bioRxiv preprint doi: https://doi.org/10.1101/527655; this version posted January 22, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Surprising impacts of crossover interference and sex-specific genetic maps on identical by descent distributions

Madison Caballero[1], Daniel N. Seidman[1], Thomas D. Dyer[2], Donna M. Lehman[3], Joanne E. Curran[2], Ravindranath Duggirala[2], John Blangero[2], and Amy L. Williams[1,*]

[1]Department of Biological Statistics and Computational Biology, Cornell University, Ithaca, NY 14853, USA
[2]South Texas Diabetes and Obesity Institute and Department of Human Genetics, University of Texas Rio Grande Valley School of Medicine, Brownsville, TX 78520, USA
[3]Department of Medicine, University of Texas Health Science Center at San Antonio, San Antonio, Texas 78229, USA

Abstract

Simulations of relatives and identical by descent (IBD) segments are common in genetic studies, yet nearly all past efforts have utilized sex averaged genetic maps while ignoring crossover interference, thus omitting factors known to affect the breakpoints of IBD segments. We developed a method for simulating relatives called Ped-sim that can utilize either sex-specific or sex averaged genetic maps and also either a model of crossover interference or the traditional Poisson distribution for inter-crossover distances. To characterize the impact of previously ignored mechanisms, we simulated data for all four possible combinations of these factors using high resolution human genetic maps and interference parameters. Modeling crossover interference heavily influences the distribution of the proportion of their genome relatives share IBD, decreasing the standard deviation by 11.2% on average for relatives ranging from full siblings to second cousins. By contrast, sex-specific maps increase the standard deviation of IBD proportion by an average of 3.37%, and also impact the number of segments relatives share, most notably producing a bimodal distribution in segment numbers shared by half-siblings. We further compared IBD sharing rates between simulated and real relatives, finding that the combination of sex-specific maps and interference modeling most accurately captures real IBD sharing rates for the relationships we considered. Under this model, pairwise IBD sharing rates depend on the sexes of the individuals through which they are related. For example, when connected only through females (and their common fifth great-grandfather), 12.8% of sixth cousins have some IBD sharing, while this rate drops to 9.05% for male-descent sixth cousins. These analyses demonstrate that sex-specific maps and interference are key factors impacting IBD sharing and underscore the necessity of incorporating these effects into simulations.

Author summary

Simulations are ubiquitous throughout statistical genetics in order to generate data with known properties. Such data are useful both to test inference methods and to understand real world processes using data that may otherwise be challenging to collect. To produce genetic data for relatives in a pedigree, methods must simulate the formation of the chromosomes parents transmit to their children. These chromosomes form as

*Correspondence: alw289@cornell.edu

bioRxiv preprint doi: https://doi.org/10.1101/527655; this version posted January 22, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

a mosaic of a given parent's two chromosomes, with the site of switches between the parent chromosomes known as crossovers. Detailed information about crossover generation based on real data from humans now exists, including the fact that men and women have overall different rates (women produce ∼1.6 times more crossovers) and that real crossovers are subject to *interference*—whereby crossovers are further apart from one another than expected under a model that ignores other crossovers in selecting their location. Our method Ped-sim can simulate pedigree data using these less commonly modeled crossover features, and we used it to compare to real data and to simulations using the more traditional crossover model. These comparisons showed that differences between the sexes and crossover interference each shape the amount of DNA two relatives share identically, better fit real data, and should be included in simulation models.

Introduction

Inferring identical by descent (IBD) segments and estimating relatedness are classical problems in human genetics [1] with much recent work motivated by the abundance of relatives in large samples [2–6]. In order to study individuals with a known relationship, many investigators have performed simulations, both to evaluate novel methods [4–8] and to characterize the properties of IBD sharing rates among relatives [9, 10]. At the same time, work to better understand the dynamics of crossovers, including crossover interference [11–13] and differences in male and female genetic maps [13–15] have yielded more precise resources for realistically simulating these recombination events. Despite this, most prior simulations as well as canonical models of IBD sharing between relatives [16] rely on sex averaged genetic maps and have ignored crossover interference.

Here, we present analyses of IBD distributions using both simulated and real human data where we performed the simulations using either sex-specific or sex averaged crossover genetic maps [14], and incorporated either a crossover interference [11–13] or a non-interference (i.e., Poisson) model. While mean IBD sharing rates are unaffected by these factors, the variance in IBD sharing proportion differs substantially between them. For example, in simulated first cousins, use of sex-specific maps increases the standard deviation of IBD proportion relative to the sex average map by 4.60% and 2.99% while using the Poisson and interference inter-crossover distributions, respectively. More significantly, incorporating crossover interference decreases this standard deviation by 12.1% and 10.8% relative to the Poisson distribution when using sex averaged and sex-specific maps, respectively.

In comparison to real data, relatives simulated using sex-specific maps and a model of crossover interference have IBD sharing rates that better fit those of relatives from the San Antonio Mexican American Family Studies (SAMAFS) [17–19] and Australian (AU) full siblings [20]. While this is expected based on the findings from studies of crossovers [13, 14], the differences between the more idealized models and those typically used are sizable and necessitate a change in current practice. Specifically, the standard deviation of IBD fraction of full siblings simulated under the sex averaged, Poisson crossover model is 0.0397, compared to 0.0362 for the more realistic model. The latter is closer to the value of 0.0374 from the SAMAFS and exact to the reported value of 0.036 for the AU siblings [20]. These differences carry over to more distant relatives, with the size of the 25th to 75th percentile range for second cousins being 0.0159 (0.0224–0.0383) using the sex averaged, Poisson model, and 0.0149 (0.0234–0.0383) under the sex-specific, interference model. Again the latter model more closely matches that of real SAMAFS second cousins who have a corresponding value of 0.0149 (0.0231–0.0380) after correcting for slightly higher than expected mean.

While IBD sharing rates are more strongly affected by including crossover interference than sex-specific maps, the number of shared segments differs noticeably when employing such maps. Most strikingly, half-siblings have a bimodal distribution in number of segments shared when simulated using sex-specific maps. This is due to the fact that half-siblings have only one parent in common, so the IBD segments they share are generated by exactly two male or two female meioses.

Considering a wider range of relationships, we simulated first through sixth cousins and examined their rate of sharing varying numbers of IBD segments with each other. Under the more realistic model, most fifth cousins and nearly all sixth cousins (67.4% and 89.0%, respectively) share no IBD segments with each

bioRxiv preprint doi: https://doi.org/10.1101/527655; this version posted January 22, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

other, consistent with most of these relatives being genetically unrelated. However, the number of nth cousins an individual has grows exponentially as n increases due to exponential population growth in the past several hundred years [21, 22], so many thousands of sixth or more distant cousins still share IBD regions (see Discussion). Interestingly, the probability of sharing IBD depends on the lineage that relates the individuals. Cousins related through female lineages—plus a common $(n-1)$th great-grandfather—have a considerably higher rate of sharing at least one IBD segment (37.6% and 12.8% of fifth and sixth cousins, respectively) compared to those related through male lineages (28.2% and 9.05%, respectively).

We conducted all simulations for this study using Ped-sim, an open source method that performs simulations of relatives using either sex-specific or sex averaged genetic maps and either a model of crossover interference [11–13] or the traditional Poisson model (Methods). Our results indicate that moving to simulations based on these non-standard models will be beneficial in evaluating IBD detection and relatedness inference methods going forward. They are also crucial in determining the probability of IBD sharing between individuals.

Results

We used Ped-sim to simulate 10,000 pairs of relatives for each of several relationship types and each of four crossover models. We compared these with real full siblings from the SAMAFS whose IBD we inferred using HAPI [23], and also compared to more distant SAMAFS relatives whose IBD we detected using Refined IBD [24]. Throughout, we abbreviate sex-specific and sex averaged as SS and SA, respectively, and refer to the four crossover models we used with Ped-sim as: SS+intf for sex-specific genetic map with interference; SS+Poiss for sex-specific map, Poisson event distribution (i.e., no interference); SA+intf for sex averaged genetic map with interference; and SA+Poiss for sex averaged map, Poisson event distribution.

Below, we first discuss the effects of map use and interference on variance in IBD sharing proportion, beginning with simulated data and then compared to real data. Following this, we consider the impact of these models on the number of IBD segments relatives share. Because of the challenges in accurately detecting IBD segments, including very short segments, we focus the latter analysis on simulated data.

Sex-specific maps and interference oppositely affect variance in IBD sharing proportion

We analyzed the impact of SS genetic maps and crossover interference on IBD sharing proportions in simulated full siblings, first cousins, first cousins once removed, and second cousins. The use of SS genetic maps increases the variance in IBD proportion compared to the SA map across all relative types, though the effect is somewhat limited. In particular, averaged among these relationships, the standard deviation increases by 2.98% under the Poisson model for crossover localization and 3.37% under the interference model (Table 1). SS maps have similar effects on the size of the interquartile (25th to 75th percentile) range that we consider in some of our real data analyses, increasing this quantity by an average of 2.89% under the Poisson model and 3.90% in the presence of crossover interference. These small differences in IBD sharing summary statistics correspond to nearly identical distributions of IBD rates between simulations using the SS versus SA maps (S1 Figure).

The factor with the strongest effect on variance in IBD sharing fraction is crossover interference, a component that decreases the standard deviation in IBD sharing compared to the Poisson model by 10.8% when simulating with the SS maps and 11.5% when simulating with the SA map (averaged over all relationships we considered; Table 1). Furthermore, interference tightens the range between the 25th and 75th percentiles by 11.5% when using the SS maps and 10.7% using the SA map. With decreased variances of these magnitudes, the distributions of IBD rates for relatives simulated with interference are noticeably more peaked near the mean, with smaller tails (Figure 1). These results highlight the importance of including interference when simulating relatives, and hint that more IBD segments may be retained between relatives simulated under a model that includes interference—a feature we analyze below.

bioRxiv preprint doi: https://doi.org/10.1101/527655; this version posted January 22, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Relationship	SS+intf	SA+intf	SS+Poiss	SA+Poiss	SAMAFS	SAMAFS corrected
	Mean 25th percentile, 75th percentile Minimum, maximum Standard deviation					
Full siblings	0.500 0.476, 0.524 0.355, 0.643 0.0363	0.500 0.477, 0.524 0.372, 0.620 0.0347	0.500 0.472, 0.528 0.338, 0.646 0.0413	0.500 0.473, 0.527 0.365, 0.660 0.0392	0.500 0.475, 0.525 0.392, 0.612 0.0374	NA
First cousins	0.125 0.111, 0.138 0.0483, 0.213 0.0202	0.125 0.112, 0.138 0.0531, 0.200 0.0194	0.125 0.109, 0.140 0.0461, 0.213 0.0227	0.125 0.110, 0.139 0.0575, 0.216 0.0221	0.127 0.112, 0.140 0.000235, 0.257 0.0231	0.125 0.111, 0.139 0, 0.255 0.0231
First cousins once removed	0.0626 0.0516, 0.0726 0.0181, 0.129 0.0156	0.0624 0.0519, 0.0723 0.0149 0.126 0.0151	0.0624 0.0505, 0.0740 0.00916, 0.135 0.0174	0.0624 0.0504, 0.0734 0.0107, 0.132 0.0170	0.0642 0.0526, 0.0751 0.000287, 0.137 0.0176	0.0625 0.0509, 0.0735 0, 0.135 0.0176
Second cousins	0.0312 0.0234, 0.0383 0.000997, 0.0811 0.0111	0.0311 0.0233, 0.0380 0.00312, 0.0836 0.0108	0.0314 0.0225, 0.0393 0.00142, 0.0906 0.0123	0.0310 0.0224, 0.0383 0.000526, 0.0974 0.0120	0.0339 0.0258, 0.0407 0.00498, 0.0854 0.0113	0.0313 0.0231, 0.0380 0.00234, 0.0827 0.0113

Table 1: IBD sharing fraction summary statistics for relatives of the indicated types. Each cell contains four lines with IBD proportion statistics as indicated at the top of the table: mean; 25th and 75th percentiles; minimum and maximum; and standard deviation. Simulation results based on various genetic map and crossover interference scenarios are as indicated: SS+intf, sex-specific maps with interference; SA+intf, sex averaged map with interference; SS+Poiss, sex-specific maps and Poisson crossover distribution; SA+Poiss, sex averaged map and Poisson crossover distribution. Sharing rates from the real SAMAFS data are based on output from HAPI for full siblings (with two outlier pairs with low IBD rates omitted) and Refined IBD for other relative types (Methods). SAMAFS corrected gives values after mean-centering the distribution separately for each relationship.

bioRxiv preprint doi: https://doi.org/10.1101/527655; this version posted January 22, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

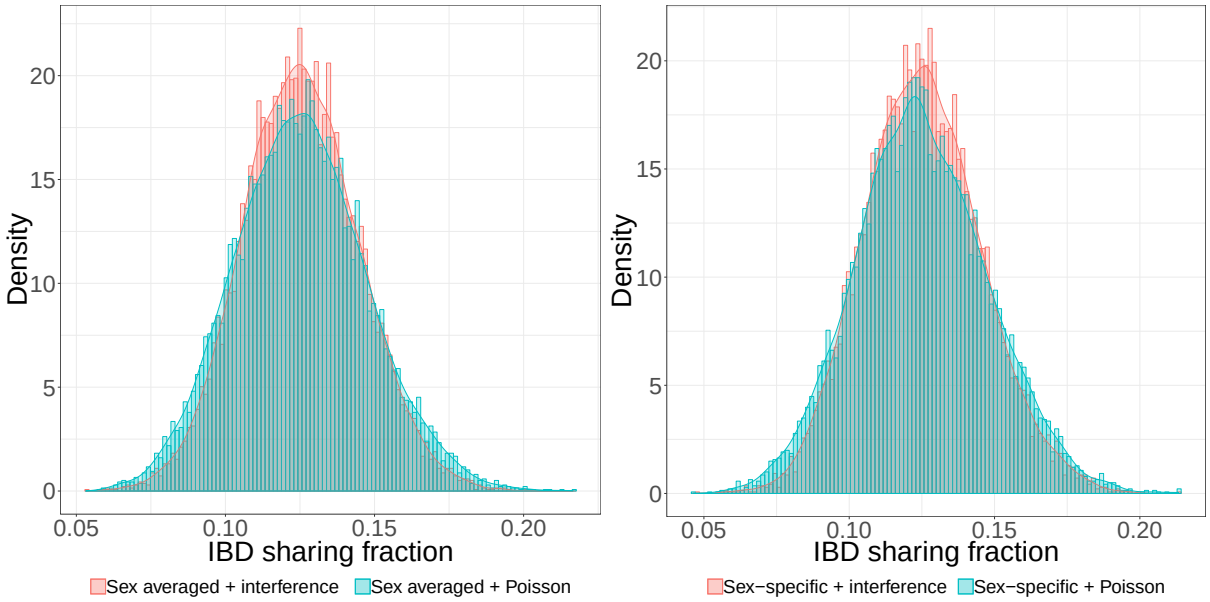

Figure 1: First cousins simulated with crossover interference have a distribution of IBD sharing proportion more concentrated near the mean. Interference decreases the variance in IBD sharing both when using sex averaged (left) and sex-specific (right) genetic maps.

Simulations including sex-specific maps and interference best fit data from real relatives

Given the substantial differences in IBD sharing proportions observed by varying the combination of map type and crossover interference among simulated relatives, we sought to understand which scenario best matches real human data. We first examined IBD sharing between pairs of full siblings in the SAMAFS, where our use of family-based phasing enables IBD detection with high precision and recall. The IBD proportions from the SAMAFS range from a minimum of 0.392 to a maximum of 0.612 (after removing two outlier pairs with proportions of 0.357 and 0.367 [Methods]). This is a much tighter range than in any simulation scenario, but only includes 1,111 SAMAFS full siblings (Table 1; S2 Figure). By contrast, the 4,401 AU siblings have a range of 0.374–0.617 [20], which is closest to the SA+intf model, but also may have fewer extreme values than the simulations due to lower sample size.

The SS+intf model produces the best fit to the standard deviation in IBD proportion with its value of 0.0363 compared to 0.0374 in the SAMAFS and 0.036 in the AU data (Table 1). This contrasts with the traditional SA+Poiss model which has a much higher standard deviation of 0.0397, and the SA+intf and SS+Poiss models which are also quite discrepant at 0.0347 and 0.0413, respectively.

The mean IBD2 sharing rates in the SAMAFS and AU siblings were both 0.248 which is slightly lower than the expected value of 0.25. (Again we removed the same two outlier points which had IBD2 rates of 0.126 and 0.134.) The standard deviation of IBD2 sharing under the SS+intf model is 0.0408, which is the closest of all the models to the SAMAFS standard deviation of 0.0426. The traditional SA+Poiss model is the next closest to the SAMAFS, with a value of 0.0448, while SS+Poiss is higher at 0.0457. The AU pairs have a slightly lower standard deviation at 0.040, and this is somewhat closer to the SA+intf quantity with its standard deviation of 0.0396. Even so, the SS+intf number is very close to the AU sibling quantity.

Turning to relationships more distant than full siblings, we focus on the IBD sharing rates between the first and third quartiles only. This is to remove potentially mislabeled relatives and to reduce the impact of false positive/negative calls that population-based IBD detection methods—which we used to analyze these relatives [24]—are susceptible to. We noted that the mean IBD rates for these real relatives are slightly elevated, potentially due to false positive IBD segments or slightly higher background relatedness in the population. For first cousins, the mean amount of IBD shared exceeds the theoretical expectation by 11.3

bioRxiv preprint doi: https://doi.org/10.1101/527655; this version posted January 22, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

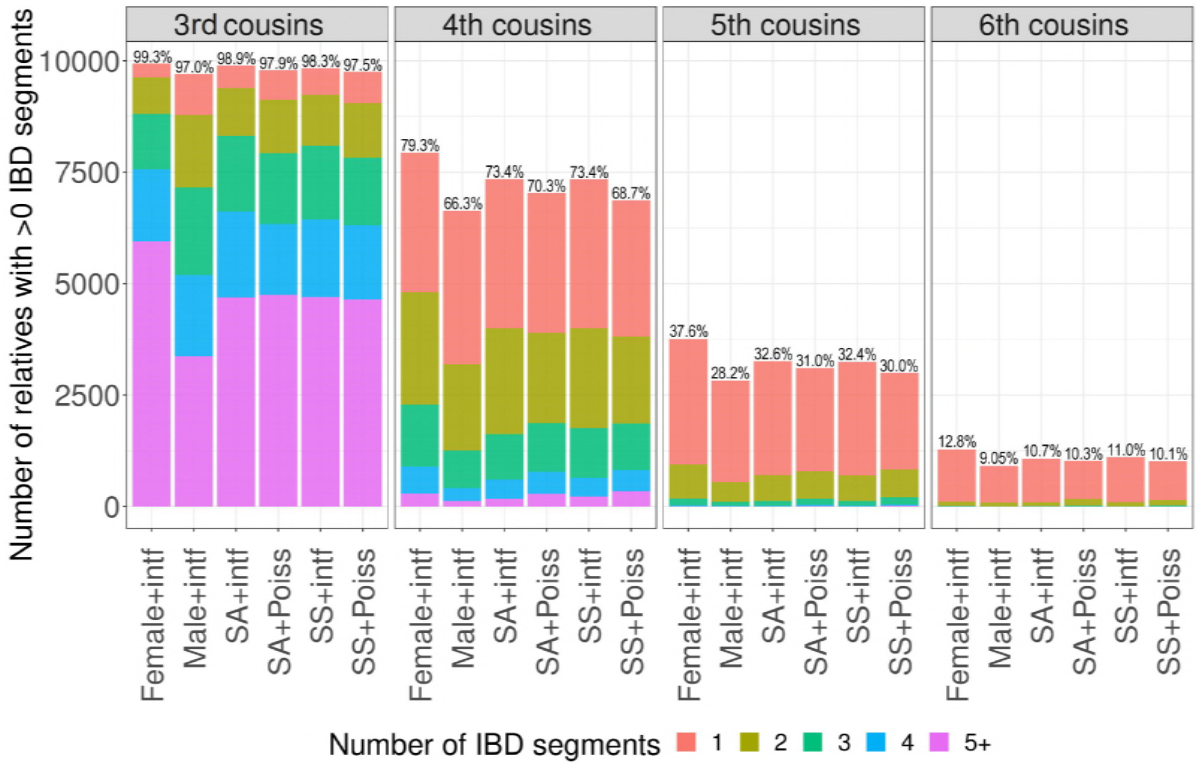

Figure 2: Number of IBD segments shared between simulated third through sixth cousins under various modeling scenarios. More distant relatives have reduced rates of sharing one or more IBD regions. Percentages above each bar indicate the fraction of simulated relatives (of 10,000 for each scenario) that have at least one segment shared. Female+intf are from simulations using sex-specific maps and interference but where the pairs are related through only female non-founders, with a male and female couple as founder common ancestors. Male+intf pairs are the same as Female+intf but with the non-founders being only male instead of female.

cM (0.17% above the expectation). For first cousins once removed and second cousins, the observed means are respectively 11.2 cM (0.34%) and 17.6 cM (0.26%) greater than suggested by theory. We therefore subtracted off these mean excesses for each relationship type in the analyses below (labeled as corrected in Table 1).

As in the full sibling analyses, use of SS genetic maps and crossover interference modeling provide a good fit to the real data across all these more distant relationship types. Notably, however, the 25th and 75th percentile quantities are similar among all the simulation models. In first cousins, the 25th and 75th percentile IBD proportions under the SS+intf model are 0.111 and 0.138—close to 0.111 and 0.139 from the mean-shifted SAMAFS data (Table 1)—but SA+Poiss is equally close, with corresponding percentiles of 0.110 and 0.139. For first cousins once removed and second cousins, none of the models provide a perfect fit to the percentile values. Considering the magnitude of difference between the 25th and 75th percentile values—the size of the interquartile range—SA+Poiss is closest among simulations of first cousin once removed, with a span of 0.0230 (0.0504–0.0734), compared to 0.0225 in the SAMAFS (0.0509–0.0735), and with SS+intf next closest at 0.0210 (0.0516–0.0726). For second cousins, the interquartile range for SS+intf is 0.0149 (0.0234–0.0383), the same as the real second cousins (0.0234–0.0380), and the next best model for second cousins is SA+intf with a percentile difference of 0.0147 (0.0233–0.0380).

bioRxiv preprint doi: https://doi.org/10.1101/527655; this version posted January 22, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Rates of sharing at least one IBD segment among distant relatives

Random assortment during meiosis commonly leads to a loss of IBD segments such that distant relatives may not share any IBD regions with each other despite having a genealogical relationship. Given the fit of the crossover model that incorporates SS maps and interference, we set out to examine the distribution of the number of IBD segments shared among full and half-siblings and first through sixth cousins. For close relatives, including full and half-siblings, and first and second cousins, all simulated pairs share some number of IBD segments with each other regardless of the crossover model. However, some proportion of third through sixth cousins share no IBD segments of any size (Figure 2). In the SS+intf simulation, 1.7% of third cousins share no IBD regions. This increases dramatically to 26.6%, 67.6%, and 89.0% of fourth, fifth, and sixth cousins, respectively. For the 1,101 (of 10,000) sixth cousins that do share IBD segments, the average total length is 5.31 cM. Unsurprisingly, most sixth cousin pairs retain only one IBD segment with very few (96) pairs sharing more than one segment (Figure 2). The total IBD length varies substantially among sixth cousins, with the top 25% of pairs that have IBD regions sharing a total of at least 10.4 cM and a maximum of 53.5 cM. Thus sixth cousins with rare extremes of IBD sharing have total shared lengths more typical of third and fourth cousins.

As already noted, crossover interference leads to a more concentrated distribution of IBD sharing rates (e.g., Figure 1). Interference also influences the number of IBD segments shared among distant relatives, resulting in a slightly larger fraction of distant cousins that share IBD segments. As an example, 32.4% of fifth cousins share one or more IBD segments under the SS+intf model compared to only 30.0% under SS+Poiss.

Sex-specific maps dramatically impact the number of IBD segments relatives share

While SS maps have a smaller effect than interference on the variance in IBD sharing proportion between two relatives, they do impact the number of segments relatives share. Specifically, females produce an average of 1.57× more crossover events per meiosis than males [14]. With such differences, females should transmit a larger number of IBD segments that are on average smaller compared to transmissions from males. This is because, without a crossover event, the probability of transmitting an IBD segment is 50%. On the other hand, when a newly generated crossover occurs within an IBD region, transmission of some portion of the IBD region (on one side or the other of the crossover) is guaranteed.

To more fully investigate the impact of SS genetic maps, we used the SS+intf model to simulate third through sixth cousins where the non-founder ancestors are all either all female or all male (with the shared founder grandparents being a male and female couple). When related primarily through females, third through sixth cousins are much more likely to share some amount of IBD than those related primarily through males. The differences are quite extreme with respectively 2.37%, 19.6%, 33.3%, and 41.4% more (in relative terms) third, fourth, fifth, and sixth female-lineage cousin pairs sharing at least one IBD region compared to the analogous male-lineage cousins (Figure 2). Consistent with intuition, the IBD regions in female-descent cousins are smaller on average than those in male-descent cousins. For example, female-lineage fifth cousins with IBD regions share an average of 1.31 segments with a mean total length of 9.01 cM compared to the male-lineage averages of 1.24 segments with total length 11.9 cM.

These differences in male and female maps impact IBD sharing between close relatives as well, with especially noticeable effects in half-siblings. In particular, maternal half-siblings share on average 1.4 times as many IBD segments as paternal half-siblings (mean segment numbers 51.9 and 37.1, respectively). The effect is substantial enough to produce a bimodal distribution with little overlap between the two types of half-siblings (Figure 3; S3 Figure). Surprisingly, the mean segment count for paternal half-siblings is less than that of first cousins with randomly assigned parent sex (who share a mean of 38.9 segments; Figure 3). However, the segments paternal half-siblings share are more than twice as long with an average length of 45.1 cM compared to 21.5 cM in first cousins.

bioRxiv preprint doi: https://doi.org/10.1101/527655; this version posted January 22, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

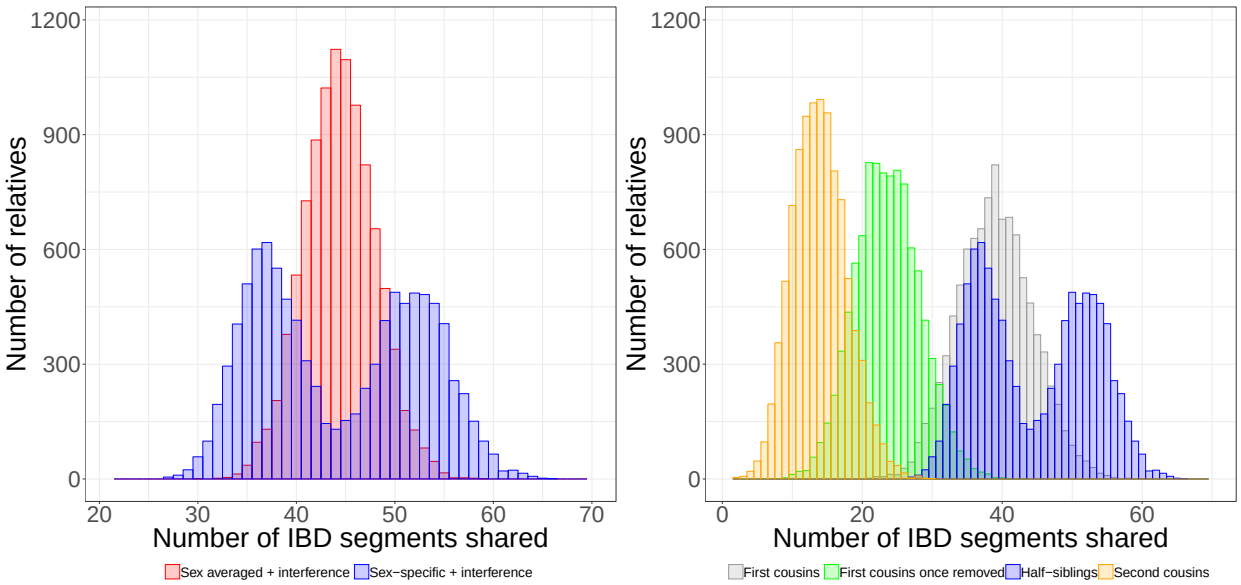

Figure 3: Sex-specific maps impact the number of segments shared between half-siblings. Half-siblings simulated with sex averaged compared to sex-specific maps have widely different shapes, with sex-specific maps producing a bimodal distribution (left). Plotted in the context of other relative types also simulated with sex-specific maps, the lower mode of half-sibling segment counts—which corresponds to IBD sharing between paternal half-siblings (S3 Figure)—is below that of first cousins (right). Includes 10,000 pairs for all relationship types.

Discussion

Modeling relatedness among individuals is more challenging than is typically appreciated due to the complexities of meiotic biology. Variable recombination rates between the sexes and the phenomenon of crossover interference both affect the quantity and size of IBD segments that individuals share. Our analyses demonstrate that use of sex-specific maps and inclusion of crossover interference provide the strongest fit to IBD sharing rates in real human data from full siblings. Likewise, the interquartile range of IBD proportion from real first through second cousins is best fit by jointly modeling sex-specific maps and interference.

In modeling the IBD sharing proportion between relatives, crossover interference has a much stronger influence than varying sex-specific versus sex averaged maps. However, sex-specific maps have a sizable impact on the number of IBD segments that both close (especially half-siblings) and distant relatives share. Therefore, both crossover interference and sex-specific maps play critical roles in the accurate representation of meiotic transmissions.

While we conducted detailed analyses of the IBD sharing rates between real full siblings, our analyses of the first through second cousins considered only the interquartile range and also corrected for excess IBD length above the theoretical mean. This and prior analyses [25] of the SAMAFS data make clear that most of its reported pedigree relationships are accurate. Yet the combination of potential mistakes in relationships and of less precise inference of IBD regions using population-based versus family-based methods necessitated focusing on the median range of relatedness quantities. Work to validate reported relationships in the SAMAFS and other studies may provide a clearer view of the tails of the distribution of IBD sharing between more distant relatives. Nevertheless, controlling for background relatedness remains an issue that may only be possible to mitigate through direct family-based modeling of IBD—i.e., explicitly tracing haplotype segments from founders to their descendants.

Given the effects on IBD sharing of the features we consider here, it is necessary to revisit the probability that a pair of relatives share any IBD with each other. A classic, influential treatment of this problem

bioRxiv preprint doi: https://doi.org/10.1101/527655; this version posted January 22, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

used an analytical approach based on Markov models and considered sex averaged maps while ignoring interference [16]. That study indicated that 10.1% of sixth cousins share IBD regions, which is close to the 10.3% we obtain using a more up-to-date sex averaged genetic map. Still, with both sex-specific maps and interference modeling, we find that 11.0% of simulated sixth cousins have non-zero IBD sharing (an increase of 8.82% compared to the older model). This factor increases to 12.8% when the sixth cousins are related primarily through females, and drops to 9.05% when they are related primarily through males.

The low fractions of IBD sharing between distant cousins contrast with the fact that the number of relatives a given person has increases as the genealogical distance between them grows. As an example, an estimate based on census and birth data of the number of fifth and sixth cousins a person in the UK has are 17,300 and 174,000, respectively [26]. Therefore, though many sixth cousins share no IBD regions, these numbers suggest that an average person in the UK shares IBD regions with over 19,000 sixth cousins. Furthermore, the tails of the IBD sharing distribution between sixth cousins can be as high as that commonly observed in third and fourth cousins.

Going forward, efforts to better understand the dynamics of crossovers, including observed "gamete effects" wherein crossover counts in a given gamete are correlated across chromosomes [27], and also potentially to model genetic variants that effect crossover rates [27] could yield models with even greater precision. Nevertheless, the effects of crossover interference and sex-specific maps are widespread and merit consideration in both crossover and relatedness models.

Methods

We analyzed data from both simulated and real relatives, the latter from the SAMAFS, generating all simulated samples using Ped-sim. Additionally, we quote summary statistics from a study of 4,401 AU full sibling pairs [20].

The IBD sharing statistic we focus on primarily is the proportion of their genome two relatives share IBD, calculated as a fraction of the diploid genome. For a given pair, this proportion is $(k^{(2)} + k^{(1)}/2)$, where $k^{(2)}$ and $k^{(1)}$ are the fraction of positions (in genetic map units) the pair shares IBD2 and IBD1, respectively. Throughout, we report numbers with three significant figures, but we calculated these values using raw data with at least six significant figures.

The Ped-sim algorithm

Ped-sim simulates relatives by tracking haplotypes—initially ignoring genetic data—as a sequence of segments that span a chromosome, each with a numerical identifier denoting the founder haplotype it descends from, and a segment end point. (The start position is implicitly either the beginning of the chromosome or the site following the end of the previous segment.) All founders have two haplotypes with only one chromosome-length segment, each with a unique identifier.

To begin, Ped-sim reads a file that defines the pedigree structure(s) it is to simulate and, for each such structure, generates haplotype segments for the founders in the first generation. For subsequent generations, it generates haplotypes for any founders and forms haplotypes for non-founders from the parents' haplotypes under a meiosis model. This model works on the two haplotypes belonging to a parent by first randomly selecting which of these begins the offspring haplotype, each having 1/2 probability of being selected. Next, Ped-sim samples the location of the crossover events, either using a model of crossover interference or a Poisson model. It then produces the offspring haplotype by copying the segments that exist on the parent's initial haplotype up to the position of the first crossover, introducing a break point in the segment at that location. The interval from this location to the next crossover consists of a copy of the segments from the parent's other haplotype across this interval, again with a break point added at the crossover. The method continues this process, switching at each crossover between the two parent haplotypes to copy segments from, and terminates after copying the region between the last crossover and the end of the chromosome.

bioRxiv preprint doi: https://doi.org/10.1101/527655; this version posted January 22, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Under the Poisson crossover model, the distance from the start of the chromosome to the first crossover and from one crossover to the next are each exponentially distributed with rate equal to 1 crossover/Morgan. This rate arises naturally from the definition of a Morgan as the distance within which an average of one crossover occurs per generation. The model sequentially samples crossovers and terminates after sampling a crossover beyond the end of the chromosome. The sampled positions are in genetic units (i.e., Morgans), and Ped-sim determines their physical location using the genetic map, storing the segment end points as physical base pair positions. When given sex-specific maps, it locates the physical positions using the map corresponding to the sex of the parent.

The crossover interference model is that of Housworth and Stahl [11, 12] which provides a good fit to real human inter-crossover distances [13]. This model uses a gamma distribution for the distances between crossovers that are subject to interference. It also includes a fraction of events that escape interference and therefore derive from a Poisson model.

By default, Ped-sim randomly assigns the sexes of parents, and can generate any number of pedigrees with a given structure (with parent sexes assigned independently in each). Allowable structures include those in which individuals marry either a new founder or another non-founder in the same generation, with multiple marriages possible. This enables simulation of a wide range of possible relationships. Ped-sim can also generate data in which all reproducing non-founders have the same sex (male or female), leading to descendants that are related to each other through nearly all male or all female relatives. In such cases, the common ancestors of those descendants will generally be a married couple, male and female, although it is possible to simulate only a single common ancestor or for individuals to be related through more than one lineage (e.g., double first cousins).

When given genetic data in the form of input haplotypes, Ped-sim randomly assigns data to each founder using one input sample with its pair of haplotypes. It then copies segments from those haplotypes to the person's descendants using the segment numerical identifiers and end points. The algorithm can also introduce genotyping errors and missing data using user-specified rates, with a uniform probability of these events at all positions.

Another way to run Ped-sim is without haplotype data—instead using only the segment numerical identifiers and end points that form simulated haplotypes. This is the way we used Ped-sim for the analyses we describe here, using version 0.99 of the tool. We detected IBD segments as regions where two individuals have one (IBD1) or two (IBD2) segments with the same numerical identifier(s). For each such IBD segment, we mapped the physical positions to genetic positions using the same sex averaged map that we used in the simulations (below) [14].

Genetic maps and crossover interference parameters

We ran Ped-sim using genetic maps produced from crossovers detected in data from over 100,000 meioses [14]. The inferred maps include those for both males and females and a sex averaged map that all span the same physical range. All simulations include the 22 autosomes but no sex chromosomes.

To simulate using the Housworth and Stahl crossover interference model, we leveraged the interference parameters ν_i and p_i for $i \in \{f, m\}$ that were inferred using, respectively, female and male data from a total of over 18,000 meioses [13]. To simulate interference using a sex averaged map, we calculated the sex averaged parameters ν_a and p_a as follows. The p parameter gives the fraction of events that escape interference and we set $p_a = (p_f + p_m)/2$. Distances between interfering crossovers in the tetrad are gamma distributed with shape and rate parameter values ν and 2ν, respectively [11, 12]. A simple average of the male and female ν parameters does not produce a distribution with summary statistics at the midpoint between the two sexes. All values of ν lead to distributions with the same expected value of 1/2 (the expectation is the shape divided by the rate parameters). We therefore calculated ν_a such that the variance of the sex averaged distribution is the mean of the variances of the male and female models: $\nu_a = (\frac{1}{2\nu_f} + \frac{1}{2\nu_m})^{-1}$. Note that the mean distance of 1/2 Morgans between events is double the number expected per chromatid. This is because the model is for events in a tetrad (all four products of meiosis). To obtain the crossovers falling on the chromatid being generated, the model randomly selects events with probability 1/2 for inclusion.

bioRxiv preprint doi: https://doi.org/10.1101/527655; this version posted January 22, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

IBD detection in the SAMAFS

We used two different methods for inferring IBD in the SAMAFS data: one applied to nuclear families and useful for analyzing IBD sharing between full siblings, and the other for analyzing IBD rates in first cousins, first cousins once removed, and second cousins. Quality control filtering of the SAMAFS data is the same as that described previously [25]. In brief, we used biallelic SNPs typed on the Illumina Human660W, Human1M, Human1M-Duo, or both the HumanHap500 and the HumanExon510S arrays, and we required the SNP probe sequences to map to a single location in the human GRCh37 build. Next, we excluded individuals and SNPs with excessive missing data (>10% and >2%, respectively) and removed duplicate SNPs. Additional SNP filters utilized information from auxiliary resources including dbSNP and the reported "accessible genome" from the 1000 Genomes Project, among others [25]. This yielded data for 2,485 samples typed at 521,184 SNPs. We further omitted 1,514 first cousin, first cousin once removed, and second cousin relative pairs that had evidence of being related through more than one lineage [25].

Family-based phasing implicitly infers IBD regions, and in the presence of data for a complete nuclear family, this inference has both high precision and recall. For this analysis, we utilized HAPI [23] version 1.87.6b, a method that performs efficient minimum recombinant phasing for nuclear families. This form of phasing is the same as that of the Lander-Green algorithm [28] when the probability of recombination between informative markers is identical at each position. To ensure reliable results, we performed this inference on 114 nuclear families for which data from both parents and three or more children were available, and we excluded likely monozygotic twins that had IBD2 rates >0.95 (three pairs). This yielded 1,111 full sibling pairs for analysis. To infer IBD regions, we parsed the inheritance vector output from HAPI to locate IBD1 and IBD2 segments, assigning genetic positions to the start and end of each such region using the same sex averaged map we used for the simulated data [14]. (The genetic map is undefined for 98 SNPs and we omitted these positions from analysis.) The exact boundaries of crossover positions are uncertain in real data due to the fact that not all sites are genotyped and homozygous positions are uninformative. We therefore estimated the start and end positions as the midpoint in genetic units between two informative sites that descend from distinct parental haplotypes and therefore bound the region in which a crossover broke an IBD segment. We also merged short regions (including non-IBD intervals) comprised of five or fewer informative SNPs with the adjacent segments so that they cover the interval. We assign these to have the same IBD type as the preceding segment (typically the two flanking segments have the same type). Finally, we removed two outlier full sibling pairs that had IBD proportions of 0.357 and 0.367.

For non-sibling relatives, we leveraged IBD segments previously inferred [25] using Refined IBD [24] version 4.1. In total, we considered 5,384 pairs of first cousins, 6,342 first cousins once removed, and 2,584 second cousins. Here as well we converted physical positions of the IBD segments to sex averaged genetic positions using the same sex averaged map as in other analyses [14].

Ethics statement

This study makes use of deidentified individuals from the SAMAFS and received exemption (#4) from IRB review from the Cornell University IRB (protocol 1408004874).

Acknowledgments

We thank Monica Ramstetter for help in testing Ped-sim. Support for this work was provided by an Alfred P. Sloan Research Fellowship and a seed grant from Nancy and Peter Meinig to A.L.W. Additionally, M.C. was supported by NIH grant T32 GM007617-37 and D.N.S. was partially supported by the NIH grant T32 GM083937. We are grateful to the participants in the SAMAFS; funding for those studies was provided by NIH grants R01 HL0113323, P01 HL045222, R01 DK047482, and R01 DK053889.

bioRxiv preprint doi: https://doi.org/10.1101/527655; this version posted January 22, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

References

[1] Bruce S Weir, Amy D Anderson, and Amanda B Hepler. Genetic relatedness analysis: modern data and new challenges. *Nature Reviews Genetics*, 7(10):771–780, 2006.

[2] Clare Bycroft, Colin Freeman, Desislava Petkova, Gavin Band, Lloyd T. Elliott, Kevin Sharp, Allan Motyer, Damjan Vukcevic, Olivier Delaneau, Jared O'Connell, Adrian Cortes, Samantha Welsh, Alan Young, Mark Effingham, Gil McVean, Stephen Leslie, Naomi Allen, Peter Donnelly, and Jonathan Marchini. The UK Biobank resource with deep phenotyping and genomic data. *Nature*, 562(7726):203–209, 2018.

[3] 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. *The American Journal of Human Genetics*, 102(5):874–889, May 2018.

[4] Jeffrey Staples, David J Witherspoon, Lynn B Jorde, Deborah A Nickerson, Jennifer E Below, Chad D Huff, University of Washington Center for Mendelian Genomics, et al. PADRE: Pedigree-aware distant-relationship estimation. *The American Journal of Human Genetics*, 99(1):154–162, 2016.

[5] Amy Ko and Rasmus Nielsen. Composite likelihood method for inferring local pedigrees. *PLOS Genetics*, 13(8):e1006963, 08 2017.

[6] Monica D. Ramstetter, Sushila A. Shenoy, Thomas D. Dyer, Donna M. Lehman, Joanne E. Curran, Ravindranath Duggirala, John Blangero, Jason G. Mezey, and Amy L. Williams. Inferring identical-by-descent sharing of sample ancestors promotes high-resolution relative detection. *The American Journal of Human Genetics*, 103(1):30–44, 2018.

[7] Jeffrey Staples, Dandi Qiao, Michael H. Cho, Edwin K. Silverman, Deborah A. Nickerson, and Jennifer E. Below. Primus: Rapid reconstruction of pedigrees from genome-wide estimates of identity by descent. *The American Journal of Human Genetics*, 95(5):553–564, Nov 2014.

[8] Michael P. Epstein, William L. Duren, and Michael Boehnke. Improved inference of relationship for pairs of individuals. *The American Journal of Human Genetics*, 67(5):1219–1231, Nov 2000.

[9] WG Hill and BS Weir. Variation in actual relationship as a consequence of Mendelian sampling and linkage. *Genetics Research*, 93(01):47–64, 2011.

[10] William G. Hill and Ian M. S. White. Identification of pedigree relationship from genome sharing. *G3: Genes, Genomes, Genetics*, 3(9):1553–1571, 2013.

[11] Karl W Broman and James L Weber. Characterization of human crossover interference. *The American Journal of Human Genetics*, 66(6):1911–1926, 2000.

[12] EA Housworth and FW Stahl. Crossover interference in humans. *The American Journal of Human Genetics*, 73(1):188–197, 2003.

[13] Christopher L. Campbell, Nicholas A. Furlotte, Nick Eriksson, David Hinds, and Adam Auton. Escape from crossover interference increases with maternal age. *Nature Communications*, 6:6260, Feb 2015.

[14] Claude Bhérer, Christopher L Campbell, and Adam Auton. Refined genetic maps reveal sexual dimorphism in human meiotic recombination at multiple scales. *Nature Communications*, 8, 2017.

[15] Augustine Kong, Gudmar Thorleifsson, Daniel F Gudbjartsson, Gisli Masson, Asgeir Sigurdsson, Aslaug Jonasdottir, G Bragi Walters, Adalbjorg Jonasdottir, Arnaldur Gylfason, Kari Th Kristinsson, et al. Fine-scale recombination rate differences between sexes, populations and individuals. *Nature*, 467(7319):1099–1103, 2010.

bioRxiv preprint doi: https://doi.org/10.1101/527655; this version posted January 22, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

[16] Kevin P. Donnelly. The probability that related individuals share some section of genome identical by descent. *Theoretical Population Biology*, 23(1):34–63, 1983.

[17] Braxton D Mitchell, Candace M Kammerer, John Blangero, Michael C Mahaney, David L Rainwater, Bennett Dyke, James E Hixson, Richard D Henkel, R Mark Sharp, Anthony G Comuzzie, et al. Genetic and environmental contributions to cardiovascular risk factors in Mexican Americans. *Circulation*, 94(9):2159–2170, 1996.

[18] Ravindranath Duggirala, John Blangero, Laura Almasy, Thomas D Dyer, Kenneth L Williams, Robin J Leach, Peter O'Connell, and Michael P Stern. Linkage of type 2 diabetes mellitus and of age at onset to a genetic location on chromosome 10q in Mexican Americans. *American Journal of Human Genetics*, 64(4):1127–1140, 1999.

[19] Kelly J Hunt, Donna M Lehman, Rector Arya, Sharon Fowler, Robin J Leach, Harald HH Göring, Laura Almasy, John Blangero, Tom D Dyer, Ravindranath Duggirala, et al. Genome-wide linkage analyses of type 2 diabetes in Mexican Americans. *Diabetes*, 54(9):2655–2662, 2005.

[20] Peter M. Visscher, Sarah E. Medland, Manuel A. R. Ferreira, Katherine I. Morley, Gu Zhu, Belinda K. Cornes, Grant W. Montgomery, and Nicholas G. Martin. Assumption-free estimation of heritability from genome-wide identity-by-descent sharing between full siblings. *PLoS Genetics*, 2(3):e41, 2006.

[21] Alex Coventry, Lara M. Bull-Otterson, Xiaoming Liu, Andrew G. Clark, Taylor J. Maxwell, Jacy Crosby, James E. Hixson, Thomas J. Rea, Donna M. Muzny, Lora R. Lewis, David A. Wheeler, Aniko Sabo, Christine Lusk, Kenneth G. Weiss, Humeira Akbar, Andrew Cree, Alicia C. Hawes, Irene Newsham, Robin T. Varghese, Donna Villasana, Shannon Gross, Vandita Joshi, Jireh Santibanez, Margaret Morgan, Kyle Chang, Walker Hale IV, Alan R. Templeton, Eric Boerwinkle, Richard Gibbs, and Charles F. Sing. Deep resequencing reveals excess rare recent variants consistent with explosive population growth. *Nature Communications*, 1:131, Nov 2010.

[22] Alon Keinan and Andrew G. Clark. Recent explosive human population growth has resulted in an excess of rare genetic variants. *Science*, 336(6082):740–743, 2012.

[23] Amy L. Williams, David Housman, Martin Rinard, and David Gifford. Rapid haplotype inference for nuclear families. *Genome Biology*, 11(10):R108, 2010.

[24] Brian L. Browning and Sharon R. Browning. Improving the accuracy and efficiency of identity-by-descent detection in population data. *Genetics*, 194(2):459–471, 2013.

[25] Monica D. Ramstetter, Thomas D. Dyer, Donna M. Lehman, Joanne E. Curran, Ravindranath Duggirala, John Blangero, Jason G. Mezey, and Amy L. Williams. Benchmarking relatedness inference methods with genome-wide data from thousands of relatives. *Genetics*, 207(1):75–82, 2017.

[26] One in 300 chance that a complete stranger is your cousin. `https://www.ancestry.com/corporate/international/press-releases/One-in-300-chance-that-a-complete-stranger-is-your-cousin`. Accessed: 13 Dec 2018.

[27] Augustine Kong, Gudmar Thorleifsson, Michael L. Frigge, Gisli Masson, Daniel F. Gudbjartsson, Rasmus Villemoes, Erna Magnusdottir, Stefania B. Olafsdottir, Unnur Thorsteinsdottir, and Kari Stefansson. Common and low-frequency variants associated with genome-wide recombination rate. *Nature Genetics*, 46:11–16, Nov 2013. Article.

[28] E. S. Lander and P. Green. Construction of multilocus genetic linkage maps in humans. *Proceedings of the National Academy of Sciences*, 84(8):2363–2367, 1987.

bioRxiv preprint doi: https://doi.org/10.1101/527655; this version posted January 22, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Supporting information legends

S1 Figure: First cousins simulated either with sex-specific or sex averaged maps have distributions of IBD sharing proportion that very similar. Sex-specific and sex averaged distributions heavily overlap both when using an interference (left) and a Poisson (right) model for inter-crossover distances.

S2 Figure: Distributions of IBD sharing fractions for simulated full siblings and the SAMAFS full siblings. Simulated distributions are as labeled: SS+intf, sex-specific maps with interference; SA+intf, sex averaged map with interference; SS+Poiss, sex-specific maps and Poisson crossover distribution; SA+Poiss, sex averaged map and Poisson crossover distribution. Each simulation includes 10,000 full sibling pairs, and the SAMAFS data are from 1,113 pairs, including the two low IBD fraction pairs excluded in the analyses described in the main text.

S3 Figure: Number of IBD segments shared between the two types of half-siblings: maternal, with female shared parents, and paternal, with male shared parents. Includes 10,000 pairs for both types of half-siblings.