IBD-based estimation of X chromosome effective population size with application to sex-specific demographic history

Ruoyi Cai,[1*] Brian L. Browning,[1,2] Sharon R. Browning[1*]

[1] Department of Biostatistics, University of Washington

[2] Division of Medical Genetics, Department of Medicine, University of Washington

[*] Address for correspondence: RC: rcai2@uw.edu, SRB: sguy@uw.edu

Abstract

Effective population size (N_e) in the recent past can be estimated through analysis of identity-by-descent (IBD) segments. Several methods have been developed for estimating N_e from autosomal IBD segments, but no such effort has been made with X chromosome IBD segments. In this work, we propose a method to estimate the X chromosome effective population size from X chromosome IBD segments. We show how to use the estimated autosome N_e and X chromosome N_e to estimate female and male effective population sizes. We demonstrate the accuracy of our autosome and X chromosome N_e estimation with simulated data, and we show that pronounced differences in female and male effective population size can be detected. In analysis of UK Biobank and the TOPMed project data, we obtain results that are consistent with equal female and male effective population sizes in the UK White British and UK Indian populations and in an African American population.

Introduction

The effective size of a population (N_e) is defined as the number of breeding individuals in an idealized randomly mating population that has the same expected value of a parameter of interest as the actual population under consideration.[1] The effective population size is a fundamental parameter in population genetics because it determines the strength of genetic drift and the efficacy of evolutionary forces such as mutation, selection, and migration.[2] Previous studies have demonstrated that estimates of recent effective population size can reveal aspects of a population's demographic history, such as past population growth or bottleneck events.[3,4]

Identity-by-descent (IBD) segments can be used to estimate effective population size in the recent past. IBD segments are haplotypes which two or more individuals have inherited from a common ancestor. IBD segments end at positions where crossovers have occurred in the meioses between the common ancestor and the descendant individuals. IBD segments for which the common ancestor is in the distant past tend to be shorter than IBD segments from a recent common ancestor because there are more meioses since the common ancestor on which crossovers can occur. The autosomes and the X chromosome are both subject to recombination, making them both amenable to IBD segment analysis.[5,6] Previous studies have developed methods for estimating recent effective population size from autosomal IBD segments.[3,4,7] However, no such effort has been made with X chromosome IBD segments.

The autosomes are equally influenced by female and male demographic processes. In contrast, the X chromosome is influenced more strongly by female demographic processes than by male demographic processes. This is because females have two copies of the X chromosome, while males have only one. Thus, comparison of statistics from the X chromosome and from autosome data can be used to estimate sex-specific parameters such as female and male effective population sizes.[6,8-16]

The standard Wright-Fisher model used to define the effective population size assumes equal numbers of breeding females and males. To define female and male effective population sizes, we consider an idealized Wright-Fisher population modified to allow for different numbers of females and males. The female and male effective sizes (N_e^f and N_e^m) of a non-idealized population are the values that would give the same rates of IBD on autosomes and sex chromosomes as the Wright-Fisher population with the corresponding numbers of females and males.[17] The effective sex ratio (ESR) is the ratio of the effective number of females to the sum of the effective number of females and the effective number of males, $N_e^f/(N_e^f + N_e^m)$.[18]

In this work, we develop an IBD-based method to estimate the X chromosome effective population size. We also show that estimated X chromosome N_e can be combined with estimated autosome N_e to estimate the female and male effective population sizes and the effective sex ratio. Through simulation studies, we validate the theoretical relationship between the autosome and X chromosome N_e, and we show that our method can accurately estimate the autosome and X chromosome N_e in simulated populations. We examine the application of X chromosome N_e to estimate sex-specific N_e and show that pronounced differences in female and male effective population sizes can be detected. We use our method to infer autosome and X chromosome N_e, as well as sex-specific N_e, for several human populations.

Methods

Probability modelling for the X chromosome

All meioses from mothers transmit an X chromosome, while approximately half of meioses from fathers transmit an X chromosome. Over many generations, approximately two-thirds of meioses from parents to offspring that transmit an X chromosome will be from females, as we show below. However, the proportion of females in a sample of individuals affects the actual proportion of meioses transmitting an

X chromosome in recent generations that are from females. For example, if the sample is solely made up of females (with two X chromosomes, one from each parent), half of the meioses in the most recent generation that transmit an X chromosome are from females. If the sample is made up solely of males (with only one X chromosome, inherited from the mother), all the meioses that directly contribute the X chromosomes in the sample are from females. If half the sampled individuals are female and half are male, then 2/3 of the meioses in the previous generation that contribute the X chromosomes in the sample will be from females. We show in the next paragraph that when half of the samples are female this 2/3 ratio will apply in all prior generations, regardless of sex-specific demographic forces.

Consider the lineage of a randomly selected haplotype at a point in the genome. Let p_g be the probability that the ancestral haplotype at generation g before present is carried by a female, where $g = 0$ corresponds to the generation of the sampled individuals, $g = 1$ to the generation of their parents, and so on. A given haplotype carried by a female has a 50% probability that its parent haplotype is carried by a female, while a given haplotype carried by a male always has its parent haplotype carried by a female. Thus, for $g \geq 0$,

$$p_{g+1} = 0.5p_g + (1 - p_g) = 1 - 0.5p_g.$$

If $p_g = 2/3$, then $p_{g+1} = 2/3$. If the sampled haplotype is randomly chosen from a set of individuals with equal numbers of females and males, then the sampled haplotype has probability $p_0 = 2/3$ of being carried by a female, and as a result $p_g = 2/3$ for all g. More generally, it can be shown by mathematical induction that

$$p_g = \frac{2}{3}\left(1 - \left(-\frac{1}{2}\right)^g\right) + \left(-\frac{1}{2}\right)^g p_0$$

which converges to $2/3$ for large g. In what follows, we assume that $p_g = 2/3$ for all g.

79 We model the probability distribution of the length of an X chromosome IBD segment resulting from a
80 coalescence event occurring g generations ago as an exponential random variable with mean length
81 $50/g$ cM on the sex-averaged recombination map. This is the same model that we use for estimating
82 IBD-based effective population size in autosomal data.[3] The model is somewhat inaccurate for small g
83 on the X chromosome because of the large difference between the female and male recombination
84 maps (with the male-specific X chromosome map having a recombination rate of 0). For larger g,
85 averaging over multiple generations ensures that the exponential distribution in this model is a good fit.
86 We investigate the performance of the method with this model through simulations.

87 IBD-based estimation of X chromosome effective population size

88 Our method for IBD-based estimation of X chromosome effective size history is based on IBDNe which
89 was designed to estimate recent effective population size from autosomal IBD segments.[3] The IBDNe
90 method calculates the expected length distribution of IBD segments exceeding a given length threshold
91 (such as 2 cM) for a given effective population size history. It finds the effective size history that equates
92 the observed and expected IBD length distributions using an iterative scheme. IBDNe applies smoothing
93 over intervals of eight generations to avoid overfitting. Although IBDNe was designed for autosomal
94 data, we show that it can also be used with X chromosome data, with some adjustments to the analysis
95 procedure that we describe in the following paragraphs.

96 The first adjustment ensures proper inference of IBD segments on the X chromosome by encoding male
97 X chromosome genotypes as haploid. This coding conforms to the VCF specification.[19] Male X
98 chromosome genotypes are frequently coded as homozygous diploid genotypes rather than haploid
99 genotypes, which typically results in duplicate reported IBD segments when using IBD detection
100 methods designed for autosomal data. The hap-ibd[20] program can correctly analyze chromosome X data
101 with haploid male genotypes. We exclude the pseudo-autosomal regions from analysis.

102 The second adjustment is to use a sex-averaged genetic map, as we also do for the autosomes. X
103 chromosomes transmitted from females are subject to recombination, while X chromosomes
104 transmitted from males are not (except for the pseudo-autosomal regions which we exclude from all
105 analyses). Genetic maps, such as the HapMap map[21] and the deCODE map[22], typically report the female-
106 specific recombination map for the X chromosome. Since an average of 2/3 of meioses transmitting an X
107 chromosome are from females, the sex-averaged X chromosome recombination map can be obtained by
108 multiplying female-specific genetic distances by 2/3. For example, a region with length 3 cM on the
109 female-specific map has length 2 cM on the sex-averaged map. Equivalently, the sex-averaged
110 recombination rates can be obtained by multiplying female-specific recombination rates by 2/3. For
111 example, an X chromosome region with female-specific recombination rate of 3×10^{-8} per base pair
112 per generation has a sex-averaged recombination rate of 2×10^{-8} per base pair per generation.

113 The third adjustment ensures equal numbers of sampled females and males. If the sample is
114 unbalanced, we selectively remove some individuals to obtain equal numbers of females and males.
115 Consequently, p_0, the proportion of sampled X chromosome haplotypes carried by females, is 2/3, and
116 hence p_g, the probability that the ancestral haplotype of a sampled X chromosome haplotype g
117 generations before the present is carried by a female is always 2/3 (see the preceding "Probability
118 Modelling for the X chromosome" section).

119 The fourth adjustment modifies the IBDNe "npairs" parameter to be equal to the number of analyzed
120 haplotype pairs. By default, IBDNe assumes that each individual contributes two haplotypes to the
121 analysis, and that all cross-individual pairs are analyzed, resulting in $(2n)(2n-2)/2$ haplotype pairs
122 when there are n individuals. On the X chromosome, with n_f females and n_m males, the number of
123 haplotype pairs is

124 $$2n_f(2n_f-2)/2 \; + \; n_m(n_m-1)/2 \; + \; 2n_f n_m.$$

125 Equation 1

126 We set the IBDNe "npairs" parameter to the value in equation 1 when analyzing X chromosome data,
127 after adjusting it for removal of close relative pairs as described below.

128 The fifth adjustment is to manually remove detected IBD segments corresponding to close relatives
129 (parent-offspring and siblings). By default, IBDNe identifies the close relatives using the input IBD
130 segments, removes the IBD segments between them, and adjusts the "npairs" parameter to account for
131 the removed sample pairs. However, this strategy does not work for the X chromosome because one
132 cannot reliably detect close relatives using only X chromosome data. We thus turn off IBDNe's filtering
133 of close relatives by setting "filtersamples=false". We can identify close relatives based on autosomal
134 data or from a pedigree file if available, and then remove IBD segments for these pairs and update the
135 "npairs" parameter accordingly in the chromosome X analysis.

136 The sixth adjustment enables calculation of confidence intervals. IBDNe obtains confidence intervals for
137 the estimated effective sizes by bootstrapping over chromosomes. We thus divide the X chromosome
138 into six pieces of equal cM length and treat these as separate "chromosomes" in the analysis with
139 IBDNe.

140 From X chromosome effective population size to sex-specific effective population size

141 We next describe how the estimated X chromosome effective population size can be used in conjunction
142 with the estimated autosome effective population size to estimate female and male effective population
143 sizes. We will write N_g^X and N_g^A for the X chromosome and autosomal effective sizes at generation g.
144 And we will write N_g^f and N_g^m for the female and male effective population sizes at generation g, which
145 can be derived from the X chromosome and autosomal effective sizes as described below.

The IBD-based effective population size (for autosomes or X chromosome) is defined in terms of the conditional coalescence probability for a Wright-Fisher population. We first consider autosomes. For a randomly selected pair of haplotypes, let G be the number of generations before present that the haplotypes coalesce. Conditional on the haplotypes not coalescing by generation $g-1$ before present, the ancestral haplotypes are distinct at that generation. For them to coalesce at generation g before present, their two parental haplotypes at generation g must be the same haplotype. If the diploid autosomal effective size is N_g^A at g generations before present, there are $2N_g^A$ autosomal haplotypes available, and the probability that the two parental autosomal haplotypes are the same is thus $1/(2N_g^A)$. That is, $P_A(G=g|G>g-1)=1/(2N_g^A)$. Thus, if we know the value of $P_A(G=g|G>g-1)$, then we can obtain the effective size g generations before present:

$$N_g^A = 1/\big(2P_A(G=g|G>g-1)\big).$$

On the X chromosome, the conditional coalescence probability can be obtained by considering that 2/3 of meioses are from female parents, while 1/3 are from male parents. For coalescence to occur, both haplotypes' parent haplotypes must be the same. This means both haplotypes must be inherited from parents that have the same sex, and the second haplotype must have the same parental haplotype as that of the first (within that sex). Thus,

$$P_X(G=g|G>g-1) = \frac{(2/3)^2}{2N_g^f} + \frac{(1/3)^2}{N_g^m}.$$

And hence,

$$N_g^X = 1/\big(2P_X(G=g|G>g-1)\big)$$

$$= \frac{9N_g^f N_g^m}{2N_g^f + 4N_g^m}$$

Equation 2

For comparison, on the autosomes, by the same reasoning but with half of the meioses from each sex and with diploid males,

$$P_A(G = g|G > g - 1) = \frac{(1/2)^2}{2N_g^f} + \frac{(1/2)^2}{2N_g^m}.$$

And hence,

$$N_g^A = \frac{1}{2}\Big/\left(\frac{(1/2)^2}{2N_g^f} + \frac{(1/2)^2}{2N_g^m}\right) = \frac{4N_g^f N_g^m}{N_g^f + N_g^m}.$$

Equation 3

From Equations 2 and 3, the ratio of X to autosomal effective size, which we denote as α, is

$$\alpha = \frac{N_g^X}{N_g^A}$$

$$= \frac{9\left(N_g^f + N_g^m\right)}{8\left(N_g^f + 2N_g^m\right)}$$

$$= \frac{9\left(1 + \left(N_g^m/N_g^f\right)\right)}{8\left(1 + 2\left(N_g^m/N_g^f\right)\right)}.$$

Equation 4

Thus $N_g^m/N_g^f \to 0$ as $\alpha \to 9/8$, and $N_g^m/N_g^f \to \infty$ as $\alpha \to 9/16$. Since α is a decreasing function of N_g^m/N_g^f, the ratio of X to autosomal effective size satisfies $9/16 < \alpha < 9/8$.

With algebra, it can be shown that Equations 2 and 3 imply that

$$N_g^f = \frac{2N_g^X N_g^A}{9N_g^A - 8N_g^X}$$

Equation 5

and

184 $$N_g^m = \frac{2N_g^X N_g^A}{16N_g^X - 9N_g^A}.$$

185 Equation 6

186 Consequently, the effective sex ratio is

187 $$\mathrm{ESR} = \frac{N_g^f}{N_g^f + N_g^m} = \frac{16N_g^X - 9N_g^A}{8N_g^X}$$

188 Equation 7

189 Thus, given estimates of N_g^X and N_g^A, one can use Equations 5, 6, and 7 to obtain estimates for N_g^f, N_g^m,
190 and the ESR. These are standard equations for estimating sex-specific effective population sizes based
191 on N_g^X and N_g^A,[23,24] although usually presented in the context of constant effective population sizes
192 across time.

193 According to Equation 4, the allowable range of X chromosome N_e is between $\frac{9}{16}$ and $\frac{9}{8}$ of the autosome
194 N_e at each generation. When the X chromosome N_e is overestimated or the autosome N_e is
195 underestimated, the estimated female effective size (Equation 5) can be negative. On the other hand,
196 underestimation of the X chromosome N_e or overestimation of the autosome N_e can result in a negative
197 estimate of the male effective size (Equation 6).

198 Analysis pipeline

199 We start with phased sequence data (true phase for the simulated data, and inferred phase for real
200 data), with males coded as haploid on the X chromosome. We use hap-ibd[20] to infer IBD segments. For
201 hap-ibd analysis on sequence data, we set the minimum seed length to 0.5 cM and the minimum
202 extension length to 0.2 cM. The relatively small minimum seed length and minimum extension length
203 increase power to detect short IBD segments. We exclude rare variants by setting the minimum minor
204 allele count filter to 100 because these lower frequency variants are less informative, have less accurate

205 phasing, and may be recent mutations. For SNP-array data, due to the lower marker density, we set the
206 minimum seed length to 1, the minimum extension length to 0.1, and the minimum number of markers
207 in a seed IBS segment to 50. Other parameters are left at their default values. The IBD analysis
208 parameters are the same for the autosomes and X chromosome.

209 We then run IBDNe on the detected IBD segments output, with one analysis for autosomes and a
210 separate analysis for the X chromosome. The genetic map file for the analysis is assumed to be a sex-
211 averaged map. For the X chromosome, this means multiplying cM positions in the female-specific map
212 by 2/3.

213 When applying IBDNe on the simulated data (autosomes or X), we set "filtersamples=false" and
214 "gmin=1" because the chromosomes are simulated independently so that a pair of individuals can share
215 ancestry one generation back (i.e. be siblings) on one chromosome without such sharing occurring on
216 other chromosomes. These settings tell IBDNe not to look for and remove close relatives, and to model
217 IBD from shared ancestry starting from one generation before present.

218 Real data often has an excess of close relatives due to the sampling scheme. Thus, in the analysis of real
219 autosomal data we allow IBDNe to detect and remove close relatives, which is the default behavior. The
220 X chromosome on its own is not sufficient to detect close relatives, so we use either available pedigree
221 information or the close relative pairs identified by IBDNe in the autosomal analysis to manually remove
222 IBD segments from close relatives in the X chromosome data. We then update the "npairs" parameter
223 accordingly (see below), and set "filtersamples=false" in the X chromosome IBDNe analysis.

224 In the IBDNe analysis of the X chromosome data, we set the "npairs" parameter equal to the number of
225 haplotype pairs for which IBD segments could be present (i.e., all pairs except for those for which we
226 have explicitly removed IBD segments). We first calculate the number of haplotype pairs based on the
227 numbers of females and males in the sample, using Equation 1. We then adjust the number of haplotype

228 pairs to account for the number of close relative pairs that were removed. Removing a male-male pair
229 decreases the count by 1. Removing a male-female pair decreases the count by 2. Removing a female-
230 female pair decreases the count by 4.

231 In order for IBDNe to obtain bootstrap confidence intervals for the X chromosome effective size
232 estimates, we divide the X chromosome into six pieces of equal cM length. We recode the chromosome
233 field of the IBD segment file using integer values between 1 and 6 according to the location of the IBD
234 segments. IBD segments that cross more than one of these "chromosomes" are split into subsegments
235 at the boundaries of the corresponding "chromosomes". We remove the centromere region from IBD
236 segments, so that each IBD segment that crosses the centromere region is replaced with two IBD
237 segments: a segment before and a segment after the centromere.

238 After obtaining the X chromosome and autosomal effective population sizes, we estimate the female
239 and male effective sizes and effective sex ratio using Equations 5, 6, and 7. We obtain bootstrap values
240 for these estimates by taking pairs of bootstrap values from the X and autosomal analyses. For example,
241 for the n-th bootstrap value of the female effective population size at generation g, we take the n-th
242 bootstrap value for the X chromosome effective size at generation g and the n-th bootstrap value for
243 the autosomal effective size at generation g, and apply Equation 5. After obtaining all the bootstrap
244 values, we use the 2.5th and 97.5th percentile to obtain an approximate 95% confidence interval for the
245 female (for example) effective population size at generation g.

246 Simulation study

247 We conducted a simulation study to evaluate the performance of our method. We used SLiM, a forward
248 simulator,[25] to simulate the demographic history for the most recent 5000 generations, and we used
249 msprime, a coalescent simulator, to complete the simulation back to full coalescence of the sample.[26,27]

250 For all scenarios, we simulated data for 30 autosomes of length 100 Mb and an X chromosome of length
251 180 Mb.

252 For the simulation in SLiM, we used a mutation rate of 10^{-8} per base pair per generation, and a
253 recombination rate of 10^{-8} per base pair per meiosis on the autosomes and for female meioses on the
254 X chromosome. We set the gene conversion initiation rate to 2×10^{-8} per base pair per meiosis on the
255 autosomes and per female meiosis on the X chromosome, with mean gene conversion tract length of
256 300 bp. We simulated populations with equal sex ratio, and we also simulated populations with 20%
257 females, 40% females, 60% females, and 80% females. We used the same total effective population size
258 ($N_g^f + N_g^m$) in each simulation. We sampled 50,000 individuals comprising 25,000 females and 25,000
259 males for each analysis.

260 We simulated data under two different demographic models. The first model is a four-stage
261 exponentially growing population with increased growth rate over time, which we call the "UK-like"
262 scenario because it approximates the demographic history of the UK population.[28] The forward
263 simulation of this model in SLiM starts from 5000 generations ago with an initial size of 3000. At 300
264 generations ago, this population starts to grow at an exponential rate of 1.4% per generation. At 60
265 generations ago, the rate of the exponential growth increases to 6%. In the most recent 10 generations,
266 the growth rate further increases to 25%, and the population size reaches around 21 million at the time
267 of sampling.

268 The second demographic model was modified from the "UK-like" model by adding a bottleneck 10
269 generations before present, which approximates the immigration of Europeans to America. This
270 bottleneck reduces the population size to 100,000. The population then continues to grow at an
271 exponential rate of 25% and reaches a final size around 1.2 million. We call this simulated population
272 the "US-like" scenario.

273 When completing the simulations with msprime, we did not apply gene conversion or sex-specific
274 population size. In the msprime simulations, mutation occurred at a rate of 10^{-8} per base pair per
275 generation, and recombination at a rate of 10^{-8} per base pair per meiosis. There is no differentiation
276 between female and male meioses in the generations prior to the starting generation of the forward
277 simulation because msprime does not have sex-specific functionality. This lack of sex-specific treatment
278 in the period more than 5000 generations ago will affect the level of variation in the data but will not
279 affect the distribution of IBD segments of length > 2 cM (the segments analyzed by IBDNe) because such
280 segments typically have ancestry within the past three hundred generations.

281 Analysis of human populations

282 We applied the above analysis pipeline on SNP-array data from the UK Biobank[29] and whole genome
283 sequence data from the TOPMed project[30] to infer the sex-specific population history of several human
284 populations.

285 The UK Biobank is a large-scale biomedical database that contains in-depth genetic, physical and health
286 data collected between 2006 and 2010 on half a million UK participants aged between 40 and 69.[31]
287 Genome-wide genetic SNP-array data were collected on every participant using the UK Biobank Axiom
288 array that assays approximately 850,000 genetic variants across genome.[29] Previous IBDNe analyses of
289 SNP-array data have found that effective population size estimates are less precise when estimating the
290 effective population size more than 50 or so generations before the present due to uncertainty in the
291 IBD-segment endpoints.[3] A previous study on UK Biobank data also showed lower false positive rates in
292 estimated IBD segments that were at least 4 cM compared to estimated IBD segments with length
293 below 4 cM.[20] Therefore, when analyzing UK Biobank data, we used IBD segments that were at least 4
294 cM to estimate population history in the past 40 generations. We used the GRCh37 European
295 recombination map developed by Bherer et al. (2017) in the analyses.[32]

296 We first analyzed the population history of the White British participants, which is the largest ethnic
297 group in the UK Biobank cohort. The UK Biobank genotype data include 221,141 White British females
298 and 187,802 White British males. Since IBDNe has limits on the number of IBD segments that it can
299 process, we randomly selected 5,000 White British females and 5,000 White British males for estimation
300 of the autosome N_e. For analysis of the X chromosome N_e, we removed 33,339 randomly selected
301 females to ensure equal numbers of females and males in the sample. For the autosomes, IBDNe
302 removed close relatives using default settings. For the X chromosome, we used the UK Biobank's kinship
303 estimates to identify and remove IBD segments from sibling pairs and parent-offspring pairs.[29]

304 The UK Biobank cohort also includes participants from several ethnic minority groups including Black
305 British, Indian, Pakistani, Asian, and Bangladeshi. Among these, we chose to analyze the effective
306 population size of the Indian group, which is the largest minority group in the UK Biobank, although this
307 group itself contains considerable diversity. There are genotype data for 2885 males and 2775 females
308 with Indian ancestry in UK Biobank. We removed 110 randomly selected males to achieve equal
309 numbers of females and males for the subsequent analysis. By default, IBDNe automatically removes
310 IBD segments from pairs of related individuals and generates a list of these related pairs. We manually
311 removed X chromosome IBD segments for these pairs of related individuals prior to estimation of X
312 chromosome N_e using IBDNe.

313 The Hypertension Genetic Epidemiology Network Study (HyperGEN) study includes individuals with
314 hypertension from Birmingham, AL and Winston-Salem, NC. Whole genome sequencing was performed
315 as part of the TOPMed project (dbGaP: phs001293.v2.p1). Haplotype phasing was performed as part of
316 the phasing of a larger set of Freeze 8 TOPMed data.[33] We estimated the effective population size from
317 the 1,586 Black non-Hispanic participants from this study. Prior to analysis, we randomly removed 416
318 Black non-Hispanic female participants to ensure equal numbers of females and males in the sample.
319 IBD segments from closely related pairs of individuals were excluded in the IBDNe analysis on the

320 autosome data. X chromosome IBD segments from close relatives were manually removed using the list
321 of relative pairs identified by IBDNe in the autosomal analysis. We used 2 cM as the IBD length threshold
322 and estimated effective population size over the past 100 generations. We used the GRCh38 deCODE
323 map developed by Halldorsson et al. (2019) in the analyses.[22]

324 Results

325 Simulation study

326 We compared N_e estimated from the simulated autosome and X chromosome data to the actual N_e for
327 the UK-like and US-like demographic models. The actual autosome or X chromosome N_e can be
328 obtained from the sex-specific effective population sizes using Equations 2 and 3.

329 The estimated N_e generally matches the true N_e closely (Figure 1 and Figures S1-S2). However, some
330 discrepancies exist between the estimated and actual N_e because IBDNe cannot localize sharp changes
331 in the population size to one exact generation and it tends to over-smooth corners of the population-
332 size trajectory.[3] For the US-like simulations with a sharp bottleneck event that happened at a single
333 generation, the estimated duration of bottleneck is spread over multiple generations (Figure 1).

334 We next used the estimated X chromosome N_e and autosome N_e to estimate the sex-specific effective
335 population sizes and the effective sex ratio. We find that the formulas for the sex-specific N_e and ESR as
336 functions of the autosome and X chromosome N_e (Equation 5, 6, 7) are sensitive to errors in N_e
337 estimation. In UK-like scenarios, with more accurate estimates, the estimated sex-specific N_e and the
338 estimated ESR are similar to the actual values, except at around 20 generations ago, where there was a
339 large change in the population growth rate (Figures 1 and 2 and Figure S1). For the US-like scenarios,
340 around the time of the bottleneck event where there is greater inaccuracy in the estimated N_e, the
341 estimated sex-specific N_e differs significantly from the actual N_e (Figure 1). As a result, the ESR

342 estimated directly from the sex-specific N_e in the US-like simulation is also quite inaccurate (Figure 2 and
343 Figure S2), and the estimated confidence band generally fails to cover the true sex ratio.

344 UK Biobank data

345 The estimated autosome N_e for the UK Biobank's White British population (Figure 3) had a high rate of
346 growth in the most recent 20 generations and reached a current population size of 40 million (95%
347 confidence interval = 35-47 million). Before that, this population had no noticeable change in its
348 effective size. The autosome N_e estimated from the genotype data of the UK Biobank Indian participants
349 (Figure 4) shows slow growth from 40 generations ago until around 10 generations ago. The increased
350 growth beginning 10 generations ago results in an estimated current effective population size of 4.2
351 million (95% confidence interval = 3.7-5.0 million).

352 The inferred X chromosome N_e has the same shape as the inferred autosome N_e for both groups.
353 Notably, the estimated X chromosome N_e is close to 75% of the inferred autosomal effective size, which
354 is the expected effective size that would be obtained from the X chromosome data if the female and
355 male effective population sizes have been equal in these populations (Figures 3 and 4). For the UK Indian
356 population, the confidence band for the X chromosome N_e overlaps that of 75% of the autosome N_e
357 over the entire 40 generation period. For the UK White British population, the confidence band of 75%
358 of the autosome N_e is also covered by that of the X chromosome N_e in the most recent 10 generations.
359 This suggests that these two populations had a balanced sex ratio in their recent history.

360 To further investigate the sex composition of the UK white British population and the UK Indian
361 population in the recent past, we applied the X chromosome and autosome N_e to estimate the female
362 and male N_e and the effective sex ratio of these two populations. For both populations, the sex-specific
363 N_e has a similar overall trend to that of the entire population (Figures 3 and 4). The estimated effective
364 sex ratio has a high level of uncertainty in both populations (Figures 3 and 4).

365 ## Trans-Omics for Precision Medicine (TOPMed) study data

366 The inferred effective population size history of the TOPMed HyperGEN African American cohort shows
367 a period of fast growth from 100 generations before present until around 40 generations ago. The
368 effective population size then stayed relatively stable for around 10 generations. A bottleneck event
369 from 32 to 11 generations ago reduced the N_e of this population from 1.4 million (95% CI: 1.1-1.7
370 million) to 112 thousand (95% CI: 105-119 thousand). A period of high growth since then has restored
371 the N_e to 3.6 million at the present time (95% CI: 2.7-5.4 million) (Figure 5). The recent bottleneck event
372 in the estimated effective-size trajectory of the HyperGEN Black non-Hispanic participants is consistent
373 with the immigration bottleneck resulting from the transatlantic slave trade in the history of the African
374 American population.

375 As in the analysis of UK Biobank data, the X chromosome effective size inferred from the HyperGEN
376 group follows the same trend as the autosome N_e of this group (Figure 5). However, we observed more
377 discrepancy between the estimated X chromosome N_e and the estimated autosomal N_e than was seen
378 in the UK Biobank data, matching the comparison of results found in the US-like and UK-like simulations.
379 Compared to the estimated autosome N_e, the X chromosome N_e estimate was high between 65 to 30
380 generations ago, which produced negative estimates of female-specific N_e during this period (Figure 5).
381 The estimated X chromosome N_e was also low compared to the estimated autosome N_e both during the
382 most recent 6 generations and 21 to 18 generations ago, which led to negative male-specific N_e (Figure
383 5). There is significant uncertainty in the ESR estimates.

384 # Discussion

385 Previous studies have shown that the X chromosome can provide information about demographic
386 processes that cannot be revealed by the analysis of autosomes alone.[9,11,12,14,34,35] In this work, we
387 focused on utilizing IBD segments on the X chromosome to infer the X chromosome effective population

388 size. We developed a framework to model X chromosome N_e and derived the relationship between X
389 chromosome and autosome N_e by considering the different coalescence rates among X chromosomes
390 and autosomes. We also showed how to apply this information to calculate the female and male
391 effective population sizes and the effective sex ratio in a population as functions of the X chromosome
392 and autosome N_e.

393 We validated the performance of our method in simulated populations with similar histories to the UK
394 and the US populations. We found that pronounced differences between female and male effective
395 population sizes can be detected. Such differences are easier to detect in populations that have not
396 undergone recent bottlenecks, because recent bottlenecks create noise in the estimates of female and
397 male effective population sizes.

398 We applied our method to estimate the X chromosome effective size for the UK British and UK Indian
399 populations in the UK Biobank, as well as an African American population from the TOPMed HyperGEN
400 study. When using the X chromosome effective population size to estimate the sex-specific effective
401 population size, we obtained results consistent with gender-balanced demographic histories for each of
402 these populations.

403 Although X chromosome IBD information has been used by previous studies for the estimation of
404 genealogical relations between individuals, especially for kinship estimation in forensic settings[5,6,36],
405 there has been a lack of studies that use X chromosome IBD segments to estimate recent effective
406 population size. Our work thus fills a gap that existed in the application of X chromosome IBD
407 information in population genetic studies.

408 Previous methods for estimating sex-specific population history or sex bias in human populations have
409 relied on comparisons of genetic diversity between autosomes and the X chromosome using allele
410 frequency differentiation, patterns of neutral polymorphism, and the site frequency

411 spectrum.[13,15,16,18,35,37] Most of these methods considered only a single estimate of the effective sex ratio
412 over the entire history of a population, although this ratio can vary over time.[13,15,18,37] Some of these
413 methods also focused on a constant overall effective size across time, although changes in effective size
414 can distort these analyses.[35,37] Recently, Musharoff et al. developed a likelihood ratio test for population
415 sex bias that considered populations of non-constant size and changing sex ratios using site frequency
416 spectrum data.[16] However, this method requires demographic parameters to be constant within time
417 epochs. In comparison, our approach for estimating the X chromosome effective population size and the
418 female and male effective population sizes requires minimal assumptions and allows the effective
419 population sizes to vary independently over time. The ability of our IBD-based analyses to infer effective
420 population sizes in the past hundred generations distinguishes our approach from other methods.

421 There are several limitations to our approach. First, IBD-based estimation of effective population size
422 requires a large sample of individuals from the population, although this may be less of an issue given
423 the continuing increase in the size of genetic studies, particularly in humans. Second, our method for
424 estimating N_e is less accurate around generations where the population experienced a drastic change in
425 population size trajectory such as a bottleneck event. Joint estimation of autosomal and X chromosome
426 effective size could improve estimation of the effective sex ratio in such situations. In addition, the
427 performance of IBDNe is affected by the number and accuracy of detected IBD segments. For example,
428 we observe wider confidence bands for the estimated N_e in the most recent generations since there
429 tend to be fewer very long IBD segments in the sample. Similarly, we observe wider confidence bands
430 for the X chromosome N_e compared to the autosomal N_e due to the smaller amount of data in the X
431 chromosome. Moreover, the estimated autosome and X chromosome effective size tend to deviate
432 more from the simulated effective size in the more distant past due to the lower relative accuracy when
433 estimating the lengths of shorter IBD segments.

434 Given these limitations, we note that our method is not suitable for rigorously testing hypotheses on
435 sex-specific N_e and the effective sex ratio in a population, because small inaccuracies in estimation of
436 autosome and X chromosome effective population sizes are magnified when transforming these to
437 estimates of sex-specific effective population size and the effective sex ratio. However, our method can
438 still serve as an exploratory tool for detecting evidence of sex-specific past demographic events.

439

440 Data and code availability

441 UK Biobank data were downloaded from the European Genome-Phenome Archive (https://ega-
442 archive.org/datasets/EGAD00010001497). TOPMed freeze 8 data for study accession phs001293.v2.p1
443 (HyperGEN) were downloaded from dbGaP (https://www.ncbi.nlm.nih.gov/gap/). The IBDNe software is
444 available from https://faculty.washington.edu/browning/ibdne.html. The hap-ibd software is available
445 from https://github.com/browning-lab/hap-ibd.

446 Acknowledgements

447 Research reported in this publication was supported by the National Human Genome Research Institute
448 of the National Institutes of Health under award number HG005701. The content is solely the
449 responsibility of the authors and does not necessarily represent the official views of the National
450 Institutes of Health.

451 This research has been conducted using the UK Biobank Resource under Application Number 19934.
452 Molecular data for the Trans-Omics in Precision Medicine (TOPMed) program was supported by the
453 National Heart, Lung and Blood Institute (NHLBI). Core support including centralized genomic read
454 mapping and genotype calling, along with variant quality metrics and filtering were provided by the
455 TOPMed Informatics Research Center (3R01HL-117626-02S1; contract HHSN268201800002I). Core

support including phenotype harmonization, data management, sample-identity QC, and general program coordination were provided by the TOPMed Data Coordinating Center (R01HL-120393; U01HL-120393; contract HHSN268201800001I). We gratefully acknowledge the studies and participants who provided biological samples and data for TOPMed. The Hypertension Genetic Epidemiology Network Study is part of the NHLBI Family Blood Pressure Program; collection of the data represented here was supported by grants U01 HL054472, U01 HL054473, U01 HL054495, and U01 HL054509; genome sequencing was funded by R01HL055673.

475 ## Literature cited

476 1. Wright S. Evolution in Mendelian populations. *Genetics.* 1931;16(2):97.

477 2. Charlesworth B. Effective population size and patterns of molecular evolution and variation.
478 *Nature Reviews Genetics.* 2009;10(3):195-205.

479 3. Browning SR, Browning BL. Accurate non-parametric estimation of recent effective population
480 size from segments of identity by descent. *Am J Hum Genet.* 2015;97(3):404-418.

481 4. Browning SR, Browning BL, Daviglus ML, et al. Ancestry-specific recent effective population size
482 in the Americas. *PLoS Genet.* 2018;14:e1007385.

483 5. Henden L, Wakeham D, Bahlo M. XIBD: software for inferring pairwise identity by descent on the
484 X chromosome. *Bioinformatics.* 2016;32(15):2389-2391.

485 6. Buffalo V, Mount SM, Coop G. A genealogical look at shared ancestry on the X chromosome.
486 *Genetics.* 2016;204(1):57-75.

487 7. Palamara PF, Lencz T, Darvasi A, Pe'er I. Length distributions of identity by descent reveal fine-
488 scale demographic history. *American Journal of Human Genetics.* 2012;91(5):809-822.

489 8. Ségurel L, Martínez-Cruz B, Quintana-Murci L, et al. Sex-specific genetic structure and social
490 organization in Central Asia: insights from a multi-locus study. *PLoS Genet.* 2008;4(9):e1000200.

491 9. Bryc K, Auton A, Nelson MR, et al. Genome-wide patterns of population structure and admixture
492 in West Africans and African Americans. *Proceedings of the National Academy of Sciences.*
493 2010;107(2):786-791.

494 10. Heyer E, Chaix R, Pavard S, Austerlitz F. Sex-specific demographic behaviours that shape human
495 genomic variation. *Mol Ecol.* 2012;21(3):597-612.

496 11. Goldberg A, Rosenberg NA. Beyond 2/3 and 1/3: the complex signatures of sex-biased admixture
497 on the X chromosome. *Genetics.* 2015;201(1):263-279.

498 12. Shringarpure SS, Bustamante CD, Lange K, Alexander DH. Efficient analysis of large datasets and
499 sex bias with ADMIXTURE. *BMC bioinformatics.* 2016;17(1):1-6.
500 13. Hammer MF, Mendez FL, Cox MP, Woerner AE, Wall JD. Sex-biased evolutionary forces shape
501 genomic patterns of human diversity. *PLoS genetics.* 2008;4(9):e1000202.
502 14. Bustamante CD, Ramachandran S. Evaluating signatures of sex-specific processes in the human
503 genome. *Nature genetics.* 2009;41(1):8-10.
504 15. Clemente F, Gautier M, Vitalis R. Inferring sex-specific demographic history from SNP data. *PLoS*
505 *genetics.* 2018;14(1):e1007191.
506 16. Musharoff S, Shringarpure S, Bustamante CD, Ramachandran S. The inference of sex-biased
507 human demography from whole-genome data. *PLoS genetics.* 2019;15(9):e1008293.
508 17. Wang J, Santiago E, Caballero A. Prediction and estimation of effective population size. *Heredity.*
509 2016;117(4):193-206.
510 18. Emery LS, Felsenstein J, Akey JM. Estimators of the human effective sex ratio detect sex biases
511 on different timescales. *The American Journal of Human Genetics.* 2010;87(6):848-856.
512 19. File Formats Task Team. The Variant Call Format Specication: VCFv4.3 and BCFv2.2. 2020.
513 http://samtools.github.io/hts-specs/VCFv4.3.pdf.
514 20. Zhou Y, Browning SR, Browning BL. A fast and simple method for detecting identity-by-descent
515 segments in large-scale data. *The American Journal of Human Genetics.* 2020;106(4):426-437.
516 21. International HapMap Consortium. A second generation human haplotype map of over 3.1
517 million SNPs. *Nature.* 2007;449(7164):851.
518 22. Halldorsson BV, Palsson G, Stefansson OA, et al. Characterizing mutagenic effects of
519 recombination through a sequence-level genetic map. *Science.* 2019;363(6425):eaau1043.
520 23. Wright S. *Evolution and the genetics of populations: Vol. 2. The theory of gene frequencies.* 1969.

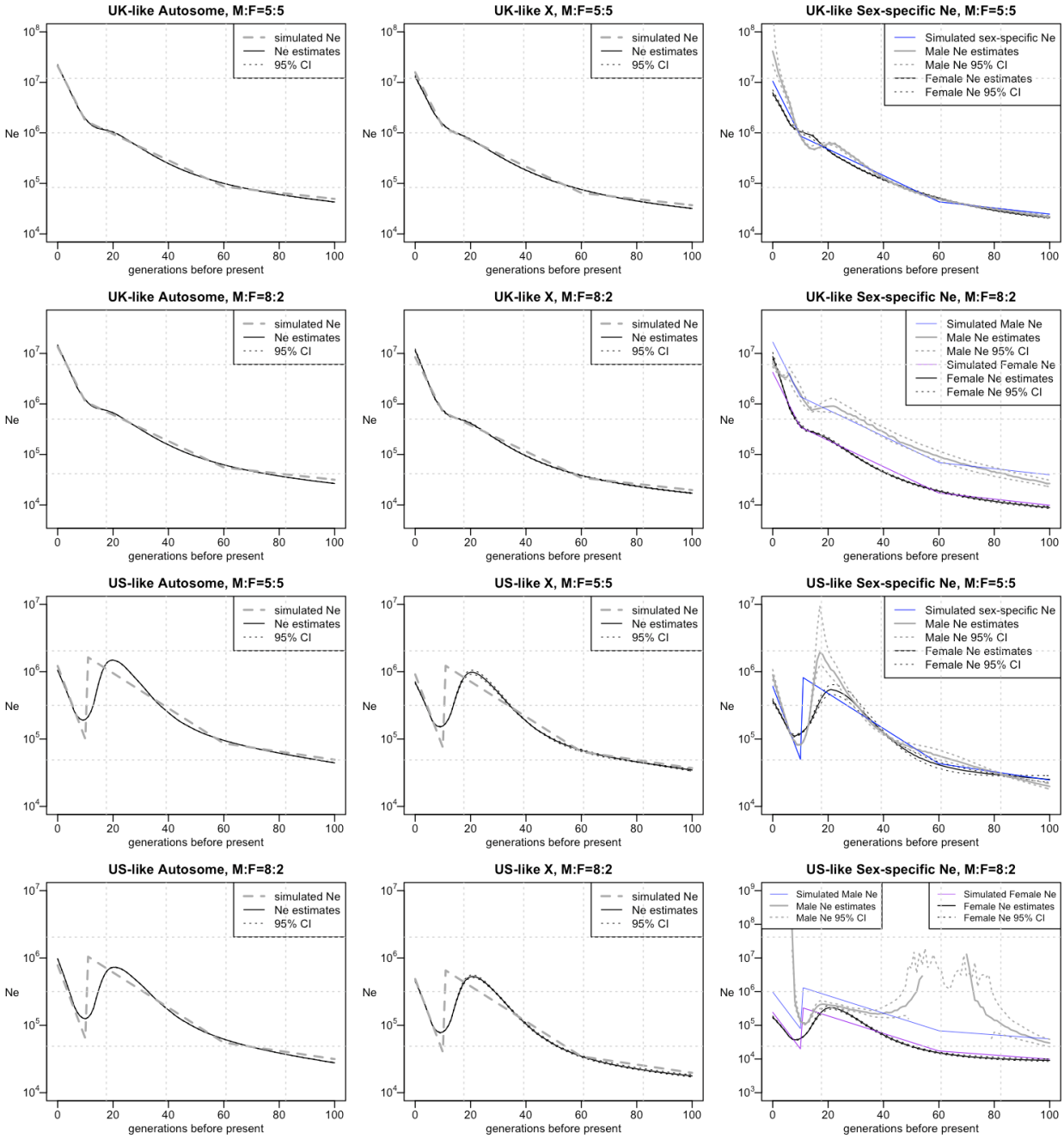
521 24. Hartl DL, Clark AG. *Principles of population genetics.* Vol 116: Sinauer associates Sunderland;
522 1997.
523 25. Haller BC, Messer PW. SLiM 3: forward genetic simulations beyond the Wright-Fisher model.
524 *Molecular biology and evolution.* 2019;36(3):632-637.
525 26. Kelleher J, Etheridge AM, McVean G. Efficient coalescent simulation and genealogical analysis
526 for large sample sizes. *PLoS computational biology.* 2016;12(5):e1004842.
527 27. Haller BC, Galloway J, Kelleher J, Messer PW, Ralph PL. Tree-sequence recording in SLiM opens
528 new horizons for forward-time simulation of whole genomes. *Molecular ecology resources.*
529 2019;19(2):552-566.
530 28. Browning BL, Browning SR. Detecting identity by descent and estimating genotype error rates in
531 sequence data. *The American Journal of Human Genetics.* 2013;93(5):840-851.
532 29. Bycroft C, Freeman C, Petkova D, et al. The UK Biobank resource with deep phenotyping and
533 genomic data. *Nature.* 2018;562(7726):203-209.
534 30. Taliun D, Harris DN, Kessler MD, et al. Sequencing of 53,831 diverse genomes from the NHLBI
535 TOPMed Program. *Nature.* 2021;590(7845):290-299.
536 31. Fry A, Littlejohns TJ, Sudlow C, et al. Comparison of sociodemographic and health-related
537 characteristics of UK Biobank participants with those of the general population. *American*
538 *journal of epidemiology.* 2017;186(9):1026-1034.
539 32. Bhérer C, Campbell CL, Auton A. Refined genetic maps reveal sexual dimorphism in human
540 meiotic recombination at multiple scales. *Nat Commun.* 2017;8(1):1-9.
541 33. Browning BL, Tian X, Zhou Y, Browning SR. Fast two-stage phasing of large-scale sequence data.
542 *The American Journal of Human Genetics.* 2021;108(10):1880-1890.
543 34. Schaffner SF. The X chromosome in population genetics. *Nature Reviews Genetics.* 2004;5(1):43-
544 51.

545 35. Ramachandran S, Rosenberg NA, Zhivotovsky LA, Feldman MW. Robustness of the inference of
546 human population structure: a comparison of X-chromosomal and autosomal microsatellites.
547 *Human genomics.* 2004;1(2):1-11.
548 36. Pinto N, Gusmão L, Amorim A. X-chromosome markers in kinship testing: a generalisation of the
549 IBD approach identifying situations where their contribution is crucial. *Forensic Science*
550 *International: Genetics.* 2011;5(1):27-32.
551 37. Keinan A, Mullikin JC, Patterson N, Reich D. Accelerated genetic drift on chromosome X during
552 the human dispersal out of Africa. *Nature genetics.* 2009;41(1):66-70.

553

554 # Figures

555 **Figure 1. Estimates of autosome, X chromosome, and sex-specific N_e in simulation studies.** Autosomal
556 N_e is shown in the left column, X chromosome N_e in the middle column, and sex-specific N_e in the right
557 column. The top two rows are the UK-like simulation, with equal sex ratio in the top row and 20%
558 females in the second row. The lower two rows are the US-like simulation, with equal sex ratio in the
559 third row and 20% females in the fourth row. Results for other choices of sex-ratio can be found in
560 Figures S1 and S2. Y-axes show N_e plotted on a log scale. In cases where the estimated sex-specific N_e is
561 negative (see Methods), it is not shown.

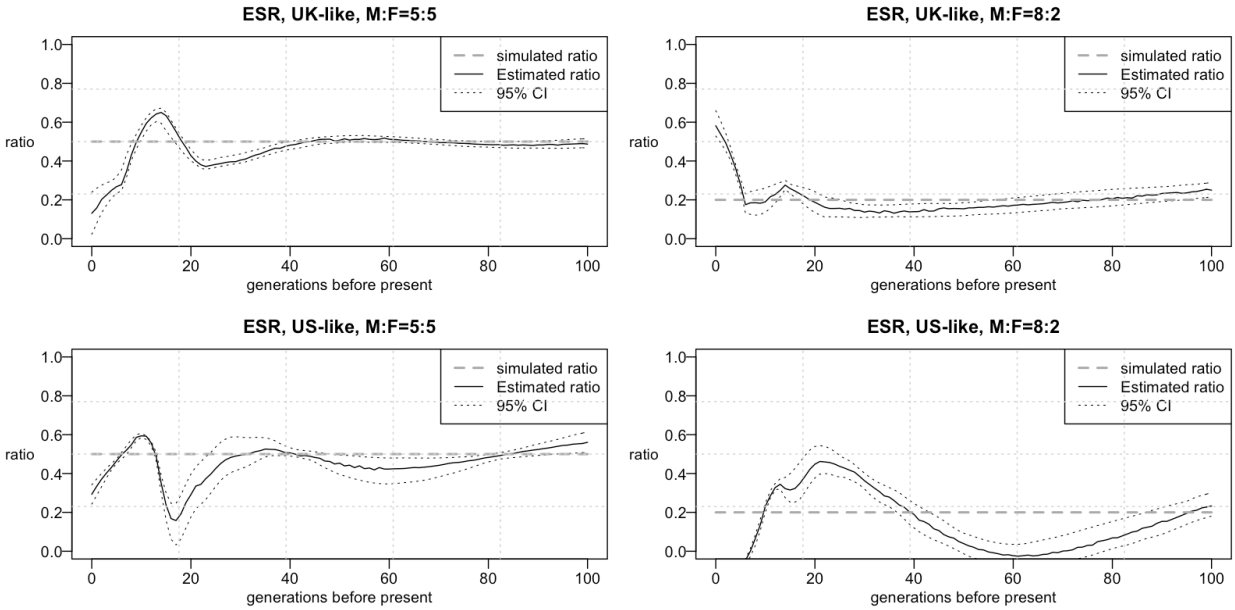

UK-like Autosome, M:F=5:5
UK-like X, M:F=5:5
UK-like Sex-specific Ne, M:F=5:5
UK-like Autosome, M:F=8:2
UK-like X, M:F=8:2
UK-like Sex-specific Ne, M:F=8:2
US-like Autosome, M:F=5:5
US-like X, M:F=5:5
US-like Sex-specific Ne, M:F=5:5
US-like Autosome, M:F=8:2
US-like X, M:F=8:2
US-like Sex-specific Ne, M:F=8:2
simulated Ne
Ne estimates
95% CI
Simulated sex-specific Ne
Male Ne estimates
Male Ne 95% CI
Female Ne estimates
Female Ne 95% CI
Simulated Male Ne
Simulated Female Ne
Ne
generations before present

562

563 **Figure 2. Estimates of population effective sex ratio (ESR) obtained from estimated sex-specific N_e.**

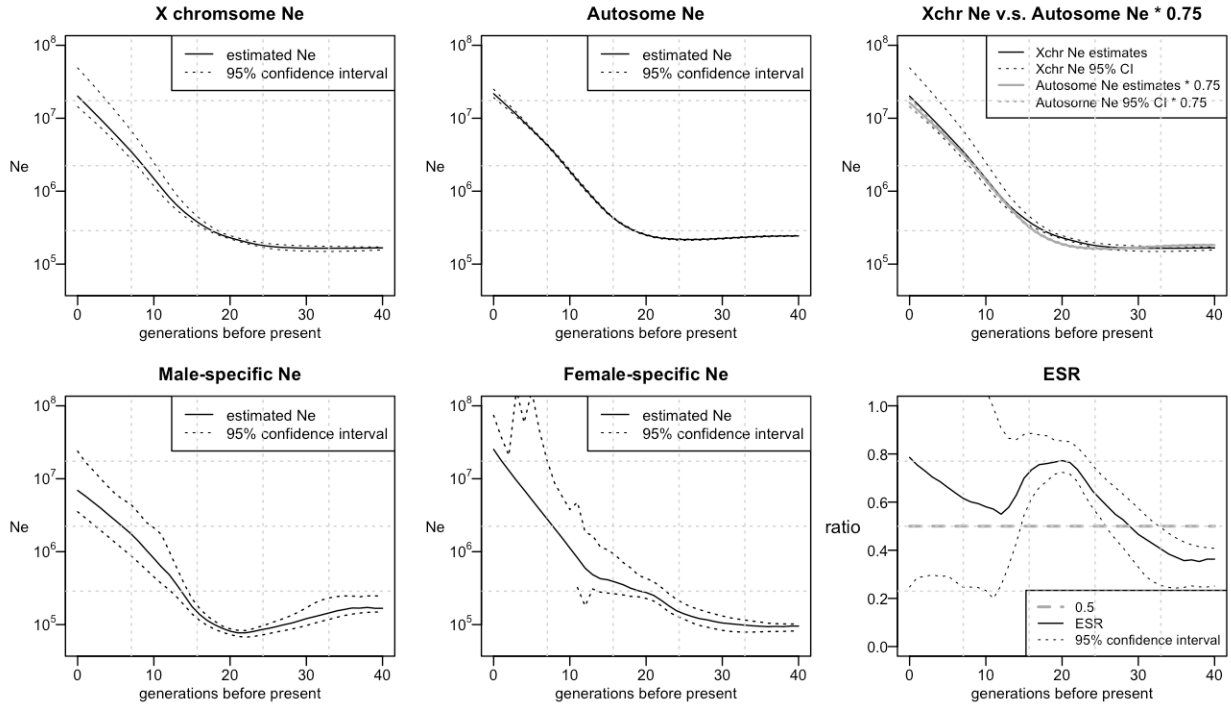
564 The top row shows the UK-like simulation, while the bottom row shows the US-like simulation. The true

565 ratio is shown by the horizontal dashed line in each plot.



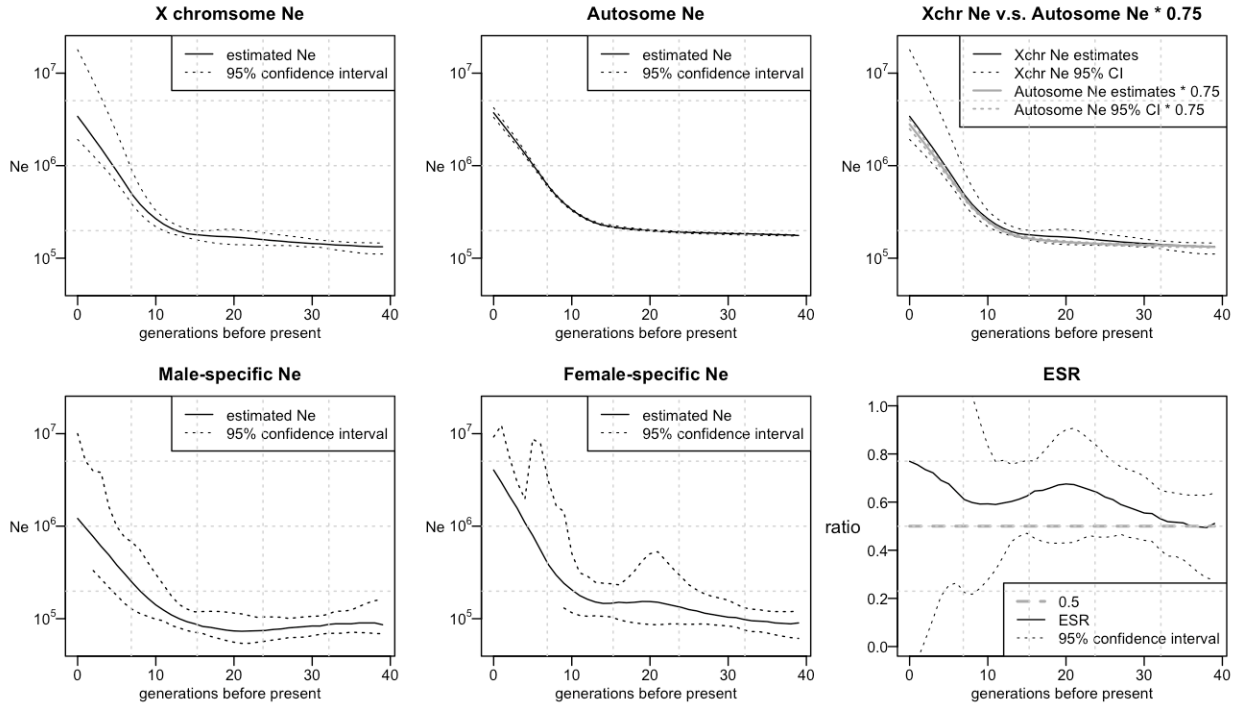
566

567 **Figure 3. Effective population size of the UK Biobank White British group.** From left to right, the top
568 row shows the estimated X chromosome N_e, the estimated autosome N_e, and a comparison of the
569 estimated X chromosome N_e with 75% of the estimated autosome N_e. The bottom row displays the
570 male-specific N_e, the female-specific N_e, and the ESR. For the N_e plots, the Y-axes show N_e on a log
571 scale. In cases where the estimated sex-specific N_e or its confidence band is negative (see Methods), the
572 negative values are not shown.



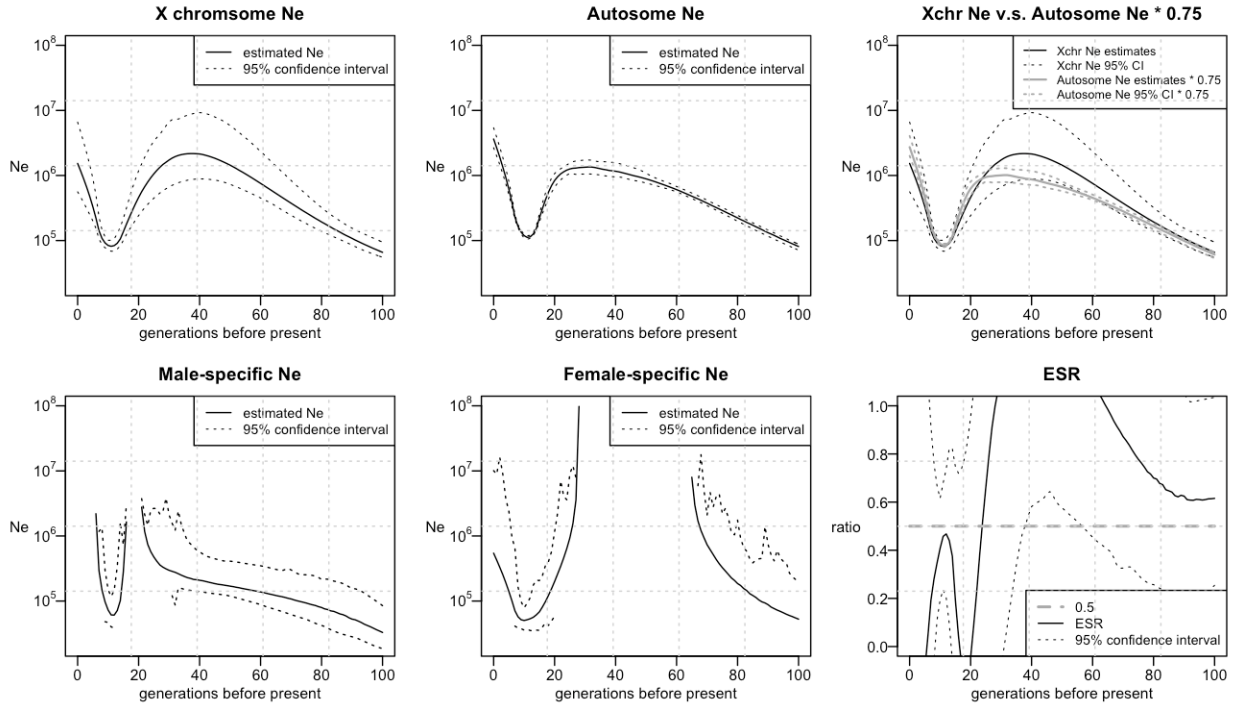
573

574 **Figure 4. Effective population size of the UK Biobank Indian group.** From left to right, the top row
575 shows the estimated X chromosome N_e, the estimated autosome N_e, and a comparison of the estimated
576 X chromosome N_e with 75% of the estimated autosome N_e. The bottom row displays the male-specific
577 N_e, the female-specific N_e, and the ESR. For the N_e plots, the Y-axes show N_e on a log scale. In cases
578 where the estimated sex-specific N_e or its confidence band is negative (see Methods), the negative
579 values are not shown.



580

581 **Figure 5. Effective population size of the Black non-Hispanic group in the HyperGEN cohort.** From left
582 to right, the top row shows the estimated X chromosome N_e, the estimated autosome N_e, and a
583 comparison of the estimated X chromosome N_e with 75% of the estimated autosome N_e. The bottom
584 row displays the male-specific N_e, the female-specific N_e, and the ESR. For the N_e plots, the Y-axes show
585 N_e on a log scale. In cases where the estimated sex-specific N_e or its confidence band is negative (see
586 Methods), the estimated values are not shown. Estimated ESR values outside the range $[0,1]$ are not
587 shown.



588

Supplementary materials

Figure S1. Estimates of sex-specific population size and effective sex ratio in UK-like simulations. From left to right, the columns display results from a UK-like simulation with 80% , 60%, and 40% females. Autosomal N_e is shown in the top row, X chromosome N_e in the second row, sex-specific N_e in the third row, and the ESR in the bottom row. For the N_e plots the Y-axes are on a log scale.

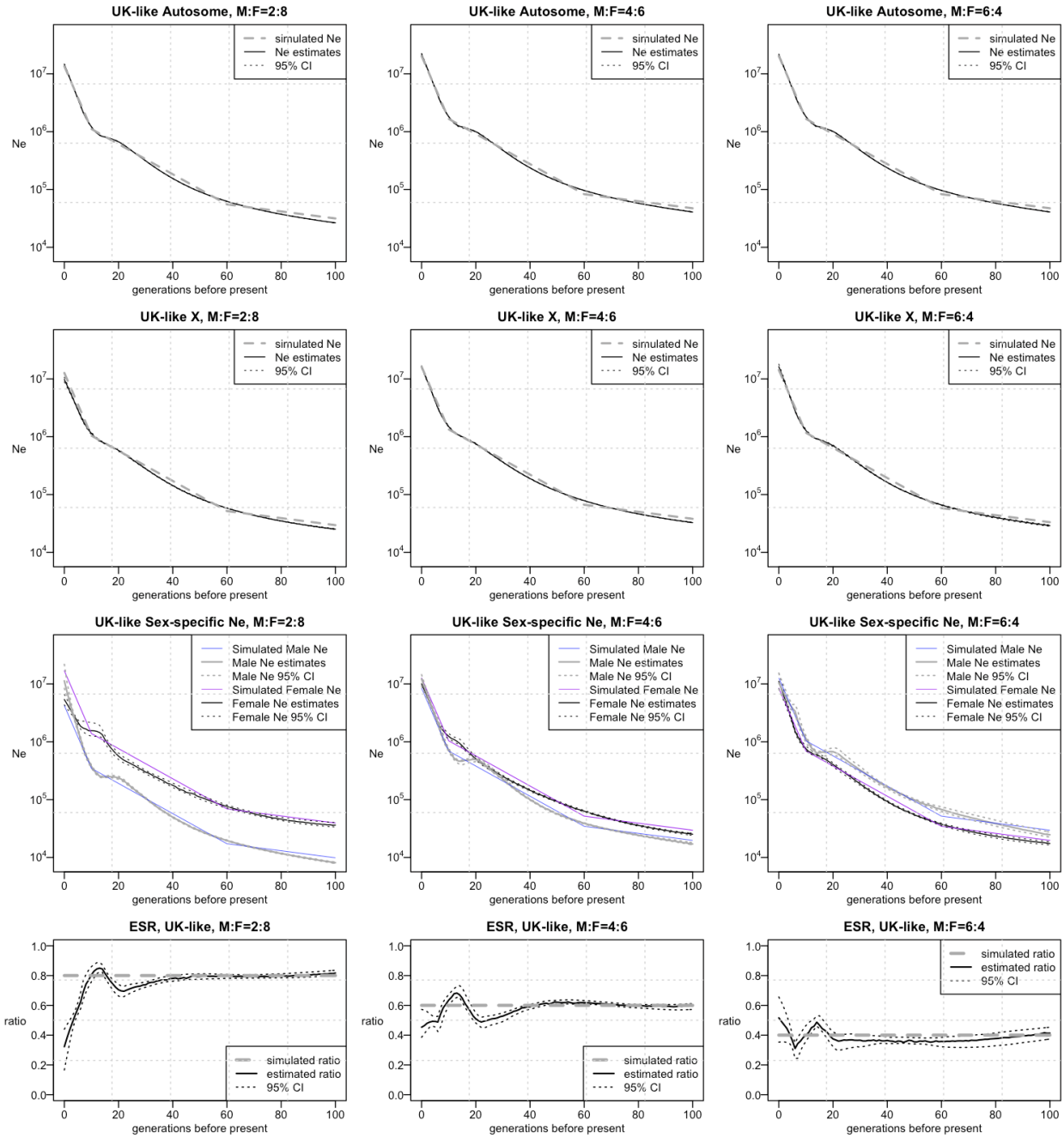
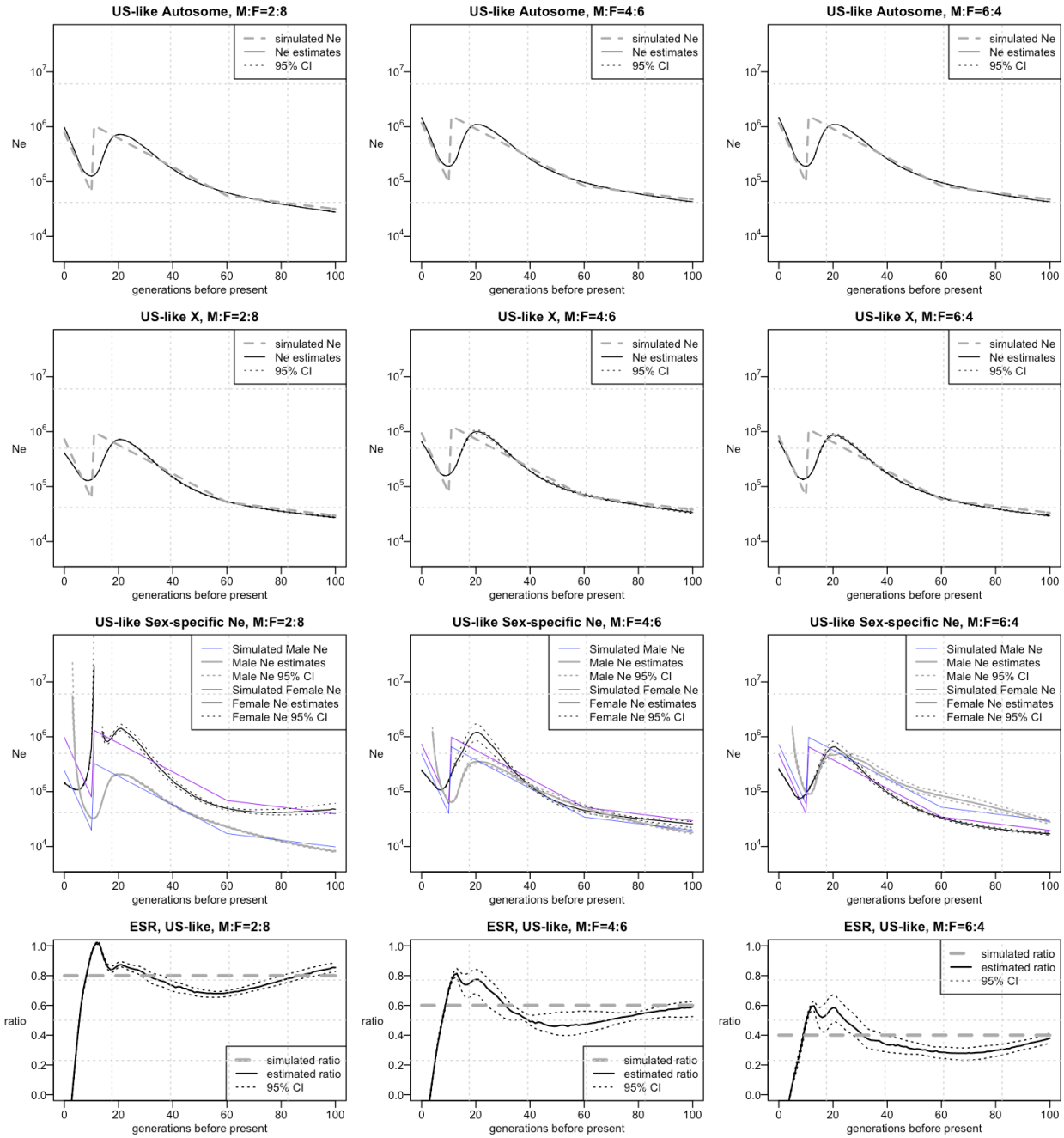



Figure S2. Estimates of sex-specific population size and effective sex ratio in US-like simulations. From left to right, the columns display results from a UK-like simulation with 80%, 60%, and 40% females. Autosomal N_e is shown in the top row, X chromosome N_e in the second row, sex-specific N_e in the third row, and the ESR in the bottom row. For the N_e plots the Y-axes are on a log scale. In cases where the estimated sex-specific N_e is negative (see Methods), it is not shown.