A Genealogical Look at Shared Ancestry on the X Chromosome

Vince Buffalo[*,1], Stephen M. Mount[†], and Graham Coop[*]

[*]Department of Evolution and Ecology, Center for Population Biology
University of California, Davis, CA 95616
[†]Department of Cell Biology and Molecular Genetics, Center for Bioinformatics and Computational Biology
University of Maryland, College Park, MD 20742
[1]Email for correspondence: vsbuffalo@ucdavis.edu

April 2, 2016

Abstract

Close relatives can share large segments of their genome identical by descent (IBD) that can be identified in genome-wide polymorphism datasets. There are a range of methods to use these IBD segments to identify relatives and estimate their relationship. These methods have focused on sharing on the autosomes, as they provide a rich source of information about genealogical relationships. We can hope to learn additional information about recent ancestry through shared IBD segments on the X chromosome, but currently lack the theoretical framework to use this information fully. Here, we fill this gap by developing probability distributions for the number and length of X chromosome segments shared IBD between an individual and an ancestor k generations back, as well as between half- and full-cousin relationships. Due to the inheritance pattern of the X and the fact that X homologous recombination only occurs in females (outside of the pseudo-autosomal regions), the number of females along a genealogical lineage is a key quantity for understanding the number and length of the IBD segments shared amongst relatives. When inferring relationships among individuals, the number of female ancestors along a genealogical lineage will often be unknown. Therefore, our IBD segment length and number distributions marginalize over this unknown number of recombinational meioses through a distribution of recombinational meioses we derive. We show how our results can be used to estimate the number of female ancestors between two relatives, giving us more genealogical details than possible with autosomal data alone.

Keywords: X chromosome, genetic genealogy, statistical genetics, identity by descent, recent ancestry

Close relatives are expected to share large contiguous segments of their genome due to the limited number of crossovers per chromosome each generation (Donnelly, 1983; Fisher, 1949, 1954). These large identical by descent (IBD) segments shared among close relatives leave a conspicuous footprint in population genomic data, and identifying and understanding this sharing is key to many applications in biology (Thompson, 2013). For example, in human genetics, evidence of recent shared ancestry is an integral part of detecting cryptic relatedness in genome-wide association studies (Gusev et al., 2009), discovering mis-specified relationships in pedigrees (Sun et al., 2002), inferring pairwise relationships (Glaubitz et al., 2003; Huff et al., 2011), and localizing disease traits in pedigrees (Thomas et al., 2008). In forensics, recent ancestry is crucial for both accounting for

population-level relatedness (Balding et al., 1994) and in familial DNA database searches (Belin et al., 1997; Sjerps et al., 1999). Additionally, recent ancestry detection methods have a range of applications in anthropology and ancient DNA to understand the familial relationships among sets of individuals (Baca et al., 2012; Fu et al., 2015; Haak et al., 2008; Keyser-Tracqui et al., 2003). In population genomics, recent ancestry has been used to learn about recent migrations and other demographic events (Palamara et al., 2012; Ralph et al., 2013). An understanding of recent ancestry also plays a large role in understanding recently admixed populations, where individuals draw ancestry from multiple distinct populations (Gravel, 2012; Liang et al., 2014; Pool et al., 2009). Finally, relative finding through recent ancestry is increasingly a key feature of direct-to-consumer personal genomics products and an important source of information for genealogists leveraging genetic information (Durand et al., 2014; Royal et al., 2010).

Approaches to infer recent ancestry among humans have often used only the autosomes, as the recombining autosomes offer more opportunity to detect a range of relationships than the Y chromosome, mitochondria, or X chromosome. However, the nature of X chromosome inheritance means that it can clarify details of the relationships among individuals and be informative about sex-specific demography and admixture histories in ways that autosomes cannot (Bryc et al., 2010; Bustamante et al., 2009; Goldberg et al., 2015; Pool et al., 2007; Ramachandran et al., 2008, 2004; Shringarpure et al., 2016).

In this paper, we look at the inheritance of chromosomal segments on the X chromosome among closely related individuals. Our genetic ancestry models are structured around biparental genealogies back in time, an approach used by many previous authors (e.g., Barton et al., 2011; Chang, 1999; Donnelly, 1983; Rohde et al., 2004). If we ignore pedigree collapse (due to inbreeding), the genealogy of a present-day individual encodes all biparental relationships back in time; e.g. the two parents, four grandparents, eight great-grandparents, 2^k great^{k-2} grandparents, and in general the 2^k ancestors k generations back; we refer to these individuals as one's *genealogical ancestors*. A genealogical ancestor of a present-day individual is said to also be a *genetic ancestor* if the present-day individual shares genetic material by descent from this ancestor. We refer to these segments of shared genetic material as being identical by descent, and in doing so we ignore the possibility of mutation in the limited number of generations separating our individuals. Throughout this paper, we will ignore the pseudo-autosomal (PAR) region of the X chromosome, which undergoes crossing over with the Y chromosome in males (Koller et al., 1934) to ensure proper disjunction in meiosis I (Hassold et al., 1991). We also ignore gene conversion which is known to occur on the X (Rosser et al., 2009).

Here, we are concerned with inheritance through the *X genealogy* embedded inside an individual's genealogy, which includes only the subset of one's genealogical ancestors who could possibly contributed to one's non-PAR X chromosome. We refer to the individuals in this X genealogy as *X ancestors*. Since males receive an X only from their mothers, a male's father cannot be an X ancestor. Consequently, a male's father and all of his ancestors are excluded from the X genealogy (Figure 1). Therefore, females are overrepresented in the X genealogy, and as we go back in one's genealogy, the fraction of individuals who are possible X ancestors shrink. This property means that genetic relationships differ on the X compared to the autosomes, a fact that changes the calculation of kinship coefficients on the X (Pinto et al., 2011, 2012) and also has interesting implications for kin-selection models involving the X chromosome (Fox et al., 2009; Rice et al., 2008).

In Section 1 we review models of autosomal identity by descent among relatives, on which we base our models of X genetic ancestry. Then, in Section 2 we look at X genealogies back in time,

Figure 1: An example genealogy with the embedded X genealogy back five generations. Males are depicted as squares and females as circles. Individuals in the X genealogy are shaded gray while unshaded individuals are ancestors that are not X ancestors. Each X ancestor is labeled with the number of recombinational meioses to the present-day female.

as the properties of X genealogies affect the transmission of X genetic material from X ancestors to a present-day individual. We develop simple approximations to the probability distributions of the number and length of X chromosome segments that will be shared IBD between a present-day female and one of her X ancestors a known number of generations back in time. These models provide a set of results for the X chromosome equivalent to those already known for the autosomes (Donnelly, 1983; Thomas et al., 1994). Then, in Section 3, we look at shared X ancestry—when two present-day cousins share an X ancestor a known number of generations back. We calculate the probabilities that genealogical half- and full-cousins are also connected through their X genealogy, and thus can potentially share genetic material on their X. We then extend our models of IBD segment length and number to segments shared between half- and full-cousins. Finally, in Section 4 we show that shared X genetic ancestry contains additional information (compared to genetic autosomal ancestry) for inferring relationships among individuals, and explore the limits of this information.

1   Autosomal Ancestry

To facilitate comparison with our X chromosome results, we first briefly review analogous autosomal segment number and segment length distributions (Donnelly, 1983; Huff et al., 2011; Thomas et al., 1994). Throughout this paper, we assume that one's genealogical ancestors k generations back are distinct, i.e. there is no pedigree collapse due to inbreeding (see Appendix D for a model of how this assumption breaks down with increasing k). Thus, an individual has 2^k *distinct* genealogical ancestors. Assuming no selection and fair meiosis, a present-day individual's autosomal genetic material is spread across these 2^k ancestors with equal probability, having been transmitted to the present-day individual solely through recombination and fair segregation.

We model the process of crossing over during meiosis as a continuous time Markov process along the chromosome, as in Thomas et al. (1994) and Huff et al. (2011), and described by Donnelly (1983). In doing so we assume no crossover interference, such that in each generation b recombinational breakpoints occur as a Poisson process running with a uniform rate equal to the total length of the genetic map (in Morgans), ν. Within a single chromosome, b breaks creates a mosaic of $b + 1$ alternating pieces of maternal and paternal segments. This alternation between maternal and paternal haplotypes creates long-run dependency between segments. We ignore these

3

dependencies in our analytic models by assuming that each chromosomal segment survives segregation independently with probability $1/2$ per generation. For d independent meioses separating two individuals, we imagine the Poisson recombination process running at rate νd, and for a segment to be shared IBD between the two ancestors it must survive $1/2^d$ segregations. Consequently, the expected number of segments shared IBD between two individuals d meioses apart in a genome with c chromosomes is approximated as (Thomas et al., 1994):

$$\mathbb{E}[N] = \frac{1}{2^d}(\nu d + c) \tag{1}$$

Intuitively, we can understand the $1/2^d$ factor as the coefficient of kinship (or path coefficient; Wright, 1922, 1934) of two individuals d meioses apart, which gives the probability that two alleles are shared IBD between these two individuals. Then, the expected number of IBD segments $\mathbb{E}[N]$ can be thought of as the average number of alleles shared between two individuals in a genome with $\nu d + c$ loci total. Under this approximation, recombination increases the number of independent loci linearly each generation (by a factor of the total genetic length). A fraction $1/2^d$ of parental alleles at these loci survive the d meioses to be IBD with the present-day individual.

By convention, we count the number of contiguous IBD segments N in the present-day individual, not the number of contiguous segments in the ancestor. For example, an individual will share exactly one block per chromosome with each parent if we count the contiguous segments in the offspring, even though these segments may be spread across the parent's two homologues. This convention, which we use throughout the paper, is identical to counting the number of IBD segments that occur in $d-1$ meioses rather than d meioses. This convention only impacts models of segments shared IBD between an individual and one of their ancestors; neither the distribution of segment lengths nor the distributions for segment number shared IBD between cousins are affected by this convention.

The distribution of IBD segments between a present-day individual and an ancestor Given that a present-day individual and an ancestor in the kth generation are separated by k meioses, the number of IBD segments can be modeled with what we call the *Poisson-Binomial* approximation. Over $d = k$ meioses, $B = b \sim \text{Pois}(\nu k)$ recombinational breakpoints fall on c independently assorting chromosomes, creating $b + c$ segments. Ignoring long-range dependencies, we assume all of these $b + c$ segments have an independent chance of surviving the k segregations to the present-day individual, and thus the probability that n segments survive given $b + c$ trials is Binomially distributed with probability $1/2^k$. Marginalizing over the unobserved number of recombinational breakpoints b, and replacing k with $k - 1$ to following the convention described above:

$$P(N = n|k) = \sum_{b=0}^{\infty} \text{Bin}(N = n \mid l = b + c, p = 1/2^{k-1}) \, \text{Pois}(B = b \mid \lambda = \nu(k - 1)) \tag{2}$$

The expected value of the Poisson-Binomial model is given by equation (1) and this model is similar to those of Donnelly (1983) and Thomas et al. (1994). We can further approximate this by assuming that we have a Poisson total number of segments with mean $(c + \nu(k - 1))$ and these segments are shared with probability $1/2^{k-1}$ as in Huff et al. (2011). This gives us a thinned Poisson distribution of shared segments:

$$P(N = n | k, \nu, c) = \text{Pois}(N = n | \lambda = (c + \nu(k-1))/2^{k-1})$$
$$= \frac{((c + \nu(k-1))/2^{k-1})^n e^{-(c+\nu(k-1))/2^{k-1}}}{n!} \tag{3}$$

This thinned Poisson model also has an expectation given by equation (1) but compared to the Poisson-Binomial model has a larger variance than the true process. This overdispersion occurs because modeling the number of segments created after b breakpoints involves incorporating the initial number of chromosomes into the Poisson rate. However, this initial number of chromosomes is actually fixed, which the Poisson-Binomial model captures but the Poisson thinning model does not (i.e. one generation back such that $k = 1$, the thinning model treats the number of segments shared IBD with one's parents is $N \sim \text{Pois}(c)$ rather than c). See Appendix A for a further comparison of these two models. A more formal description of this approximation as a continuous-time Markov process is given in Thomas et al. (1994).

The distribution of IBD segments between cousins Similarly, we can derive the distribution for the number of IBD segments shared between two half-cousins with an ancestor in the kth generation. Two half-cousins are separated by $2k$ meioses, thus the distribution for number of segments is:

$$P(N = n \mid k, \nu, c) = \text{Pois}\left(N = n \mid \lambda = (c + 2k\nu)/2^{2k-1}\right) \tag{4}$$

Since either of the shared parent's haplotypes can be shared IBD between the two cousins, the Poisson process rate is doubled. Full-cousins can share segments via either of their two shared ancestors, leading the distribution to be:

$$P(N = n \mid k, \nu, c) = \text{Pois}\left(N = n \mid \lambda = (c + 2k\nu)/2^{2k-2}\right) \tag{5}$$

The distribution of autosome segment lengths In addition to the number of IBD segments, the length of segments is also informative about ancestry (e.g. Palamara et al., 2012). As we model crossing over as a Poisson process, a one Morgan region will experience on average d recombination events over d meioses. Therefore, the probability density of segment lengths shared IBD between two individuals d meioses apart is exponential with rate d:

$$p(U = u | d) = de^{-du} \tag{6}$$

Equations (6) and (4) specify a model of the number and lengths of segments shared between various degree relatives. Various authors have used these types of results to derive likelihood-based models for classifying the genealogical relationship between pairs of individuals using autosome IBD data (Durand et al., 2014; Henn et al., 2012; Huff et al., 2011).

We will use similar models as these in modeling the length and number of X chromosome segments shared been relatives. However, the nature of X genealogies (which we cover in the next section) requires we adjust these models. Specifically, while one always has k recombinational meioses between an autosomal ancestor in the kth generation, the number of recombinational meioses varies across the lineages to an X ancestor with the number of females in a lineage, since X homologous

5

recombination only occurs in females (Figure 1). This varying number of recombinational meioses across lineages leads to a varying-rate Poisson recombination process, with the rate depending on the specific lineage to the X ancestor. After we take a closer look at X genealogies in the next section, we adapt the models above to handle the varying-rate Poisson process needed to model IBD segments in X genealogies.
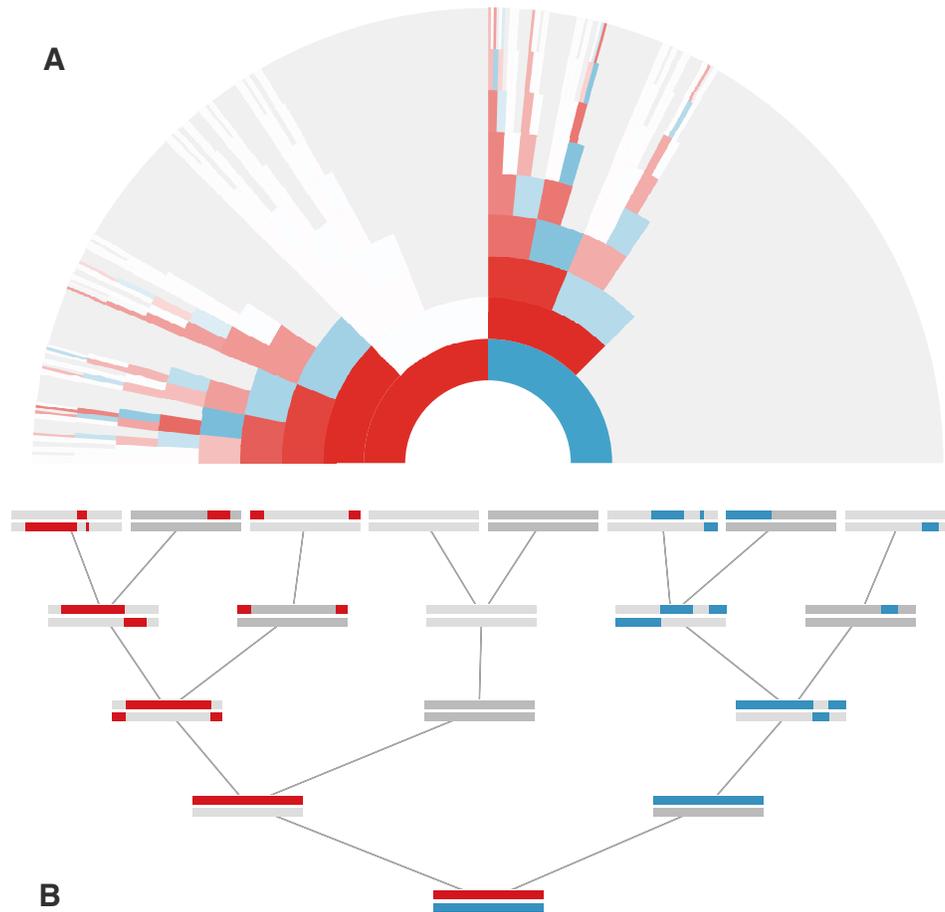

Figure 2: A: Simulated X genealogy of a present-day female, back nine generations. Each arc is an ancestor, with female ancestors colored red, and male ancestors colored blue. The transparency of each arc reflects the genetic contribution of this ancestor to the present-day female. White arcs correspond to an X genealogical ancestor that shares no genetic material with the present-day female, and gray arcs are genealogical ancestors that are not X ancestors. B: The X segments of the simulation in (A), back five generations. The maternal X lineage's segments are colored red, and the paternal X segments are colored blue. A male ancestor's sex chromosomes are colored dark gray (and include the Y) and a female ancestor's sex chromosomes are colored light gray.

2   X Ancestry
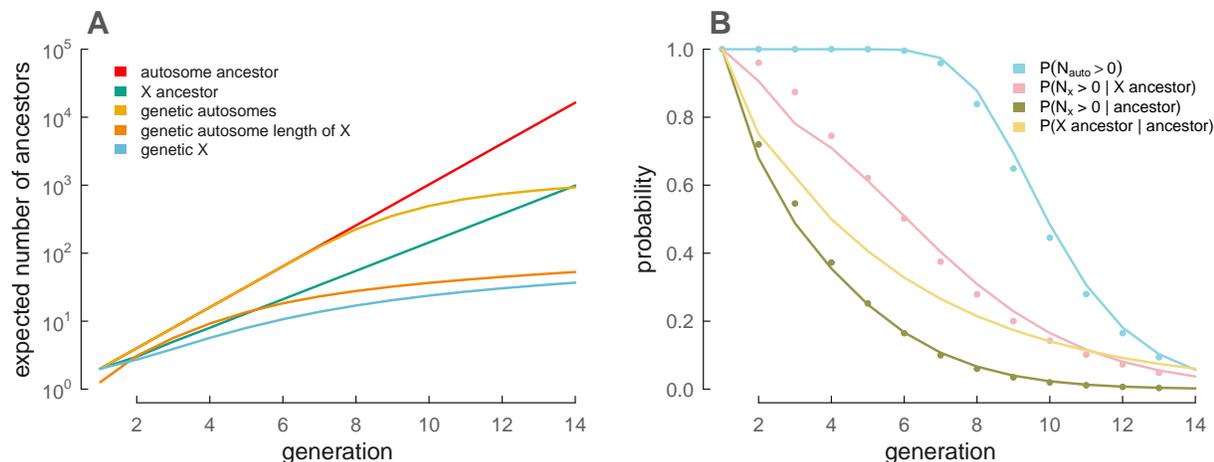
Figure 3: A: A present-day female's number of genealogical and genetic ancestors, for X chromosomes and autosomes. For comparison to an X chromosome, the number of genetic ancestors of an autosome of length equal to the X is included. B: The probability of genealogical and genetic ancestry for a variety of cases. $P(\text{N}_{\text{auto}} > 0)$ is derived from equation (3), $P(\text{N}_{\text{X}} > 0 \mid \text{X ancestor})$ from equation (12), $P(\text{N}_{\text{X}} > 0 \mid \text{ancestor})$ from equations (12) and (7), and $P(\text{X ancestor} \mid \text{ancestor})$ from equation (7). Points show simulated results.

Number of Genealogical X Ancestors   While a present-day individual can potentially inherit autosomal segments from any of its 2^k genealogical ancestors k generations back, only a fraction of these individuals can possibly share segments on the X chromosome. In contrast to biparental genealogies, males have only one genealogical X ancestor—their mothers—if we ignore the PAR. This constraint (which we refer to through as the *no two adjacent males condition*) shapes both the number of X ancestors and the number of females along an X lineage. For example, consider a present-day female's X ancestors one generation back: both her father and mother contribute X chromosome material. Two generations back, she has three X genealogical ancestors: her father only inherits an X from her paternal grandmother, while her mother can inherit X material from either of her maternal grandparents. Continuing this process, this individual has five ancestors three generations back and eight ancestors four generations back (Figure 1).

In general, a present-day female's X genealogical ancestors is growing as a Fibonacci series (Basin, 1963; Laughlin, 1920), such that k generations back she has $\mathcal{F}_{k+2}$ X genealogical ancestors, where $\mathcal{F}_k$ is the kth Fibonacci number (OEIS A000045; Sloane, 2010). We can demonstrate that one's number of X genealogical ancestors grows as a Fibonacci series by encoding the X inheritance rules for the number of males and females (m_k and f_k, respectively) in the kth generation as a set of recurrence relations:

$$f_k = n_{k-1} \qquad \text{every individual receives an X chromosome from his/her mother}$$
$$m_k = f_{k-1} \qquad \text{every female receives an X chromosome from her father}$$
$$n_k = f_k + m_k$$

Rearranging these recurrence equations gives us $n_k = n_{k-1} + n_{k-2}$, which is the Fibonacci recurrence. Starting with a female in the $k = 0$ generation, we have initial values $n_0 = 1$ and $n_1 = 2$, which gives us the Fibonacci numbers offset by two, $\mathcal{F}_{k+2}$. For a present-day male, his

7

number of X ancestors is $\mathcal{F}_{k+1}$, i.e. offset by one to count the number of X ancestors his mother has. To simplify our expressions, we will assume throughout the paper that all-present day individuals are female since a simple offset can be made to handle males.

In Figure 3A we show the increase in the number of X genealogical and genetic ancestors and compare these to the growth of all of one's genealogical ancestors and autosomal genetic ancestors. The closed-form solution for the kth Fibonacci number is given by Binet's formula, which shows that the Fibonacci sequence grows at an exponential rate slower than 2^k.

Consequently, the fraction of ancestors who can contribute to the X chromosome is declining. Given that a female has $\mathcal{F}_{k+2}$ X ancestors and 2^k genealogical distinct ancestors, her proportion of X ancestors is:

$$P(\text{X ancestor} \mid \text{ancestor}) = \frac{\mathcal{F}_{k+2}}{2^k} \tag{7}$$

This fraction can also be interpreted as the probability that a randomly chosen genealogical ancestor k generations ago is also an X genealogical ancestor. We show this probability as a function of generations into the past in Figure 3B (yellow line).

From our recurrence equations we can see that a present-day female's $\mathcal{F}_{k+2}$ ancestors in the kth generation are composed of $\mathcal{F}_{k+1}$ females and $\mathcal{F}_k$ males. Likewise for a present-day male, his $\mathcal{F}_{k+1}$ ancestors in the kth generation are composed of $\mathcal{F}_k$ females and $\mathcal{F}_{k-1}$ males. We will use these results when calculating the probability of a shared X ancestor.

Ancestry Simulations In the next sections, we use stochastic simulations to verify the analytic approximations we derive; here we briefly describe the simulation methods. We have written a C and Python X genealogy simulation procedure (source code available in the supplement and at https://github.com/vsbuffalo/x-ancestry/). We simulate a female's X chromosome genetic ancestry back through her X genealogy. Each simulation begins with two present-day female X chromosomes, one of which is passed to her mother and one to her father. Segments transmitted to a male ancestor are simply passed directly back to his mother (without recombination). For segments passed to a female ancestor, we place a Poisson number of recombination breakpoints (with mean ν) on the X chromosome and the segment is broken where it overlaps these recombination events. The first segment along the chromosome is randomly drawn to have been inherited from either her mother/father, and we alternate this choice for subsequent segments. This procedure repeats until the target generation back k is reached. The segments in the k-generation ancestors are then summarized as either counts (number of IBD segments per individual) or lengths. These simulations are necessarily approximate as they ignore crossover interference. However, unlike our analytic approximations, our simulation procedure maintains long-run dependencies created during recombination, allowing us to see the extent to which assuming independent segment survival adversely impacts our analytic results.

Figure 2A depicts the X genetic ancestors of one simulated example X genealogy back nine generations to illustrate this process. Each arc represents a single ancestor of a present-day female; red arcs are female X ancestors and blue arcs are male X ancestors. Ancestors that are not X ancestors are shaded gray. The saturation indicates the genetic contribution of this individual to the present-day female's X (with white indicating an X ancestor that made no genetic contribution on the X). Figure 2B shows the underlying X chromosome IBD segments for this same simulation back four generations.

The number of recombinational meioses along an unknown X lineage If we pick an ancestor at random k generations ago, the probability that they are an X genealogical ancestor is given by equation (7). We can now extend this logic and ask: having randomly sampled an X genealogical ancestor, how many recombinational meioses (i.e. females) lie in the lineage between a present day individual and this ancestor? Since IBD segment number and length distributions are parameterized by a rate proportional to the number of recombination events, this quantity is essential to our further derivations.
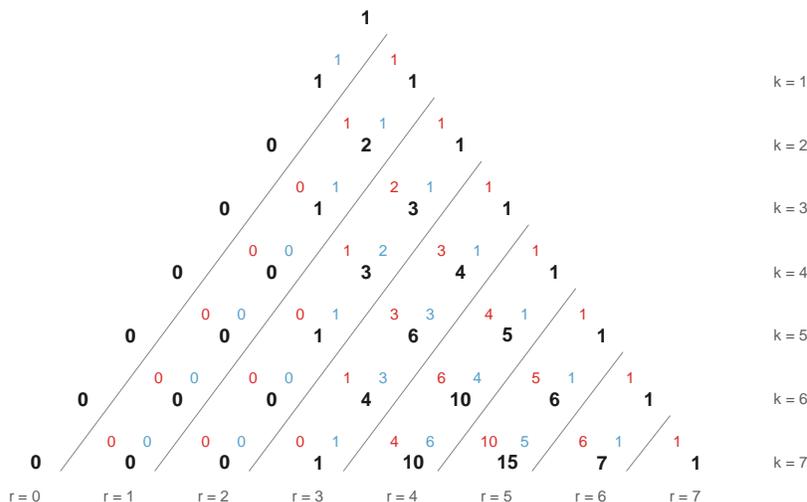
Figure 4: An triangle of $\binom{r+1}{k-r}$ binomial coefficients, similar to an offset version of Pascal's triangle. Black numbers indicate the number of genealogical ancestors at generation k whose X chromosome would have been transmitted to the current generation by a route involving r recombinant meioses. Each value is further broken down into female (red value, upper left) and male (blue value, upper right) ancestors. Values for a new row in the triangle are generated by summing the black numbers to the left in the row above (these are the mothers of all ancestors in the next generation) and two rows directly above (these are male ancestors). The sum of each row (fixed k) is a Fibonacci number and the values in the diagonal corresponding to a fixed value of r are binomial coefficients.

Specifically, if there's uncertainty about the particular lineage between a present-day female and one of her X ancestors k generations back (such that all of the $\mathcal{F}_{k+2}$ lineages to an X ancestor are equally probable), the number of females (thus, recombinational meioses) that occur is a random variable R. By the no two adjacent males condition, the possible number of females R is constrained; R has a lower bound of $\lfloor k/2 \rfloor$, which corresponds to a male alternating each generation to an ancestor in the kth generation. Similarly, the upper bound of R is k, since it is possible every individual along one X lineage is a female. Noting that an X genealogy extending back k generations enumerates every possible way to arrange r females such that none of the $k - r$ males are adjacent, we find that the number of ways of arranging r such females this way is

$$\binom{r+1}{k-r}. \tag{8}$$

For some readers, it may be useful to visualize the relationship between the numbers of recombinational meioses across the generations using a variant of Pascals triangle (Figure 4). The

9

sequence of recombinational meioses is related to a known integer sequence; see OEIS A030528 (Sloane, 2010) for a description of this sequence and its other applications.

If we pick an X genealogical ancestor at random k generations ago the probability that there are r female meioses along the lineage leading to this ancestor is

$$P_R(R = r|k) = \frac{\binom{r+1}{k-r}}{\mathcal{F}_{k+2}}. \tag{9}$$

In Appendix B, we derive a generating function for the number of recombinational meioses. We can use this generating function to obtain properties of this distribution such as the expected number of recombinational meioses. We can show that the expected number of recombinational meioses converges rapidly to $\mathbb{E}[R] \approx \varphi/\sqrt{5}\, k$ with increasing k, where φ is the Golden Ratio, $\frac{1+\sqrt{5}}{2}$.

The Distribution of Number of Segments Shared with an X Ancestor Using the distribution of recombinational meioses derived in the last section, we now derive a distribution for the number of IBD segments shared between a present-day individual and an X ancestor in the kth generation. For clarity, we first derive the number of IBD segments counted in the *parents* (i.e. not following the convention described in Section 1), but we can adjust this simply by replacing k with $k - 1$.

First, we calculate the probability of a present-day individual sharing N IBD segments with an X genealogical ancestor k generations in the past, where it is *known* that there are $R = r$ females (and thus recombinational meioses) along the lineage to this ancestor. This probability uses the Poisson-Binomial model described in Section 1:

$$P(N = n|r, k) = \sum_{b=0}^{\infty} \text{Bin}(N = n \mid l = b + 1, p = 1/2^r)\, \text{Pois}(B = b \mid \lambda = \nu r) \tag{10}$$

If we consider a single X genealogical ancestor k generations back, this individual could be any of the present-day female's $\mathcal{F}_{k+2}$ X ancestors. Since the particular lineage to this ancestor is unknown, we marginalize over all possible numbers of recombinational meioses that could occur:

$$P(N = n \mid k, \nu) = \sum_{r=\lfloor k/2 \rfloor}^{k} \sum_{b=0}^{\infty} \text{Bin}(N = n \mid l = b + 1, p = 1/2^r)\, \text{Pois}(B = b \mid \lambda = \nu r) \frac{\binom{r+1}{k-r}}{\mathcal{F}_{k+2}}$$

$$= \sum_{r=\lfloor k/2 \rfloor}^{k} \sum_{b=0}^{\infty} \binom{b+1}{n} 1/2^n (1 - 1/2^r)^{b-n+1} \frac{(\nu r)^b e^{-\nu r}}{b!} \frac{\binom{r+1}{k-r}}{\mathcal{F}_{k+2}} \tag{11}$$

Note that once we've conditioned on the number of recombinational meioses r, the lineages to an X ancestor are interchangeable; the specific X lineage affects recombination (and thus the IBD number and length distributions) only through the number of recombinational meioses along the lineage.

For the distribution of number of IBD segments counted in the offspring, we substitute $k - 1$ for k

$$P(N = n \mid k, \nu) = \sum_{r=\lfloor (k-1)/2 \rfloor}^{k-1} \sum_{b=0}^{\infty} \text{Bin}(n \mid l = b + 1, p = 1/2^r)\, \text{Pois}(B = b \mid \lambda = \nu r) \frac{\binom{r+1}{k-r-1}}{\mathcal{F}_{k+1}}. \tag{12}$$
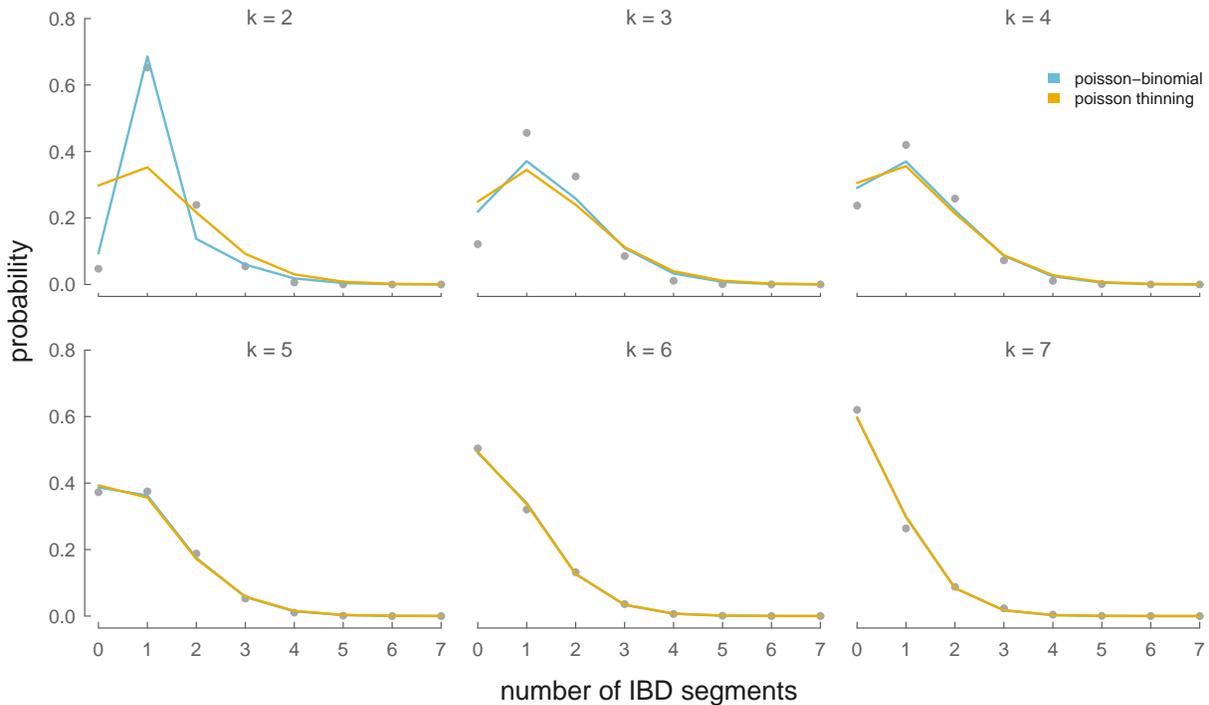
Figure 5: The Poisson thinning (yellow lines) and Poisson-Binomial (blue lines) analytic distributions of IBD segment number between an X ancestor in the kth generation (each panel) and a present-day female. Simulation results averaged over 5,000 simulations are the gray points.

In this formulation, if $k = 1$, $r = 0$. In this case, the lack of recombinational meioses implies $b = 0$, such that a present-day female shares $n = 1$ X chromosomes with each of her two parents in the $k = 1$ generation with certainty.

We can use our equation (12) to obtain $P(N > 0)$, the probability that a genealogical X ancestor k generations ago is a *genetic* ancestor. This probability over $k \in \{1, 2, \ldots, 14\}$ generations is shown in Figure 3B. For comparison, Figure 3B also includes the probability of a genealogical ancestor in the kth generation being an autosomal genetic ancestor and the probability of being a genetic X ancestor unconditional on being an X genealogical ancestor.

We have also assessed the Poisson thinning approach to modeling X IBD segment number. As with the Poisson-Binomial model, we marginalize over R:

$$P(N = n \mid k, \nu) = \sum_{r=\lfloor (k-1)/2 \rfloor}^{k-1} \text{Pois}(B = b \mid \lambda = (1 + \nu r)/2^r) \frac{\binom{r+1}{k-r-1}}{\mathcal{F}_{k+1}} \tag{13}$$

In Figure 5 we have compared the Poisson-Binomial and Poisson-thinning approximations for the number of IBD segments (counted in the offspring) shared between an X-ancestor in the kth generation and a present-day female. Overall, the analytic approximations are close to the simulation results, with the Poisson-Binomial model a closer approximation for small k and both models' accuracy improving quickly with increasing k. The Poisson-Binomial model is more accurate than

11

the Poisson thinning approximation due to the overdispersion discussed in Section 1. As k increases, this extra variance quickly decreases. Note that since the total genetic length of the X is relatively short compared to the autosomes, the extra variance makes a relatively small contribution to the total variance (see Appendix equation (36)); this is why the Poisson thinning approach is a better approximation for autosomal IBD segments. See Appendix A for details. Throughout the paper, we use the more accurate Poisson-Binomial model rather than this Poisson thinning model. If only X ancestry more than 3 generations back is of interest, the Poisson thinning approach may be used without much loss of accuracy.

The Distribution of IBD Segment Lengths with an X Ancestor The distribution of IBD segment lengths between a present-day female and an unknown X genealogical ancestor in the kth generation is similar to the autosomal length distribution (equation 6). However, with uncertainty about the particular lineage to the X ancestor, the number of recombinational meioses can vary between $\lfloor k/2 \rfloor \leq r \leq k$; we marginalize over the unknown number of recombinational meioses using the distribution equation (9). Our length density function is:

$$p(U = u|k) = \sum_{r=\lfloor k/2 \rfloor}^{k} r e^{-ru} \frac{\binom{r+1}{k-r}}{\mathcal{F}_{k+2}} \tag{14}$$

In Figure 6, we compare our analytic length density to an empirical density of X segment lengths calculated from 5,000 simulations. As with our IBD segment number distributions, our analytic model is close to the simulated data's empirical density, and converges rapidly with increasing k.

Note that both the IBD segment length and number distributions marginalize over an unobserved number of recombinational meioses (R) that occur along the lineage between individuals. As the IBD segments shared between two individuals is a function of the number breakpoints B, and thus recombinational meioses, neither the length nor number distributions $P(N = n)$ and $p(U = u)$ (which separately marginalize over both R and B) are independent of each other. Intuitively, we can understand this dependence in the following way: if we observe one long segment nearly the entire length of the X, this makes observing a large number of additional segments in the remaining region of the X unlikely.

3 Shared X Ancestry

Because only a fraction of one's genealogical ancestors are X ancestors (and this fraction rapidly decreases with k; see equation (7)), two individuals sharing X segments IBD from a recent ancestor considerably narrows the possible ancestors they could share. In this section, we describe the probability that a genealogical ancestor is an X ancestor, and the distributions for IBD segment number and length across full- and half-cousin relationships. For simplicity we concentrate on the case where the cousins share a genealogical ancestor k generations ago in both of their pedigrees, i.e. the individuals are $k - 1$ degree cousins. The formulae could be generalized to ancestors of unequal generational-depths (e.g. second cousins once removed) but we do not pursue this here.

Probability of a Shared X Ancestor Given that two individuals share their first common genealogical ancestor in the kth generation, we can calculate the probability that this single ancestor is also an X genealogical ancestor. We can think of this first shared ancestor in the kth generation
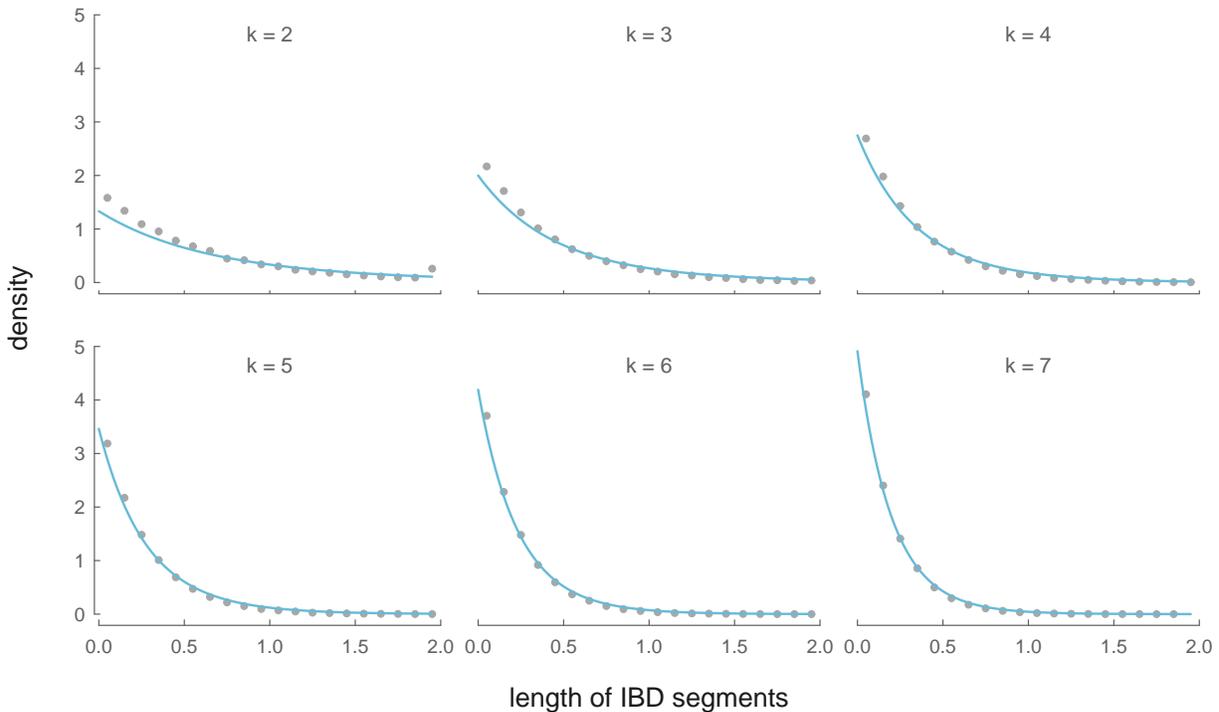
Figure 6: The analytic distributions of IBD segment length between an ancestor in the kth generation (for $k \in \{2, \dots, 7\}$) and a present-day female (blue lines), and the binned average over 5,000 simulations (gray points).

as the same individual in both of the two sets of 2^k ancestors of each present-day individual. Since this shared ancestor must be of the same sex in each of the two present-day individuals' genealogies, we condition on the ancestor's sex (with probability $1/2$ each) and then calculate the probability that this individual is also an X ancestor (with the same sex). Let us define $N_{\female}$ and $N_{\male}$ as the number of genealogical female and male ancestors, and $N_{\female}^X$ and $N_{\male}^X$ as the number of X female and male ancestors of a present-day individual in the kth generation. Then:

$$
\begin{aligned}
P(\text{shared X ancestor} \mid \text{shared ancestor } k \text{ generations ago}) &= \frac{N_{\female}}{2^k}\left(\frac{N_{\female}^X}{N_{\female}}\right)^2 + \frac{N_{\male}}{2^k}\left(\frac{N_{\male}^X}{N_{\male}}\right)^2 \\
&= \frac{1}{2}\left(\frac{\mathcal{F}_{k+1}}{2^{k-1}}\right)^2 + \frac{1}{2}\left(\frac{\mathcal{F}_k}{2^{k-1}}\right)^2 \quad (15)
\end{aligned}
$$

Thus, the probability that a shared genealogical ancestor is also a shared X ancestor is decreasing at an exponential rate. By the 8th generation, a shared genealogical ancestor has less than a five percent chance of being a shared X ancestor of both present-day individuals.

The Sex of Shared Ancestor Unlike genealogical ancestors, which have equal numbers of female and male ancestors, recent X genealogical ancestors are predominantly female. Since a

13

present-day female has $\mathcal{F}_{k+1}$ female ancestors and $\mathcal{F}_k$ male ancestors k generations ago, the ratio of female to male X genealogical ancestors converges to the Golden Ratio $\varphi = \frac{1+\sqrt{5}}{2}$ (Simson, 1753; Wells, 1997):

$$\lim_{k \to \infty} \frac{\mathcal{F}_{k+1}}{\mathcal{F}_k} = \varphi \tag{16}$$

In modeling the IBD segment number and length distributions between present day individuals, the sex of the shared ancestor k generations ago affects the genetic ancestry process in two ways. First, a female shared ancestor allows the two present-day individuals to share segments on either of her two X chromosomes while descendents of a male shared ancestor share IBD segments only through his single X chromosome. Second, the no two adjacent males condition implies a male shared X genealogical ancestor constrains the X genealogy such that the present-day X descendents are related through his two daughters. Given that the ratio of female to male X ancestors is skewed, our later distributions require an expression for the probability that a shared X ancestor in the kth generation is female, which we work through in this section.

As in equation (15), an ancestor shared in the kth generation of two present-day individuals' genealogies must have the same sex in each genealogy. Assuming both present-day cousins are females, in each genealogy there are $\mathcal{F}_k$ possible male ancestors and $\mathcal{F}_{k+1}$ female ancestors that could be shared. Across each present-day females' genealogies there are $(\mathcal{F}_k)^2$ possible male ancestor combinations and $(\mathcal{F}_{k+1})^2$ possible female ancestor combinations. Thus, if we let ♀$_X$ and ♂$_X$ denote that the sex of the shared is female and male respectively, the probability of a female shared ancestor is:

$$P(♀_X) = \frac{(\mathcal{F}_{k+1})^2}{(\mathcal{F}_k)^2 + (\mathcal{F}_{k+1})^2} \tag{17}$$

The probability that the shared ancestor is male is simply $1 - P(♀_X)$. One curiosity is that as $k \to \infty$, $P(♀_X) = \frac{\varphi}{\sqrt{5}} = \frac{5+\sqrt{5}}{10} \approx 0.7236$, where φ is the Golden Ratio.

Partnered Shared Ancestors Thus far, we have only looked at two present-day individuals sharing a single X ancestor k generations back. In monogamous populations, most shared ancestry is likely to descend from *two* ancestors; we call such relationships *partnered* shared ancestors. In this section, we look at full-cousins descending from two shared genealogical ancestors that may also be X ancestors. Two full-cousins could either (1) both descend from two X ancestors such that they are X full-cousins, (2) share only one X ancestor, such that they are X half-cousins, or (3) share no X ancestry. We calculate the probabilities associated with each of these events here.

Two individuals are full-cousins if the great^{k-2} grandfather and the great^{k-2} grandmother in one individual's genealogy are the same couple as in the other individual's genealogy. For these two full-cousins to be X full-cousins, this couple must also be a couple in both individuals' X genealogies. In every X genealogy, the number of couples in generation k is the number of females in generation $k-1$, as every female has two X ancestors in the next generation (while males only have one). Thus, the probability two female $k-1$ degree full-cousins are also X full-cousins is:

$$P(\text{X full-cousins} \mid \text{full-cousins}) = \left(\frac{\mathcal{F}_k}{2^{k-1}}\right)^2 \tag{18}$$

14

Now, we consider the event that two genealogical full-cousins are X half-cousins. Being X half-cousins implies that the partnered couple these full-cousins descend from includes a single ancestor that is in the X genealogies of both full-cousins. This single X ancestor must be a female, as a male X ancestor's female partner must also be an X ancestor (since mothers must pass an X). For a female to be an X ancestor but not her partner, one or both of her offspring must be male. Either of these events occurs with probability:

$$P(\text{X half-cousins} \mid \text{full-cousins}) = \frac{\mathcal{F}_{k-1}^2 + 2\mathcal{F}_{k-1}\mathcal{F}_k}{2^{2(k-1)}} \tag{19}$$

The Distribution of Recombinational Meioses between Two X Half-Cousins   To find distributions for the number and lengths of IBD segments shared between two half-cousins on the X chromosome, we first need to find the distribution for the number of females between two half-cousins with a shared ancestor in the kth generation. We refer to the individuals connecting the two cousins as a *genealogical chain*. As we'll see in the next section, the number of IBD X segments shared between half-cousins depends on the sex of the shared ancestor; thus, we also derive distributions in this section for the number of recombinational meioses along a genealogical chain, conditioning on the sex of the shared ancestor. As earlier, our models assume two present-day female cousins but are easily extended to male cousins.

First, there are $2k - 1$ ancestral individuals separating two present-day female $(k - 1)^{\text{th}}$ degree cousins. These X ancestors in the genealogical chain connecting the two present-day female cousins follow the no two adjacent male condition; thus the distribution of females follows the approach used in equation (9) with k replaced with $2k - 1$:

$$P_{HC}(R = r|k) = \frac{\binom{r+1}{2k-r-1}}{\mathcal{F}_{2k+1}} \tag{20}$$

where the HC (for half-cousin) subscript differentiates this equation from equation (9) and k is the generation of the shared ancestor. Similarly to equation (9), R is bounded such that $\lfloor (2k-1)/2 \rfloor \leq R \leq 2k - 1$.

Now, we derive the probability of $R = r$ females conditional on the shared ancestor being female, $♀_X$. This conditional distribution differs from equation (20) since it eliminates all genealogical chains with a male shared ancestor.

We find the distribution of recombinational meioses conditional on a female shared ancestor by placing the other $R' = r'$ females (the prime denotes we do not count the shared female ancestor here) along the two lineages of $k - 1$ individuals from the shared female ancestor down to the present-day female cousins. These $R' = r'$ females can be placed in both lineages by positioning s females in the first lineage and $r' - s$ females in the second lineage, where s follows the constraint $\lfloor (k-1)/2 \rfloor \leq s \leq (k-1)/2$. Our equation (9) models the probability of an X genealogical chain having r females in k generations; here, we use this distribution to find the probabilities of s females in $k - 1$ generations in one lineage and $r' - s$ females in $k - 1$ generations in the other lineage. As the number of females in each lineage is independent, we take the product of these probabilities and sum over all possible s; this is the discrete convolution of the number of females in two lineages $k - 1$ generations long. Finally, we account for the shared female ancestor, by the transform $R = R' + 1 = r$:

$$P(R = r | ♀_X) = \sum_{s=\lfloor (k-1)/2 \rfloor}^{k-1} \frac{\binom{s+1}{k-s-1}\binom{r-s}{k+s-r}}{(\mathcal{F}_{k+1})^2} \tag{21}$$

In general, this convolution approach allows us find the distribution of females in a genealogical chain under various constraints, and can easily be extended to the case of a shared male X ancestor (with necessarily two daughters).

Finally, note that we have modeled the number of *females* in a genealogical chain of $2k - 1$ individuals. Thus far in our models, the number of females has equaled the number of recombinational meioses. However, when considering the number of recombinational meioses between half-cousins, *two* recombinational meioses occur if the shared ancestor is a female (as she produced two independent gametes she transmits to her two offspring). Thus, for a single shared X ancestor, the number of recombinational meioses ρ is

$$\rho = \begin{cases} r+1 & \text{if } ♀_X \\ r & \text{if } ♂_X \end{cases} \tag{22}$$

which we use when parameterizing the rate of recombination in our IBD segment number distributions. Furthermore, since a shared female ancestor has two X haplotypes that present-day cousins could share segments IBD through, the binomial probability $1/2^\rho$ is doubled. Further constraints are needed to handle full-cousins; we will discuss these in section 3.

Half-Cousins In this section we calculate the distribution of IBD X segments shared between two present-day female X half-cousins with a shared ancestor in the kth generation. We imagine we do not know any details about the lineages to this shared ancestor nor the sex of the shared ancestor, so we marginalize over both. Thus, the probability of two $(k-1)^{\text{th}}$ degree X half-cousins sharing $N = n$ segments is:

$$P(N = n | k) = \sum_{r=\lfloor (2k-1)/2 \rfloor}^{2k-1} P_S(R = r | k) \left[P(N = n | ♀_X, R = r) P(♀_X | R = r) + \right.$$
$$\left. P(N = n | ♂_X, R = r) P(♂_X | R = r) \right] \tag{23}$$

As discussed in the previous section, the total number of recombinational meioses along the genealogical chain between half-cousins depends on the unobserved sex of the shared ancestor (i.e. equation (22)). Likewise, the binomial probability also depends on the shared ancestor's sex. Accounting for these adjustments, the probabilities $P(N = n | ♀_X, R = r)$ and $P(N = n | ♂_X, R = r)$ are:

$$P(N = n | ♀_X, R = r) = \sum_{b=0}^{\infty} \text{Pois}(B = b | \lambda = (r+1)\nu) \, \text{Bin}(N = n | l = b+1, p = 1/2^r) \tag{24}$$

$$P(N = n | ♂_X, R = r) = \sum_{b=0}^{\infty} \text{Pois}(B = b | \lambda = r\nu) \, \text{Bin}(N = n | l = b+1, p = 1/2^r) \tag{25}$$

16

Since the sex of the shared ancestor depends on the number of females in the genealogical chain between the two cousins (e.g. if $r = 2k - 1$, the shared ancestor is a female with certainty), we require an expression for the probability of the shared ancestor being male or female given $R = r$. Using Bayes' theorem, we can invert the conditional probability $P(R = r | \female_X)$ to find that the probability that a shared X ancestor is female conditioned on R females in the genealogical chain is

$$P(\female_X | R = r) = \frac{\mathcal{F}_{2k-1}}{\binom{r+1}{2k-r-1}\left((\mathcal{F}_{k+1})^2 + (\mathcal{F}_k)^2\right)} \sum_{s=\lfloor (k-1)/2 \rfloor}^{k-1} \binom{s+1}{k-s+1}\binom{r-s}{k+s-r} \qquad (26)$$

and $P(\male_X | R = r)$ can be found as the complement of this probability.
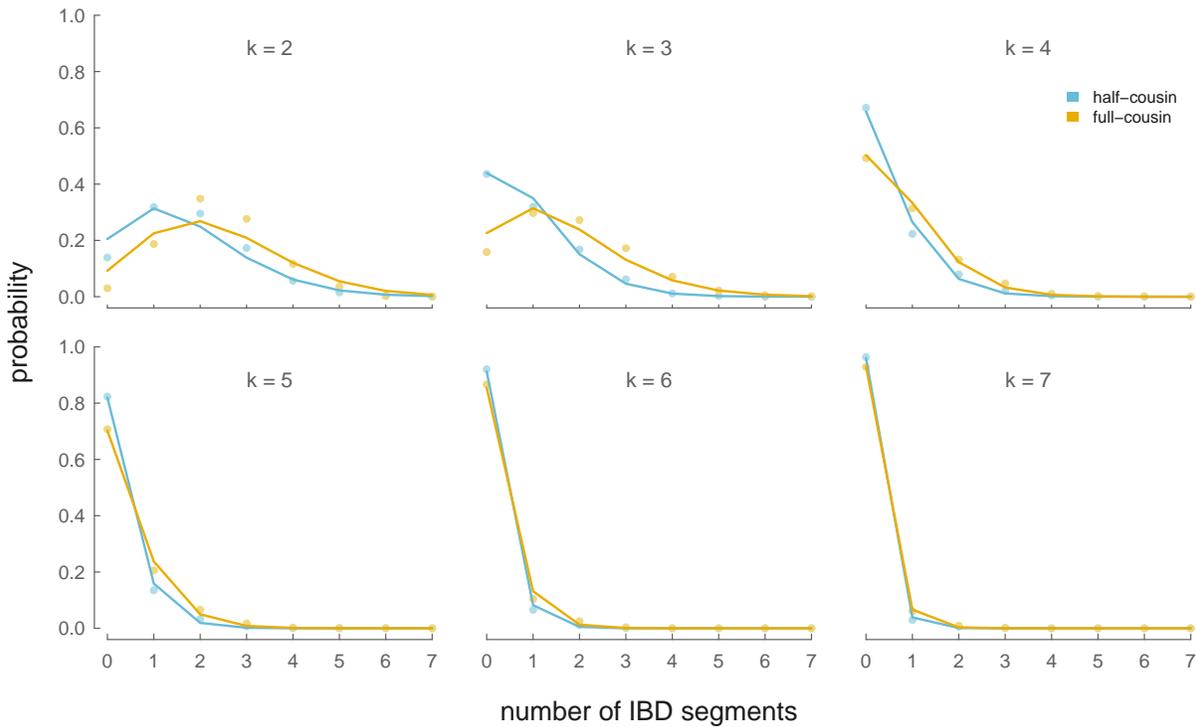
Figure 7: Distributions of X IBD segment number for X half- (blue) and X full-cousins (yellow). Lines show the analytic approximations (equations (23) and (28)) and gray points show the probability averaged over 5,000 simulations.

Inserting equations (25), (24), and (26) into (23) gives us an expression for the distribution of IBD segment numbers between two half-cousins with a shared ancestor k generations ago. Figure 7 compares the analytic model in equation (23) with the IBD segments shared between half-cousins over 5,000 simulated pairs of X genealogies.

The density function for IBD segment lengths between X cousins (either half- or full-cousins; length distributions are only affected by the number of recombinations in the genealogical chain) is equation (14) but marginalized over the number of recombinational meioses between two cousins (equation (20)) rather than the number of recombinational meioses between a present-day individual

17

and a shared ancestor. Simulations show the length density closely matches simulation results (see Figure 11 in the appendix).

Full-Cousins Full-cousin relationships allow descendents to share IBD autosomal segments from either their shared maternal ancestor, shared paternal ancestral, or both. In contrast, since males only pass an X chromosome to daughters, only full-sibling relationships in which both offspring are female (due to the no to adjacent males condition) are capable of leaving X genealogical descendents. We derive a distribution for the number of IBD segments shared between $(k-1)^{\text{th}}$ degree full X cousins by conditioning on this familial relationship and marginalizing over the unobserved number of females from the two full-sibling daughters to the present-day female full-cousins.

First, we find the number of females (including the two full-sibling daughters in the $(k-1)^{\text{th}}$ generation) in the genealogical chain between the two full-cousins (omitting the shared male and female ancestors, which we account for separately). Like equation (21), this is a discrete convolution:

$$P_{FC}(R = r) = \sum_{s=\lfloor (k-2)/2 \rfloor}^{k-2} \frac{\binom{s+1}{k-s-2}\binom{r-s-1}{k-r+s}}{(\mathcal{F}_k)^2} \tag{27}$$

This probability is valid for $2\lfloor (k-2)/2 \rfloor + 2 \leq r \leq 2k-2$ and is 0 elsewhere. For $N = n$ segments to be shared between two full-cousins, z segments can be shared via the maternal shared X ancestor (where $0 \leq x \leq n$) and $n - z$ segments can be shared through the paternal shared X ancestor. We marginalize over all possible values of z, giving us another discrete convolution:

$$P(N = n) = \sum_{r=2\lfloor (k-2)/2 \rfloor+2}^{2k-2} \sum_{z=0}^{n} P(N = z|\female_X)P(N = n - z|\male_X)P_{FC}(R = r) \tag{28}$$

where

$$P(N = n|\female_X, R = r) = \sum_{b=0}^{\infty} \text{Bin}(N = n|l = b + 1, p = 1/2^{r+1})\text{Pois}(B = b|\lambda = v(r + 2)) \tag{29}$$

$$P(N = n|\male_X, R = r) = \sum_{b=0}^{\infty} \text{Bin}(N = n|l = b + 1, p = 1/2^{r})\text{Pois}(B = b|\lambda = vr) \tag{30}$$

are the probabilities of sharing n segments through the shared female and male X ancestors respectively. For the female shared ancestor, we account for two additional recombinational meioses (one for each of the two gametes she passes to her two daughters), and the fact she can share segments through either of her X chromosomes (hence, why the binomial probability is $1/2^{r+1}$). We compare our analytic X full-cousin IBD segment number results to 5,000 genealogical simulations in Figure 7.

4   Inference

With our IBD X segment distributions, we now turn to how these can be used to infer recent X ancestry. In practice, inferring the number of generations back to a common ancestor (k) is best accomplished through the signature of recent ancestry from the 22 autosomes, rather than through the short X chromosome. Therefore, we concentrate on questions about the extra information that

the X provides conditional on k being known with certainty. A number of methods are available for the task of estimating k through autosomal IBD segments (Durand et al., 2014; Henn et al., 2012; Huff et al., 2011).

Here, we focus on two separate questions: (1) what is the probability of being an X genealogical ancestor given that no IBD segments are observed, and (2) can we infer details about the X genealogical chain between two half-cousins? These questions address how informative the number of segments shared between cousins is about the precise relationship of cousins. We assume that segments of X chromosome IBD come only from the kth generation, and not from deeper relationships or from false positives. In practice, inference from the X IBD segments would have to incorporate both of these complications, and as such our results represent best case scenarios.
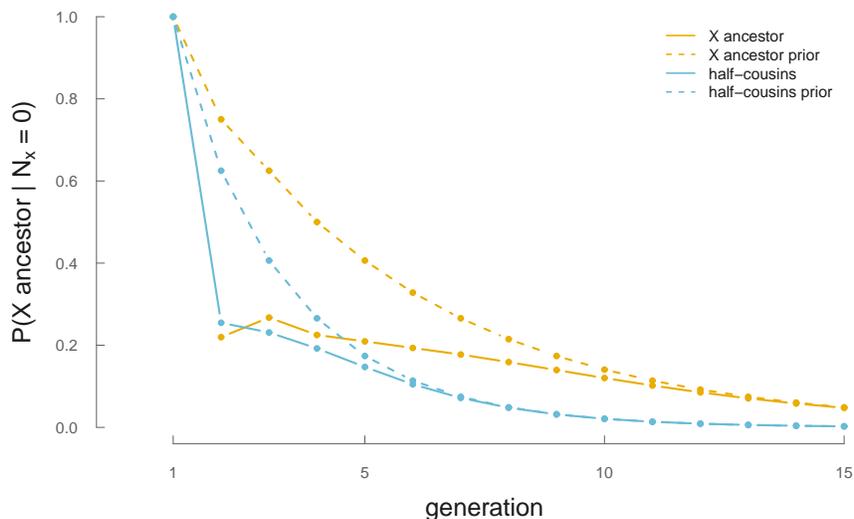
Figure 8: The probabilities of an individual being an X ancestor to a present-day female (yellow solid line) or single shared X ancestor (blue solid line) to two present-day female cousins conditional on not sharing any X IBD segments, k generations back. Dashed lines indicate the prior probabilities.

It's possible that k generations back, an individual is a genealogical X ancestor but shares no X genetic material with a present-day descendent. To what extent is the lack of sharing on the X chromosome with an ancestor informative about our relationship to them? How does the lack of sharing of the X chromosome between $k-1$ cousins change our views as to their relationship? To get at these issues, we can use our analytic approximations to calculate the probability that one is an X ancestor given that no segments are observed, $P(\text{X ancestor} \mid N = 0)$:

$$P(\text{X ancestor} \mid N = 0) = \frac{P(\text{X ancestor})P(N = 0 \mid \text{X ancestor})}{P(N = 0 \mid \text{X ancestor})P(\text{X ancestor}) + P(\text{not X ancestor})} \quad (31)$$

Here, $P(N = 0 \mid \text{X ancestor})$ is given by equation (12) and $P(\text{X ancestor})$ is given by equation (7). This function is shown in Figure 8 (yellow lines). We can derive an analogous expression for the probability of two female half-cousins sharing an X ancestor but not having any X segments IBD by replacing $P(\text{X ancestor} \mid N = 0)$ with equation (23), and replacing $P(\text{X ancestor})$ with $P(\text{shared X ancestor})$ which is given by equation (15) (blue lines). We also plot the prior distributions to show the answer if no information about the X chromosome was observed. In both cases,

the lack of sharing on the X chromosome does narrow the field of ancestors, making those who are not X genealogical ancestors more likely, beyond our prior. This is especially true for close relationships ($k < 5$), where the X is likely to be shared if the ancestor was an X genealogical ancestor.

X IBD segments can be used to infer genealogical details that are not possible looking only at autosome IBD segments. Unlike the X, the autosomes undergo recombinational meioses in every individual each generation. Consequently, while IBD autosome segments leave a signature of recent ancestry between two individuals, the uniformity of recombinational meioses across every lineage to the shared ancestor leaves no signal of *which* genealogical lineage connects two present-day cousins.

In contrast, X genealogical chains have two properties that leave a signature of which genealogical lineage connects two present-day cousins. First, recombinational meioses only occur in females. Consequently, more females along an X genealogical chain increase the rate of recombination and alter the IBD segments length and number distributions in a lineage-specific manner. Second, the number of females varies across X lineages; for example the number of females between two present-day half-cousins varies by a factor of two or more. These two properties allow us to find a posterior distribution for the number of females that separate two present-day cousins. This number of females R is our only inferable summary of the genealogical chain and constrains the possible X genealogical chains between these two cousins by varying amounts dependent on R and k.

Our approach to inference is through the posterior distribution of R given an observed number of IBD segments N and conditioning on k. We calculate this posterior conditional on the cousins sharing an X ancestor; we do this to separate it from the question of whether a pair share an X ancestor (as derived in equation (31)). Our posterior probability is given by Bayes' theorem

$$P(R|N = n) = \frac{P(N = n|R)P(R)}{P(N = n)} \tag{32}$$

where the prior $P(R)$ is readily calculable through equation (20) and $P(N = n)$ given by equation (23). The data likelihood $P(N = n|R)$ is given by equation (10).

In Figure 9, we show the posterior distributions over the number of recombinational meioses, given an observed number of IBD segments between two females known to be X half-cousins. Again, these posterior distributions condition on knowing how many generations have occurred since the shared ancestor, k. With an increasing number of generations to the shared ancestor, fewer segments survive to be IBD between the present-day cousins. Consequently, observing IBD segments increases the likelihood of fewer females (and thus fewer recombinational meioses) between the cousins. For example, for $k = 6$, observing (the admittedly unlikely) six or more IBD segments leads to a posterior mode over the fewest possible number of females in the genealogical chain ($\lfloor (2k - 1)/2 \rfloor = 5$; Figure 9). Similarly, observing between three and five segments places the posterior mode over six females in the genealogical chain. For $k > 4$, seeing zero segments provides little information over the prior about the relationship between the cousins, as sharing zero segments is the norm.

5   Discussion

Detecting and inferring the nature of recent ancestry is important to a range of applications and the nature of such relationships are often of inherent interest. As the sample sizes of population
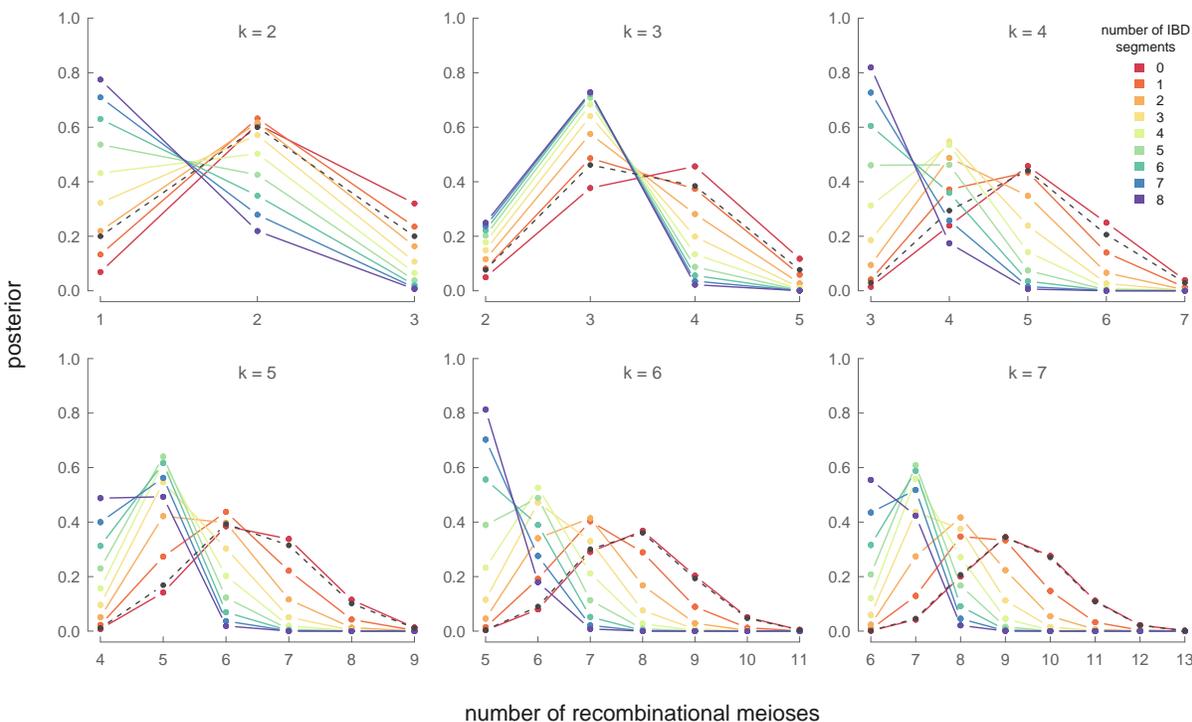
Figure 9: Posterior probability distribution $P(R = r | N = n, k)$ for different generations (each panel), and the observed number of IBD segments (each colored line). The prior distribution of recombinational meioses given k is indicated by a gray dashed line.

genomic data sets increase, so will the probability of sampling individuals that share recent ancestry. In particular, the very large data sets being developed in human genetics will necessitate taking a genealogical view of recent relatedness. Our methods extend existing methods for the autosomes by accounting for the special inheritance pattern of the X. Specifically, recent ancestry on the X differs from the autosomes since males only inherit an X from their mothers, and fathers pass an unrecombined (ignoring the PAR) X to their daughters. Consequently, the number of recombinational meioses, which determine the length and number of IBD segments, varies across the X genealogy. Since in most cases the number of females between two individuals in a genealogical chain is often unknown, we derive a distribution for recombinational meioses (equation (9)).

We derive distributions for the length and number of IBD X segments by marginalizing over the unknown number of recombinational meioses that can occur between two individuals connected through a genealogical chain. In both cases, we condition on knowing k (the generations back to a shared ancestor) which can be inferred from the autosomes (Huff et al., 2011). Our models for IBD segment number and length use a Poisson-Binomial approximation to the recombination process, which match simulation results closely.

Our results here not only allow X IBD segments to be used to model recent ancestry, but are in fact more informative about which genealogical ancestors individuals share than autosomal data alone. This additional information occurs through two avenues. First, sharing IBD segments on the X immediately reduces the potential genealogical ancestors two individuals share, since one's

21

X ancestors are only a fraction of their possible genealogical ancestors (i.e. $\mathcal{F}_{k+2}/2^k$ in the case of a present-day female). Second, the varying number of females in an X genealogy across lineages combined with the fact that recombinational meioses only occur in females to some extent leave a lineage-specific signature of ancestry.

Unfortunately, the X chromosome is short, such that the chance of any signal of recent ancestry on the X decays rather quickly. However, growing sample sizes will increase both the detection of the pairwise relatedness and cases of relatedness between multiple individuals. In these large data sets, overlapping pairwise relationships (e.g. a present-day individual that shares X segments with two distinct other individuals) could be quite informative about the particular ancestors that individuals share.

Our results should also be of use in understanding patterns of admixture on the X chromosome. In particular our results about the posterior information from the number and length of X segments shared with a genealogical ancestor can help us understand what can be learned from the presence (or absence) of particular segments of particular ancestry on the X chromosome. While this information for an individual decays somewhat quickly after a small number of generations, models of X chromosome segment ancestry will be useful at a population-level for understanding sex-biased admixture (Bryc et al., 2010; Goldberg et al., 2015; Shringarpure et al., 2016).

6   Acknowledgements

References

Baca, Mateusz, Karolina Doan, Maciej Sobczyk, Anna Stankovic, and Piotr Wgleski (2012). "Ancient DNA reveals kinship burial patterns of a pre-Columbian Andean community". In: *BMC genetics* 13.1, p. 1.

Balding, David J and Richard A Nichols (1994). "DNA profile match probability calculation: how to allow for population stratification, relatedness, database selection and single bands". In: *Forensic Science International* 64.2-3, pp. 125–140.

Barton, Nicholas H and Alison M Etheridge (2011). "The Relation Between Reproductive Value and Genetic Contribution". In: *Genetics* 188.4, pp. 953–973.

Basin, S. L. (1963). "The Fibonacci sequence as it appears in nature". In: *The Fibonacci sequence as it appears in nature* 1.1, pp. 53–56.

Belin, Thomas R, David W Gjertson, and Ming-Yi Hu (1997). "Summarizing DNA evidence when relatives are possible suspects". In: *Journal of the American Statistical Association* 92.438, pp. 706–716.

Bryc, Katarzyna et al. (2010). "Genome-wide patterns of population structure and admixture in West Africans and African Americans". In: *Proceedings of the National Academy of Sciences* 107.2, pp. 786–791.

Bustamante, C D and S Ramachandran (2009). "Evaluating signatures of sex-specific processes in the human genome". In: *Nature Genetics* 41.1, pp. 8–10.

Chang, Joseph T (1999). "Recent common ancestors of all present-day individuals". In: *Advances in Applied Probability*, pp. 1002–1026.

Donnelly, K P (1983). "The probability that related individuals share some section of genome identical by descent". In: *Theoretical Population Biology* 23.1, pp. 34–63.

Durand, Eric Y, Chuong B Do, Joanna L Mountain, and J. Michael Macpherson (2014). "Ancestry Composition: A Novel, Efficient Pipeline for Ancestry Deconvolution". In: *bioRxiv*. DOI: 10.1101/010512. eprint: http://biorxiv.org/content/early/2014/10/18/010512.full.pdf. URL: http://biorxiv.org/content/early/2014/10/18/010512.

Feller, Willliam (1950). *An Introduction to Probability Theory and its Applications*. Vol. 1. John Wiley & Sons.

Fisher, Ronald Aylmer et al. (1949). "The theory of inbreeding." In: *Edinburgh: Oliver & Boyd*.

— (1954). "A fuller theory of" junctions" in inbreeding". In: *Heredity* 8, pp. 187–197.

Fox, M, R Sear, and J Beise (2009). "Grandma plays favourites: X-chromosome relatedness and sex-specific childhood mortality". In: *Proceedings of the Royal Society B*.

Fu, Qiaomei et al. (2015). "An early modern human from Romania with a recent Neanderthal ancestor". In: *Nature* 524.7564, pp. 216–219.

Glaubitz, Jeffrey C, O Eugene Rhodes, and J Andrew DeWoody (2003). "Prospects for inferring pairwise relationships with single nucleotide polymorphisms". In: *Molecular Ecology* 12.4, pp. 1039–1047.

Goldberg, Amy and Noah A Rosenberg (2015). "Beyond 2/3 and 1/3: The Complex Signatures of Sex-Biased Admixture on the X Chromosome." In: *Genetics* 201.1, pp. 263–279.

Gravel, Simon (2012). "Population genetics models of local ancestry." In: *Genetics* 191.2, pp. 607–619.

Gusev, Alexander, Jennifer K Lowe, Markus Stoffel, Mark J Daly, David Altshuler, Jan L Breslow, Jeffrey M Friedman, and Itsik Pe'er (2009). "Whole population, genome-wide mapping of hidden relatedness". In: *Genome research* 19.2, pp. 318–326.

Haak, Wolfgang et al. (2008). "Ancient DNA, Strontium isotopes, and osteological analyses shed light on social and kinship organization of the Later Stone Age". In: *Proceedings of the National Academy of Sciences* 105.47, pp. 18226–18231.

Hassold, TJ, SL Sherman, D Pettay, DC Page, and PA Jacobs (1991). "XY chromosome nondisjunction in man is associated with diminished recombination in the pseudoautosomal region." In: *American journal of human genetics* 49.2, p. 253.

Henn, Brenna M, Lawrence Hon, J Michael Macpherson, Nick Eriksson, Serge Saxonov, Itsik Pe'er, and Joanna L Mountain (2012). "Cryptic Distant Relatives Are Common in Both Isolated and Cosmopolitan Genetic Samples". In: *PloS one* 7.4, e34267–13.

Huff, Chad D et al. (2011). "Maximum-likelihood estimation of recent shared ancestry (ERSA)". In: *Genome Research* 21.5, pp. 768–774.

Keyser-Tracqui, Christine, Eric Crubezy, and Bertrand Ludes (2003). "Nuclear and mitochondrial DNA analysis of a 2,000-year-old necropolis in the Egyin Gol Valley of Mongolia". In: *The American Journal of Human Genetics* 73.2, pp. 247–260.

Koller, P Ch and CD Darlington (1934). "The genetical and mechanical properties of the sex-chromosomes". In: *Journal of Genetics* 29.2, pp. 159–173.

Laughlin, Harry H (1920). "Calculating Ancestral Influence in Man: A Mathematical Measure of the Facts of Bisexual Heredity". In: *Genetics* 5.5, p. 435.

Liang, Mason and Rasmus Nielsen (2014). "The lengths of admixture tracts." In: *Genetics* 197.3, pp. 953–967.

Palamara, Pier Francesco, Todd Lencz, Ariel Darvasi, and Itsik Peer (2012). "Length distributions of identity by descent reveal fine-scale demographic history". In: *The American Journal of Human Genetics* 91.5, pp. 809–822.

Pinto, Ndia, Leonor Gusmo, and Antnio Amorim (2011). "X-chromosome markers in kinship testing: a generalisation of the IBD approach identifying situations where their contribution is crucial". In: *Forensic Science International: Genetics* 5.1, pp. 27–32.

Pinto, Ndia, Pedro V Silva, and Antnio Amorim (2012). "A general method to assess the utility of the X-chromosomal markers in kinship testing". In: *Forensic Science International: Genetics* 6.2, pp. 198–207.

Pool, John E and Rasmus Nielsen (2007). "Population size changes reshape genomic patterns of diversity". In: *Evolution* 61.12, pp. 3001–3006.

— (2009). "Inference of historical changes in migration rate from the lengths of migrant tracts". In: *Genetics* 181.2, pp. 711–719.

R Core Team (2015). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria. URL: https://www.R-project.org/.

Ralph, Peter and Graham Coop (2013). "The Geography of Recent Genetic Ancestry across Europe". In: *PLoS Biol* 11.5, e1001555. DOI: 10.1371/journal.pbio.1001555. URL: http://dx.doi.org/10.1371%2Fjournal.pbio.1001555.

Ram, Karthik and Hadley Wickham (2015). *wesanderson: A Wes Anderson Palette Generator*. R package version 0.3.2. URL: http://CRAN.R-project.org/package=wesanderson.

Ramachandran, Sohini, Noah A Rosenberg, Marcus W Feldman, and John Wakeley (2008). "Population differentiation and migration: coalescence times in a two-sex island model for autosomal and X-linked loci". In: *Theoretical population biology* 74.4, pp. 291–301.

Ramachandran, Sohini, Noah A Rosenberg, Lev A Zhivotovsky, and Marcus W Feldman (2004). "Robustness of the inference of human population structure: a comparison of X-chromosomal and autosomal microsatellites". In: *Human genomics* 1.2, p. 1.

Rice, William R, Sergey Gavrilets, and Urban Friberg (2008). "Sexually antagonistic zygotic drive of the sex chromosomes". In: *PLoS Genet* 4.12, e1000313.

Rohde, DLT, S Olson, and J T Chang (2004). "Modelling the recent common ancestry of all living humans". In: *Nature*.

Rosser, Z H, P Balaresque, and M A Jobling (2009). "Gene conversion between the X chromosome and the male-specific region of the Y chromosome at a translocation hotspot". In: *The American Journal of Human* . . .

Rossum, Guido (1995). *Python Reference Manual*. Tech. rep. Amsterdam, The Netherlands, The Netherlands.

Royal, Charmaine D, John Novembre, Stephanie M Fullerton, David B Goldstein, Jeffrey C Long, Michael J Bamshad, and Andrew G Clark (2010). "Inferring genetic ancestry: opportunities, challenges, and implications". In: *The American Journal of Human Genetics* 86.5, pp. 661–673.

Shringarpure, Suyash S., Carlos D. Bustamante, Kenneth L. Lange, and David H. Alexander (2016). "Efficient analysis of large datasets and sex bias with ADMIXTURE". In: *bioRxiv*. DOI: 10.

1101/039347. eprint: http://biorxiv.org/content/early/2016/02/10/039347.full.pdf. URL: http://biorxiv.org/content/early/2016/02/10/039347.

Sjerps, Marjan and Ate D Kloosterman (1999). "On the consequences of DNA profile mismatches for close relatives of an excluded suspect". In: *International journal of legal medicine* 112.3, pp. 176–180.

Sloane, N.J.A. (2010). "Online Encyclopaedia of Integer Sequences". In: *Published electronically at http://www.oeis.org*.

Sun, Lei, Kenneth Wilder, and Mary Sara McPeek (2002). "Enhanced pedigree error detection". In: *Human heredity* 54.2, pp. 99–110.

Thomas, Alun, Nicola J Camp, James M Farnham, Kristina Allen-Brady, and Lisa A Cannon-Albright (2008). "Shared genomic segment analysis. Mapping disease predisposition genes in extended pedigrees using SNP genotype assays". In: *Annals of human genetics* 72.2, pp. 279–287.

Thomas, Alun, Mark H Skolnick, and Cathryn M Lewis (1994). "Genomic mismatch scanning in pedigrees". In: *Mathematical Medicine and Biology* 11.1, pp. 1–16.

Thompson, Elizabeth A (2013). "Identity by descent: variation in meiosis, across genomes, and in populations". In: *Genetics* 194.2, pp. 301–326.

Wachter, Kenneth W, Eugene A Hammel, and Peter Laslett (1979). *Statistical studies of historical social structure*. Academic Press Inc.

Wells, David (1997). *The Penguin dictionary of curious and interesting numbers*. Penguin.

Wickham, Hadley (2009). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. ISBN: 978-0-387-98140-6. URL: http://had.co.nz/ggplot2/book.

— (2016a). *purrr: Functional Programming Tools*. R package version 0.2.0. URL: http://CRAN.R-project.org/package=purrr.

— (2016b). *tidyr: Easily Tidy Data with 'spread()' and 'gather()' Functions*. R package version 0.4.0. URL: http://CRAN.R-project.org/package=tidyr.

Wickham, Hadley and Romain Francois (2015). *dplyr: A Grammar of Data Manipulation*. R package version 0.4.3. URL: http://CRAN.R-project.org/package=dplyr.

Wilf, Herbert S (2013). *generatingfunctionology*. Elsevier.

Wright, Sewall (1922). "Coefficients of inbreeding and relationship". In: *The American Naturalist* 56.645, pp. 330–338.

— (1934). "The method of path coefficients". In: *The Annals of Mathematical Statistics* 5.3, pp. 161–215.

A Convergence of the Thinned Poisson Process to Poisson-Binomial Model

We compared the Poisson thinning approximation and the Poisson-Binomial models. One can show using the law of total expectation that the Poisson-Binomial and Poisson model have the same expected value:

$$\mathbb{E}[N] = \sum_{b=0}^{\infty} \mathbb{E}[N|B=b]P(B=b) \tag{33}$$

$$\mathbb{E}[N] = 1/2^d \left(\sum_{b=0}^{\infty} b\mathrm{Pois}(B=b) + c\sum_{b=0}^{\infty} \mathrm{Pois}(B=b) \right) \tag{34}$$

$$\mathbb{E}[N] = \frac{1}{2^d}(\nu d + c) \tag{35}$$

This is the same expected value as the thinned Poisson process with rate $(\nu d + c)/2^d$. However, the Poisson thinning and Poisson-Binomial models differ in their variance. Using Eve's law, we can show the Poisson-Binomial model has variance

$$\mathrm{V}[N] = \mathbb{E}_B[\mathrm{V}[N|B]] + \mathrm{V}_B[\mathbb{E}[N|B]] \tag{36}$$

$$\mathrm{V}[N] = \frac{d\upsilon + 1}{2^d} - \frac{1}{2^{2d}} \tag{37}$$

This differs from the thinned Poisson process variance by the term $1/2^{2d}$, which grows smaller with increasing d. Finally, we numerically show these two distributions (here, we label the two distributions for k generations $\mu_k(x)$ and $\nu_k(x)$, where x is the number of segments) converge quickly in total variational distance $(d_{TV}(\mu_k, \nu_k) = 1/2 \sum_{n=0}^{\infty} |\mu_k(n) - \nu_k(n)|)$ as k increases, in Figure 10.

B Generating Function for Recombinational Meioses

We also develop a generating function $g(x, k)$ that encodes the number of recombinational meioses in the kth generation as the coefficient for the term x^k. This generating function can also be used in approximations and finding moments of the distribution $p_k(r)$.

Lemma 1. *An expansion of the generating function below encodes the number of lineages with r females in a genealogical chain k generations long $(n_{k,r})$ as the coefficient of the term x^k*

$$g(x, k) = \frac{1}{2^{k+1}\sqrt{x}\sqrt{x+4}} \left[x\left((R^+)^k - (R^-)^k \right) + \sqrt{x}\sqrt{x+4}\left((R^-)^k + (R^+)^k \right) - 2(R^-)^k + 2(R^+)^k \right] \tag{38}$$

where

$$R^- = x - \sqrt{x}\sqrt{x+4} \tag{39}$$

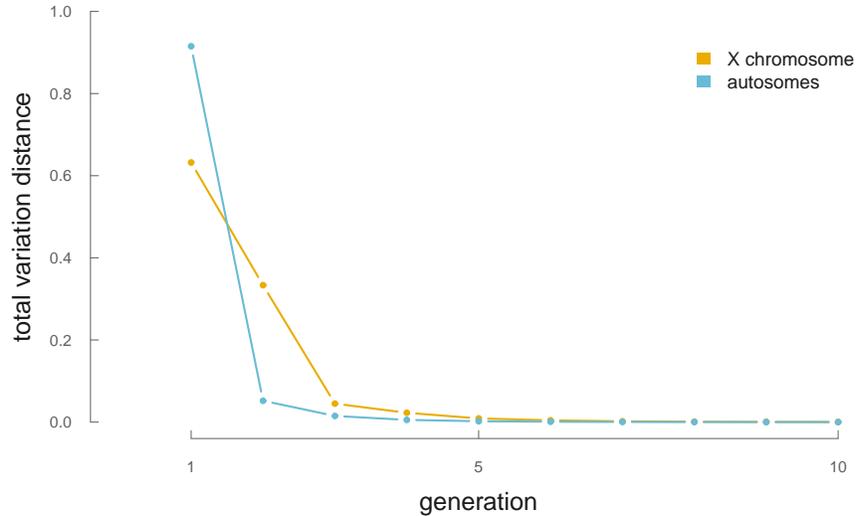$$R^+ = x + \sqrt{x}\sqrt{x+4} \tag{40}$$

Figure 10: The total variation distance between the Poisson thinning and the Poisson-Binomial model for IBD segment number for X segments (yellow) and the autosomal segments (blue).

Proof. We begin by stating some recurrences that occur from the inheritance pattern of X ancestry:

1. $n_{k,r} = m_{k,r} + f_{k,r}$

2. $m_{k,r} = f_{k-1,r}$

3. $f_{k,r} = f_{k-1,r-1} + m_{k-1,r-1}$

Starting from (1):

$$n_{k,r} = m_{k,r} + f_{k,r} \tag{41}$$
$$n_{k,r} = f_{k-1,r} + f_{k,r} \tag{42}$$
$$n_{k,r} = f_{k-1,r} + f_{k-1,r-1} + m_{k-1,r-1} \tag{43}$$
$$n_{k,r} = f_{k-1,r} + n_{k-1,r-1} \tag{44}$$
$$n_{k,r} = f_{k-2,r-1} + m_{k-2,r-1} + n_{k-1,r-1} \tag{45}$$
$$\tag{46}$$

finally, substituting (1) again gives us the desired recurrence relation for $n_{k,r}$:

$$n_{k,r} = n_{k-2,r-1} + n_{k-1,r-1} \tag{47}$$

We can now use generating functions (Wilf, 2013) to tackle this recurrence. Define:

$$A_k(x) = \sum_{r \geq 0} n_{k,r} x^r \tag{48}$$

then, multiply both sides of (47) by x^r and sum over r. On the LHS:

27

$$= \sum_{r \geq 0} n_{k-2,r-1}x^r + \sum_{r \geq 0} n_{k-1,r-1}x^r \tag{49}$$

Note that $n_{k,r} = 0$ if $r < 0$. Multiplying and dividing the first term by x yields:

$$x(n_{k-1,0}x + n_{k-1,1}x^2 + n_{k-1,2}x^3 + \ldots)/x = xA_{k-1}(x) \tag{50}$$

An identical derivation works for the second term. We find:

$$A_k(x) = xA_{k-1}(x) + xA_{k-2}(x) \tag{51}$$

This generating function is in the form of another recurrence.

We can solve this recurrence with the initial conditions below (which can be derived from (47) and its initial conditions):

1. $A_0(x) = 1$

2. $A_1(x) = 1 + x$

The solution to this recurrence with these initial conditions is our desired generating function.

$\square$

We can see that our generating function works via an expansion and verification the coefficients match known numbers of recombinational meioses for some k. For example, let's expand $g(x, k)$ at $k = 5$:

$$x^2 + 6x^3 + 5x^4 + x^5$$

which matches the $n_{k,r}$ values found via computational calculation.

C   Half-Cousins IBD Length Distribution Simulation Results

Figure 11 show the concordance between our cousin IBD segment length analytic distributions and the binned average (1.98cM bin intervals) of 5,000 simulations.

D   An Approximation of X Pedigree Collapse

Since our models of recent X ancestry omit the possibility of pedigree collapse, it is worthwhile to see when this assumption breaks down. To see how pedigree collapse becomes an increasing problem further generations back, we look at the probability that all of a single individual's $\mathcal{F}_{k+2}$ X ancestors, when sampled from a population of N individuals with replacement, are distinct. We treat generations as discrete and non-overlapping, and look at the probability that all $\mathcal{F}_{k+2}$ are distinct individuals as a function of how many generations we go back. This problem is similar to the celebrated birthday problem, but with two rooms of participants: one room of females and another of males. Assuming random mating, each generation, one's X ancestors must be randomly
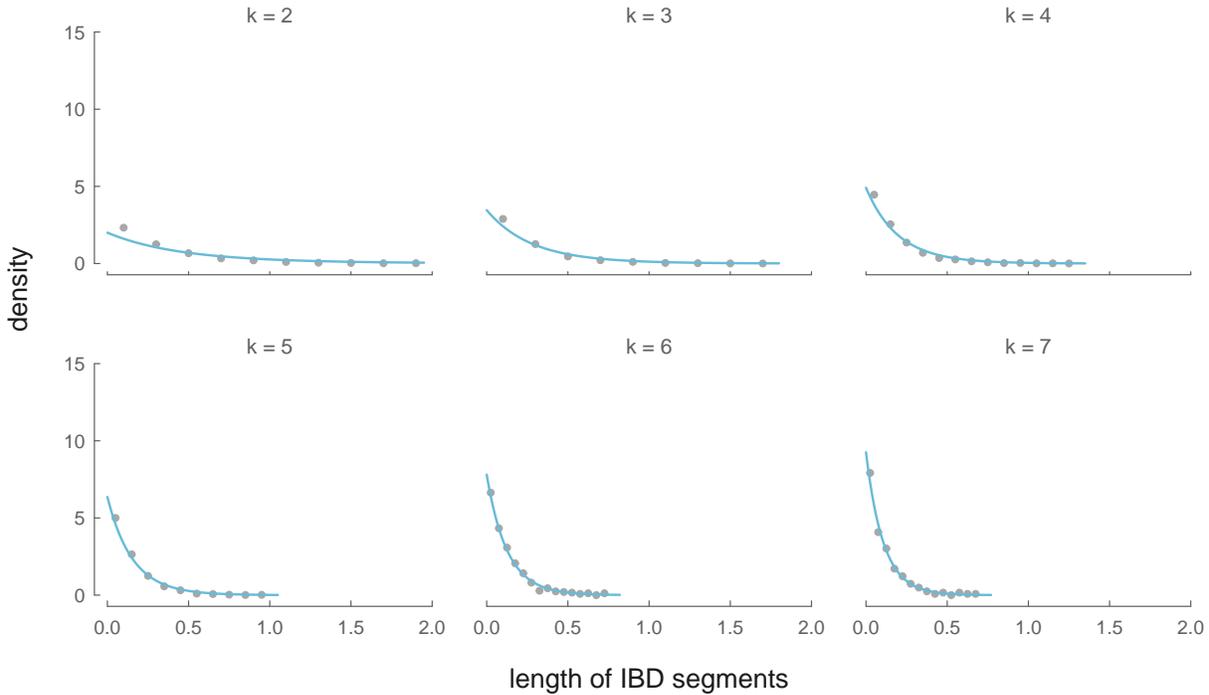
28

Figure 11: The analytic distributions of IBD segment length (blue lines) between two-present-day female half-cousins with a shared ancestor in the kth generation (each panel), and the binned average 5,000 simulations (gray points).

selected with replacement from a population of N individuals. For all ancestors to be distinct, all $\mathcal{F}_k$ male ancestors selected from a pool of $N/2$ and $\mathcal{F}_{k+1}$ female ancestors selected from a pool of $N/2$ must be unique:

$$P(\text{X ancestors all distinct}) = P(\text{male X ancestors unique})P(\text{female X ancestors unique}) \quad (52)$$

$$= \prod_{i=1}^{\mathcal{F}_k} \left(1 - \frac{2i}{N}\right) \prod_{j=1}^{\mathcal{F}_{k+1}} \left(1 - \frac{2j}{N}\right) \quad (53)$$

This probability as a function of k is plotted in Figure 12. For X ancestors, the probability that at least two individuals are not distinct only becomes a significant problem after around 12 generations. Note that this is a very conservative account of how pedigree collapse could affect our calculations; even if two ancestors were to be non-distinct, this is unlikely to affect our calculations greatly. For pedigree collapse to affect our IBD segment models, an individual has to both be a genealogical ancestor *and* a genetic ancestor of the present-day individual; pedigree collapse has no genetic affect if non-distinct individuals are not genetic ancestors.

For other pedigree collapse related quantities (e.g., what's the average number of distinct ancestors k generations back), see Wachter et al. (1979) approximations which use Feller's (1950) occupancy models.
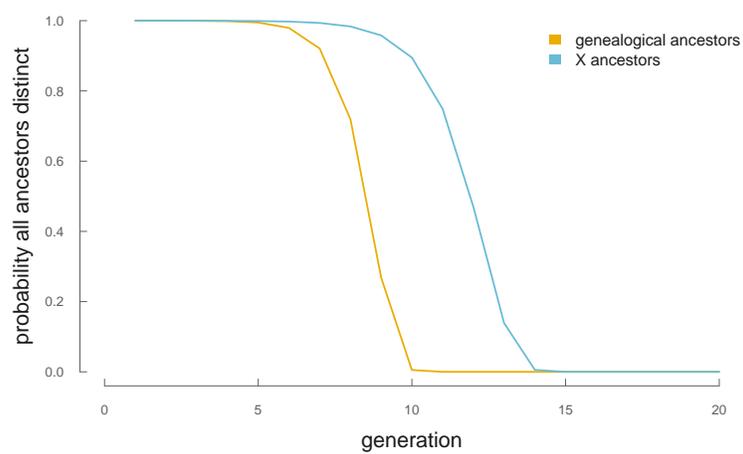
Figure 12: The probability that all genealogical and X ancestors are distinct in a population of $N = 100,000$.