CORRECTING MODEL MISSPECIFICATION IN RELATIONSHIP ESTIMATES

Ethan M. Jewett* and the 23andMe Research Team.

23andMe, Inc. Sunnyvale, CA., 94086.

*Address correspondence to: ejewett@23andme.com

1. Abstract

The datasets of large genotyping biobanks and direct-to-consumer genetic testing companies contain many related individuals. Until now, it has been widely accepted that the most distant relationships that can be detected are around fifteen degrees (approximately 8^{th} cousins) and that practical relationship estimates have a ceiling around ten degrees (approximately 5^{th} cousins). However, we show that these assumptions are incorrect and that they are due to a misapplication of relationship estimators. In particular, relationship estimators are applied almost exclusively to putative relatives who have been identified because they share detectable tracts of DNA identically by descent (IBD). However, no existing relationship estimator conditions on the event that two individuals share at least one detectable segment of IBD anywhere in the genome. As a result, the relationship estimates obtained using existing estimators are dramatically biased for distant relationships, inferring all sufficiently distant relationships to be around ten degrees regardless of the depth of the true relationship. Existing relationship estimators are derived under a model that assumes that each pair of related individuals shares a single common ancestor (or mating pair of ancestors). This model breaks down for relationships beyond 10 generations in the past because individuals share many thousands of cryptic common ancestors due to pedigree collapse. We first derive a corrected likelihood that conditions on the event that at least one segment is observed between a pair of putative relatives and we demonstrate that the corrected likelihood largely eliminates the bias in estimates of pairwise relationships and provides a more accurate characterization of the uncertainty in these estimates. We then reformulate the relationship inference problem to account for the fact that individuals share many common ancestors, not just one. We demonstrate that the most distant relationship that can be inferred using IBD may be 200 degrees or more, rather than ten, extending the time-to-common ancestor from approximately 300 years in the past to approximately 3,000 years in the past or more. This dramatic increase in the range of relationship estimators makes it possible to infer relationships whose common ancestors lived before historical events such as European settlement of the Americas, the Transatlantic Slave Trade, and the rise and fall of the Roman Empire.

2. Introduction

A genetic relationship inference method is an algorithm that takes as input the genotyped or sequenced DNA of a putative pair of relatives and potentially other information such as ages and sexes and returns an estimate of their relationship. These algorithms are commonly applied in the context of direct-to-consumer genetic testing in order to identify relatives and infer genealogies [Henn et al., 2012, Ball et al., 2016, Jewett et al., 2021] and they are applied in the context of medical genetic studies to identify and leverage or control for cryptic relatedness [Voight and Pritchard, 2005, Staples et al., 2018, Howe et al., 2022].

All relationship inference methods rely on probability distributions that describe how much IBD is observed between a pair of individuals as a function of their relationship. Common statistics include the total length of observed IBD in centimorgans [Henn et al., 2012, Ball et al., 2016, Jewett et al., 2021] or equivalent quantities such as the kinship coefficient [Staples et al., 2014, 2016, Manichaikul et al., 2010, Ramstetter et al., 2018]. Other common quantities include the number of observed IBD segments and their lengths [Huff et al., 2011].

Regardless of the approach, all existing methods rely implicitly or explicitly on distributions of IBD statistics that were obtained without conditioning on the event that IBD was observed between the two putative relatives. Likelihood methods like the ERSA method of Huff et al. [2011] rely on the probability distribution of one or more observed IBD statistics. Estimators commonly used in direct-to-consumer (DTC) genetic testing rely on empirical versions of these probability distributions that can be obtained using simulations [Henn et al., 2012, Ball et al., 2016]. Other methods rely on analytically-derived bounds that delineate regions of "IBD space" in which the observed values of IBD statistics are most consistent with different relationships [Manichaikul et al., 2010, Ramstetter et al., 2018].

All of these distributions are unconditional on IBD sharing. Although Huff et al. [2011] derive a conditional version of the probability distribution of observed segment lengths conditional on the event that two putative relatives are ascertained because they share IBD at a particular locus, this distribution is not the same as the distribution conditional on observing at least one segment of IBD anywhere in the genome.

Estimates made without conditioning on observing at least one IBD segment are appropriate in scenarios in which pairs of putative relatives were ascertained in a manner that does not depend on the amount of IBD they share; for example, to verify a self-reported relationship that was identified based on previous genealogical knowledge. However, in most contexts in which relationship inference is applied, pairs of putative relatives are ascertained by first detecting shared IBD. In this context, it is inappropriate to apply estimators that do not condition on the event that IBD is observed.

Failure to condition on the observation of IBD has relatively little effect on close relationships because the probability that close relatives share IBD is high. However, when relationships are distant, failure to condition on the event that IBD is observed has profound consequences resulting in dramatically biased relationship estimates as we will demonstrate.

Here, we derive the probability distribution of the observed number of IBD segments and their lengths as a function of the genealogical relationship that gave rise to the segments, conditional on the event that at least one segment of IBD was observed. We show that the corrected estimator no

longer has the profound bias observed in the unconditional estimator and that it allows relationship estimates that extend into the distant past.

We also derive a relationship estimator that explicitly accounts for the fact that pairs of individuals have many thousands of common ancestors. This model of relatedness is arguably more realistic than the most prevalent existing model of relatedness in which each pair of individuals has exactly one common ancestor or mating pair of common ancestors.

Finally, we derive a approximate formula for the expected number of ancestors in each generation in the past who contributed a detectable IBD segment longer than a minimum segment length to a given pair of present-day individuals. Using this formula, we demonstrate that the number of detectable-IBD-contributing common ancestors living more than 100 generations (∼3,000 years) in the past is likely to be non-negligible.

3. Results

3.1. The expected fraction of relationships that are beyond the range of existing estimators.

Before investigating the bias in existing relationship estimators, it's informative to consider how often we might expect to encounter distant IBD-sharing relationships in the first place.

If each each pair of relatives had exactly one common ancestor (or mating pair of common ancestors), then the probability of sharing IBD with a distant relative would indeed be very small. Caballero et al. [2019] found that simulated sixth cousins with just one pair of common ancestors typically shared no IBD segment with one another, and a related analysis found that simulated 8th cousins and beyond (individuals who shared a pair of common ancestors nine generations or more in the past) were exceedingly unlikely to share any detectable IBD segments at all [Williams, 2024]. For instance, the probability that 8th cousins with a single pair of common ancestors shared at least one segment was less than 1%. Henn et al. [2012] predicted that IBD from simulated 9th cousins and beyond would be undetectable (Table 2 of Henn et al. [2012]).

Although it is very unlikely for two distant relatives to share detectable IBD through a particular common ancestor, each pair of individuals has many common genealogical ancestors (Figure 1). Moreover, due to pedigree collapse, each modern individual can have multiple semi-independent lineages back to each of their common ancestors. Thus, the chances of observing a very distant IBD-sharing relationship may actually be quite high.

Although a pair of present-day people may have many genealogical common ancestors, only some of these ancestors contributed detectable segments of IBD longer than the minimum observable segment length τ cM. Figure 2A shows the expected number of distinct detectable-IBD-transmitting common ancestors in generation g in the past that are shared between two individuals in the present day. The analytical expectation is compared with the mean number of common ancestors observed in simulations for several population sizes that are small enough to be computationally efficient. The approximation is relatively good except for very small and very large generation times.

At small generation times, the analytical expectation is an overestimate because the coalescent is not restricted by the pedigree; it therefore allows segments to find common ancestors amongst the full population even one generation in the past. However the approximation becomes fairly accurate even a few generations in the past. One could correct for this discrepancy between the analytical

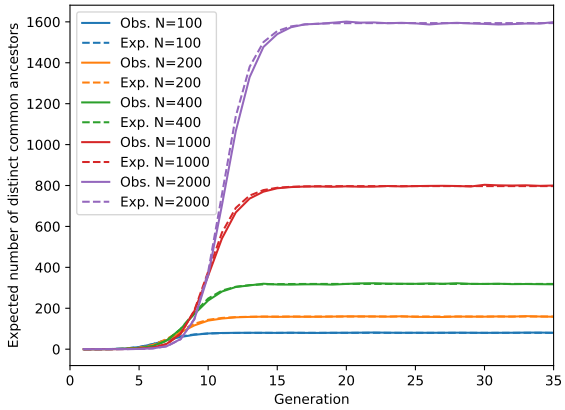

FIGURE 1. The expected number of distinct common ancestors shared between two present-day people. The simulated values are compared with analytical values obtained using Equation (29) for various population sizes for which simulation is fast.

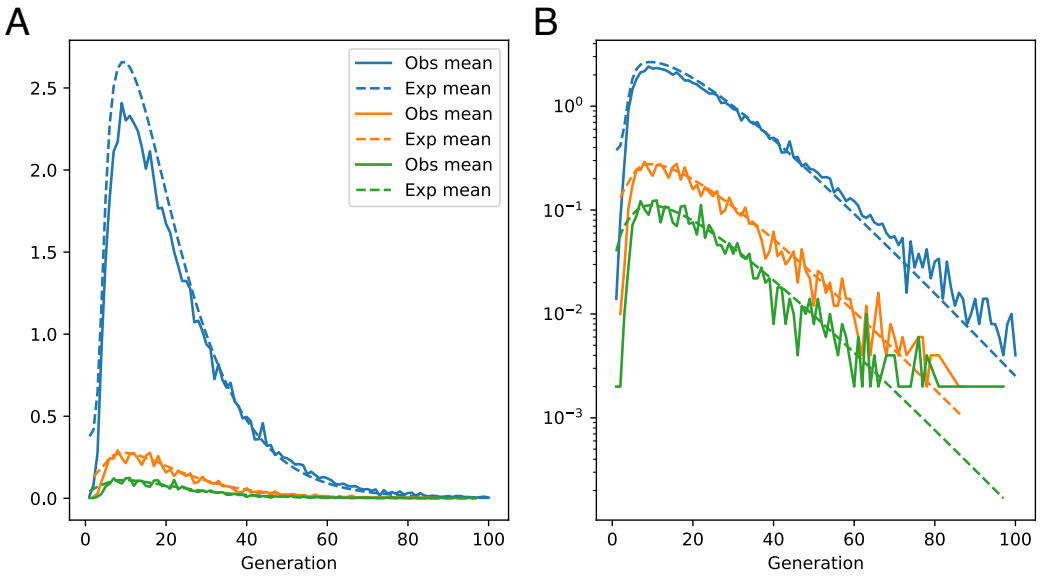

FIGURE 2. (A) The expected number of distinct detectable-IBD-transmitting common ancestors at each generation in the past. Curves are shown for a minimum segment length of $\tau = 5$ cM and for three different effective population sizes: $N = 1,000$, $N = 2,000$ and $N = 5,000$ individuals. (B) Same as (A) in log scale.

and simulated values for small generation times; however, the discrepancy is unlikely to have much effect on relationship inference because the information for inferring close relatives is so strong that it overcomes the fact that the prior is slightly misspecified for close relationships (Figure 3D).

At large generation times and for large population sizes, the simulations predict more shared ancestors than the analytical approximation, perhaps because there is a non-negligible probability

of observing two adjacent segments. Such adjacent segments have a higher probability of being long enough to be observed. As in the case of small coalescent times, the discrepancy for large coalescent times is unlikely to have much effect on estimates because the peak of the prior occurs around 10 generations and forces most Bayesian estimates to be recent (Figure 3D).

From Figure 2, it can be seen that the probability of observing a detectable IBD segment longer than 5 cM from an ancestor who lived many generations in the past is non-negligible. For a particular pair of present-day individuals, Table 1 quantifies the expected number of detectable-IBD-contributing common ancestors living at least g generations in the past in a population with effective population size N. From Table 1, it can be seen that the expected number of ancestors arising at least 100 generations in the past is at least 10^{-4}, even in populations with relatively large effective sizes.

Although pairs of individuals are not independent, the values in Table 1 suggest that each person can have thousands of distant detectable-IBD-contributing common ancestors with others in a database containing millions of individuals. Moreover, since the analytical approximation underestimates the expected number of distant ancestors, the expected number of distant ancestors is probably even higher.

TABLE 1. Expected number of detectable-IBD-contributing common ancestors between two particular present-day individuals. Expectations are shown for various values of g and effective population sizes N.

	Generation					
N	1	10	20	50	75	100
1,000	14.4159	10.9097	5.9837	0.5863	0.0673	0.0071
5,000	2.9688	2.2043	1.2128	0.1202	0.0139	0.0015
10,000	1.4921	1.1035	0.6074	0.0603	0.0070	0.0007
50,000	0.3000	0.2209	0.1216	0.0121	0.0014	0.0001
100,000	0.1501	0.1105	0.0608	0.0060	0.0007	0.0001

The take-home message of Figure 2 and Table 1 is that IBD-sharing relationships greater than ten degrees are likely to occur frequently. These relationships are all inferred as tenth-degree relationships by existing relationship estimators. Therefore, we must use relationship estimators that are capable of inferring these distant relationships.

3.2. **The degree of bias in existing estimators.** To understand the bias that results when unconditional relationship estimators are applied to relatives ascertained on the basis of IBD sharing, we simulated IBD between relatives of various degrees, conditional on the event that they shared at least one detectable segment of IBD with one another [Jewett, 2024]. This sampling scheme reflects what we would expect to see in most real data applications involving relative detection in direct-to-consumer genetic testing or biobank data.

We sampled IBD for 1,000 relative pairs for each degree of relationship between one and forty degrees. For each simulation replicate, we inferred the degree of the relationship between the pair of individuals by maximizing the unconditional likelihood [Huff et al., 2011] and separately by maximizing the conditional likelihood derived in Section 5.1.

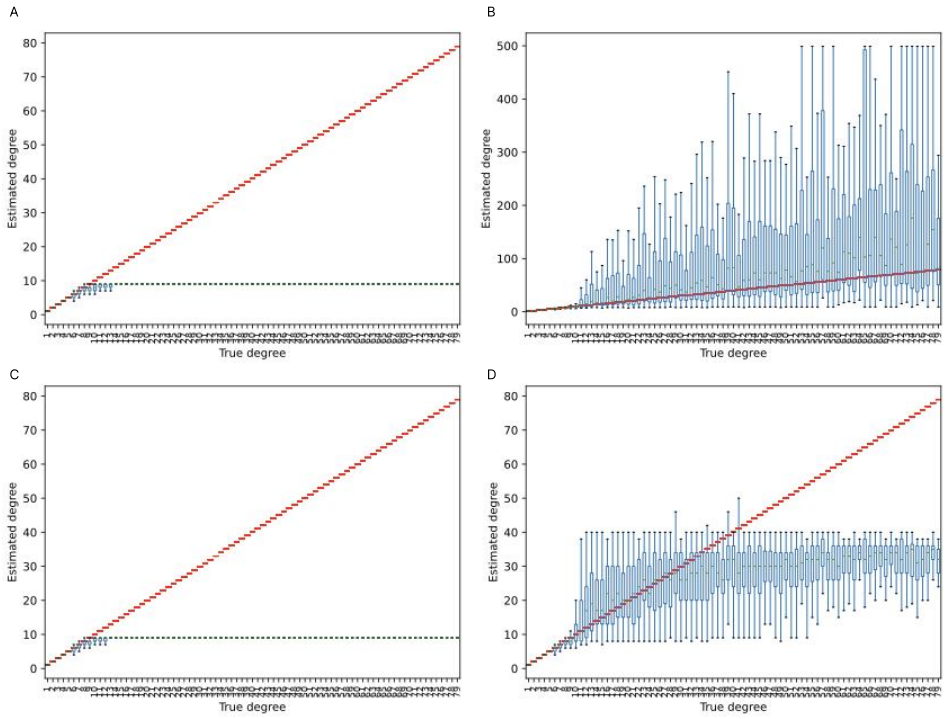

FIGURE 3. Inferred degree using unconditional and conditional estimators for relationships between 1 and 79 degrees. (A) The unconditional likelihood. (B) The conditional likelihood (Equation 8). (C) The unconditional likelihood together with the prior obtained by normalizing Equation (19) with $N = 10,000$. (D) The conditional likelihood together with the prior obtained by normalizing Equation (19) with $N = 10,000$.

From Figure 3A, it can be seen that the unconditional relationship estimates are fairly accurate for close relationships up to approximately ten degrees (approximately fourth cousins), but they begin to diverge sharply from the true degree for relationships beyond ten degrees. Moreover, as the degree of the true relationship increases, the estimated relationships become increasingly tightly grouped around ten degrees. The ceiling at ten degrees is a property of the unconditional likelihood, as we discuss in Section 3.4. Figure 3 shows estimates when all IBD segments are detectable. However, setting a threshold on the minimum detectable segment length has little effect on the estimates, given that IBD is observed in the first place (Figure S1).

In contrast to the unconditional estimates, the conditional estimates shown in Figure 3B have considerably less bias. For these estimates, the mean inferred degree tracks reasonably well with the true degree (red line). The trade-off for reduced bias is increased variance, which can be seen in the range of inferred values shown in the boxplot. This increased variance is due to the fact that there is typically only one segment shared between distant relatives and all information about the degree of the relationship is contained in the length of that segment. As noted in Caballero et al. [2019], segment lengths for distant relationships do not carry much information about the true relationship degree, as the segment length distributions for different degrees overlap considerably.

3.3. **Bayesian relationship estimates.** The approximate prior distribution for the generation in which an IBD-contributing ancestor lived (Equation 19 and Figures 2C and 2D) can be used to obtain a Bayesian estimate of the relationship. The accuracy of the Bayesian estimate using the unconditional estimator is shown in Figure 3C and the accuracy of the conditional Bayesian estimator is shown in Figure 3D for an effective population size of $N = 10,000$.

From Figures 3C and 3D, it can be seen that the unconditional estimator continues to have a ceiling at ten or eleven degrees. The prior introduces considerable bias into the conditional estimator as well, with the counterpoint that the variability in the estimates is dramatically reduced and all estimates are constrained to lie within a range that is more likely for human populations.

Ultimately, the prior has a considerable effect on the estimates so it is important to be fairly confident that the prior captures the true range of possible degrees of relationship. As we have noted, the prior in Equation (19) underestimates the number of ancestors observed at very deep timescales; however, even without this bias it seems unlikely that the prior would admit deep estimates beyond 50 generations or so. Thus, there is potentially an argument for employing the likelihood estimator for certain applications because it allows for deep estimates when the true relationship is distant, whereas the Bayesian estimator does not. It may therefore provides a less biased picture of overall relationship.

3.4. **The total length of IBD.** Figure 3 shows the inferred relationship using the full set of observed segment lengths. However, it is also common for relationship estimators to use the total sum $L = \sum_{i=1}^{n} \ell_i$ of lengths of observed IBD [Ball et al., 2016, Jewett et al., 2021] or a related statistic such as the kinship coefficient [Staples et al., 2014, 2016, Manichaikul et al., 2010, Ramstetter et al., 2018]. In Section 5.2, we derive a formula for the total observed length L of IBD for both the conditional case in which IBD was observed and the unconditional case in which IBD may or may not have been observed.

Figure 4A shows the distribution of the total length L of IBD for $a = 1$ common ancestors and several small values of m, the number of meioses in the lineage connecting two putative relatives. For small values of m, it can be seen that the unconditional and conditional distributions are nearly identical. This is because the probability of observing at least one segment of IBD is nearly one when m is small and the correction term $1 - e^{-\eta_{a,m}\tau}$ in Equation (14) is approximately one.

The difference between the conditional and unconditional distributions can be seen by comparing Figures 4B and 4C. For values of m in the range 6 to 14, the conditional and unconditional distributions begin to diverge from one another and begin to take on qualitatively different behavior.

Figure 4B explains why the likelihood estimator in Figure 3A tops out at $d = m - a + 1 \approx 10$ degrees. In particular, for $a = 1$ the density for all $m > 10$ is uniformly lower than the density for $m = 10$ in the region $L > 0$. This property of the density is made possible by the fact that the unconditional distribution has a point mass at $L = 0$, allowing the density for $L > 0$ to integrate to less than one. This property implies that the greatest possible value of m that can be inferred by maximum likelihood is $m = 10$ when $a = 1$ and $m = 11$ when $a = 2$ since the likelihood surface for all higher values of m is uniformly lower. Thus, the asymptote in Figures 3A and 3C at $d = m - a + 1 = 10$ degrees is a fundamental property of the likelihood. Moreover, all existing

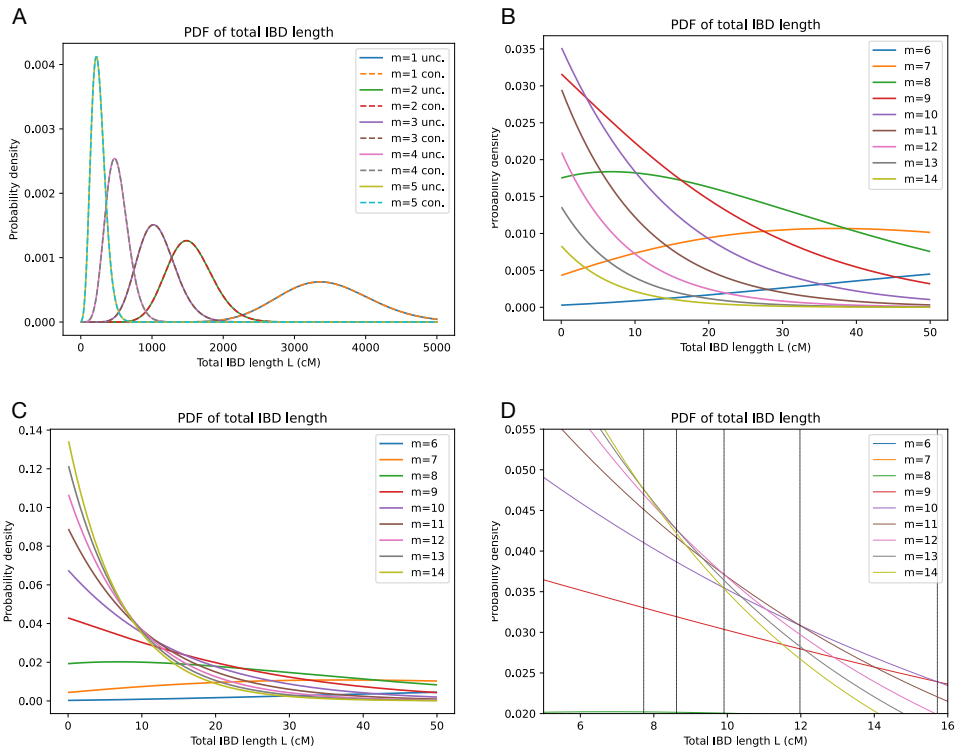

FIGURE 4. The distribution of the total length of IBD. (A) The distribution of the total length of IBD for $a = 1$ ancestors and several values of the number m of meioses separating two putative relatives. Both the unconditional (unc.) and conditional (con.) distributions are shown. (B) The unconditional distribution for values of m in the range $m \in \{6, ..., 14\}$. (C) The conditional distribution for values of m in the range $m \in \{6, ..., 14\}$. (D) Close up of the conditional distribution along with points L_d (black vertical lines) marking transition points where the likelihood surface for $d = m - a + 1$ is greater than the likelihood surface for $d = m + 1 - a + 1$.

estimators do something similar to maximizing the unconditional likelihood, which results in a ceiling at $d = 10$ degrees for existing estimators.

In contrast to the unconditional likelihood, an estimator based on the conditional likelihood (Figures 4C and 4D) does not have a ceiling. The reason is that for any degree $d > 0$, there is always a region $(L_{d+1}, L_d]$ such that the likelihood is maximized at degree d whenever the total sum of IBD lengths L is within $(L_{d+1}, L_d]$. The bounds L_d of these regions (black vertical lines) are shown in the close-up of the conditional distribution shown in Figure 4D.

3.5. **Regions where the likelihood is maximized.** A relationship estimator can search for the values of a and m that maximize the likelihood of the observed value of L; however, for a particular value of a it is also possible to precompute regions $(L_{a,m+1}, L_{a,m}]$ such that the likelihood of L is maximized by a and m for $L \in (L_{a,m+1}, L_{a,m}]$. This is the approach taken by some genetic testing companies, where the regions $(L_{a,m+1}, L_{a,m}]$ are determined empirically using simulations [Henn et al., 2012, Ball et al., 2016]. Other methods use similar bounds obtained from kinship coefficients [Manichaikul et al., 2010, Ramstetter et al., 2018]. In general, because there is more information for

inferring the compound parameter $d = m - a + 1$ and it is difficult to resolve a and m for the same value of d, estimators typically express ranges in terms of d rather than a and m. Specifically, they use the ranges $(L_{d+1}, L_d]$, which can be approximated by setting a to a fixed value.

Note that because the simulations that are used to obtain the regions $(L_{d+1}, L_d]$ are performed unconditionally on the event O that an IBD segment is observed, the resulting estimator is equivalent to the maximum likelihood estimator based on the unconditional distribution (Figure 3A). Kinship coefficients are also unconditional on IBD sharing. In Section 5.3, we use the conditional distribution of L to derive the bounds $(L_{d+1}, L_d]$ under the conditional likelihood. These are shown in Table 2.

TABLE 2. Bounds L_d on the regions in which the likelihood is maximized for degree d whenever $L \in (L_{d+1}, L_d]$. Values are shown for the case $a = 1$ and $\tau = 0$.

m	Unconditional	Conditional	m	Unconditional	Conditional
1	2267.06	2267.06	24	-	3.92
2	1275.18	1275.18	25	-	3.77
3	743.12	743.12	26	-	3.64
4	357.57	357.57	27	-	3.51
5	172.78	172.82	28	-	3.39
6	82.83	83.44	29	-	3.28
7	38.59	41.65	30	-	3.17
8	16.28	23.52	31	-	3.08
9	3.39	15.72	32	-	2.99
10	0.	11.98	33	-	2.90
11	-	9.91	34	-	2.82
12	-	8.62	35	-	2.74
13	-	7.72	36	-	2.67
14	-	7.06	37	-	2.60
15	-	6.53	38	-	2.53
16	-	6.10	39	-	2.47
17	-	5.74	40	-	2.41
18	-	5.42	41	-	2.35
19	-	5.13	42	-	2.30
20	-	4.88	43	-	2.25
21	-	4.65	44	-	2.20
22	-	4.45	45	-	2.15
23	-	4.26	46	-	2.11
24	-	4.08	...	...	...

From Table 2, we can see that $d = m = 10$ is the most likely degree in the region $L \in [0, 3.39]$ for the unconditional estimator and the full region $L > 0$ is covered by regions corresponding to degrees 1 through 10. In comparison, for the conditional likelihood, there is a region of L-space in which each degree d is the most likely degree.

4. A RELATIONSHIP ESTIMATOR THAT ACCOUNTS FOR MULTIPLE ANCESTORS

So far, we have considered the problem of updating existing relationship estimators to condition on the event that at least one segment of IBD is observed. However, these estimators make use of a conceptual model of relatedness in which each pair of individuals is connected through a single

common ancestor or mating pair of common ancestors (5A). Individuals i and j in Figure 5A may have other very distant ancestors (grey circles) that give rise to occasional small segments of "background IBD," but in this conceptual model, "background IBD" reflects very distant relationships that we aren't interested in.

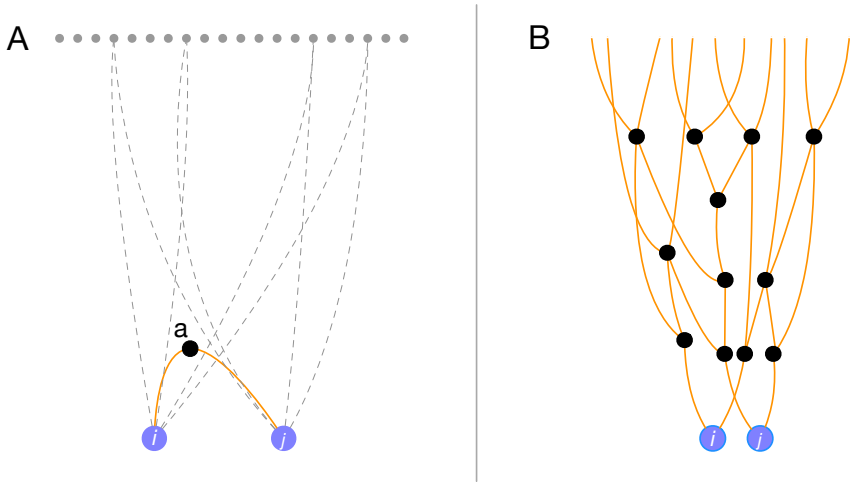

FIGURE 5. Conceptual models for the development of relationship estimators. Panel A shows the conceptual model that underlies existing relationship estimators. In this model, each pair of individuals, i and j (purple dots), shares a single common ancestor, a, or a single mating pair of common ancestors (a_1, a_2) (black circle). Other genealogical ancestors (grey dots) exist in the very distant past, but any IBD these ancestors contribute amounts to background noise. Panel B shows a conceptual model that more accurately describes genealogical relatedness at distant timescales. In this model, each pair of individuals shares many common ancestors in each generation in the past. Some of these ancestors contribute detectable IBD to the pair and some do not.

The assumption that two people are connected through a single close relationship may be true when we restrict our attention to the very recent past (perhaps within the most recent five to ten generations), but this assumption quickly breaks down as the degree of the relationship increases.

As we discussed in Section 3.1, the number of common genealogical ancestors between two people can be large even in the not-too-distant past and each individual has many lineages connecting them to each ancestor due to pedigree collapse. Therefore, for distant relationships, it is more appropriate to conceptualize the relationship inference problem in the form shown in Figure 5B.

The goal of relationship inference under the model in Figure 5B is not to infer "the relationship" between i and j since there is not just one relationship. Instead, the goal is to infer any one of several quantities of interest such as (1) the most recent genealogical relationship, (2) the most recent genealogical relationship that resulted in detectable shared IBD, (3) the number of genealogical ancestors at each generation in the past, and (4) the number of genealogical ancestors at each generation in the past who contributed observed IBD, or some other suitable quantity that reflects the fact that individuals share many common ancestors through many different relationships.

To derive a relationship estimator for deep relationships, we conceptualize inheritance under the model in Figure 5B and our goal is to infer a statistic that captures this kind of multi-ancestral relationship. Of the statistics above, Statistics 3 and 4 do the most comprehensive job of reflecting the reality of relatedness. However, compared with Statistics 1 and 2, Statistics 3 and 4 pertain to a much larger state space and are therefore more computationally challenging. They are also more susceptible to the statistical problem of non-identifiability or near-nonidentifiability. In particular, many genealogical relationships can give rise to similar observed IBD patterns, so we may be unable to say with high certainty which ancestral relationships gave rise to the observed patterns.

We investigate Statistics 2 and 4 in Section 5.5. There, we show how to infer the number of detectable-IBD-transmitting common ancestors, along with which generation in the past they are from. This is effectively Statistic 4. If we take the minimum generation of such a common ancestor, we get Statistic 3.

Figure 6A shows a comparison of Statistic 3 with the true degree of relationship for individuals who are truly related through only a single common ancestors. Because the estimator of Section 5.5 has a very large state space when the number of observed segments is large, it is computationally taxing for the inference of close relationships that share many IBD segments. Therefore, Figure 6A excludes close relationships. From Figure 6 it can be seen that Statistic 3, the degree induced by the most recent detectable-IBD-transmitting common ancestor, tracks reasonably well with the true degree of relationship, although it is noisy.

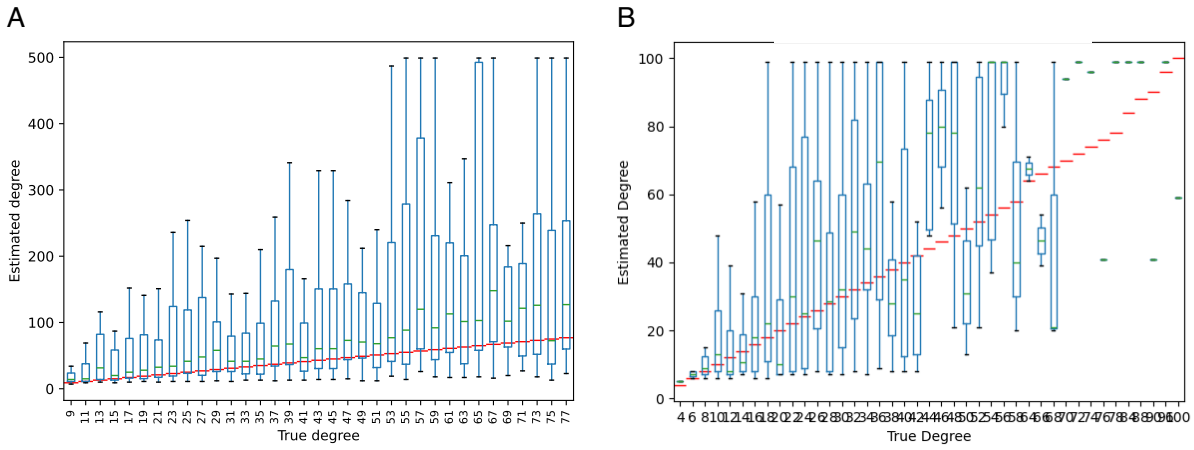

FIGURE 6. (A) Concordance between Statistic 3 (the degree induced by the most recent detectable-IBD-transmitting common ancestor) and the true degree for relationships in which two people were truly related through a single pair of common ancestors. (B) Concordance between Statistic 3 and the shortest degree among individuals related through multiple common ancestors.

For a more realistic application of the multi-ancestor estimator, we simulated 500 relative pairs in a population with effective size $N = 5,000$, leading to multiple common ancestors, on average, contributing more than 5 cM. Figure 6B shows a comparison of Statistic 3 (the inferred degree of the most recent detectable-IBD-contributing common ancestor) with the true degree of the most

recent detectable-IBD-contributing common ancestor. For these estimates, we restricted the state space to include at most three common ancestors and generation depths of at most 100 generations. Despite these restrictions, Figure 6B shows that the estimator tracks reasonably well with the true degree, although the estimates are quite noisy as expected.

Estimators based on the model in Figure 6B have an important advantage over estimators based on the model in Figure 6A. Specifically, estimators based on Figure 6A, model background IBD segments as noise that confounds the signal of the close relationship (orange curve) on which we are focusing. The conceptualization of background IBD as noise effectively imposes a ceiling on the depth of relationships that can be inferred because a short segment whose length is close to the expected background length cannot be distinguished from background noise. The likelihood is maximized by assigning such segments to the "noise" category, rather than assigning them to distant ancestors. In contrast, estimators like the one shown in Figure 6 treat all segments as real and they do not have such a ceiling.

5. Methods

5.1. **Updating existing relationship estimators by conditioning on observed IBD.** Here, we examine how existing relationship estimators (operating under the model in Figure 5A) are affected by conditioning on the event that at least one segment is observed. This allows the direct comparison of existing estimators with versions that condition on observing IBD (Figure 3). We are specifically interested in deriving the distribution of the number and lengths of observed IBD segments, conditional on observing at least one segment.

Following the notation of Ko and Nielsen [2017], let R denote a particular relationship between individuals, i and j, where $R = (u, v, a)$ indicates that i and j are related through $a \in \{1, 2\}$ common ancestor(s) with u meioses separating i from the ancestor(s) and v meioses separating j from the ancestor(s). The total number of meioses is $m = u + v$ and the degree is $d = m - a + 1$.

Let n denote the number of segments shared between relatives i and j arising through relationship R. In the conceptual model underlying the relationship inference problem (Figure 5A), some of these segments come from the common ancestor who contributed detectable IBD to i and j, giving rise to the relationship that we are attempting to infer. Other segments come from other ancestors. Let n_d denote the number of segments that came from the common ancestor of interest and let n_b denote the number of segments that came from other ancestors.

Let $\{\ell_1, ..., \ell_n\}$ denote the lengths of the $n = n_d + n_b$ IBD segments observed between i and j in units of centimorgans. Let O be the event that i and j are observed to share at least one segment of IBD. Our goal is to compute the probability $\mathbb{P}(\ell_1, ...\ell_n | O; m, a)$ of the observed IBD, conditional on the event, O, that at least one IBD segment is observed. Assuming that the n_d segments were transmitted through the relationship R, this probability is a function of the number, a, of most-recent common genealogical ancestors and the number, m, of meioses that separate i and j.

We closely follow the derivation in Huff et al. [2011], who derived the corresponding probability distribution in the unconditional case. As in Huff et al. [2011], we make the simplifying assumption that the n_d segments coming from the most-recent IBD-contributing common ancsestor(s) are the

longest segments. This assumption allows us to avoid conditioning on the subset of IBD segments that arose from this ancestor, which allows us to avoid a summation over all subsets of segments that could have come from the common ancestor. Given this simplifying assumption, the distribution of segment lengths is

$$\begin{aligned}
&\mathbb{P}(\ell_1, ..., \ell_n | O; a, m) \\
&\approx \sum_{n_d=1}^{n} \mathbb{P}(\ell_1, ..., \ell_n | n_d = i, n_b = n - i, O; a, m) \mathbb{P}(n_d = i, n_b = n - i | O; a, m) \\
&= \sum_{i=1}^{n} \mathbb{P}(\ell^{(1)}, ..., \ell^{(n_d)}; a, m) \mathbb{P}_b(\ell^{(n_d+1)}, ..., \ell^{(n)}) \mathbb{P}(n_d = i, n_b = n - i | O; a, m) \\
&= \sum_{i=1}^{n} \mathbb{P}(\ell^{(1)}, ..., \ell^{(n_d)}; a, m) \mathbb{P}_b(\ell^{(n_d+1)}, ..., \ell^{(n)}) \frac{\mathbb{P}(n_d = i, n_b = n - i, O; a, m)}{\mathbb{P}(O; a, m)} \\
&= \sum_{i=1}^{n} \mathbb{P}(\ell^{(1)}, ..., \ell^{(n_d)}; a, m) \mathbb{P}_b(\ell^{(n_d+1)}, ..., \ell^{(n)}) \frac{\mathbb{P}(n_d = i; a, m) \mathbb{P}(n_b = n - i)}{\mathbb{P}(O; a, m)} \qquad (1)
\end{aligned}$$

where $\mathbb{P}_b(\cdot)$ denotes the probability distribution of background IBD segment lengths.

The terms in Equation (1) can be obtained using equations from Huff et al. [2011] and are as follows:

$$\mathbb{P}(\ell^{(1)}, ..., \ell^{(n_d)}; a, m) = \left[\prod_{j=1}^{n_d} \lambda_{a,m} e^{-\lambda_{a,m}(\ell^{(j)} - \tau)} \right] \qquad (2)$$

where $\lambda_{a,m}$ is the inverse of the expected length of an IBD segment between two people separated by m meioses and τ is the minimum detectable segment length in centimorgans [Huff et al., 2011]. As in Huff et al. [2011], the segment lengths are modeled as independent, which is likely to be an accurate approximation when m is moderate to large [Huff et al., 2011, Caballero et al., 2019]. We are primarily concerned here with distant relationships[1] so we will make use of the approximation $\lambda_{a,m} \approx m/100$, in which case the right-hand side of Equation (2) doesn't depend on a.

Similarly, the term $\mathbb{P}_b(\ell^{(n_d+1)}, ..., \ell^{(n)})$ can be modeled as the product of $n - n_d$ independent exponential distributions:

$$\mathbb{P}_b(\ell^{(n_d+1)}, ..., \ell^{(n)}) = \left[\prod_{j=1}^{n-n_d} \lambda_\ell e^{-\lambda_\ell(\ell^{(j)} - \tau)} \right] \qquad (3)$$

where λ_ℓ can be found empirically by assuming that most pairs of individuals in a large database are only distantly-related and collecting the lengths of IBD segments shared between many randomly sampled pairs.

[1]The approximation begins to break down when $a = 2$ and $m \leq 3$ (avuncular relationships and closer), but in this region of the parameter space, the information coming from the shared IBD is typically so strong that relationships can be inferred accurately even when the likelihood is misspecified.

The distribution $\mathbb{P}(n_b = i)$ of the number, n_b, of background segments can be found empirically as well. In particular, it is reasonable to model n_b as a Poisson random variable

$$\mathbb{P}(n_b = i) = \frac{\eta_b^i e^{-\eta_b}}{i!} \tag{4}$$

with mean η_b equal to the average number of IBD segments observed between a randomly-sampled pair of individuals in a dataset.

Following Huff et al. [2011], we model the number of observed segments from the genealogically most recent IBD-contributing common ancestor(s) as Poisson, with mean $\eta_{a,m,\tau}$:

$$\mathbb{P}(n_d = i) = \frac{\eta_{a,m,\tau}^i e^{-\eta_{a,m,\tau}}}{i!}, \tag{5}$$

where, for moderate-to-large m, $\eta_{a,m,\tau}$ can be approximated [Thomas et al., 1994, Huff et al., 2011] as

$$\eta_{a,m,\tau} \approx 2^{1-m} a(rm + c) e^{-m\tau/100} \tag{6}$$

where r is the expected number of recombination events per meiosis in the genome, c is the number of chromosomes, and τ is the minimum detectable segment length. For the autosomal genome in humans, we have $r \approx 35$ and $c = 22$ [McVean et al., 2004, Huff et al., 2011].

Finally, the probability of observing any segments at all is:

$$\mathbb{P}(O; a, m) = \mathbb{P}(n_d + n_b \geq 1; a, m) = 1 - e^{-\eta_b - \eta_{a,m,\tau}}, \tag{7}$$

which comes from the fact that n_d and n_b are each Poisson, so their sum is Poisson with mean equal to the sum of the individual means. Equation (7) is one minus the probability that the sum $n_d + n_b$ is zero.

All together, Equation (1) becomes

$$\begin{aligned} &\mathbb{P}(\ell_1, ..., \ell_n | O; a, m) \\ &\approx \sum_{i=1}^{n} \left[\prod_{j=1}^{i} \frac{m e^{-m(\ell^{(j)} - \tau)/100}}{100} \right] \left[\prod_{j=i+1}^{n} \lambda_\ell e^{-\lambda_\ell (\ell^{(j)} - \tau)} \right] \frac{\eta_b^{n-i} e^{-\eta_b}}{i!(n-i)!} \frac{\eta_{a,m,\tau}^i e^{-\eta_{a,m,\tau}}}{1 - e^{-\eta_b - \eta_{a,m,\tau}}} \end{aligned} \tag{8}$$

where λ_ℓ and η_b are found empirically and $\eta_{a,m,\tau} \approx 2^{1-m} a(rm + c) e^{-m\tau/100}$.

Note that Equation (8) is nearly identical to its unconditional version presented in Equation (9) of Huff et al. [2011]. The only difference is the normalizing factor $1 - e^{-\eta_b - \eta_{a,m,\tau}}$ and the fact that the summation starts at 1 rather than at 0. This provides a simpler alternative derivation, as the conditional distribution is effectively obtained from the unconditional expression by ignoring the mass at zero and renormalizing the mass above zero.

5.2. **The distribution of the total length of IBD.** We can also obtain the distribution of the total length of observed IBD. We make the simplifying assumption that there are no background IBD segments in order to obtain a formula that yields values that are comparable the distributions obtained using simulations in existing methods [Henn et al., 2012, Ball et al., 2016]. The joint distribution of the total length of IBD L and the number of segments n can be obtained by noting that the sum of exponential random variables follows a gamma distribution. Conditioning on the

event O that at least one segment is observed, we obtain

$$f_{L,n}|O(L,n|O) = \frac{\lambda_{a,m}^n}{\Gamma(n)}(L-n\tau)^{n-1}e^{-\lambda_{a,m}(L-n\tau)}\frac{\eta_{a,m,\tau}^n}{\Gamma(n+1)}e^{-\eta_{a,m,\tau}}\frac{1}{1-e^{-\eta_{a,m,\tau}}} \tag{9}$$

where $\eta_{a,m,\tau} = 2^{1-m}a(rm+c)e^{-m\tau/100}$ and $\lambda_{a,m} \approx m/100$.

We can also obtain the distribution of L alone by summing over n:

$$\begin{aligned}
&f_{L|O}(L|O)\\
&= \sum_{n=1}^{\infty} f_{L,n|O}(L,n|O)\\
&= \sum_{n=1}^{\infty} \frac{\lambda_{a,m}^n}{\Gamma(n)}(L-n\tau)^{n-1}e^{-\lambda_{a,m}(L-n\tau)}\frac{\eta_{a,m,\tau}^n}{\Gamma(n+1)}e^{-\eta_{a,m,\tau}}\frac{1}{1-e^{-\eta_{a,m,\tau}}}.
\end{aligned} \tag{10}$$

For the case in which the minimum segment length is $\tau = 0$, this equation has a closed form, which can be obtained by rearranging the terms to resemble the summation in a modified Bessel function:

$$\begin{aligned}
&f_{L|O}(L|O)\\
&= \sum_{n=1}^{\infty} \frac{\lambda_{a,m}^n}{\Gamma(n)}L^{n-1}e^{-\lambda_{a,m}L}\frac{\eta_{a,m,\tau}^n}{\Gamma(n+1)}e^{-\eta_{a,m,\tau}}\frac{1}{1-e^{-\eta_{a,m,\tau}}}\\
&= \frac{e^{-\eta_{a,m,\tau}}}{1-e^{-\eta_{a,m,\tau}}}e^{-\lambda_{a,m}L}\sum_{n=1}^{\infty}\frac{\eta_{a,m,\tau}^n\lambda_{a,m}^n}{\Gamma(n)}L^{n-1}\frac{1}{\Gamma(n+1)}\\
&= \frac{e^{-\eta_{a,m,\tau}}}{1-e^{-\eta_{a,m,\tau}}}e^{-\lambda_{a,m}L}\frac{1}{L}\sum_{n=1}^{\infty}\frac{[L\eta_{a,m,\tau}\lambda_{a,m}]^n}{\Gamma(n)\Gamma(n+1)}\\
&= \frac{e^{-\eta_{a,m,\tau}}}{1-e^{-\eta_{a,m,\tau}}}e^{-\lambda_{a,m}L}\frac{1}{L}\sum_{k=0}^{\infty}\frac{[L\eta_{a,m,\tau}\lambda_{a,m}]^{k+1}}{\Gamma(k+1)\Gamma(k+1+1)}\\
&= \frac{e^{-\eta_{a,m,\tau}}}{1-e^{-\eta_{a,m,\tau}}}e^{-\lambda_{a,m}L}\frac{1}{L}\sum_{k=0}^{\infty}\frac{\sqrt{L\eta_{a,m,\tau}\lambda_{a,m}}^{2k+2}}{\Gamma(k+1)\Gamma(k+1+1)}\\
&= \frac{e^{-\eta_{a,m,\tau}}}{1-e^{-\eta_{a,m,\tau}}}e^{-\lambda_{a,m}L}\frac{\sqrt{L\eta_{a,m,\tau}\lambda_{a,m}}}{L}\sum_{k=0}^{\infty}\frac{\sqrt{L\eta_{a,m,\tau}\lambda_{a,m}}^{2k+1}}{\Gamma(k+1)\Gamma(k+1+1)}\\
&= \frac{e^{-\eta_{a,m,\tau}}}{1-e^{-\eta_{a,m,\tau}}}e^{-\lambda_{a,m}L}\frac{\sqrt{L\eta_{a,m,\tau}\lambda_{a,m}}}{L}\sum_{k=0}^{\infty}\frac{\left(\frac{2\sqrt{L\eta_{a,m,\tau}\lambda_{a,m}}}{2}\right)^{2k+1}}{\Gamma(k+1)\Gamma(k+1+1)}\\
&= \frac{e^{-\eta_{a,m,\tau}}}{1-e^{-\eta_{a,m,\tau}}}e^{-\lambda_{a,m}L}\frac{\sqrt{L\eta_{a,m,\tau}\lambda_{a,m}}}{L}I_1(2\sqrt{L\eta_{a,m,\tau}\lambda_{a,m}}),
\end{aligned} \tag{11}$$

where $I_1(\cdot)$ is the modified Bessel function of the first kind.

Note that the only aspect of Equation (11) that corresponds to conditioning on the event O is the factor $1-e^{-\eta_{a,m,\tau}}$, in the denominator. Thus, to obtain the unconditional distribution of the

total length, we simply remove this factor:

$$f_L(L) = e^{-\eta_{a,m,\tau}} e^{-\lambda_{a,m} L} \frac{\sqrt{L\eta_{a,m,\tau}\lambda_{a,m}}}{L} I_1(2\sqrt{L\eta_{a,m,\tau}\lambda_{a,m}}). \tag{12}$$

Equation (12) provides the analytical version of the of the empirical distribution that is used by genetic testing companies for relationship inference (for example, see Figure 3 of Henn et al. [2012] and the figure in Section 5.2 of Ball et al. [2016]). Plots of Equations (11) and (12) are shown in Figure 4.

Genetic testing companies typically obtain Equation (12) empirically using simulations. However, because there is a lot of noise in empirical simulations, the distributions obtained by genetic testing companies are noisy. In comparison, the analytical distribution has no noise.

5.3. **Regions where the likelihood is maximized.** From Equation (12), we can analytically obtain the thresholds on L that are used for relationship inference by some genetic testing companies. These thresholds are points L_d that partition the space of the total IBD length into regions $(L_1, L_0]$, $(L_2, L_1]$, $(L_3, L_2]$, etc. such that the most likely degree of relationship in the range $(L_d, L_{d-1}]$ is d. We can also extend these regions to the case of conditional likelihoods using Equation (11).

The range $(L_d, L_{d-1}]$ depends only on the degree $d = m - a + 1$, rather than explicitly on a and m. For distant relationships, this dependence on d alone is reasonable because a relationship of degree d with two common ancestors $a = 2$ has a very similar pattern of shared IBD compared with a relationship of degree d and one common ancestor $a = 1$. Although age information can help to determine the number of ancestors a for close relationships, for very distant relationships, there is almost no available information to determine whether the number of ancestors is 2 or 1. Hence, the degree d is often the most relevant quantity to infer for distant relatives.

Suppose for simplicity that a is equal to 1. The case $a = 2$ yields similar values of L_d for large d. Under this assumption we have $m = d$ and Equation (12) becomes

$$f_L(L) = e^{-\eta_{d,\tau}} e^{-\lambda_d L} \frac{\sqrt{L\eta_{d,\tau}\lambda_d}}{L} I_1(2\sqrt{L\eta_{d,\tau}\lambda_d}), \tag{13}$$

and Equation (11) becomes

$$f_L(L) = \frac{e^{-\eta_{d,\tau}}}{1 - e^{-\eta_{d,\tau}}} e^{-\lambda_d L} \frac{\sqrt{L\eta_{d,\tau}\lambda_d}}{L} I_1(2\sqrt{L\eta_{d,\tau}\lambda_d}), \tag{14}$$

where $\lambda_d = d/100$ and $\eta_{d,\tau} = 2^{1-d}(rd + c)e^{-d\tau/100}$. Using Equation (13), we can analytically obtain the points L_d by solving for the value of L such that $f_L(L; d+1) = f_L(L; d)$. Using Equation (14) we can obtain these bounds in the conditional case in which at least one segment of IBD is observed. Solving these equations gives the bounds L_d shown in Table (2).

5.4. **The prior distribution of the generation in which a detectable-IBD-transmitting common ancestor lived.** When the minimum observable segment length is τ cM, we say that an ancestor is a detectable-IBD-contributing ancestor for a pair of individuals i and j if they contributed at least one IBD segment longer than τ cM to i and j. To obtain the expected number of detectable-IBD-contributing common ancestors at each generation in the past, we first obtain a more fundamental quantity, which is the expected number $E[S_g]$ of detectable transmitted IBD

segments arising in generation g. To obtain $E[S_g]$, note that all of the DNA in a person living in the present day came from their set of ancestors living in generation g in the past. Moreover, g generations in the past, each copy of a present-day individual's genome existed in approximately H_g tiny semi-independent haplotype chunks that would ultimately combine over the generations to produce the linear genome of the present-day person. The expected number $E[H_g]$ of haplotypes chunks is

$$E[H_g] = rg + c, \tag{15}$$

where c is the number of autosomes under consideration (e.g., $c = 22$ for humans) and r is the expected number of recombination events per meiosis per genome (e.g., $r \approx 35$ cM for humans). Equation (15) comes from the fact that a genome has, on average, $rg + c - 1$ breaks introduced into it over g generations including the obligatory breaks at chromosome endpoints, yielding $rg + c - 1 + 1 = rg + c$ pieces on average. This result comes from the backward-in-time version of the reasoning of Thomas et al. [1994], who considered the number of segments a chromosome breaks into as it is transmitted forward in time.

Now, if we consider the overlapping chunks of two linear genomes g generations in the past, the $rc + g$ chunks from one copy of i's genome will overlap the $rc + g$ chunks from a copy of j's genome in an average of $2(rc + g)$ overlapping regions. The length one of these overlapping chunks between two people found anywhere in the genome is exponentially distributed with mean $100/2g$. Therefore, the probability that a shared segment anywhere in the genome is longer than τ cM is then $e^{-2g\tau/100} = e^{-g\tau/50}$. Finally, under the coalescent model, the probability that a shared segment between two individuals coalesces g generations in the past in a population with N diploid individuals is $\frac{1}{2N}e^{-g/2N}$. Putting all of this together, we get

$$E[S_g] \approx 4 \times 2(rg + c)e^{-g\tau/50}\frac{1}{2N}e^{-g/2N}, \tag{16}$$

where the factor of 4 comes from the fact that individuals i and j each have two copies of the linear genome that can coalesce with one another.

We now use Equation (16) to find the expected number of distinct detectable-IBD-transmitting ancestors in generation g. If i and j share s segments through common ancestors living g generations ago, then the probability they came from exactly k distinct common ancestors can be obtained as follows: let c be the number of common ancestors. The number of ways of placing s indistinguishable segments into c individuals is $\binom{c+s-1}{c-1}$ [Ross, 2002]. The number of ways of choosing k specific individuals is $\binom{c}{k}$. Finally, the number of ways of placing s indistinguishable segments into k specific individuals so that each individual has at least one segment is $\binom{s-1}{k-1}$ [Ross, 2002]. Note that each of the s shared segments must have come from a common ancestor. So, all together, the probability that there are k distinct ancestors in generation g, given that there are s shared segments is

$$Pr(K = k|S = s) = \frac{\binom{c}{k}\binom{s-1}{k-1}}{\binom{c+s-1}{c-1}}. \tag{17}$$

Thus, the expected number of common ancestors, given that there are s segments is

$$E[K|S=s] = \sum_{k=0}^{c} k \frac{\binom{c}{k}\binom{s-1}{k-1}}{\binom{c+s-1}{c-1}} = \frac{cs}{c+s-1}. \tag{18}$$

Modeling the number of shared segments S_g arising in generation g as Poisson distributed with mean $E[S_g]$, we find that the expected number of distinct common ancestors K_g is

$$\begin{aligned} E[K_g] &= E[E[K_g|S_g]] \\ &= \sum_{s=0}^{\infty} Pr(S_g = s) E[K_g|S_g = s] \\ &= \sum_{s=0}^{\infty} \frac{\mu_g^s e^{-\mu_g}}{s!} \frac{c_g s}{c_g + s - 1}, \end{aligned} \tag{19}$$

where $\mu_g \equiv E[S_g] \approx 8(rg+c)e^{-g\tau/50}\frac{1}{2N}e^{-g/2N}$.

5.5. **Multi-ancestor relationship inference.** We noted in Section 4 that the existing modeling framework is not particularly appropriate for inferring deep relationships because this framework assumes that people have just one common ancestor, or one mating pair of ancestors who transmitted IBD. However, the prior distributions shown in Figure 2 demonstrate that it is not unusual for a pair of individuals to have multiple detectable-IBD-transmitting ancestors at various generations in the past.

As we noted, several statistics are more appropriate for capturing the complexity of true relationships, such as 1) the time to the most recent genealogical ancestor, 2) the time to the most recent IBD-contributing common ancestor, 3) the number of genealogical ancestors in each generation in the past, and 4) the number of detectable-IBD-transmitting common ancestors at each generation in the past. Statistics 2 and 4 are perhaps easier to infer because they pertain to ancestors who contributed observed IBD. We focus on these statistics to avoid the extra layer of complexity involved in inferring ancestors who left no genetic traces of themselves in the putative relative pair. Statistics 1 and 3 are likely to depend largely on the demographic history of the population and are left for future work.

Suppose we observe n segments with a total of L cM shared between a pair of individuals. What is our best estimate for the number of ancestors who transmitted these segments and the generations in which they lived? To answer this question, we can write down the probability of the observed IBD, given the full, complex ancestral relationship of the two individuals.

Suppose the two people share a total of K common ancestors who lived in generations $g \in \mathcal{G}$ in the past for some set of generations $\mathcal{G}$. Let k_g be the number of ancestors in generation g, so that we have $\sum_{g\in\mathcal{G}} k_g = K$. Let d_g be the degree separating a pair of individuals with a common ancestor in generation g. For example if the two people are contemporaneous, then $d_g = 2g$, but they need not be contemporaneous.

We can find the joint distribution of n and L by summing over the number of segments attributable to ancestors in each generation. Let n_g be the number of segments inherited from the ancestors in generation g. The generation from which a segment was inherited specifies the expected length θ_g of the segment. Since the segment lengths are exponentially distributed, the expected total length

L_g of the n_g segments arising in generation g follows a Gamma distribution with shape parameter n_g and scale parameter θ_g.

The total observed length $L = \sum_{g\in\mathcal{G}} L_g$ can be approximated as the sum of $|G|$ independent but not identically distributed gamma-distributed random variables. L can be approximated by a single gamma-distributed random variable whose mean and variance match the mean and variance of the sum $\sum_{g\in\mathcal{G}} L_g$ [Covo and Elalouf, 2014]. Since the L_g are independent, the mean and variance of L are

$$E[L] = \sum_{g\in\mathcal{G}} E[L_g] = \sum_{g\in\mathcal{G}} n_g\theta_g \tag{20}$$

$$var(L) = \sum_{g\in\mathcal{G}} var(L_g) = \sum_{g\in\mathcal{G}} n_g\theta_g^2, \tag{21}$$

where we have used the fact that the parameters α and β of a gamma-distributed random variable X satisfy $E[X] = \alpha\beta$ and $var(X) = \alpha\beta^2$. It follows that the parameters α_L and β_L of L, given $\{n_g\}_{g\in\mathcal{G}}$ satisfy

$$\alpha_L = \frac{E[L]^2}{var(L)} = \frac{\left[\sum_{g\in\mathcal{G}} n_g\theta_g\right]^2}{\sum_{g\in\mathcal{G}} n_g\theta_g^2} \tag{22}$$

$$\beta_L = \frac{var(L)}{E[L]} = \frac{\sum_{g\in\mathcal{G}} n_g\theta_g^2}{\sum_{g\in\mathcal{G}} n_g\theta_g}, \tag{23}$$

where the mean segment length θ_g from an ancestor in generation g is $100/d_g$ and d_g is the degree of the relationship between the two relatives i and j that passes through an ancestor in generation g.

Given that $\{n_g\}_{g\in\mathcal{G}} \equiv \mathbf{n}$ segments are transmitted from each generation $g \in \mathcal{G}$ and that the degrees of the induced relationships are $\{d_g\}_{g\in\mathcal{G}}$, we find that the observed number and total length of IBD has the distribution

$$f_{L|\mathbf{n}}(L|\mathbf{n}) = \frac{1}{\Gamma(\alpha_L)\beta_L^{\alpha_L}} L^{\alpha_L-1} e^{-L/\beta_L}, \tag{24}$$

where α_L and β_L are given by Equations (22) and (23).

We now need to know the probability of $\mathbf{n} \equiv \{n_g\}_{g\in\mathcal{G}}$ conditional on $\mathbf{k} \equiv \{k_g\}_{g\in\mathcal{G}}$ and the fact that each ancestor in $\mathbf{k}$ contributed at least one segment longer than τ. Since there are $E[K_g]$ ancestors on average in generation g and an expected number $E[S_g]$ of IBD segments, that means that each ancestor contributes $E[S_g]/E[c_g]$ segments on average, where c_g is the number of genealogical common ancestors in generation g. If there are k_g ancestors in generation g then the expected number of contributed segments, beyond the k_g obligatorily contributed segments is $k_g E[S_g]/E[c_g]$, where c_g is the expected number of distinct common ancestors in generation g.

Putting this all together, we find that

$$f(n, L; \mathbf{k}) \quad = \quad \sum_{\mathbf{n}\ :\ n_g \geq k_g,\ n=\sum_{g\in\mathcal{G}} n_g} f(L|\mathbf{n}; \mathbf{k})\mathbb{P}(\mathbf{n}; \mathbf{k})$$

$$= \sum_{\mathbf{n}\,:\,n_g \geq k_g,\ n=\sum_{g\in\mathcal{G}} n_g} \frac{1}{\Gamma(\alpha_L)\beta_L^{\alpha_L}} L^{\alpha_L-1} e^{-L/\beta_L} \prod_{g\in\mathcal{G}} \mathbb{P}(n_g; k_g)$$

$$= \sum_{\mathbf{n}\,:\,n_g \geq k_g,\ n=\sum_{g\in\mathcal{G}} n_g} \frac{1}{\Gamma(\alpha_L)\beta_L^{\alpha_L}} L^{\alpha_L-1} e^{-L/\beta_L} \prod_{g\in\mathcal{G}} \frac{\gamma_g^{n_g} e^{-\gamma_g}}{n_g!} \frac{1}{1-e^{-\gamma_g}\sum_{j=1}^{k_g}\gamma_g^j/j!} \quad (25)$$

where we have modeled the n_g segments from the k_g ancestors in generation g as Poisson distributed with mean γ_g, conditional on the event that $n_g \geq k_g$. Here, γ_g is given by

$$\gamma_g = k_g 2^{1-2g}(2rg + c)e^{-g\tau/50}. \quad (26)$$

We can also use a value of γ_g that more accurately accounts for the fact that each common ancestor is actually connected to their descendants through multiple lineages. Such a value of γ_g is useful when working with populations with background IBD in which the expected number of segments transmitted from an ancestor in generation g can be much higher than in a large population with low amounts of background IBD. In this case, each common ancestor in generation g contributes $E[S_g]/E[c_g]$ segments, on average, where c_g is the number of common ancestors shared between i and j in generation g in the past. Since there are k_g ancestors in our estimator, this gives

$$\gamma_g = k_g \frac{E[S_g]}{E[c_g]}. \quad (27)$$

The quantity c_g is derived in Appendix A.

Given counts $\mathbf{k} = \{k_g\}_{g\in\mathcal{G}}$ of IBD-contributing ancestors in each generation in the past, Equation (25) gives the probability of observing n segments of total length L. By viewing Equation (25) as a likelihood, we can infer the most likely set of ancestral relationships that gave rise to the observed IBD and by employing the prior in Equation (19), we can obtain a Bayesian estimate. Since the parameter space of possible relationships is enormous, this approach is likely to be infeasible for relationships sharing more than a few segments. We leave it to future work to make versions or approximations of this estimator that are more efficient.

6. Discussion

In this paper, we have shown that existing relationship estimators that do not condition on the event that IBD is observed between a pair of relatives produce biased estimates, inferring all sufficiently distant relationships to be ten degrees. This is a fundamental property of the likelihood of the segment data and it affects all estimators that are explicitly or implicitly based on unconditional distributions of segment lengths or the total IBD. We have also demonstrated that IBD-sharing relationships of degree greater than ten are ubiquitous, amounting to a large percentage of all relationships that are detectable in the population.

Because relationship estimators all demonstrate this bias and because it was supposed that distantly-related individuals are very unlikely to share IBD at all, it has generally been assumed that most relationships are within ten degrees (5^{th} cousins) and that relationships beyond seventeen degrees (8^{th} cousins) are simply undetectable. The belief that relationships beyond 8^{th} cousins were undetectable is evidenced by the fact that relationship estimators used by direct-to-consumer genetic testing companies do not attempt to detect more distant relationships. For example, 23andMe

considers IBD sharing down to 20 cM, calling all other relationships as "distant cousins" while the estimators used by AncestryDNA are trained only for 10^{th} cousins and closer [Ball et al., 2016]. The fact that existing estimators did not detect very distant relationships was taken as evidence of a fundamental truth about relatedness, rather than raising suspicions about the estimators themselves.

The distribution we obtain for the expected number of detectable-IBD-contributing common ancestors at each generation in the past suggests that detectable ancestral relationships sharing a common ancestor 100 or more generations in the past are likely to be common, especially in large genotyping datasets with millions of sampled individuals. The fact that shared segments can be detected from deep relationships is probably not news to the community of researchers working on problems involving deep coalescence. After all, methods such as the PSMC [Li and Durbin, 2011] have leveraged these kinds of deep relationships for years. However, there has been a disconnect between this coalescent-style research and pedigree inference.

The implication of this work is that a large fraction of distant relationship estimates reported by direct-to-consumer genetic testing companies are simply incorrect. Individuals do indeed have many thousands of fifth cousins as these platforms report. However, a large proportion of relatives that are reported as fifth cousins are in fact much more distant.

The reality is perhaps more interesting than the current incorrect estimates imply: we can detect relationships with common ancestors who lived before major world migrations such as European contact in the Americas and the Transatlantic slave trade and before events such as the rise of the Roman Empire and perhaps even the building of the Pyramids at Giza. This finding is particularly important for projects connecting present day people of African ancestry in the United states with relatives living in Africa with whom they share a common ancestor prior to or during the Transatlantic Slave Trade [David, 2023, 2024]. These studies provide a means of uncovering a genealogical history that was lost due to slavery.

Although the variance in distant estimates is high, estimates can be used in aggregate to obtain a more accurate picture of the ancestral connections of an individual, as well as the interrelationships among populations over a timespan of several thousand years. The tapestry of relationships that we can infer may in fact be quite rich.

7. Acknowledgements

I would like to thank the employees and research participants of 23andMe who made this research possible. I am especially grateful to Amy L. Williams, William A. Freyman and David A. Hinds for their insightful comments and thoughtful reviews of this manuscript and Peter R. Wilton for insightful points raised in discussions. Funding for this work was provided by NIH grant R35 GM133805 and by 23andMe, Inc. Members of the 23andMe Research Team are Stella Aslibekyan, Adam Auton, Elizabeth Babalola, Robert K. Bell, Jessica Bielenberg, Ninad S. Chaudhary, Zayn Cochinwala, Sayantan Das, Emily DelloRusso, Payam Dibaeinia, Sarah L. Elson, Nicholas Eriksson, Chris Eijsbouts, Teresa Filshtein, Pierre Fontanillas, Davide Foletti, Will Freyman, Zach Fuller, Julie M. Granka, Chris German, Éadaoin Harney, Alejandro Hernandez, Barry Hicks, David A. Hinds, M. Reza Jabalameli, Ethan M. Jewett, Yunxuan Jiang, Sotiris Karagounis, Lucy Kaufmann, Matt Kmiecik, Katelyn Kukar, Alan Kwong, Keng-Han Lin, Yanyu Liang, Bianca A. Llamas, Aly Khan,

Steven J. Micheletti, Matthew H. McIntyre, Meghan E. Moreno, Priyanka Nandakumar, Dominique T. Nguyen, Jared O'Connell, Steve Pitts, G. David Poznik, Alexandra Reynoso, Shubham Saini, Morgan Schumacher, Leah Selcer, Anjali J. Shastri, Jingchunzi Shi, Suyash Shringarpure, Keaton Stagaman, Teague Sterling, Qiaojuan Jane Su, Joyce Y. Tung, Susana A. Tat, Vinh Tran, Xin Wang, Wei Wang, Catherine H. Weldon, and Peter Wilton.

References

C.A. Ball, M.J. Barber, J. Byrnes, P. Carbonetto, K.G. Chahine, R.E. Curtis, J.M. Granka, E. Han, E.L. Hong, A.R. Kermany, N.M. Myres, K. Noto, J. Qi, K. Rand, Y. Wang, and L. Willmore. Rapid forward-in-time simulation at the chromosome and genome level. *https://www.ancestry.com/dna/resource/whitePaper/AncestryDNA-Matching-White-Paper.pdf*, 2016.

M. Caballero, D.N. Seidman, Y. Qiao, J. Sannerud, T.D. Dyer, D.M. Lehman, J.E. Curran, R. Duggirala, J. Blangero, S. Carmi, and Williams A.L. Crossover interference and sex-specific genetic maps shape identical by descent sharing in close relatives. *PLoS Genet.*, 15:e1007979, 2019.

Shai Covo and Amir Elalouf. A novel single-gamma approximation to the sum of independent gamma variables, and a generalization to infinitely divisible distributions. *Electronic Journal of Statistics*, 8(1):894 – 926, 2014. doi: 10.1214/14-EJS914. URL `https://doi.org/10.1214/14-EJS914`.

L. T. David. Addressing the feasibility of people of african descent finding living african relatives using direct-to-consumer genetic testing. *American Journal of Biological Anthropology*, 181(2): 163–165, 2023.

L.T. David. Supporting the use of genetic genealogy in restoring family narratives following the transatlantic slave trade. *Am Anthropol.*, 126:153–157, 2024.

B.M. Henn, L. Hon, J.M. Macpherson, N. Eriksson, S. Saxonov, I. Pe'er, and J.L. Mountain. Cryptic distant relatives are common in both isolated and cosmopolitan genetic samples. *PLOS One.*, 7: e34267, 2012.

L.J. Howe, M.G. Nivard, T.T. Morris, A.F. Hansen, and H. et al. Rasheed. Within-sibship genome-wide association analyses decrease bias in estimates of direct genetic effects. *Nature Genetics*, 54: 592A–592M, 2022. doi: 10.1038/s41588-022-01062-7.

C.D. Huff, D.J. Witherspoon, T.S. Simonson, J. Xing, W.S. Watkins, Y. Zhang, T.M. Tuohy, D.W. Neklason, R.W. Burt, S.L. Guthery, S.R. Woodward, and L.B. Jorde. Maximum-likelihood estimation of recent shared ancestry (ersa). *Genome Research*, 21:768–774, 2011.

E.M. Jewett. Simulating pedigrees ascertained on the basis of observed ibd sharing. 2024.

E.M. Jewett, K.F. McManus, W.A. Freyman, and A. Auton. Bonsai: An efficient method for inferring large human pedigrees from genotype data. *Am. J. Hum, Genet.*, 108:2052–2070, 2021.

A. Ko and R. Nielsen. Composite likelihood method for inferring local pedigrees. *PLOS Genet.*, 13: e1006963, 2017.

Heng Li and Richard Durbin. Inference of human population history from individual whole-genome sequences. *Nature*, 475(7357):493–496, 2011. doi: 10.1038/nature10231. URL `https://doi.org/10.1038/nature10231`.

A. Manichaikul, J.C. Mychaleckyj, S.S. Rich, K. Daly, M. Sale, and W.M. Chen. Robust relationship inference in genome-wide association studies. *Bioinformatics*, 26:2867–2873, 2010.

G. A. T. McVean, S. R. Myers, S. Hunt, P. Deloukas, D. R. Bentley, and P. Donnelly. The fine-scale structure of recombination rate variation in the human genome. *Science*, 304:581–584, 2004.

M.D. Ramstetter, S.A. Shenoy, T.D. Dyer, D.M. Lehman, J.E. Curran, R. Duggirala, J. Blangero, J.G. Mezey, and A.L. Williams. Inferring identical-by-descent sharing of sample ancestors promotes

high-resolution relative detection. *Am. J. Hum. Genet.*, 103:30–44, 2018.

S.M. Ross. *A First Course in Probability*. Prentice Hall, 2002. ISBN 9780130338518. URL https://books.google.com/books?id=hHgpAQAAMAAJ.

J. Staples, D. Qiao, M.H. Cho, E.K. Silverman, University of Washington Center for Mendelian Genomics, D.A. Nickerson, and J.E. Below. PRIMUS: Rapid reconstruction of pedigrees from genome-wide estimates of identity by descent. *Am. J. Hum. Genet.*, 95:553–564, 2014.

J. Staples, D.J. Witherspoon, L.B. Jorde, D.A. Nickerson, University of Washington Center for Mendelian Genomics, J.E. Below, and C.D. Huff. PADRE: Pedigree-aware distant-relationship estimation. *Am. J. Hum. Genet.*, 0:https://doi.org/10.1101/2020.02.25.965376, 2016.

Jeffrey Staples, Evan K. Maxwell, Nehal Gosalia, Claudia Gonzaga-Jauregui, Christopher Snyder, Alicia Hawes, John Penn, Ricardo Ulloa, Xiaodong Bai, Alexander E. Lopez, Cristopher V. Van Hout, Colm O'Dushlaine, Tanya M. Teslovich, Shane E. McCarthy, Suganthi Balasubramanian, H. Lester Kirchner, Joseph B. Leader, Michael F. Murray, David H. Ledbetter, Alan R. Shuldiner, George D. Yancoupolos, Frederick E. Dewey, David J. Carey, John D. Overton, Aris Baras, Lukas Habegger, and Jeffrey G. Reid. Profiling and leveraging relatedness in a precision medicine cohort of 92,455 exomes. *Am. J. Hum. Genet.*, 102(5):874–889, 2018. ISSN 0002-9297. doi: https://doi.org/10.1016/j.ajhg.2018.03.012. URL https://www.sciencedirect.com/science/article/pii/S0002929718301010.

A. Thomas, M. H. Skolnick, and C. M. Lewis. Genomic mismatch scanning in pedigrees. *Math Med Biol*, 11:1–16, 1994.

B.F. Voight and J.K. Pritchard. Confounding from Cryptic Relatedness in Case-Control Association Studies. *PLoS Genet.*, 2005. URL https://doi.org/10.1371/journal.pgen.0010032.

A.L. Williams. 2024. URL https://hapi-dna.org/2020/11/how-often-do-two-relatives-share-dna-2/.

Appendix A. The expected number of distinct common ancestors

Here, we derive the distribution of the number of distinct common ancestors that two present-day people share in generation g in the past. Let a_g denote the number of *distinct* ancestors in generation g. Assuming a randomly-mating population in which the two parents of each individual in generation $g-1$ are chosen uniformly at random from among the N individuals in the population in generation g, we model a_g as the number of distinct draws from a population of N distinct individuals when $2a_{g-1}$ samples are taken with replacement. We now develop a recurrence relation that approximates the expected number of distinct common ancestors shared between two people in generation g in the past.

Let δ_g denote the number of distinct ancestors of individual i in generation g that are not common ancestors of individual j. Let c_g denote the number of distinct common ancestors shared between individuals i and j in generation g. Finally, let $\tilde{\delta}_g$ denote the number of distinct ancestors of the δ_{g-1} individual ancestors of i in generation $g-1$ before considering mergers with ancestors of j and let $\tilde{c}_g$ denote the number of distinct ancestors of the c_{g-1} common ancestors before considering mergers among individuals. Thus $\tilde{\delta}_g$ is the number of distinct objects when drawing $2\delta_{g-1}$ samples

with replacement from a population of N individuals and $\tilde{c}_g$ is the number of distinct individuals selected when drawing $2c_{g-1}$ objects with replacement from a population of N individuals.

The quantity $\tilde{a}_g$ can be obtained from the quantity a_{g-1} by noting that the probability that any given one of the N individuals in generation g is a parent of one of the a_{g-1} individuals in generation $g-1$ with probability $1-(1-1/N)^{2a_{g-1}}$. This probability comes from the fact that the individual is selected in a single draw from the population with probability $1/N$. Thus the probability that they are not selected in any of the $2a_{g-1}$ draws is $(1-1/N)^{2a_{g-1}}$. It follows that $\tilde{a}_g$ is binary with probability $1-(1-1/N)^{2a_{g-1}}$ and expectation $E[\tilde{a}_g|a_{g-1}] = N[1-(1-1/N)^{2a_{g-1}}]$. Similarly, $E[\tilde{\delta}_g] = N[1-(1-1/N)^{2\delta_{g-1}}]$ and $E[\tilde{c}_g] = N[1-(1-1/N)^{2c_{g-1}}]$.

When the population size is large, the quantities δ_g, $\tilde{\delta}_g$, c_g and $\tilde{c}_g$ behave quasi-deterministically and we will assume that they follow their expected values. Thus, in generation g, the $\tilde{\delta}_g$ distinct non-common ancestors of i overlap with one of the $\tilde{c}_g$ distinct common ancestors at rate $\tilde{\delta}_g\tilde{c}_g/N^2$ and they overlap with one of the $\tilde{\delta}_g$ distinct ancestors of individual j at rate $\tilde{\delta}_g^2/N^2$. Therefore, the expected number of distinct and common ancestors at generation g is

$$\begin{aligned} E[\delta_g] &\approx E[\tilde{\delta}_g] - N\frac{E[\tilde{\delta}_g]}{N}\left(\frac{E[\tilde{c}_g]}{N} + \frac{E[\tilde{\delta}_g]}{N}\right) \\ E[c_g] &\approx E[\tilde{c}_g] + N\left(\frac{E[\tilde{\delta}_g]}{N}\right)^2, \end{aligned} \tag{28}$$

where

$$\begin{aligned} E[\tilde{\delta}_g] &\approx N\left[1-\left(1-\frac{1}{N}\right)^{2E[\delta_{g-1}]}\right] \\ E[\tilde{c}_g] &\approx N\left[1-\left(1-\frac{1}{N}\right)^{2E[c_{g-1}]}\right]. \end{aligned} \tag{29}$$

Repeatedly applying this recursion gives an approximation of the number of distinct common ancestors at each generation in the past.