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 8th cousins) and that practical relationship estimates have a ceiling around ten degrees (approximately 5th 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. Moreover, 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 may be forty degrees or more, rather than ten, extending the time-to-common ancestor from approximately 200 years in the past to approximately 600 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 and the Transatlantic Slave Trade, and possibly much earlier.

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.

Failure to condition on observing at least one IBD segment is 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.

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 individual has many ancestors at each generation in the not-too-distant past (Figure 1A) and each pair of individuals has many common ancestors at each generation in the past (Figure 1B). 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 are actually quite high.

Figure 1C shows an approximation of the fraction of distinct IBD-transmitting common ancestors shared by two individuals who lived less than g generations in the past for various values of g and various effective population sizes. This distribution is similar to that obtained by Wilton [2022], who derived the probability that two n^{th} cousins share IBD through the ancestor of that relationship, rather than from an older ancestor. The distribution shown in Figure 1C provides an approximation of the fraction of an individual's IBD-sharing relatives whose IBD arose from a common ancestor living in each past generation.

From Figure 1C, we see that for effective population sizes N that are reasonable for human populations (e.g., $1,000 \lesssim 2N \lesssim 10^5$) [Terhorst, 2024] between 20% and 60% of all IBD-sharing relationships are through an ancestor who lived more than ten generations in the past (approximately 20 degrees). In reality, the distribution in Figure 1 is likely to be shifted more to the right, perhaps considerably so, because the distribution shown does not take into account the fact that each individual can have many thousands of lineages connecting them to each of their common ancestors. Moreover, the actual human population may have contained millions of individuals in the past, allowing many more ancestors than the effective population size. Therefore, we show curves for population sizes of $2N = 10^6$ and $2N = 10^7$ as well, which may be closer to the historical census population size.

The take-home message of Figure 1 is that IBD-sharing relationships greater than 20 degrees are likely to occur frequently. These relationships are all inferred as tenth-degree relationships by

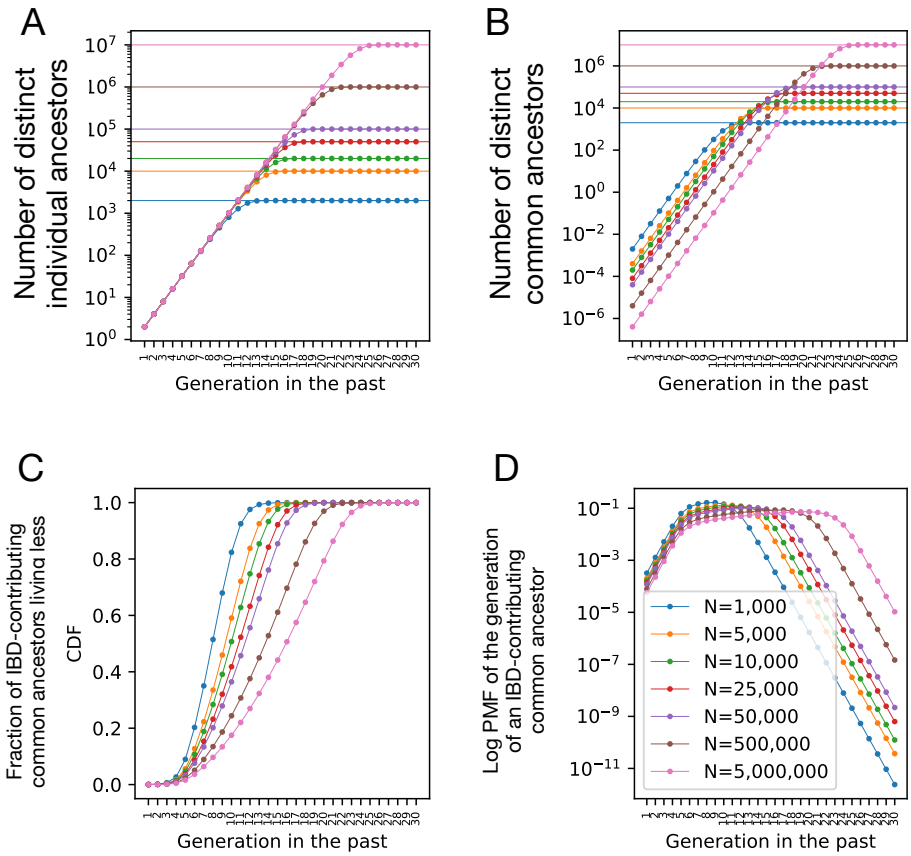

FIGURE 1. (A) The expected number of distinct ancestors of an individual at each generation in the past for various population sizes N. (B) The expected number of distinct common ancestors shared by two people. (C) The cumulative distribution function of the generation in which an IBD-transmitting ancestor lived relative to two putative relatives living in the present generation. (D) The of the fraction of IBD-contributing common ancestors in each generation in the past. Each panel shows several different population sizes that are plausible for different human populations ($2,000 < 2N < 10^5$) as well as larger sizes of $2N = 10^6$ and $2N = 10^7$ that may be closer to historical census population sizes.

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 4.1.

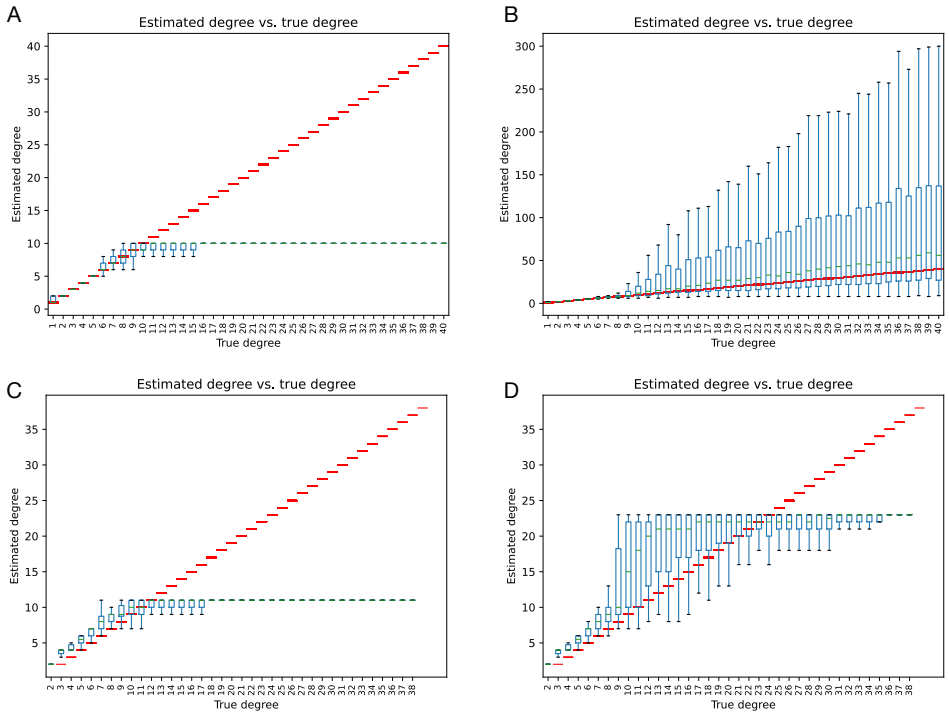

FIGURE 2. Inferred degree using unconditional and conditional estimators for relationships between 1 and 40 degrees. (A) The unconditional likelihood. (B) The conditional likelihood (Equation 8). (C) the unconditional likelihood together with the prior in Equation (19), setting $N = 10,000$. (D) The conditional likelihood together with the prior in Equation (19), setting $N = 10,000$.

From Figure 2A, 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 2 shows estimates when all IBD segments are detectable. However, setting a threshold on the minimum observable 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 2B 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 1C and 1D) 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 2C and the accuracy of the conditional Bayesian estimator is shown in Figure 2D for an effective population size of $2N = 20,000$.

From Figures 2C and 2D, it can be seen that the unconditional estimator continues to have a ceiling at ten or eleven degrees, while the estimates for lower degrees become biased upward. The prior introduces considerable bias into the conditional estimator as well, with the counterpoint that the mean squared error is dramatically reduced and all estimates are constrained to lie within a range that is more plausible for human populations. In particular, the highest degree estimate is around 25 degrees rather than several hundred degrees.

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) is likely to be biased towards lower values of g because it does not account for the many possible lineages that can connect each modern person to each of their ancestors and because it constrains the total number of ancestors to the effective population size rather than the full census population size.

3.4. **The total length of IBD.** Figure 2 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 4.2, 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 no IBD was observed.

Figure 3A shows the distribution of the total length L of IBD for $a = 1$ common ancestors and several small values of m. For these 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 for these values of m so 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 3B and 3C. For values of m in this range, the conditional and unconditional distributions begin to diverge from one another and begin to take on qualitatively different behavior.

Figure 3B explains why the likelihood estimator in Figure 2A 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

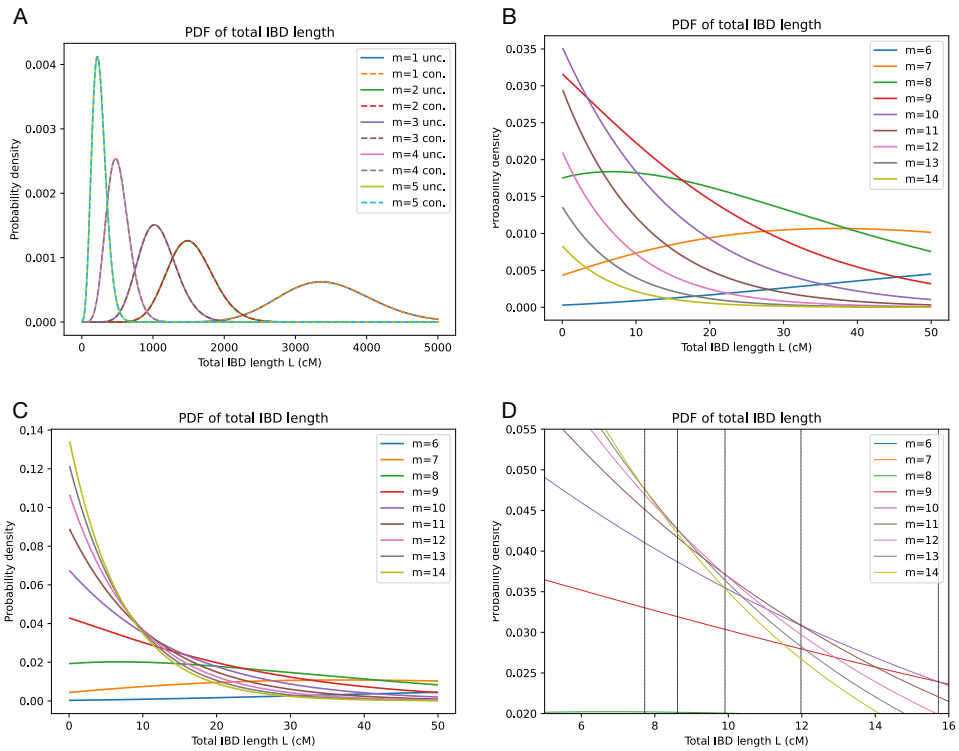

FIGURE 3. The distribution of the total length of IBD. (A) The distribution of the total length of IBD for $a = 1$ ancestors and several values of the number m of meioses separating two putative relatives. Both the unconditional (unc.) and conditional (con.) distributions are shown. (B) The unconditional 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 $d = m + 1 - a + 1$.

surface for all higher values of m is uniformly lower. Thus, the asymptote in Figures 2A and 2C at $d = m - a + 1 = 10$ degrees is a fundamental property of the likelihood. Moreover, all existing 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 3C and 3D) 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 3D.

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 use the ranges $(L_{d+1}, L_d]$, which can be computed 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 2A). Kinship coefficients are also unconditional on IBD sharing. In Section 4.3, 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 1.

TABLE 1. 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	0.00	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 1, 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 in which each degree d is the most likely degree.

4. Updating the maximum likelihood relationship estimator

Before deriving a new conditional relationship estimator, it is useful to clarify what we mean by inferring "the relationship" between a pair of individuals. Existing relationship estimators are derived under the conceptual model shown in Figure 4A. In this model, two individuals, i and j are related through a single recent common ancestor, a, or through a mating pair of recent common ancestors (a_1, a_2). Individuals i and j may have other very distant ancestors (grey circles in 4A) 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. Instead, background IBD segments produce noise that confounds the signal of the close relationship (orange curve) on which we are focusing.

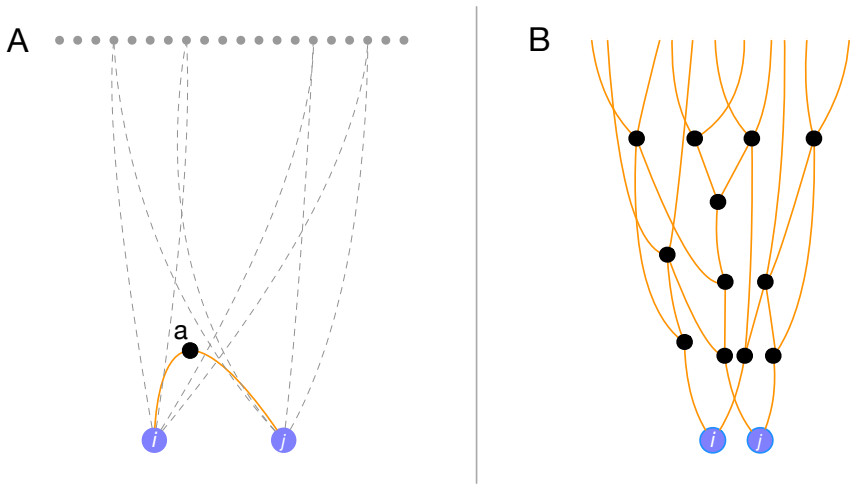

Figure 4. Conceptual models for the development of relationship estimators. Panel A shows the conceptual model that underlies existing relationship estimators. In this model, each pair of individuals, i and j (purple dots), shares a single common ancestor, a, or a single mating pair of common ancestors (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 observable 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 4B.

The goal of relationship inference under the model in Figure 4B 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 observable shared IBD, (3) the number of genealogical ancestors at each generation in the past, or some other suitable quantity that reflects the fact that individuals share many common ancestors through many different relationships.

In this manuscript, we conceptualize inheritance under the model in Figure 4B and our goal is to infer Statistic (2): the genealogically closest relationship that contributed observable IBD. We have chosen this quantity to infer because it seems conceptually close to the idea of inferring "the relationship" between a pair of IBD-sharing individuals.

4.1. **The conditional likelihood.** We now derive the conditional 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. 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

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

$$= \sum_{i=1}^{n} \mathbb{P}(\ell^{(1)}, ..., \ell^{(n_d)}; a, m)\mathbb{P}_b(\ell^{(n_d+1)}, ..., \ell^{(n)}) \frac{\mathbb{P}(n_d = i, n_b = n - i, O; a, m)}{\mathbb{P}(O; a, m)}$$

$$= \sum_{i=1}^{n} \mathbb{P}(\ell^{(1)}, ..., \ell^{(n_d)}; a, m)\mathbb{P}_b(\ell^{(n_d+1)}, ..., \ell^{(n)}) \frac{\mathbb{P}(n_d = i; a, m)\mathbb{P}(n_b = n - i)}{\mathbb{P}(O; a, m)} \tag{1}$$

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:

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

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 observable 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:

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

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.

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

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

[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.

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 observable 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.

One could also imagine writing Equation (8) in a different way that would be even more consistent with the concept that two individuals share many recent ancestors. In particular, we could assign each observed IBD segment to a different ancestor. However, Equation (8) allows us to make comparisons directly to existing likelihoods.

4.2. **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:

$$\begin{aligned} &f_{L|O}(L|O) \\ &= \sum_{n=1}^{\infty} f_{L,n|O}(L, n|O) \end{aligned}$$

$$= \sum_{n=1}^{\infty} \frac{\lambda_{a,m}^n}{\Gamma(n)} (L - n\tau)^{n-1} e^{-\lambda_{a,m}(L-n\tau)} \frac{\eta_{a,m,\tau}^n}{\Gamma(n+1)} e^{-\eta_{a,m,\tau}} \frac{1}{1 - e^{-\eta_{a,m,\tau}}}. \tag{10}$$

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 3.

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.

4.3. **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 (1).

4.4. **A prior distribution of the number of generations to the most recent segment-contributing common ancestor.** Equation (8) allows us to obtain a maximum likelihood estimate of the relationship, given the observed IBD segments. However, it may also be of interest to perform inference in a Bayesian context because the relative probability of observing a segment-contributing common ancestor at each generation in the past can be highly informative.

The prior distribution on the number of generations to the most recent IBD-contributing common ancestor of a pair of individuals can be obtained by considering the number of ancestors each person has in each prior generation. This prior distribution is also useful because it provides us with a sense of the most distant relationships we are likely to encounter in practice. Moreover, it can be used to gain a sense of the number of relatives of varying degrees with whom each person shares IBD.

Suppose the population has a constant number $2N$ of individuals in each generation, half of whom are male and half of whom are female. Ignoring pedigree collapse, the total number of female ancestors in generation g is $a_{g,f} = 2^{g-1}$ and the total number of male ancestors is $a_{g,m} = 2^{g-1}$. However, because of pedigree collapse, some of these ancestors are the same individuals.

Let $\bar{a}_{g,f}$ and $\bar{a}_{g,m}$ respectively denote the number of *distinct* female and male ancestors in generation g. Assuming a randomly-mating population, half of whom are female, we model $\bar{a}_{g,f}$ as

the number of distinct draws from a population of N distinct objects when $a_{g,f}$ samples are taken with replacement.

Considering all N female ancestors in generation g, let S_i^g be the event that individual ancestor i (out of N) is sampled when drawing $a_{g,f} = 2^{g-1}$ individuals. Under 2^{g-1} draws with replacement, the probability that a given one of the N possible ancestral individuals is chosen is

$$\mathbb{P}(S_i^g) = 1 - \left(1 - \frac{1}{N}\right)^{2^{g-1}}. \tag{15}$$

Therefore, the expected number of distinct female (or male) ancestors a person has g generations in the past is

$$E[a_{g,f}] = E[a_{g,m}] = NP(S_i^g) = N\left[1 - \left(1 - \frac{1}{N}\right)^{2^{g-1}}\right]. \tag{16}$$

The probability that an individual female ancestor in generation g is the common ancestor of two specified present-day individuals is $\mathbb{P}(S_i)^2$. So the expected number $c_{g,f} = c_{g,m}$ of shared common male or female ancestors between two specified present-day individuals is

$$E[c_{g,f}] = E[c_{g,m}] = NP(S_i^g)^2 = N\left[1 - \left(1 - \frac{1}{N}\right)^{2^{g-1}}\right]^2. \tag{17}$$

As a special case of Equation (7), the probability that any one of these distinct common ancestors in generation g was a source of a detectable IBD segment (i.e., a segment longer than the minimum observable segment length τ) observed between the present-day pair is approximately

$$\mathbb{P}(O; g) \approx 1 - e^{-\eta_{g,\tau}}, \tag{18}$$

where, using Equation (6), the expected number of shared segments is $\eta_{g,\tau} = 2^{1-2g}(2rg + c)e^{-g\tau/50}$. Equation (18) is approximate because it doesn't account for the fact that there are multiple lineages between each modern person an each common ancestor. These additional lineages will tend to increase the number of segments transmitted to ancestors in the past.

The probability that two specified present-day individuals share an IBD-contributing common female (or male) ancestor g generations in the past is approximately proportional to the expected number ($\mathbb{P}(O; g)E[c_{g,f}]$) of IBD-contributing shared common female (or male) common ancestors in that generation. Let C denote the generation in which two people share a common ancestor, then

$$\mathbb{P}(C = g) \approx \frac{\mathbb{P}(O; g)(E[c_{g,f}] + E[c_{g,m}])}{\sum_{g=1}^{\infty} \mathbb{P}(O; g)(E[c_{g,f}] + E[c_{g,m}])} = \frac{\mathbb{P}(O; g)E[c_{g,f}]}{\sum_{g=1}^{\infty} \mathbb{P}(O; g)E[c_{g,f}]}. \tag{19}$$

In practice, we can compute this distribution by taking the summation in the denominator up to some suitably large number of generations $g \gg 1$.

The approximation in Equation (19) is biased towards lower values of g compared to the true distribution because it does not account for the fact that multiple lineages connect each descendant and ancestor and we have restricted the number of possible ancestors to the effective population size, whereas the number of ancestors can be as large as the census population size.

Note that the distribution in Equation (19) assumes that the present-day individuals exist in the same generation and that they have the same number of generations separating them from

their common ancestor. It is straightforward to update Equation (19) to allow modern individuals from different generations. Suppose that one individual lived Δg generations before the other, then counting g from the generation of the most recent individual we have

$$\begin{aligned} E[c_{g,\Delta g,f}] = E[c_{g,\Delta g,m}] &= NP(S_i^g)P(S_i^{g-\Delta g}) \\ &= N\left[1-\left(1-\frac{1}{N}\right)^{2^{g-1}}\right]\left[1-\left(1-\frac{1}{N}\right)^{2^{g-\Delta g-1}}\right]. \end{aligned} \tag{20}$$

Moreover, the probability that the common ancestor(s) in generation g contributed a detectable segment is

$$\mathbb{P}(O; g, \Delta g) = 1 - e^{-\eta_{g,\Delta g,\tau}}, \tag{21}$$

where the expected number of shared segments is $\eta_{g,\Delta g,\tau} = 2^{1-2g+\Delta g}(r(2g-\Delta g)+c)e^{-(2g-\Delta g)\tau/100}$. Thus, in the more general case in which the two putative relatives are in different generations we have

$$\mathbb{P}(C = g; \Delta g) = \frac{\mathbb{P}(O; g, \Delta g)E[c_{g,\Delta g,f}]}{\sum_{g=1}^{\infty}\mathbb{P}(O; g, \Delta g)E[c_{g,\Delta g,m}]}. \tag{22}$$

Typically, we do not know the value of Δg; however, we can estimate it from the ages of the putative relatives by assuming a fixed number of years per generation, or we can apply Equation (19) if the two individuals are in approximately the same generation.

5. 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 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 (5th cousins) and that relationships beyond seventeen degrees (8th cousins) are simply undetectable. The belief that relationships beyond 8th 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 10th 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 fraction of IBD-contributing common ancestors who lived in different generations in the past should be interpreted as a somewhat conservative approximation, the purpose of which is to demonstrate that IBD-sharing distant relationships are ubiquitous. It

should be used with caution as a prior for purposes of inference because it has a profound influence on estimates. The prior distribution does not account for multiple lineages between each modern individual and each of their ancestors and it assumes that the number of ancestral individuals is bounded by the effective population size, when in fact it is bounded by the ancestral census size. Thus, it almost certainly underestimates the relative fraction of distant IBD-sharing relationships.

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. The latter fact 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 approximately 500 years. The tapestry of relationships that we can infer may in fact be quite rich.

6. 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

Soheil Baharian, Maxime Barakatt, Christopher R. Gignoux, Suyash Shringarpure, Jacob Errington, William J. Blot, Carlos D. Bustamante, Eimear E. Kenny, Scott M. Williams, Melinda C. Aldrich, and Simon Gravel. The great migration and african-american genomic diversity. *PLOS Genetics*, 12(5):1–27, 05 2016. doi: 10.1371/journal.pgen.1006059. URL `https://doi.org/10.1371/journal.pgen.1006059`.

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.

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.

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.

S.J. Micheletti, K. Bryc, S.G. Ancona Esselmann, W.A. Freyman, M.E. Moreno, G.D. Poznik, A.J. Shastri, The 23andMe Research Team, S. Beleza, and J.L. Mountain. Genetic Consequences of the Transatlantic Slave Trade in the Americas. *Am J Hum Genet.*, 107:265–277, 2020.

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.

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`.

J. Terhorst. Accelerated bayesian inference of population size history from recombining sequence data. *bioRxiv*, 2024. URL `https://www.biorxiv.org/content/10.1101/2024.03.25.586640v1.full.pdf`.

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/`.

P. Wilton. Attributing ibd to cousin relationships. unpublished, 2022.