Browning. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *Am J Hum Genet*, 81:1084–1097, 2007.

[8] SR Browning and BL Browning. Identity by descent between distant relatives: Detection and applications. *Ann Rev Genet*, 46:617–633, 2012.

[9] G Casella and RL Berger. *Statistical Inference*. Duxbury, 2002.

[10] GK Chen, P Marjoram, and JD Wall. Fast and flexible simulation of dna sequence data. *Genome Res*, 19:136–142, 2009.

[11] Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, 447:661–678, 2007.

[12] A Gusev, LK Lowe, M Stoffel, MJ Daly, D Altshuler, JL Breslow, JM Friedman, and I Pe'er. Whole population, genome-wide mapping of hidden relatedness. *Genome Res*, 19:318–326, 2009.

[13] L Han and M Abney. Identity by descen estimation with dense genome-wide genotype data. *Genet Epi*, 35:557–567, 2011.

[14] J Hein, MK Schierup, and C Wiuf. *Gene Genealogies, Variation, and Evolution*. Oxford University Press, 2005.

[15] F Kaper, S Swamy, B Klotzle, S Munchel, J Cottrell, M Bibikova, HY Chuang, S Kruglyak, M Ronaghi, MA Eberle, and JB Fan. Whole-genome haplotyping by dilution, amplification, and sequencing. *Proc Natl Acad Sci U S A*, 14:5552–5557, 2013.

[16] A Kong, G Masson, ML Frigge, A Gylfason, P Zusmanovich, G Thorleifsson, PI Olason, A Ingason, S Steinberg, T Rafnar, P Sulem, M Mouy, F Jonsson, U Thorsteinsdottir, DF Gudbjartsson, H Stefansson, and K Stefansson. Detection of sharing by descent and long-range phasing and haplotype imputation. *Nat Genet*, 40:1068–1075, 2008.

[17] I Moltke, A Albrechtsen, TV Hansen, FC Nielsen, and R Nielsen. A method for detecting ibd regions simultaneously in multiple individuals–with applications to disease genetics. *Genome Res*, 21:1168–1180, 2011.

[18] MR Nelson, D Wegmann, MG Ehm, D Kessner, P St. Jean, C Verzilli, J Shen, Z Tang, SA Bacanu, D Fraser, L Warren, J Aponte, M Zawistowski, X Liu, H Zhang, Y Zhang, J Li, Y Li, L Li, P Woollard, S Top, MD Hall, K Nangle, J Wang, G Abecasis, LR Cardon, S Zoellner, JC Whittaker, SL Chissoe, J Novembre, and V Mooser. An

abundance of rare functional variants in 202 drug target genes sequenced in 14,002 people. *Science*, 337:100–104, 2012.

[19] K Oldham, J Myland, and J Spanier. *An Atlas of Functions*. Springer, 2008.

[20] PF Palamara, T Lencz, A Darvasi, and I Pe'er. Length distributions of identity by descent reveal fine-scale demographic history. *Am J Hum Genet*, 91:809–822, 2012.

[21] BA Peters, BG Kermani, AB Sparks, O Alferov, P Hong, A Alexeev, Y Jiang, F Dahl, YT Tang, J Haas, K Robasky, AW Zaranek, JH Lee, MP Ball, JE Peterson, H Perazich, G Yeung, J Liu, L Chen, MI Kennemer, K Pothuraju, K Konvicka, M Tsoupko-Sitnikov, KP Pant, JC Ebert, GB Nilsen, J Baccash, AL Halpern, GM Church, and R Drmanac. Accurate whole-genome sequencing and haplotyping from 10 to 20 human cells. *Nature*, 487:190–195, 2012.

[22] P Purcell, B Neale, K Todd-Brown, L Thomas, MAR Ferreira, D Bender, J Maller, P Sklar, PIW de Bakker, MJ Daly, and PC Sham. Plink: A tool set for whole-genome association and population-based linkage analysis. *Am J Hum Genet*, 81:559–575, 2007.

[23] SY Su, J Kasberger, S Baranzini, W Byerley, W Liao, J Oksenberg, E Sherr, and E Jorgenson. Detection of identity by descent using next-generation whole genome sequencing data. *BMC Bioinformatics*, 13:121–128, 2012.

[24] JA Tennessen, AW Bigham, TD O'Connor, W Fu, EE Kenny, S Gravel, S McGee, R Do, X Liu, G Jun, HM Kang, D Jordan, SM Leal, S Gabriel, MJ Rieder, G Abecasis, D Altshuler, DA Nickerson, E Boerwinkle, S Sunyaev, CD Bustamante, MJ Bamshad, JM Akey, Broad GO, and Seattle GO. Evolution and functional impact of rare coding variation from deep sequencing of human exomes. *Science*, 337:64–69, 2012.