bioRxiv preprint doi: https://doi.org/10.1101/2020.01.31.928291; this version posted July 22, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity.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-NC-ND 4.0 International license.

Detecting sex-linked genes using genotyped individuals sampled in natural populations

Jos Käfer[1], Nicolas Lartillot[2] , Gabriel A.B. Marais[2], Franck Picard[2]
Université de Lyon; Université Lyon 1; CNRS; UMR 5558; Laboratoire de Biométrie et Biologie Evolutive; F-69622, Villeurbanne, France

Abstract We propose a method, SDpop, able to infer sex-linkage caused by recombination suppression typical of sex chromosomes. The method is based on the modeling of the allele and genotype frequencies of individuals of known sex in natural populations. It is implemented in a hierarchical probabilistic framework, accounting for different sources of error. It allows to statistically test for the presence or absence of sex chromosomes, and to infer sex-linked genes based on the posterior probabilities in the model. Furthermore, for gametologous sequences, the haplotype and level of nucleotide polymorphism of each copy can be inferred, as well as the divergence between both. We test the method using simulated and human sequencing data, and show that, for most cases, robust predictions are obtained with 5 to 10 individuals per sex.

1 Introduction

Sex chromosomes, which are found in many species with genetic sex determination, are key elements of the biology of the sexes (e.g. between-sexes differences in development, morpho-anatomy, physiology) and are consequently studied by different branches of biology (e.g. genetics, genomics, developmental biology, physiology, evolutionary biology, medical research, agronomy). While sex chromosomes share striking features even between independently evolved systems, there is also much diversity (reviewed in Bachtrog et al., 2014).

Sex chromosomes are thought to evolve from a pair of autosomes (Lahn and Page, 1999; Charlesworth et al., 2005). One sex is characterized by a pair of sex chromosomes for which recombination is suppressed over part of the length, while the other has two identical chromosomes that recombine normally. Thus, the sex chromosome which is limited to one of the two sexes (the so-called "sex limited chromosome", SLC) contains a part that never recombines, and evolves independently from the homologous part on the other sex chromosome. These systems are termed according to the sex that is heterogametic; when the males are the heterogametic sex, the sex chromosomes are named X and Y, and when the females are heterogametic, they are named Z and W; Y and W are the SLCs in these systems, respectively. The recombining part of the sex chromosomes is called the pseudo-autosomal region.

The complete lack of homologous recombination in a part of the SLC has dramatic consequences for its evolution (Charlesworth et al., 2005; Lynch, 2007): it accumulates deleterious mutation (Muller's ratchet) and transposable elements, which deteriorate gene function and finally chromosome integrity. Once most genes have lost their function, parts of the chromosome can be lost without negative effects on the individual's fitness, leading to situation that is considered paradigmatic of a large X chromosome and a small Y chromosome with only few functional genes, as in humans. However, in many sex chromosome systems, the SLC does not differ significantly in size from its homologous counterpart, and recombination is suppressed in only a small region of these chromosomes. As more sex chromosome systems are being studied in phylogenetically distant lineages, more questions arise on their similarities and differences (e.g. Charlesworth, 2016; Muyle et al., 2017; Ponnikas et al., 2018).

Even though sequencing technologies have been improving constantly, sex chromosomes have remained difficult to study for a long time. In particular, SLC sequences used to be scarce for two reasons (Tomaszkiewicz et al., 2017). First, genome projects have been typically using whole-genome shotgun (WGS), in which DNA is fragmented and sequenced using short read technologies and then re-assembled in contigs. The assembly of the non-recombining regions of the SLC, which are full of repeats, using short reads is virtually impossible (Hughes and Rozen, 2012). Second, in the heterogametic sex, read coverage for the sex chromosomes is half the coverage for the autosomes, making the assembly of both sex chromosomes more difficult, while sequencing of the homogametic sex only yields one of the chromosomes.

Recently, the development of new methods to study sex chromosomes has boosted their study. The first Y sequence of humans and other species (e.g. papaya) have been obtained with a specific methods based on BACs,

[1]To whom correspondance should be sent: Jos Käfer, Laboratoire de Biométrie et Biologie Evolutive; CNRS; UMR 5558; Université Lyon 1; Bat. Grégor Mendel, 43 bd du 11 novembre 1918, 69622 Villeurbanne cedex, E-mail: jos.kafer@univ-lyon1.fr.
[2]Equal contributions; authors are listed in alphabetical order

bioRxiv preprint doi: https://doi.org/10.1101/2020.01.31.928291; this version posted July 22, 2020. 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-NC-ND 4.0 International license.

such as the SHIMS, which are very time-consuming and costly (Skaletsky et al., 2003; Hughes and Rozen, 2012; Wang et al., 2012; Bellott et al., 2014; for example, the human Y sequencing costed 6 million dollars). Costs have decreased in a new version of SHIMS, using Illumina short-read sequencing of pooled BACs instead of Sanger sequencing of individual BACs (Bellott et al., 2018), but they are still substantial for any Y chromosome larger than a few megabases. More efficient and cheaper methods have thus been developed (reviewed in Muyle et al., 2017; Palmer et al., 2019).

A first alternative approach was to compare male and a female genomes obtained using short reads (WGS). Absence in the female genome is enough to identify the Y contigs (Carvalho and Clark, 2013; Cortez et al., 2014). The X contigs can be identified by computing the read depth in male and female: the male to female read depth ratio is expected to be 1 for autosomes and 0.5 for the X, as males have one X copy and females two (Vicoso and Bachtrog, 2011). Variants of this approach have been widely used (e.g. Vicoso et al., 2013a,b; Hall et al., 2013; Li et al., 2018). However, this approach works only with strongly diverged, old sex chromosomes, in which the reads map uniquely to the X or the Y chromosomes but not to both. In more recently evolved, weakly diverged systems, the X and the Y still share a high level of sequence similarity, and reads cannot be unambiguously assigned to one of the sex chromosomes with alignment methods.

To study weakly divergent sex chromosomes, a second category of empirical methods uses segregation patterns in genotyping data from natural populations. In XY systems, an excess of heterozygotes is expected in XY males compared to XX females at the sex-linked genes. Such patterns could be identified using male versus female F_{ST} (e.g. Zhou et al., 2017) or differences in heterozygote frequencies (e.g. Picq et al., 2014; Kirkpatrick and Guerrero, 2014). GWAS using sex as the studied trait is another option which will work based on different allele frequencies in the sexes, but a precise model of the expected association is preferable (Galichon et al., 2012). Indeed, the patterns these empirical approaches look for can be caused by other processes than sex linkage, such as sex-antagonistic selection (e.g. Qiu et al., 2013). Such methods thus have high false positive rates and either must use conservative cut-offs, which significantly reduces their sensibility, or are applicable to species with well-assembled genomes (e.g. Picq et al., 2014; Mathew et al., 2014). An empirical approach somewhat intermediate between the former and the latter categories is to identify sex-specific sequences in the genomic reads, but these mostly only allow a first characterization of the sex chromosome system, and further sequencing is needed to study the sex-linked regions (Scharmann et al., 2017; Torres et al., 2018).

Recently, the SEX-DETector statistical framework, relying on the modeling of the transmission of alleles that account for the observed genotypes, was introduced (Muyle et al., 2016). The method has been shown to yield high power and high sensitivity with as few as five offspring of each sex, and has successfully been used on a number of species (Muyle et al., 2018; Veltsos et al., 2019; Martin et al., 2019; Prentout et al., 2020). It can be used for several purposes, using a single dataset: testing whether a species has sex chromosomes and of which kind (XY or ZW), identifying sex-linked genes, reconstructing the haplotypes of the gametolog copies, and estimating expression of each of the copies if RNAseq data are used. However, its requirement to produce a controlled cross for sequencing limits its use to species than can be easily bred or cultivated in controlled conditions, and hinders its application to species with long generation times.

A method that can be applied to sequencing data from individuals sampled in natural populations, DETSEX, was introduced by Gautier (2014). While it offers some attractive features, like automatic sexing of some individuals in the sample, it relies on ploidy levels to detect sex-linked sequences, and can thus only be used for sufficiently divergent sex chromosomes. The method has, to our knowledge, not been used in any study that aimed to identify sex-linked sequences.

We here introduce a new hierarchical framework based on the modeling of genotype and allele frequencies in a population, according to autosomal and sex-linked segregation types. As SEX-DETector statistically analyses Mendelian segregation and its deviations due to sex-linkage, the new method does so based on the Hardy-Weinberg equilibrium and its deviations. The method we present here is "SEX-DETector in a population", and is thus termed "SDpop". It can be applied to any sample collected from natural populations, provided the sex of the individuals can be determined morphologically. Despite the similarity of its goals with those of SEX-DETector (notably, applicability to sex chromosome systems of any age, statistical testing, likelihood-based inferences, prediction of gametolog haplotypes), it is an entirely different model due to the data it handles, the underlying population genetic expectations, and some of the predictions it can produce.

2 Materials and methods

2.1 Model

2.1.1 Terminology

A sex-linked gene (i.e. a gene that is present on the non-recombining region of the sex chromosomes) has two independent copies. These copies will accumulate mutations, be subject to selection and drift, and accumulate allele fixations independently from one another. These copies thus diverge after the suppression of recombination,

bioRxiv preprint doi: https://doi.org/10.1101/2020.01.31.928291; this version posted July 22, 2020. 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-NC-ND 4.0 International license.

in a way that is reminiscent of the divergence of paralogous genes after gene duplication; the two copies of a sex-linked gene are thus termed "gametologs" (Garcia-Moreno and Mindell, 2000). Eventually, due to neo- or sub-functionalization, or the accumulation of deleterious mutations, one of the copies can get lost, most often the copy present on the Y or W chromosome. The remaining gene, on the X or Z, is termed "X-" or "Z-hemizygous".

Gametology and hemizygosity cause differences in the relation between genotype and allele frequencies with respect to autosomal segregation. More precisely, these differences occur in one sex, i.e. the heterogametic one. If one considers a sexual, random mating population of diploid individuals, most autosomal genes will be at Hardy-Weinberg equilibrium. The deviations from this equilibrium in the heterogametic sex are themselves equilibria, that can be described assuming panmixia, and that are the basis of the present model. For gametologs, two independent copies are present in the heterogametic sex, while for hemizygous genes, only one copy is present. For non-sex-linked genes, the presence of two independent copies can also occur for both sexes under paralogy, and the presence of only one copy for the haploid mitochondrial and chloroplastic genes, or when only one allele is expressed in transcriptome data. We thus distinguish five possible segregation types in a diploid population: diploid autosomy (hereafter just called "autosomy"), haploidy, paralogy, hemizygosity, and gametology.

We model genotypes that are obtained by mapping sequencing data on a reference sequence. As the organisms are supposed to be diploid, the input data at all positions are diploid genotypes, and deviations from diploidy will be detected afterwards. We consider polymorphisms consisting of two alleles; these can be single nucleotide polymorphisms (SNPs) or structural variants (indels and length polymorphisms). For paralogs and gametologs, which are the result of the mapping of two non-recombining copies, we assume that cases in which both copies are polymorphic at the same position and for the same two alleles are very rare and can be neglected.

2.1.2 Population genetic equilibria

We adopt a hierarchical probabilistic framework in which the distribution of alleles is modeled given a segregation type for each polymorphism. A technical presentation and derivations are given in the Appendix; here, we describe the general principles of the model.

Sex-linkage produces different genotype-allele equilibria for each sex. We consider sites with two alleles, a and b, and three possible genotypes $g' \in \{aa, ab, bb\}$. The genotype frequencies are denoted $p(g')$ if they are equal for both sexes, and indexed with ♂ and ♀ symbols when different. In the following, the allele frequency f is the frequency of allele a, unless otherwise stated.

For **autosomal segregation**, we can simply write

$$p(aa) = f^2 \quad ; \quad p(ab) = 2f(1-f) \quad ; \quad p(bb) = (1-f)^2,$$

which is the Hardy-Weinberg equilibrium.

Haploid genes correspond to haploid genotypes, but genotyping will have given diploid genotypes. No sex-specific difference is expected, so

$$p(aa) = f \quad ; \quad p(ab) = 0 \quad ; \quad p(bb) = 1-f.$$

Paralogy is caused by the mapping of the reads of two more or less recently duplicated genes on one locus in the reference. Thus, paralogous genes have tetraploid genotypes, which again are considered diploid by the genotyper. There is no recombination between the copies, that thus evolve independently. Only biallelic sites are modeled here; for simplicity, we assume that one of the copies is fixed for one of the alleles. Thus, if one copy is fixed for allele a, the tetraploid genotypes are $aaaa$, $aaab$ and $aabb$; the latter two will yield the diploid genotype ab. We define the allele frequency f as the frequency of allele b in the polymorphic copy. Again, no difference between the sexes is expected, and we thus obtain the diploid genotype frequencies

$$p(aa) = (1-f)^2 \quad ; \quad p(ab) = f^2 + 2f(1-f) \quad ; \quad p(bb) = 0.$$

A choice has to be made about which allele to consider fixed, as this is not always clear *a priori*. The previous segregation types, autosomal and haploid, are symmetrical with respect to the choice of the allele to model, *i.e.*, the expected genotype frequencies are strictly identical whether the frequency of allele a was used or the frequency of allele b. The paralogous segregation type is asymmetrical with respect to this choice: considering allele a or allele b as fixed does not lead to identical genotype frequencies. The details about the calculations are specified in the Appendix.

For **hemizygously segregating genes**, the members of one sex are haploid while the others are diploid. Thus, in one sex, the Hardy-Weinberg equilibrium is expected, while in the other, the expectations are the same as for the haploid segregation type. In the case of an XY system, expectations are thus

$$p_{♀}(aa) = f^2 \quad ; \quad p_{♀}(ab) = 2f(1-f) \quad ; \quad p_{♀}(bb) = (1-f)^2$$

$$p_{♂}(aa) = f \quad ; \quad p_{♂}(ab) = 0 \quad ; \quad p_{♂}(bb) = 1-f.$$

bioRxiv preprint doi: https://doi.org/10.1101/2020.01.31.928291; this version posted July 22, 2020. 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-NC-ND 4.0 International license.

Gametologous segregation is characterized by the presence of two independent copies in one sex, and one copy in the other. As for the paralogous case, we assume that an allele is fixed in at least one of the copies. We have to distinguish two cases: either the X copy is polymorphic, or the Y copy. In either case, we have to choose which allele we consider fixed on one of the copies, thus leading to four different allele-genotype equilibria. Here, one case of X-polymorphism and one case of Y-polymorphism are described; details about the four equilibria are given in the Appendix.

In the case of polymorphism on the X, female genotypes are modeled by the Hardy-Weinberg equilibrium, while male genotypes show a specific deviation. If we consider allele b to be fixed on the Y chromosome, and define f as the frequency of a on the X chromosome, we get

$$p_{♀}(aa) = f^2 \quad ; \quad p_{♀}(ab) = 2f(1-f) \quad ; \quad p_{♀}(bb) = (1-f)^2$$
$$p_{♂}(aa) = 0 \quad ; \quad p_{♂}(ab) = f \quad ; \quad p_{♂}(bb) = 1-f.$$

For a polymorphism on the Y chromosome, considering allele b to be fixed on the X-chromosome and defining f to be the frequency of allele a on the Y chromosome yields

$$p_{♀}(aa) = 0 \quad ; \quad p_{♀}(ab) = 0 \quad ; \quad p_{♀}(bb) = 1$$
$$p_{♂}(aa) = 0 \quad ; \quad p_{♂}(ab) = f \quad ; \quad p_{♂}(bb) = 1-f.$$

The X-hemizygous and XY segregation described here can easily be converted to Z-hemizygous and ZW segregation (see Appendix for details).

2.1.3 Parameters and optimization

SDpop provides a complete probabilistic model for the posterior inference of segregation types at the level of polymorphisms. The input of SDpop consists in the observed genotype (g) frequencies at each biallelic site, that can differ from the (unobserved) true genotypes g' with probability e. Then we introduce the segregation type, modeled as an unobserved variable S with multinomial *prior* distribution of parameter π. The conditional probabilities of the true genotypes under each segregation type, $p(g')$, are then obtained through the population genetic equations described above, using the empirical allele frequencies. This yields directly the conditional observed genotype probability $p(g|S) = p(g'|S)\,p(g|g')$. Our model also considers that under gametologous segregation, either one of the gametolog copies could be polymorphic, as well as the fact that in the gametologous and paralogous segregation types, either one of the alleles can be fixed in one of the copies, with equal prior probability (see Appendix). Thanks to an Expectation-Maximization (EM) algorithm we infer the parameters of the model by maximum likelihood, and provide the posterior probability of each segregation type at each polymorphic, $p(S|g)$, site using Bayes' rule. We also propose a contig-wise version of this posterior, by aggregating the site-wise posterior probabilities using geometric means of site-wise posteriors. All the technical derivations of SDpop are provided in the Appendix.

SDpop can investigate several models of sexual systems: absence of sex-linkage, sex-linkage of the XY type, sex-linkage of the ZW type, and sex-linkage of both types. Apart from the sex-linkage, models can be run with or without haploid and paralogous segregation types. The XY and ZW model include both hemizygous and gametologous segregation types. The likelihood of the model when parameter value convergence has been attained is used to calculate the Bayesian Information Criterion (BIC), which allows to compare models with different numbers of parameters.

2.1.4 Haplotypes and population genetics inferences

For the sequences inferred as gametologous, several statistics can be calculated from the optimized values. The posterior probabilities of the segregation subtypes indicate which one of the copies (X or Y) is polymorphic, and which one of the alleles is fixed in one of the copies. We can thus infer the frequency of the alleles in the X and Y copies ($\widehat{f}_X$ and $\widehat{f}_Y$), which is the mean of the empirical allele frequencies, weighted by the posterior probability for each subtype. These allow to reconstruct the X and Y haplotypes, by using only alleles that are fixed (or nearly so) in the X and Y copies.

Furthermore, the inferred allele frequencies allow to calculate the nucleotide diversity of the X and Y copies (π_X and π_Y) and divergence between both gametolog copies (D_{XY}) as the means of the diversity and the divergence over the whole contig:

$$\pi_X = \left\langle 2\widehat{f}_X\left(1-\widehat{f}_X\right)\right\rangle ; \quad \pi_Y = \left\langle 2\widehat{f}_Y\left(1-\widehat{f}_Y\right)\right\rangle ; \quad D_{XY} = \left\langle \widehat{f}_X\left(1-\widehat{f}_Y\right) + \widehat{f}_Y\left(1-\widehat{f}_X\right)\right\rangle$$

2.1.5 Implementation and availability

SDpop is written in C with some C++ functionalities. It is available on https://gitlab.in2p3.fr/sex-det-family/sdpop.

bioRxiv preprint doi: https://doi.org/10.1101/2020.01.31.928291; this version posted July 22, 2020. 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-NC-ND 4.0 International license.

2.2 Simulations

Sets of K contigs of n individuals per sex (i.e., $2n$ individuals in total) were generated using `ms` (Hudson, 2002), that simulates samples drawn from a population under the neutral Wright-Fisher model of genetic variation. Samples of autosomal contigs were generated by simulating $4n$ haploid sequences and combining them arbitrarily into $2n$ diploid samples. The parameter $\theta = 4N_e\mu = 4N_euL$ gives the average number of segregating sites. Here, μ is the overall mutation rate for the sequence with length L and per base mutation rate u. The heterozygosity rate (per site) is thus $H = \frac{\theta/L}{1+\theta/L}$, which is equal to the level of polymorphism π. Thus, for example, to simulate a sequence of $L = 500$ base pairs with a level of polymorphism $\pi = 0.002$, we use the parameter $\theta = \frac{\pi L}{1-\pi} = 1.002$.

Gametologous contigs were generated from a sample of $3n$ X-linked sequences and n Y-linked sequences, that were simulated by supposing two populations that split at time t from a population with size N_e into one population of size $\frac{3}{4}N_e$ (X-linked sequences) and one of size $\frac{1}{4}N_e$ (Y-linked sequences). In `ms`, the time t is scaled by $4N_e$, such that, according to coalescent theory, $t = 1$ is the average time to fixation for a neutral mutation. X-hemizygous sequences were simulated similarly, except that no Y-linked sequences were simulated; male "diploid" genotypes were obtained from the haploid X sequence for each individual.

Errors were introduced after the coalescent simulations, such that each individual at each position had a probability e to have an erroneous genotype. Two types of errors were introduced: either homozygous genotypes were rendered heterozygous, or heterozygous genotypes homozygous, both with the same rate. As there are typically many more homozygous than heterozygous genotypes, many more errors occur on monomorphic sites than on polymorphic sites. SDpop estimates parameters using polymorphic sites only, and in this subset of all sites, the fraction of sites with errors is higher than in the whole genome, which includes monomorphic sites as well. Thus, the dataset of polymorphic sites on which SDpop runs is enriched in sites with errors; we refer to the error rate among polymorphic sites as the "effective error rate", e_e. As an example, for a series of simulations with 2, 5, 10, 20, 50, and 100 individuals per sex, which were all conducted with an error rate of 0.0001 and a level of polymorphism (π) of 0.002, e_e is 0.015, 0.011, 0.0088, 0.0068, 0.0045, and 0.0030, respectively.

The four models of SDpop (without sex chromosomes, with XY chromosomes, with ZW chromosomes, and with both XY and ZW chromosomes) were run on each simulation. The best model was chosen based on the BIC values. The contigs were classified as sex-linked or not based on SDpop's posterior probabilities of the contigs, using a posterior probability threshold of 0.8, and this classification was used to count the number of true positives (TP), false positives (FP), and false negatives (FN). To test the power and precision of the classification of contigs as sex-linked or not is quantified by the true positive rate ($\mathrm{TPR} = \mathrm{TP}/(\mathrm{TP}+\mathrm{FN})$) and the positive predictive value ($\mathrm{PPV} = \mathrm{TP}/(\mathrm{TP}+\mathrm{FP})$). Nucleotide polymorphism and divergence were calculated using the observed allele frequencies before errors were added.

2.3 Human data

Exome-targeted sequence data for Finnish individuals (Auton et al., 2015) were downloaded from the 1000 genomes project website. To facilitate file handling in the analysis pipeline, only individuals for which all reads were present in two fastq files (forward and reverse) were retained, i.e. a total of 66 individuals (44 females and 22 males). These reads were mapped on the human genome reference GRCh37 after removal of the Y sequence was removed (the reference used includes the primary assembly, i.e. chromosomal plus unlocalized and unplaced contigs, and the rCRS mitochondrial sequence, Human herpesvirus 4 type 1 and the concatenated decoy sequences) using BWA (Li and Durbin, 2009) with standard parameters. The individual alignments to the reference were merged, and the genotypes were called using bcftools (mpileup & call), and the targeted exons (file provided by the 1000 genome project) were extracted using bedtools. Exons were grouped by gene using the exon list for the reference GRCh37 retrieved through Ensembl's biomart tool.

3 Results

3.1 Tests on simulations

3.1.1 Model choice: detecting the presence of sex chromosomes

We used simulated data to test the range of validity of SDpop with a controlled population genetic background. We used varying numbers of individuals, different fractions of sex-linked sequences in the genome and varying times since recombination suppression, excluding biologically implausible scenarios (*i.e.* a very small fraction of sex-linked sequences and long times since recombination suppression, or the inverse). The results are shown in Fig. 1 and Supp. Fig. 1.

The behavior of the model when only 2 individual per sex are used is somewhat erratic, and leads to type I errors (Supp. Fig. 1), which is perhaps not surprising given the information that is present in a sample of only

bioRxiv preprint doi: https://doi.org/10.1101/2020.01.31.928291; this version posted July 22, 2020. 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-NC-ND 4.0 International license.

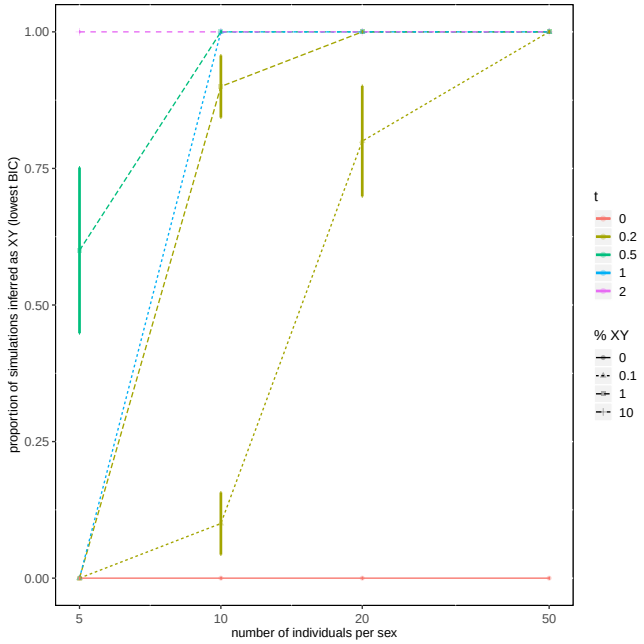


Figure 1: Model choice by SDpop on simulated data. The proportion of simulations for which the XY model had the lowest BIC is indicated; each combination of simulation parameter values was repeated 10 times from a random seed. Colors represent the time since recombination suppression (t) and the line types the percentage of genes that was simulated as gametologous. For clarity, a subset of the data, notably only with $e = 0.0001$, is shown; more results are shown in Supp. Fig. 1.

4 individuals. With 5 or more individuals per sex, no type I errors were observed, and the power to detect sex-linkage increases with the proportion of sex-linked sequences and the time since recombination suppression, as expected (Fig. 1). Indeed, SDpop relies on polymorphic sites that show evidence for sex linkage: for gametologs that have stopped recombining $4N_e$ generations ago ($t = 1$; this is the average time to fixation in a neutral model), we expect and observe that most "true" polymorphisms (i.e. those without errors) in sex-linked genes are polymorphic in only one of the gametologs, or have different alleles fixed on both gametologs. However, for recombination suppression much less than $4N_e$ generations ago, most polymorphisms in gametologs are either derived from ancestral polymorphism, or due to recent mutations with low alternative allele frequency in one of the copies, making detection of sex-linkage much harder. Furthermore, the number of sex-linked genes is typically low in such situations. Our simulations indicate that even with time since recombination suppression as low as $0.1 \times 4N_e$ generations and 0.1% of sex-linked sequences, the method can nevertheless select the appropriate model in most cases when 50 or more individuals per sex are used (Supp. Fig. 1).

3.1.2 Assignment of genes to segregation types

Contigs are considered sex-linked when their posterior probability to be either XY or X-hemizygous exceeds the threshold value of 0.8. We measure the precision of this assignment, i.e. the fraction of contigs assigned as sex-linked that were indeed simulated as sex-linked contigs, as quantified by the positive predictive value (PPV), and the power of the assignment, i.e. the fraction of contigs that were simulated as sex-linked that we are able to detect, as quantified by the true positive rate (TPR). Results are shown in Fig. 2 and Supp. Fig. 2.

For XY gametolog pairs, precision and power increase with time since recombination suppression and the size of the non-recombining region. When the time since recombination suppression exceeds $4N_e$ generations (i.e. $t \geq 1$) and the non-recombining region is sufficiently large (i.e. more than 1% of the genome), the precision is larger than 95% and the power larger than 70% with as less as 5 individuals per sex (Fig. 2). Even with relatively recent recombination suppression ($2N_e$ generations ago, i.e. $t = 0.5$), the method has a precision close to 100% and a power close to 70% with 20 individuals per sex. For shorter time since recombination suppression, both decrease rapidly. Indeed, in these cases, X- and Y-linked SNPs will still have similar frequencies, and many individuals are needed to test whether the observed allele frequency differences are due to sampling or not. For almost all most cases, the precision is higher than the power, meaning that the type I error (false positives) is lower than the type II error (false negatives).

The time since recombination suppression has no clear effect on the detection of X-hemizygous sequences, understandably, as it is only the nucleotide polymorphism in X-linked sequences that creates a signal for detection, and this level of polymorphism only depends on N_e. Of course, when the error rate increases (Supp.

bioRxiv preprint doi: https://doi.org/10.1101/2020.01.31.928291; this version posted July 22, 2020. 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-NC-ND 4.0 International license.

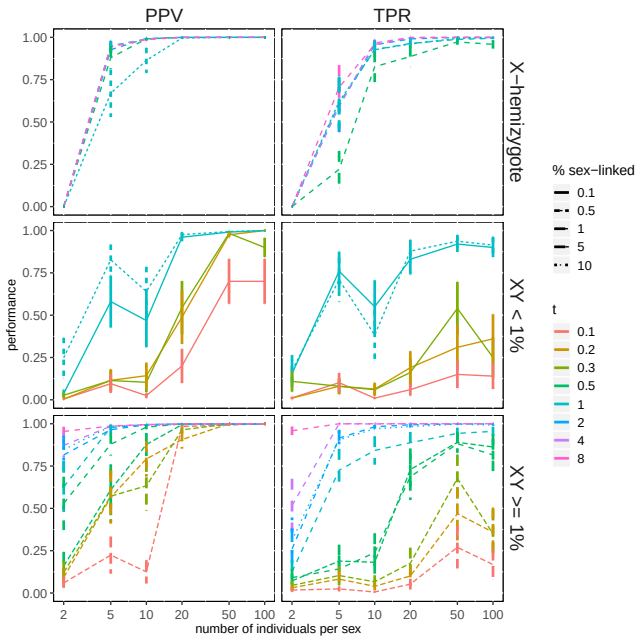


Figure 2: Precision (Positive Predictive Value, left) and power (True Positive Rate, right) of the detection of sex-linked contigs in simulated data, using a threshold for the posterior probability of 0.8. Top graphs: X-hemizygous genes. Middle graphs: XY gametologs, simulations with a small fraction of gametologs in the genome ($< 1\%$); bottom graphs: XY gametologs, simulations with a larger fraction of gametologs in the genome ($\geq 1\%$). Each point is the average of 100 simulations, with the bars representing the standard error. For all cases shown here, the simulated error rate was 0.0001; for $e = 0.001$, see Supp. Fig. 2.

Fig. 2), power and precision decrease, as expected.

3.1.3 Population genetic inferences

An original feature of SDpop is that for genes inferred as gametologs, the parameter estimates and posterior probabilities can be used to estimate the allele frequencies on the X and Y copies. We use these to calculate the level of diversity in X and Y sequences, as well as their divergence. In Fig. 3, the estimates of nucleotide diversity and divergence calculated directly from the simulations, before errors were added, are compared to the estimates based on SDpop's output.

D_{XY} is expected to increase with time since recombination suppression, and for $t \geq 1$, the values based on SDpop's inference correlate well with the simulated values. For more recent recombination suppression, SDpop slightly overestimates D_{XY}; indeed, genotyping and sequencing errors have a larger influence here. Importantly, for $t < 1$, as for π_X and π_Y the effect of random variation is large, as expected due to stochastic effects (Takahata and Nei, 1985).

3.2 Performance on human data

The human genome has a strongly heteromorphic XY chromosome pair, with the Y chromosome being much smaller than the X. This difference is due to the strong degeneration the Y chromosome has been subject to: it has lost many genes. Recombination suppression has occurred several times in the lineage leading to humans, leading to distinct strata characterized by different levels of degeneration (Lahn and Page, 1999). Only in the youngest stratum, genes have retained both X and Y copies, while in the older strata, Y copies have mostly been lost. Note that recombination was suppressed much longer than $4N_e$ generations ago. The most recent stratum is estimated to have stopped recombining between 30×10^6 and 50×10^6 years ago (Lahn and Page, 1999), while a gross estimate of $4N_e$ would be 10^6 years (the human N_e has varied greatly, but is in the order of magnitude of 10^4 (Auton et al., 2015), and the generation time is about 25 years); t is thus much larger than in the cases we've simulated.

SDpop clearly identifies the XY chromosome pair in human sequencing data (Fig. 4; Supp. Fig. 4). Here, the reference consists of the 22 chromosomes and the X (excluding the Y). The larger part of the X chromosome is detected as X-hemizygous, as most of the genes on the X have lost their Y copy. Genes on the extremities of the X chromosome have a high probability to be autosomal, which is again expected as these regions are pseudo-autosomal and do recombine between X and Y. Only one small region with XY gametologs is detected,

bioRxiv preprint doi: https://doi.org/10.1101/2020.01.31.928291; this version posted July 22, 2020. 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-NC-ND 4.0 International license.

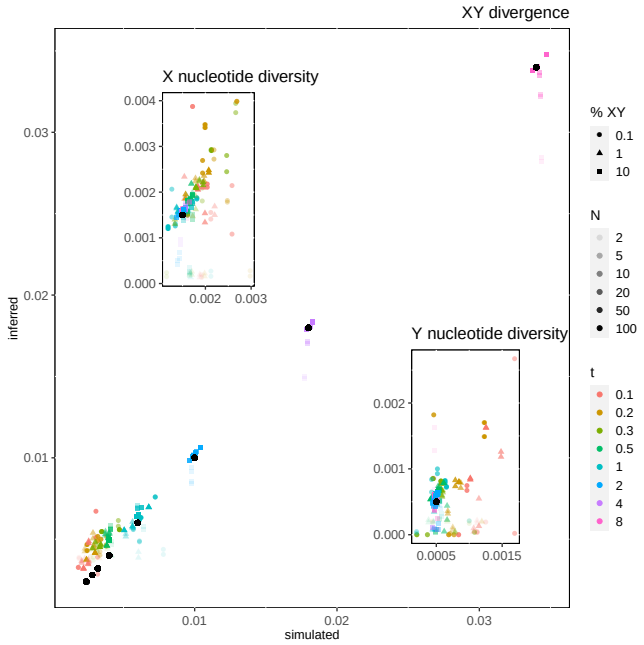


Figure 3: Population genetic inferences of SDpop. The values of nucleotide diversity and divergence calculated directly from the simulation results are compared to the values inferred from SDpop's output. Comparisons are based on SDpop's assignment of the genes (i.e. all genes with a posterior probability > 0.8 were used). Main graph: gametolog divergence (D_{XY}); left inset: X nucleotide diversity (π_X); right inset: Y nucleotide diversity (π_Y). N is the number of individuals per sex, t the time since recombination suppression, and %XY the simulated proportion of gametologs. The black points indicate the theoretical values (D_{XY}: one for each t; π_X and π_Y: one value for all runs). Here, $e = 0.0001$; for higher error rates, see Supp. Fig. 3.

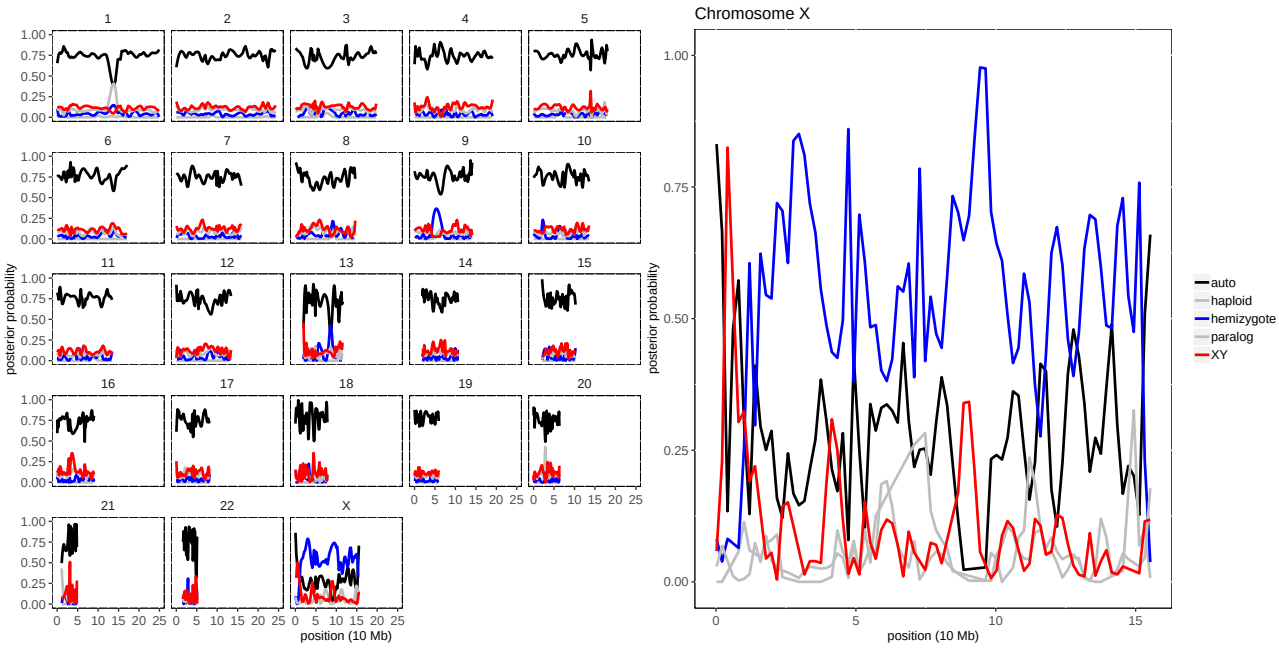


Figure 4: Test of SDpop's performance on the human exome-targeted sequencing data from the 1000 genome project, using 10 individuals per sex. Smoothed gene-level posterior probabilities for autosomal (black), X-hemizygote (blue) and XY (red) segregation are shown; haploid and paralogous posterior probabilities are indicated in grey. The right panel show the results on the X chromosome: the extremities corresponding to the pseudo-autosomal regions are predicted to be autosomal by SDpop, while most XY gametologous genes are found close to the pseudoautosomal region on the left arm, which represents the youngest stratum where Y copies have not yet been lost. The rest of the chromosome consists of X-hemizygous genes, that lack a Y copy. Results obtained with 5 and 20 individuals per sex are shown in Supp. Fig. 4.

bioRxiv preprint doi: https://doi.org/10.1101/2020.01.31.928291; this version posted July 22, 2020. 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-NC-ND 4.0 International license.

near the left pseudo-autosomal region, at the position of the youngest stratum with XY gametologs. The genes in the older strata, for which the Y copies have been lost, are detected as X-hemizygous by SDpop. The Y also has several genes without X homologs, that probably resulted from transpositions or translocations from other autosomes, the so-called ampliconic genes (Skaletsky et al., 2003). Indeed, the autosomal gene DAZL, known to have given rise to the Y-ampliconic family DAZ, had a very high probability to be sex-linked (> 0.99). Other genes are also shown be homologous with sequences on the sex chromosomes (Galichon et al., 2012), and two of these (PPP1R12B and TPTE2) harbor the majority of sex-linked SNPs detected outside chromosome X.

As Y-specific genes are often present in a few copies, which allows gene conversion to rescue these sequences that suffer from the lack of recombination, Y-specific diversity and XY-divergence are not correctly estimated from the outputs of SDpop. Indeed, we observe that the parameter ρ, the proportion of X-polymorphism in XY gametologs, is 0.12 when using 20 individuals per sex, indicating that 88% of polymorphism in XY gametologs is due to Y polymorphism. This could be due to the mapping of several copies of the Y gene to the same position on the X chromosome.

4 Discussion

SDpop is a probabilistic framework for the detection of sex-linked sequences, that relies on the modeling of the expected equilibrium between allele and genotype frequencies under sex-linked segregation. It combines the principles that are at the basis of several methods for detection of sex-linkage, such as increased frequencies of heterozygotes or allele frequency deviations, in a specific framework. As such, it requires less individuals than methods that are based on such allele frequencies or genotype frequencies alone: e.g., GWAS usually needs > 50 individuals, as do studies using tests for heterozygote frequencies (Picq et al., 2014). Of course, SDpop's power depends on the size and age of the non-recombining region, as would be the case for any method.

The likelihood-based framework of SDpop allows to compare models with and without sex linkage. It is thus possible, with a moderate sequencing effort, to determine the sex chromosome system and to obtain the sex-linked sequences for any species that can be sexed and sampled in the field. The approach can be used on any kind of individual-based sequencing data, such as RNA-seq, DNA-reseq and RAD-seq. The functionality to calculate gene-level posterior probabilities is of course only useful for gene-based sequencing, such as RNA-seq and exome capture. For DNA-reseq, posterior probabilities could be calculated when the scaffold size is small, or by splitting the chromosomes into windows of fixed size.

The underlying population genetics model uses the classical assumptions of the Hardy-Weinberg principle, notably random mating between the sexes in an infinitely large population. The model proposed here will thus perform best when used on a sample of individuals taken from a single, panmictic population. Any population structure will weaken the performance of the model. Indeed, when females and males are sampled from separate populations, this will lead to type I errors (false positives), as the population structure mimics the deviations from Hardy-Weinberg equilibrium that are expected for sex linkage. However, if the individuals sampled from different populations comprise both females and males from each population, the expected influence on SDpop's performance is mainly a loss of power (i.e., an increased proportion of false negatives).

We model different kinds of sources of error in SDpop. First, we explicitly model haploid and paralogous sequences. These are more or less frequently encountered in NGS data; haploid sequences could result from contamination (e.g. mitochondria, chloroplasts, or bacteria), monoallelic expression in RNAseq data, or redundancy in the reference (i.e. different alleles of a gene are split into several contigs). Paralogous sequences could result from recent paralogs that were not recognized as such in the reference genome or transcriptome, or from contamination between samples. Furthermore, SDpop incorporates an error parameter, to account for other sequencing or genotyping errors. These sources of errors are modeled solely to improve the detection of sex-linkage: failing to take them into account would lead to haploid sequences being inferred as X-hemizygotes, paralog sequences as XY gametologs, while sequencing and genotyping errors would penalize these sex-linked segregation types more than the inference of autosomal segregation. The power, sensitivity and precision of SDpop's inferences of haploid and paralogous segregation as well as the error rate have not been tested extensively, and thus one should not use the method to estimate these segregation types and the error rate from the data.

SDpop's sex-linked segregation types include both gametologous segregation and hemizygote segregation. In the first case, both gametologous copies are present, while in the second, there is no information about the sequence from the sex-limited chromosome. X-hemizygous loci thus correspond loss of the Y copy of a gene, presumably through Y degeneration. However, apparent X-hemizygosity can also be caused by artifacts that are more or less difficult to control. First, the X and Y copies might be incorporated as distinct genes in the genome or transcriptome assembly. In species with known and well described sex chromosomes, such as humans, the Y assembly could simply be excluded from mapping, as we did here. In species with more recently evolved or less well described sex chromosomes, one should thus preferably use the homogametic sex for preparing a mapping reference. The case that is harder to solve arises when Y sequences are too divergent to map on the X

bioRxiv preprint doi: https://doi.org/10.1101/2020.01.31.928291; this version posted July 22, 2020. 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-NC-ND 4.0 International license.

copy in the reference; thus, when some sex-linked genes have high XY divergence values, one should be aware of the fact that some genes that have been inferred as X-hemizygous might be artifacts.

As shown here and by others (e.g. Crowson et al., 2017), XY gametologs are much easier to detect than X-hemizygous genes, as, first, XY gametologs will have more SNPs than X-hemizygous genes, and second, the information contained in a fixed XY SNP is much less ambiguous than the one of a X-hemizygous SNP. These are additional reasons to try to reduce the number of artifactual X-hemizygous as much as possible, in order to obtain more reliable inferences.

Importantly, there is further information to be obtained from XY gametologs. We use the inferred allele frequencies on the X and the Y copies to calculate the nucleotide diversity in the X and Y copies, and their divergence. Thus, using the information on which alleles are fixed in each copy, SDpop is also able to reconstruct the haplotypes of the X and Y copies, even when the input data have not previously been phased. We've shown that the estimation of population genetic parameters comes quite close to the simulated values, even though these estimations are based on empirical allele frequencies (i.e., they do not take the error rate into account). However, even if the estimate would be perfectly unbiased, we expect that the variance of these parameters on a per-gene basis (cf Takahata and Nei, 1985) is much larger than the bias introduced by not taking the error rate into account. Thus, obtaining reliable estimates of these parameters (i.e., estimates that reflect population processes and not mere stochasticity) requires many independent samples, and cannot be addressed solely by modeling.

Although it has similarities with available methods, SDpop is unique in the combination of input data it requires and the predictions it can produce. As a consequence, we cannot compare its performance directly to any of previously published methods. Our simulations and tests show that reasonable performance can be achieved with as few as 5 individuals per sex. This is close to the sequencing effort required for SEX-DETector (Muyle et al., 2016), which relies on a controlled cross to infer sex-linkage. Given the fact that usage of SDpop puts less constraint on the input data (kinship does not need to be known, and is assumed to be absent), it is, at the first sight, somewhat surprising that SDpop can work with as few individuals as can SEX-DETector. The reason for this is that not all sites are informative for SEX-DETector (e.g. when both parents are heterozygous) and are discarded, whereas SDpop considers all polymorphic sites as informative; the amount of information supplied by a site varies depending on the difference of the conditional likelihood for all segregation sites.

Apart form the main practical advantage that SDpop does not require controlled breeding of the study organism, it has a few more advantages compared to SEX-DETector. First, as SEX-DETector considers parents and their F1 offspring, there has been little recombination between the homologous chromosomes. For the sex chromosomes, this implies that genes from the pseudo-autosomal region are genetically linked to the non-recombining region, and will have more or less distorted segregation. This leads SEX-DETector to overestimate the size of the non-recombining region, especially when the pseudo-autosomal region is large and the sex-linked region small (with a large non-recombining and a small pseudo-autosomal region, this effect is less important, as the sparse recombination events of the sex chromosomes will be located in the small pseudo-autosomal region, and linkage disequilibrium will decrease rapidly). Thus, we expect SDpop to yield a more precise indication of the pseudo-autosomal boundary. Note that, in the case of a very small and recently evolved non-recombining region that will be difficult to detect by SDpop, SEX-DETector's behavior of overestimating the non-recombining region might be beneficial as it increases the capacity to detect the sex chromosome; in such cases, model choice could be done with SEX-DETector on a controlled cross, and delimitation of the non-recombining region with SDpop on population data.

Second, an advantage of the use of population data in SDpop is that estimates of population genetic parameters are possible. In a cross, there will be three X chromosomes and one Y, so it will be impossible to distinguish fixed substitutions from polymorphism on the Y chromosome.

DETSEX (Gautier, 2014) uses a Bayesian framework that can be used on samples collected in natural populations to infer whether markers (SNPs) are sex-linked or not. Despite these obvious similarities with the model proposed here, there are several important differences. First, SDpop allows to compare of the total likelihoods of the models, and thus can be used as a statistical test for the presence or absence of sex chromosomes, while in DETSEX, the presence of sex chromosomes is assumed, but the individual's sex does not need to be known. Second, in DETSEX, X and Y-linked sequences are expected to map to different positions, which can be a safe assumption in well-studied species with an old sex chromosomes system (such as humans), but not for more recently evolved sex chromosomes.

Thus, SDpop fills a gap in the current panel of methodological approaches for the identification and study of sex chromosomes. It requires input data that have now become classical: short reads sequencing of genomes or transcriptomes of ten to twenty individuals, collected in any population. It uses standard vcf files as input, thus allowing integration in existing genotyping pipelines. Its probabilistic framework and its implementation in a widely used and efficient programming language (C/C++) furthermore allow future developments (including, but not limited to, inference of individual's sex, corrections for population structure).

bioRxiv preprint doi: https://doi.org/10.1101/2020.01.31.928291; this version posted July 22, 2020. 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-NC-ND 4.0 International license.

Acknowledgements

We thank Laurent Duret for help with the treatment of the 1000 genomes data, François Gindraud for help with programming, and Sylvain Mousset and Thibault Latrille for discussions on population genetics issues. This work was performed using the computing facilities of the CC LBBE/PRABI, and was supported by funding from the Agence Nationale de la Recherche (grant number ANR-14-CE19-0021-01).

Author contributions

GM conceived the project and acquired funding; JK, NL, GM and FP conceived the model; JK, NL and FP developed the methodology and the formal analysis; JK developed the software, performed simulations and data analysis, and prepared the first draft manuscript; JK, GM and FP wrote the current version of the manuscript; all authors approve of the manuscript.

References

Auton, A., L. D. Brooks, R. M. Durbin, E. P. Garrison, H. M. Kang, et al. (2015). "A global reference for human genetic variation." In: *Nature* 526.7571, pp. 68–74.

Bachtrog, D., J. E. Mank, C. L. Peichel, M. Kirkpatrick, S. P. Otto, et al. (2014). "Sex determination: why so many ways of doing it?" In: *PLoS Biol.* 12.7, e1001899.

Bellott, D. W., T. J. Cho, J. F. Hughes, H. Skaletsky, and D. C. Page (2018). "Cost-effective high-throughput single-haplotype iterative mapping and sequencing for complex genomic structures." In: *Nat. Protoc.* 13.4, pp. 787–809.

Bellott, D. W., J. F. Hughes, H. Skaletsky, L. G. Brown, T. Pyntikova, et al. (2014). "Mammalian Y chromosomes retain widely expressed dosage-sensitive regulators." In: *Nature* 508.7497, pp. 494–499.

Carvalho, A. B. and A. G. Clark (2013). "Efficient identification of Y chromosome sequences in the human and *Drosophila* genomes." In: *Genome Res.* 23.11, pp. 1894–1907.

Charlesworth, D. (2016). "Plant sex chromosomes." In: *Annu. Rev. Plant Biol.* 67, pp. 397–420.

Charlesworth, D., B. Charlesworth, and G. Marais (2005). "Steps in the evolution of heteromorphic sex chromosomes." In: *Heredity* 95.2, pp. 118–128.

Cortez, D., R. Marin, D. Toledo-Flores, L. Froidevaux, A. Liechti, P. D. Waters, F. Grutzner, and H. Kaessmann (2014). "Origins and functional evolution of Y chromosomes across mammals." In: *Nature* 508.7497, pp. 488–493.

Crowson, D., S. C. H. Barrett, and S. I. Wright (2017). "Purifying and Positive Selection Influence Patterns of Gene Loss and Gene Expression in the Evolution of a Plant Sex Chromosome System." In: *Mol. Biol. Evol.* 34.5, pp. 1140–1154.

Galichon, P., L. Mesnard, A. Hertig, B. Stengel, and E. Rondeau (2012). "Unrecognized sequence homologies may confound genome-wide association studies." In: *Nucleic Acids. Res.* 40.11, pp. 4774–4782.

Garcia-Moreno, J. and D. P. Mindell (2000). "Rooting a phylogeny with homologous genes on opposite sex chromosomes (gametologs): a case study using avian CHD." In: *Mol. Biol. Evol.* 17.12, pp. 1826–1832.

Gautier, M. (2014). "Using genotyping data to assign markers to their chromosome type and to infer the sex of individuals: a Bayesian model-based classifier." In: *Mol. Ecol. Resour.* 14.6, pp. 1141–1159.

Hall, A. B., Y. Qi, V. Timoshevskiy, M. V. Sharakhova, I. V. Sharakhov, and Z. Tu (2013). "Six novel Y chromosome genes in *Anopheles* mosquitoes discovered by independently sequencing males and females." In: *BMC Genomics.* 14, p. 273.

Hudson, R. R. (2002). "Generating samples under a Wright-Fisher neutral model of genetic variation." In: *Bioinformatics* 18.2, pp. 337–338.

Hughes, J. F. and S. Rozen (2012). "Genomics and genetics of human and primate y chromosomes." In: *Annu. Rev. Genomics. Hum. Genet.* 13, pp. 83–108.

Kirkpatrick, M. and R. F. Guerrero (2014). "Signatures of sex-antagonistic selection on recombining sex chromosomes." In: *Genetics* 197.2, pp. 531–541.

Lahn, B. T. and D. C. Page (1999). "Four evolutionary strata on the human X chromosome." In: *Science* 286.5441, pp. 964–967.

Li, H. and R. Durbin (2009). "Fast and accurate short read alignment with Burrows-Wheeler transform." In: *Bioinformatics* 25.14, pp. 1754–1760.

Li, S., M. Ajimura, Z. Chen, J. Liu, E. Chen, et al. (2018). "A new approach for comprehensively describing heterogametic sex chromosomes." In: *DNA Res.* 25.4, pp. 375–382.

Lynch, M. (2007). *The Origins of Genome Architecture.* Sunderland, MA, USA: Sinauer Associates, Inc.

bioRxiv preprint doi: https://doi.org/10.1101/2020.01.31.928291; this version posted July 22, 2020. 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-NC-ND 4.0 International license.

Martin, H., F. Carpentier, S. Gallina, C. Gode, E. Schmitt, A. Muyle, G. A. B. Marais, and P. Touzet (2019). "Evolution of Young Sex Chromosomes in Two Dioecious Sister Plant Species with Distinct Sex Determination Systems." In: *Genome Biol. Evol.* 11.2, pp. 350–361.

Mathew, L. S., M. Spannagl, A. Al-Malki, B. George, M. F. Torres, et al. (2014). "A first genetic map of date palm (*Phoenix dactylifera*) reveals long-range genome structure conservation in the palms." In: *BMC Genomics.* 15, p. 285.

Muyle, A., J. Käfer, N. Zemp, S. Mousset, F. Picard, and G. A. Marais (2016). "SEX-DETector: a probabilistic approach to study sex chromosomes in non-model organisms." In: *Genome Biol. Evol.* 8.8, pp. 2530–2543.

Muyle, A., R. Shearn, and G. A. Marais (2017). "The evolution of sex chromosomes and dosage compensation in plants." In: *Genome Biol. Evol.* 9.3, pp. 627–645.

Muyle, A., N. Zemp, C. Fruchard, R. Cegan, J. Vrana, et al. (2018). "Genomic imprinting mediates dosage compensation in a young plant XY system." In: *Nat. Plants.* 4.9, pp. 677–680.

Palmer, D. H., T. F. Rogers, R. Dean, and A. E. Wright (2019). "How to identify sex chromosomes and their turnover." In: *Mol. Ecol.* 28.21, pp. 4709–4724.

Picq, S., S. Santoni, T. Lacombe, M. Latreille, A. Weber, et al. (2014). "A small XY chromosomal region explains sex determination in wild dioecious *V. vinifera* and the reversal to hermaphroditism in domesticated grapevines." In: *BMC. Plant Biol.* 14, p. 229.

Ponnikas, S., H. Sigeman, J. K. Abbott, and B. Hansson (2018). "Why do sex chromosomes stop recombining?" In: *Trends Genet.* 34.7, pp. 492–503.

Prentout, D., O. Razumova, B. Rhoné, H. Badouin, H. Henri, C. Feng, J. Käfer, G. Karlov, and G. A. Marais (2020). "A high-throughput segregation analysis identifies the sex chromosomes of *Cannabis sativa*". In: *Genome Res.* in press.

Qiu, S., R. Bergero, and D. Charlesworth (2013). "Testing for the footprint of sexually antagonistic polymorphisms in the pseudoautosomal region of a plant sex chromosome pair." In: *Genetics* 194.3, pp. 663–672.

Scharmann, M., T. U. Grafe, F. Metali, and A. Widmer (2017). "Sex-determination and sex chromosomes are shared across the radiation of dioecious *Nepenthes* pitcher plants". In: *bioRxiv.*

Skaletsky, H., T. Kuroda-Kawaguchi, P. J. Minx, H. S. Cordum, L. Hillier, et al. (2003). "The male-specific region of the human Y chromosome is a mosaic of discrete sequence classes." In: *Nature* 423.6942, pp. 825–837.

Takahata, N. and M. Nei (1985). "Gene genealogy and variance of interpopulational nucleotide differences." In: *Genetics* 110.2, pp. 325–344.

Tomaszkiewicz, M., P. Medvedev, and K. D. Makova (2017). "Y and W Chromosome Assemblies: Approaches and Discoveries." In: *Trends Genet.* 33.4, pp. 266–282.

Torres, M. F., L. S. Mathew, I. Ahmed, I. K. Al-Azwani, R. Krueger, et al. (2018). "Genus-wide sequencing supports a two-locus model for sex-determination in *Phoenix*." In: *Nat. Commun.* 9.1, p. 3969.

Veltsos, P., K. E. Ridout, M. A. Toups, S. C. Gonzalez-Martinez, A. Muyle, et al. (2019). "Early Sex-Chromosome Evolution in the Diploid Dioecious Plant *Mercurialis annua*." In: *Genetics* 212.3, pp. 815–835.

Vicoso, B. and D. Bachtrog (2011). "Lack of global dosage compensation in *Schistosoma mansoni*, a female-heterogametic parasite." In: *Genome Biol. Evol.* 3, pp. 230–235.

Vicoso, B., J. J. Emerson, Y. Zektser, S. Mahajan, and D. Bachtrog (2013a). "Comparative sex chromosome genomics in snakes: differentiation, evolutionary strata, and lack of global dosage compensation." In: *PLoS Biol.* 11.8, e1001643.

Vicoso, B., V. B. Kaiser, and D. Bachtrog (2013b). "Sex-biased gene expression at homomorphic sex chromosomes in emus and its implication for sex chromosome evolution." In: *Proc. Natl. Acad. Sci. U.S.A.* 110.16, pp. 6453–6458.

Wang, J., J. K. Na, Q. Yu, A. R. Gschwend, J. Han, et al. (2012). "Sequencing papaya X and Yh chromosomes reveals molecular basis of incipient sex chromosome evolution." In: *Proc. Natl. Acad. Sci. U.S.A.* 109.34, pp. 13710–13715.

Zhou, Y., M. Massonnet, J. S. Sanjak, D. Cantu, and B. S. Gaut (2017). "Evolutionary genomics of grape (*Vitis vinifera ssp. vinifera*) domestication." In: *Proc. Natl. Acad. Sci. U.S.A.* 114.44, pp. 11715–11720.

Appendix

Full description of the model

Observed and hidden variables

The data consist of genotyped individuals for genes $k \in \{1, 2, ..., K\}$, each with biallelic sites $t \in \{1, 2, ..., T_k\}$. For each biallelic site, the observed alleles are randomly named a and b, such that three genotypes are possible: homozygote for allele a ("aa"), heterozygote ("ab"), and homozygote for allele b ("bb"). At each biallelic site t of gene k, and for each individual i, $\mathrm{OG}^{tk}_{i_h g}$ is an indicator of the individual having observed genotype $g \in$

bioRxiv preprint doi: https://doi.org/10.1101/2020.01.31.928291; this version posted July 22, 2020. 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-NC-ND 4.0 International license.

$\{1 = aa, 2 = ab, 3 = bb\}$ and sex $h \in \{1, 2\}$. $h = 1$ for females, $h = 2$ males; for convenience, we will write ♀ and ♂. N^{kt}_{gh} is the number of individuals with sex h and observed genotype g; note that the total number of observations (i.e., genotyped individuals) can vary between sites. The vector $\mathbf{OG}^{kt}$ describes all observations at a site.

We seek under which segregation type S_j, $j \in 1...7$, these observed genotypes are most likely. These segregation types will be listed in fixed order, and a specific number corresponds to each of them:

1. Diploid autosomal segregation
2. Haploid sequences
3. Paralogs
4. X-hemizygous segregation
5. XY gametologous segregation
6. Z-hemizygous segregation
7. ZW gametologous segregation

For some segregation types, as will be detailed below, several sub-types of segregation have to be specified; these are denoted A_l with $l \in \{1..L\}$. The conditional likelihood of observing the genotypes under each segregation type depends on the allele frequencies and a genotyping error rate.

We introduce a hidden variable $\mathrm{TG}^{kt}_{i_h g'}$ which is an indicator for the true genotype $g' \in \{aa, ab, bb\}$ of an individual i with sex h. The conditional probabilities of observing a true genotype for an individual, given the fully specified segregation type, are

$$\mathbf{TG}^{kt}_{i_h} | S_j, A_l \sim \mathcal{M}\left(1; P^{kt}_{1hjl}, P^{kt}_{2hjl}, P^{kt}_{3hjl}\right) .$$

$\mathbf{P}^{kt}_{hjl}$ is the vector of the probabilities $\left(P^{kt}_{1hjl}, P^{kt}_{2hjl}, P^{kt}_{3hjl}\right)$ for each genotype at a site, given the sex of the individual and the segregation type and subtype. The genotype probabilities P^{kt}_{ghjl} are calculated from the empirical allele frequencies $\widehat{f}^{kt}_{jl}$ using the following population genetic expectations.

1. For autosomal segregation, the Hardy-Weinberg equilibrium should hold in both sexes. Thus,

$$\mathbf{P}^{kt}_{♀,j=1} = \mathbf{P}^{kt}_{♂,j=1} = \begin{pmatrix} (\widehat{f}^{kt}_{j=1})^2 \\ 2\widehat{f}^{kt}_{j=1}\left(1 - \widehat{f}^{kt}_{j=1}\right) \\ \left(1 - \widehat{f}^{kt}_{j=1}\right)^2 \end{pmatrix}$$

   where

$$\widehat{f}^{kt}_{j=1} = \frac{2N^{kt}_{aa} + N^{kt}_{ab}}{2N^{kt}} .$$

2. Haploid segregation is modeled by

$$\mathbf{P}^{kt}_{♀,j=2} = \mathbf{P}^{kt}_{♂,j=2} = \begin{pmatrix} \widehat{f}^{kt}_{j=2} \\ 0 \\ 1 - \widehat{f}^{kt}_{j=2} \end{pmatrix}$$

   and

$$\widehat{f}^{kt}_{j=2} = \frac{N^{kt}_{aa}}{N^{kt}_{aa} + N^{kt}_{bb}} .$$

3. Paralogy is caused by the mapping of the reads of two more or less recently duplicated genes on one locus in the reference. There is no recombination between the copies, that thus evolve independently. For simplicity, we assume that one of the copies is fixed for one of the alleles. The genotype probabilities depend on which allele is considered fixed in one of the copies, and two sub-types have to be modeled.

   (a) First, we consider that allele a is fixed in one of the copies. $\widehat{f}^{kt}_{j=3,l=1}$ is the frequency of allele b in the other copy. In reality, such sites have four copies; thus, the genotypes are $aaaa$, $aaab$, are $aabb$, with frequencies $\left(1 - \widehat{f}^{kt}_{j=3,l=1}\right)^2$, $2\widehat{f}^{kt}_{j=3,l=1}\left(1 - \widehat{f}^{kt}_{j=3,l=1}\right)$ and $(\widehat{f}^{kt}_{j=3,l=1})^2$. Genotypes $aaab$ and $aabb$

bioRxiv preprint doi: https://doi.org/10.1101/2020.01.31.928291; this version posted July 22, 2020. 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-NC-ND 4.0 International license.

will probably be considered as ab by the genotyper that expects only diploids; thus, the genotype probabilities are:

$$\mathbf{P}^{kt}_{♀,j=3,l=1} = \mathbf{P}^{kt}_{♂,j=3,l=1} = \begin{pmatrix} \left(1-\widehat{f}^{kt}_{j=3,l=1}\right)^2 \\ (\widehat{f}^{kt}_{j=3,l=1})^2 + 2\widehat{f}^{kt}_{j=3,l=1}\left(1-\widehat{f}^{kt}_{j=3,l=1}\right) \\ 0 \end{pmatrix}$$

To estimate the empirical allele frequency, note that the ab genotype counts that are obtained from the genotyper results will be a mixture of $aaab$ and $aabb$, in the expected proportions that depend on the frequency that we concisely denote $\widehat{f}$ here: $2\widehat{f}\left(1-\widehat{f}\right)/(\widehat{f}^2+2\widehat{f}\left(1-\widehat{f}\right))$ and $\widehat{f}^2/(\widehat{f}^2+2\widehat{f}\left(1-\widehat{f}\right))$. While in reality, $\widehat{f} = 0.5(N_{aaab}+2N_{aabb})/\left(N_{aaaa}+N_{aaab}+N_{aabb}\right)$, we instead calculate

$$\widehat{f} \quad = \quad \frac{1}{2\left(N_{aa}+N_{ab}\right)}\left(\frac{2\widehat{f}\left(1-\widehat{f}\right)N_{ab}}{\widehat{f}^2+2\widehat{f}\left(1-\widehat{f}\right)} + \frac{2\widehat{f}^2 N_{ab}}{\widehat{f}^2+2\widehat{f}\left(1-\widehat{f}\right)}\right)$$

This yields

$$\widehat{f}^{kt}_{j=3,l=1} = 1-\sqrt{1-\frac{N^{kt}_{ab}}{N^{kt}_{aa}+N^{kt}_{ab}}}\,.$$

(b) Alternatively, allele b could be fixed in one of the copies. $\widehat{f}^{kt}_{j=3,l=2}$ is the frequency of allele a in the other copy. The genotype probabilities and empirical allele frequency are

$$\mathbf{P}^{kt}_{♀,j=3,l=2} = \mathbf{P}^{kt}_{♂,j=3,l=2} = \begin{pmatrix} 0 \\ (\widehat{f}^{kt}_{j=3,l=2})^2 + 2\widehat{f}^{kt}_{j=3,l=2}\left(1-\widehat{f}^{kt}_{j=3,l=2}\right) \\ \left(1-\widehat{f}^{kt}_{j=3,l=2}\right)^2 \end{pmatrix}$$

$$\widehat{f}^{kt}_{j=3,l=2} = 1-\sqrt{1-\frac{N^{kt}_{ab}}{N^{kt}_{ab}+N^{kt}_{bb}}}\,.$$

4. For X-hemizygously segregating genes, the members of one sex are haploid while the others are diploid.

$$\mathbf{P}^{kt}_{♀,j=4} = \begin{pmatrix} (\widehat{f}^{kt}_{j=4})^2 \\ 2\widehat{f}^{kt}_{j=4}\left(1-\widehat{f}^{kt}_{j=4}\right) \\ \left(1-\widehat{f}^{kt}_{j=4}\right)^2 \end{pmatrix} \quad ; \quad \mathbf{P}^{kt}_{♂,j=4} = \begin{pmatrix} \widehat{f}^{kt}_{j=4} \\ 0 \\ 1-\widehat{f}^{kt}_{j=4} \end{pmatrix}$$

$$\widehat{f}^{kt}_{j=4} = \frac{2N^{kt}_{aa,♀}+N^{kt}_{ab,♀}+N^{kt}_{aa,♂}}{2\left(N^{kt}_{aa,♀}+N^{kt}_{ab,♀}+N^{kt}_{bb,♀}\right)+N^{kt}_{aa,♂}+N^{kt}_{bb,♂}}$$

5. XY gametologous segregation is characterized by the presence of two independent copies in males, and two copies of the X gene in females. We assume that an allele is fixed in at least one of the copies.

(a) X-polymorphism, allele 1 fixed on Y. f is the frequency of allele 2 on X.

$$\mathbf{P}^{kt}_{♀,j=5,l=1} = \begin{pmatrix} \left(1-\widehat{f}^{kt}_{j=5,l=1}\right)^2 \\ 2\widehat{f}^{kt}_{j=5,l=1}\left(1-\widehat{f}^{kt}_{j=5,l=1}\right) \\ (\widehat{f}^{kt}_{j=5,l=1})^2 \end{pmatrix} \quad ; \quad \mathbf{P}^{kt}_{♂,j=5,l=1} = \begin{pmatrix} 1-\widehat{f}^{kt}_{j=5,l=1} \\ \widehat{f}^{kt}_{j=5,l=1} \\ 0 \end{pmatrix}$$

$$\widehat{f}^{kt}_{j=5,l=1} = \frac{2N^{kt}_{bb♀}+N^{kt}_{ab♀}+N^{kt}_{ab♂}}{2\left(N^{kt}_{aa♀}+N^{kt}_{ab♀}+N^{kt}_{bb♀}\right)+N^{kt}_{aa♂}+N^{kt}_{ab♂}}$$

(b) X-polymorphism; allele 2 fixed on Y. f is the frequency of allele 1 on X :

$$\mathbf{P}^{kt}_{♀,j=5,l=2} = \begin{pmatrix} (\widehat{f}^{kt}_{j=5,l=2})^2 \\ 2\widehat{f}^{kt}_{j=5,l=2}\left(1-\widehat{f}^{kt}_{j=5,l=2}\right) \\ \left(1-\widehat{f}^{kt}_{j=5,l=2}\right)^2 \end{pmatrix} \quad ; \quad \mathbf{P}^{kt}_{♂,j=5,l=2} = \begin{pmatrix} 0 \\ \widehat{f}^{kt}_{j=5,l=2} \\ 1-\widehat{f}^{kt}_{j=5,l=2} \end{pmatrix}$$

bioRxiv preprint doi: https://doi.org/10.1101/2020.01.31.928291; this version posted July 22, 2020. 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-NC-ND 4.0 International license.

$$\widehat{f}^{kt}_{j=5,l=2} = \frac{2N^{kt}_{aa♀} + N^{kt}_{ab♀} + N^{kt}_{ab♂}}{2\left(N^{kt}_{aa♀} + N^{kt}_{ab♀} + N^{kt}_{bb♀}\right) + N^{kt}_{bb♂} + N^{kt}_{ab♂}}$$

(c) Y-polymorphism, allele 1 fixed on X. f is the frequency of allele 2 on Y:

$$\mathbf{P}^{kt}_{♀,j=5,l=3} = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} \quad ; \quad \mathbf{P}^{kt}_{♂,j=5,l=3} = \begin{pmatrix} 1 - \widehat{f}^{kt}_{j=5,l=3} \\ \widehat{f}^{kt}_{j=5,l=3} \\ 0 \end{pmatrix}$$

$$\widehat{f}^{kt}_{j=5,l=3} = \frac{N^{kt}_{ab♂}}{N^{kt}_{aa♂} + N^{kt}_{ab♂}}$$

(d) Y-polymorphism, allele 2 fixed on X. f is the frequency of allele 1 on Y:

$$\mathbf{P}^{kt}_{♀,j=5,l=4} = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} \quad ; \quad \mathbf{P}^{kt}_{♂,j=5,l=4} = \begin{pmatrix} 0 \\ \widehat{f}^{kt}_{j=5,l=4} \\ 1 - \widehat{f}^{kt}_{j=5,l=4} \end{pmatrix}$$

$$\widehat{f}^{kt}_{j=5,l=4} = \frac{N^{kt}_{ab♂}}{N^{kt}_{bb♂} + N^{kt}_{ab♂}}$$

6. Z-hemizygous segregation is similar to X-hemizygous segregation:

$$\mathbf{P}^{kt}_{♀,j=6} = \begin{pmatrix} \widehat{f}^{kt}_{j=6} \\ 0 \\ 1 - \widehat{f}^{kt}_{j=6} \end{pmatrix} \quad ; \quad \mathbf{P}^{kt}_{♂,j=6} = \begin{pmatrix} (\widehat{f}^{kt}_{j=6})^2 \\ 2\widehat{f}^{kt}_{j=6}\left(1 - \widehat{f}^{kt}_{j=6}\right) \\ \left(1 - \widehat{f}^{kt}_{j=6}\right)^2 \end{pmatrix}$$

$$\widehat{f}^{kt}_{j=6} = \frac{2N^{kt}_{aa,♂} + N^{kt}_{ab,♂} + N^{kt}_{aa,♀}}{2\left(N^{kt}_{aa,♂} + N^{kt}_{ab,♂} + N^{kt}_{bb,♂}\right) + N^{kt}_{aa,♀} + N^{kt}_{bb,♀}}$$

7. ZW gametologous segregation is modeled similar to XY gametologous segregation, for both Z and W polymorphism, and two asymmetrical cases for each.

(a) Z-polymorphism, allele 1 fixed on W. f is the frequency of allele 2 on Z.

$$\mathbf{P}^{kt}_{♀,j=7,l=1} = \begin{pmatrix} 1 - \widehat{f}^{kt}_{j=7,l=1} \\ \widehat{f}^{kt}_{j=7,l=1} \\ 0 \end{pmatrix} \quad ; \quad \mathbf{P}^{kt}_{♂,j=7,l=1} = \begin{pmatrix} \left(1 - \widehat{f}^{kt}_{j=7,l=1}\right)^2 \\ 2\widehat{f}^{kt}_{j=7,l=1}\left(1 - \widehat{f}^{kt}_{j=7,l=1}\right) \\ (\widehat{f}^{kt}_{j=7,l=1})^2 \end{pmatrix}$$

$$\widehat{f}^{kt}_{j=7,l=1} = \frac{2N^{kt}_{bb♂} + N^{kt}_{ab♂} + N^{kt}_{ab♀}}{2\left(N^{kt}_{aa♂} + N^{kt}_{ab♂} + N^{kt}_{bb♂}\right) + N^{kt}_{aa♀} + N^{kt}_{ab♀}}$$

(b) Z-polymorphism; allele 2 fixed on W. f is the frequency of allele 1 on Z :

$$\mathbf{P}^{kt}_{♀,j=7,l=2} = \begin{pmatrix} 0 \\ \widehat{f}^{kt}_{j=7,l=2} \\ 1 - \widehat{f}^{kt}_{j=7,l=2} \end{pmatrix} \quad ; \quad \mathbf{P}^{kt}_{♂,j=7,l=2} = \begin{pmatrix} ()\widehat{f}^{kt}_{j=7,l=2})^2 \\ 2\widehat{f}^{kt}_{j=7,l=2}\left(1 - \widehat{f}^{kt}_{j=7,l=2}\right) \\ \left(1 - \widehat{f}^{kt}_{j=7,l=2}\right)^2 \end{pmatrix}$$

$$\widehat{f}^{kt}_{j=7,l=2} = \frac{2N^{kt}_{aa♂} + N^{kt}_{ab♂} + N^{kt}_{ab♀}}{2\left(N^{kt}_{aa♂} + N^{kt}_{ab♂} + N^{kt}_{bb♂}\right) + N^{kt}_{bb♀} + N^{kt}_{ab♀}}$$

(c) W-polymorphism, allele 1 fixed on Z. f is the frequency of allele 2 on W:

$$\mathbf{P}^{kt}_{♀,j=7,l=3} = \begin{pmatrix} 1 - \widehat{f}^{kt}_{j=7,l=3} \\ \widehat{f}^{kt}_{j=7,l=3} \\ 0 \end{pmatrix} \quad ; \quad \mathbf{P}^{kt}_{♂,j=7,l=3} = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}$$

$$\widehat{f}^{kt}_{j=7,l=3} = \frac{N^{kt}_{ab♀}}{N^{kt}_{aa♀} + N^{kt}_{ab♀}}$$

bioRxiv preprint doi: https://doi.org/10.1101/2020.01.31.928291; this version posted July 22, 2020. 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-NC-ND 4.0 International license.

(d) W-polymorphism, allele 2 fixed on Z. f is the frequency of allele 1 on W:

$$\mathbf{P}^{kt}_{♀,j=7,l=4} = \begin{pmatrix} 0 \\ \widehat{f}^{kt}_{j=7,l=4} \\ 1-\widehat{f}^{kt}_{j=7,l=4} \end{pmatrix} \quad ; \quad \mathbf{P}^{kt}_{♂,j=7,l=4} = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}$$

$$\widehat{f}^{kt}_{j=7,l=4} = \frac{N^{kt}_{ab♀}}{N^{kt}_{bb♀} + N^{kt}_{ab♀}}$$

In some cases, calculation of $\widehat{f}^{kt}_{jl}$ might lead to division by 0. To avoid this problem, counts that are expected to be 0 under a segregation type are added to the numerator and the denominator.

Genotyping errors (whether they are due to sequencing errors, read mapping errors, or violations of the assumptions of the method for genotyping) cause the observed genotype g to be different from the true genotype g'. We define $q_{gg'} = \mathbb{P}\left(\mathrm{OG}^{kt}_{i_h g}|\mathrm{TG}^{kt}_{i_h g'}\right)$, *i.e.*, the probability to observe genotype g when the true genotype is g', and $\mathbf{Q}$ is the matrix of all $q_{gg'}$, such that

$$\mathbf{Q} = \begin{pmatrix} q_{1,1} & q_{1,2} & q_{1,3} \\ q_{2,1} & q_{2,2} & q_{2,3} \\ q_{3,1} & q_{3,2} & q_{3,3} \end{pmatrix}$$

We can now directly calculate the probabilities of the observed genotypes for each segregation type:

$$\begin{aligned} \mathbb{P}\left(\mathrm{OG}^{kt}_{i_h g}|S_j, A_l\right) &= \sum_{g'} \mathbb{P}\left(\mathrm{TG}^{kt}_{i_h g'}|S_j, A_l\right) \mathbb{P}\left(\mathrm{OG}^{kt}_{i_h g}|\mathrm{TG}^{kt}_{i_h g'}\right) \\ &= \sum_{g'} P^{kt}_{g'hjl} q_{gg'} \end{aligned}$$

We rename the quantity $\sum_{g'} P^{kt}_{g'hjl} q_{gg'}$ as $\widetilde{P}^{kt}_{ghjl}$; it is the expected frequency of the observed genotype given the segregation type and a certain genotyping error rate. For each sex, **OG** follows a multinomial distribution $\mathcal{M}\left(N^{kt}_h; \widetilde{P}^{kt}_{1hjl}, \widetilde{P}^{kt}_{2hjl}, \widetilde{P}^{k}_{3hjl}\right)$. Thus, the conditional likelihood of the data (given the segregation type) at each site, that we name M^{kt}_{jl}, is

$$M^{kt}_{jl} = \mathbb{P}\left(\mathbf{OG}^{kt}|S_j, A_l\right) = \prod_{gh} \left(\widetilde{P}^{kt}_{ghjl}\right)^{N^{kt}_{gh}} \tag{1}$$

Parameters

The error rates $q_{gg'}$ depend on one error parameter e. We assume all genotyping errors to occur with the same frequency, so $q_{g,g'\neq g} = e$ and $q_{g,g'=g} = 1-2e$, which gives the error matrix

$$\mathbf{Q} = \begin{pmatrix} 1-2e & e & e \\ e & 1-2e & e \\ e & e & 1-2e \end{pmatrix}$$

Two more series of parameters are required to model the data; these indicate the proportion of the genome that segregates under each type. There are a maximum of seven segregation types S_j, each occupying a proportion π_j of the genome, such that $\sum_j \pi_j = 1$. $\boldsymbol{\pi}$ is the vector containing all π_j. The segregation types S^{kt}_j are distributed multinomially, thus

$$\mathbf{S} \sim \mathcal{M}(1, \boldsymbol{\pi})$$

Several biologically relevant segregation types (S) have several "subtypes" (A), depending on the fixation of one of the alleles on either of the copies. Thus, for a segregation type S_j, there are L subtypes, and each subtype A_{jl} applies to a proportion α_{jl} of the proportion π_j of the genome (corresponding to the segregation type S_j). For each segregation type with subtypes, $\sum_{l=1}^{L} \alpha_{jl} = 1$, and

$$\mathbf{A}_j|S_j \sim \mathcal{M}(1, \boldsymbol{\alpha}_j)$$

For the paralogs, the subtype depends uniquely on the choice of what allele is called a, which is random. Thus, no parameter is needed, and

$$\boldsymbol{\alpha}_3 = \left(\frac{1}{2}, \frac{1}{2}\right)$$

bioRxiv preprint doi: https://doi.org/10.1101/2020.01.31.928291; this version posted July 22, 2020. 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-NC-ND 4.0 International license.

For the XY and ZW types, more sites can be polymorphic on one chromosome than on the other. The proportion of XY or ZW sites that are polymorphic on X or on Z is described by the parameter ρ_j, which takes a single value for each segregation type. The (random) choice what allele is called a affects both X (or Z) and Y (or W) polymorphisms, leading to four subtypes

$$\boldsymbol{\alpha}_5 = \left(\frac{\rho_5}{2}, \frac{\rho_5}{2}, \frac{1-\rho_5}{2}, \frac{1-\rho_5}{2}\right)$$

When aggregating the segregation subtypes (A) to biologically relevant types (S), we get

$$\mathbb{P}\left(\mathbf{OG}^{kt}|S_j\right) = B_j^{kt} = \sum_l \alpha_{jl} M_{jl}^{kt} \tag{2}$$

Expectation-Maximization algorithm

The full log-likelihood of the model is given by

$$\begin{aligned}\log \mathbb{P}\left(\mathbf{OG}, \mathbf{TG}, \mathbf{S}, \mathbf{A}\right) &= \log \mathbb{P}\left(\mathbf{OG}|\mathbf{TG}\right) \\ &+ \log \mathbb{P}\left(\mathbf{TG}|\mathbf{S}, \mathbf{A}\right) \\ &+ \log \mathbb{P}\left(\mathbf{A}|\mathbf{S}\right) \\ &+ \log \mathbb{P}\left(\mathbf{S}\right)\end{aligned}$$

This likelihood is maximized through an Expectation-Maximization (EM) algorithm.

E-step The posterior segregation types are given by

$$\mathbb{E}\left(\log \mathbb{P}\left(\mathbf{S}\right)|\mathbf{OG}\right) = \sum_{jkt} \mathbb{E}(S_j^{kt}|\mathbf{OG}^{kt}) \log \pi_j$$

with

$$\mathbb{E}(S_j^{kt}|\mathbf{OG}^{kt}) = \widehat{S}_j^{kt} = \frac{\pi_j B_j^{kt}}{\sum_{j'} \pi_{j'} B_{j'}^{kt}} \tag{3}$$

The posteriors for the subtypes are calculated by

$$\begin{aligned}\mathbb{E}\left(\log \mathbb{P}\left(\mathbf{A}^{kt}|\mathbf{S}^{k(t)}\right)|\mathbf{OG}^{kt}\right) &= \sum_{jlkt} \mathbb{E}\left(A_l^{kt} S_j^{k(t)}|\mathbf{OG}^{kt}\right) \log \alpha_{jl} \\ &= \sum_{jlkt} \mathbb{E}\left(A_l^{kt}|\mathbf{OG}^{kt}, S_j^{k(t)}\right) \mathbb{E}\left(S_j^{k(t)}|\mathbf{OG}^{k(t)}\right) \log \alpha_{jl} \\ &= \sum_{jlkt} \widehat{A}_{lj}^{kt} \widehat{S}_j^{k(t)} \log \alpha_{jl}\end{aligned}$$

$$\widehat{A}_{lj}^{kt} = \frac{\alpha_{jl} M_{jl}^{kt}}{\sum_{l'} \alpha_{jl'} M_{jl'}^{kt}}$$

For the true expected true genotypes, we calculate

$$\begin{aligned}\mathbb{E}\left(\log \mathbb{P}(\mathbf{TG}|\mathbf{S}, \mathbf{A})|\mathbf{OG}\right) &= \sum_{(kt)(jl)(i_h g')} \mathbb{E}\left(S_j^{kt} A_{jl}^{kt} \mathrm{TG}_{i_h g'}^{kt}|\mathbf{OG}_{i_h}^{kt}\right) \log P_{g'hjl}^{kt} \\ &= \sum_{(kt)(jl)(i_h g')} \mathbb{E}\left(\mathrm{TG}_{i_h g'}^{kt}|\mathbf{OG}_{ih}^{kt}, S_j^{kt}, A_{jl}^{kt}\right) \mathbb{E}\left(S_j^{kt}, A_{jl}^{kt}|\mathbf{OG}\right) \log P_{g'hjl}^{kt} \\ &= \sum_{(kt)(jl)(i_h g')} \mathbb{E}\left(\mathrm{TG}_{i_h g'}^{kt}|\mathbf{OG}_{ih}^{kt}, S_j^{kt}, A_{jl}^{kt}\right) \widehat{A}_{lj}^{kt} \widehat{S}_j^{kt} \log P_{g'hjl}^{kt}\end{aligned}$$

$$\widehat{\mathrm{TG}}_{i_h g' jl}^{kt} = \frac{P_{g'hjl}^{kt} \prod_g q_{gg'}^{\mathrm{OG}_{i_h g}^{kt}}}{\sum_{g''} P_{g''hjl}^{kt} \prod_g q_{gg''}^{\mathrm{OG}_{i_h g}^{kt}}}$$

bioRxiv preprint doi: https://doi.org/10.1101/2020.01.31.928291; this version posted July 22, 2020. 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-NC-ND 4.0 International license.

As individuals are defined uniquely by their sex and their observed genotype, $\widehat{\mathrm{TG}}^{kt}_{i_h g' jl}$ is the same for two individuals having the same sex and genotype. Thus, we write $\widehat{\mathrm{TG}'}^{kt}_{hgg'jl} = \frac{P^{kt}_{g'hjl} q_{gg'}}{\sum_{g''} P^{kt}_{g''hjl} q_{gg''}}$.

Finally, the conditional likelihood of the observed genotypes is given by

$$\begin{aligned}
\mathbb{E}\left(\log \mathbb{P}\left(\mathbf{OG}|\mathbf{TG}\right)|\mathbf{OG}\right) &= \sum_{(kt)(jl)(i_h g')g} \mathrm{OG}^{kt}_{i_h g} \mathbb{E}\left(\mathrm{TG}^{kt}_{i_h g'}|\mathbf{OG}^{kt}\right) \log q_{gg'} \\
&= \sum_{(kt)(jl)(i_h g')g} \mathrm{OG}^{kt}_{i_h g} \widehat{\mathrm{TG}}^{kt}_{i_h g' jl} \widehat{A}^{kt}_{lj} \widehat{S}^{kt}_{j} \log q_{gg'}
\end{aligned}$$

M-step The key quantity to be used in the M-step is the conditional expectation of the complete-data likelihood:

$$\begin{aligned}
\mathbb{E}\left(\log \mathbb{P}\left(\mathbf{OG},\mathbf{TG},\mathbf{S},\mathbf{A}\right)|\mathbf{OG}\right) &= \sum_{(kt)(jl)(i_h g')g} \mathrm{OG}^{kt}_{i_h g} \widehat{\mathrm{TG}}^{kt}_{i_h g' jl} \widehat{A}^{kt}_{lj} \widehat{S}^{k(t)}_{j} \log q_{gg'} \\
&+ \sum_{(kt)(jl)(i_h g')} \widehat{\mathrm{TG}}^{kt}_{i_h g' jl} \widehat{A}^{kt}_{lj} \widehat{S}^{k(t)}_{j} \log P^{kth}_{g'jl} \\
&+ \sum_{(kt)(jl)} \widehat{A}^{kt}_{lj} \widehat{S}^{k(t)}_{j} \log \alpha_{jl} + \sum_{(k(t))j} \widehat{S}^{k(t)}_{j} \log \pi_j \\
&= \sum_{(kt)(hg)} N^{kth}_{g} \left(\sum_{(jl)} \widehat{A}^{kt}_{lj} \widehat{S}^{k(t)}_{j} \left(\sum_{g'} \widehat{\mathrm{TG}'}^{kt}_{hgg'jl} \left(\log q_{gg'} + \log P^{kth}_{g'jl}\right) \right) \right) \\
&+ \sum_{(kt)(jl)} \widehat{A}^{kt}_{lj} \widehat{S}^{k(t)}_{j} \log \alpha_{jl} + \sum_{(k(t))j} \widehat{S}^{k(t)}_{j} \log \pi_j
\end{aligned}$$

Parameters to estimate are $\boldsymbol{\pi}, \boldsymbol{\alpha}$ and error rate e. These parameters only involve

$$\begin{aligned}
\mathbb{E}\left(\log \mathbb{P}\left(\mathbf{OG}|\mathbf{TG}\right)|\mathbf{OG}\right) &= \sum_{(kt)} \sum_{(jl)(i_h)} \mathrm{OG}^{kt}_{i_h g} \widehat{\mathrm{TG}}^{kt}_{i_h g' jl} \widehat{A}^{kt}_{lj} \widehat{S}^{k(t)}_{j} \log q_{gg'} \\
&= \sum_{(kt)} \sum_{(jl)(gh)} N^{kth}_{g} \widehat{\mathrm{TG}'}^{kt}_{hgg'jl} \widehat{A}^{kt}_{lj} \widehat{S}^{k(t)}_{j} \log q_{gg'}
\end{aligned}$$

To simplify notations, let us denote

$$\begin{aligned}
\widehat{U}_{gg'} &= \sum_{(kt)} \sum_{(jl)(i_h)} \mathrm{OG}^{kt}_{i_h g} \widehat{\mathrm{TG}}^{kt}_{i_h g' jl} \widehat{A}^{kt}_{lj} \widehat{S}^{k(t)}_{j} \\
&= \sum_{(kt)} \sum_{(jl)(h)} N^{kth}_{g} \widehat{\mathrm{TG}'}^{kt}_{hgg'jl} \widehat{A}^{kt}_{lj} \widehat{S}^{k(t)}_{j}
\end{aligned}$$

Thus,

$$\mathbb{E}\left(\log \mathbb{P}\left(\mathbf{OG}|\mathbf{TG}\right)|\mathbf{OG}\right) = \sum_{(gg')} \widehat{U}_{gg'} \log q_{gg'}$$

We find the new values of e by $\frac{\partial \mathbb{E}(\log \mathbb{P}(\mathbf{OG}|\mathbf{TG})|\mathbf{OG})}{\partial e} = 0$, which gives:

$$\begin{aligned}
\widehat{e} &= \frac{\widehat{U}_{ab} + \widehat{U}_{13} + \widehat{U}_{21} + \widehat{U}_{23} + \widehat{U}_{31} + \widehat{U}_{32}}{2(\widehat{U}_{ab} + \widehat{U}_{bb} + \widehat{U}_{32} + \widehat{U}_{ab} + \widehat{U}_{13} + \widehat{U}_{21} + \widehat{U}_{23} + \widehat{U}_{31} + \widehat{U}_{32})} \\
&= \frac{\sum_{(g \neq g')} \widehat{U}_{gg'}}{2 \sum_{(gg')} \widehat{U}_{gg'}}
\end{aligned}$$

$$\widehat{\alpha} = \frac{\sum_{kt} \widehat{S}^{kt}_{3} \left(\widehat{A}^{kt}_{13} + \widehat{A}^{kt}_{23} \right)}{\sum_{kt} \widehat{S}^{kt}_{3}}$$

bioRxiv preprint doi: https://doi.org/10.1101/2020.01.31.928291; this version posted July 22, 2020. 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-NC-ND 4.0 International license.

Monitoring and convergence The likelihood of the data in the model is

$$\begin{aligned}\log \mathbb{P}(\mathbf{OG}) &= \sum_{kt} \log \left(\sum_{jl} \left(\mathbb{P}\left(A_l^{kt}\right) \mathbb{P}\left(S_j^{kt}\right) \prod_i \mathbb{P}\left(\mathrm{OG}_{i_h g}^{kt} | S_j^{kt}, A_l^{kt}\right) \right) \right) \\ &= \sum_{kt} \log \left(\sum_{jl} \pi_j \alpha_l M_{jl}^{kt} \right)\end{aligned}$$

Convergence is evaluated as a function of the relative change in parameter value estimations. Optimization is halted when the largest relative change of all parameters has been less than 10^{-4} for 10 iterations, except for the error rate parameter, which is not considered for convergence.

There are $J-1$ free parameters for the segregation types, one parameter α for each of the XY and ZW types, and one parameter for the error rate. If the number of parameters is ν, we calculate the Bayesian Information Criterion (BIC) as follows:

$$\mathrm{BIC} = \log \mathbb{P}\left(\mathbf{OG}; \hat{\theta}\right) - \frac{1}{2} \log \left(\sum_{kt} 1 \right) \nu$$

The model with the lowest BIC has the best fit.

Site- and contig-wise probabilities The posterior probabilities per site, as given in Eq. [3], are calculated using the priors π_j. The smaller π_j, the higher the conditional likelihood B_j^{kt} should be to produce a high posterior probability. For the sex-linked segregation types π_j can easily be very small. If, say, 0.1% of the sites are inferred as gametologous and 99.9% as autosomal, the conditional likelihood for the gametologous segregation types should be 1000 times higher than the one for autosomal segregation to obtain comparable posterior likelihoods with this formula. In order to avoid excessive biases against rare segregation types, for inference purposes at the end of the optimization, we calculate the posterior probabilities without priors, which amount to using a uniform prior,. Thus, for the output, we compute

$$\widehat{S}_j^{kt} = \frac{B_j^{kt}}{\sum_{j'} B_{j'}^{kt}} \tag{4}$$

At the contig level, the goal is to estimate the posterior probability to be sex-linked, autosomal, or not informative (*i.e.*, haploid or paralogous), given the observed data for each of its sites and the optimal parameter values. This probability is the expectation of each segregation type, $\widehat{S}_j^k$. It can be calculated directly given the expectations per site, $\widehat{S}_j^{kt}$ from Eq. [3]:

$$\widehat{S}_j^k = \frac{1}{T_k} \sum_t \widehat{S}_j^{kt} = \frac{1}{T_k} \sum_t \frac{\pi_j B_j^{kt}}{\sum_{j'} \pi_{j'} B_{j'}^{kt}} \tag{5}$$

For robustness, geometric mean calculations of the posterior probabilities per contig are also calculated. The geometric mean of a series of observations x_i, $i \in 1, ..., n$ is $\left(\prod_i x_i\right)^{\frac{1}{n}} = e^{\frac{1}{n} \sum_i \log x_i}$. The geometric mean is more robust than the arithmetic mean since the relative value of the contig-level geometric means is independent of site-level normalizations, which is not true for arithmetic means. Furthermore, the ratio of the geometric means of two series is equal to the geometric mean of the ratios of the elements of the series. The posterior probability based on the geometric mean is

$$\widehat{S_{\mathrm{G}}}_j^k = \frac{\pi_j \mathrm{GM}\left(B_j^{kt}\right)}{\sum_{j'} \pi_{j'} \mathrm{GM}\left(B_{j'}^{kt}\right)} = \frac{\pi_j \exp\left(\frac{1}{T_k} \sum_t \log B_j^{kt}\right)}{\sum_{j'} \pi_{j'} \exp\left(\frac{1}{T_k} \sum_t \log B_{j'}^{kt}\right)} \tag{6}$$

Finally, we calculate the geometric mean of the site-wise conditional likelihoods without taking the proportion π_j for each segregation type as a prior. Indeed, when the sex-linked region is very small, containing few fixed differences, these proportions will be very low for the sex-linked segregation types, and the influence of the prior would be too strong. Thus, for the output, we calculate:

$$\widehat{S_{\mathrm{N}}}_j^k = \frac{\mathrm{GM}\left(B_j^{kt}\right)}{\sum_{j'} \mathrm{GM}\left(B_{j'}^{kt}\right)} = \frac{\exp\left(\frac{1}{T_k} \sum_t \log B_j^{kt}\right)}{\sum_{j'} \exp\left(\frac{1}{T_k} \sum_t \log B_{j'}^{kt}\right)} \tag{7}$$

We recommend that inferences of segregation types should be based on the posterior probabilities that were calculated without the priors, *i.e.* Eq. [4] for sites and Eq. [7] for contigs. Other posteriors are nevertheless provided for expert users.

bioRxiv preprint doi: https://doi.org/10.1101/2020.01.31.928291; this version posted July 22, 2020. 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-NC-ND 4.0 International license.

Population genetic predictions From the allele frequencies and segregation subtypes, it is possible to calculate the expected diversity and divergence of the gametologous copies. For each site, the frequency of allele a on chromosome $v \in \{W, X, Y, Z\}$ is

$$\widehat{f}_v^{\,kt} = \widehat{A}^{kt}_{j,l=1}\left(1 - \widehat{f}^{kt}_{j,l=1}\right) + \widehat{A}^{kt}_{j,l=2}\widehat{f}^{kt}_{j,l=2} + \widehat{A}^{kt}_{j,l=3} \quad \text{for} \quad v \in \{X, Z\},$$
$$\widehat{f}_v^{\,kt} = \widehat{A}^{kt}_{j,l=1} + \widehat{A}^{kt}_{j,l=3}\left(1 - \widehat{f}^{kt}_{j,l=3}\right) + \widehat{A}^{kt}_{j,l=4}\widehat{f}^{kt}_{j,l=4} \quad \text{for} \quad v \in \{W, Y\}.$$

A different way of predicting the allele frequency on both sex chromosomes is to assign it to be the frequency corresponding to the most probable subtype.

This information can be used to infer the consensus sequences of the X and Y sequences. For a given contig (that can be chosen on the basis of $\widehat{S}^k_j$, but not necessarily if we have other reasons to believe the contig is sex-linked), each polymorphic site can be considered fixed for an allele if $\widehat{f}_X$ or $\widehat{f}_Y$ are above a threshold U_f ($0.5 \leq U_f \leq 1$) or below $1 - U_f$. A further threshold can be applied to genotype non-fixed sites: if $\widehat{f}_X$ or $\widehat{f}_Y$ are above a threshold u_f ($0.5 \leq u_f \leq U_f$) or below $1 - u_f$.

Nucleotide diversity can be calculated as

$$\pi^k_v = \frac{1}{\tau_k}\sum_t 2\widehat{f}_v^{\,kt}\left(1 - \widehat{f}_v^{\,kt}\right)$$

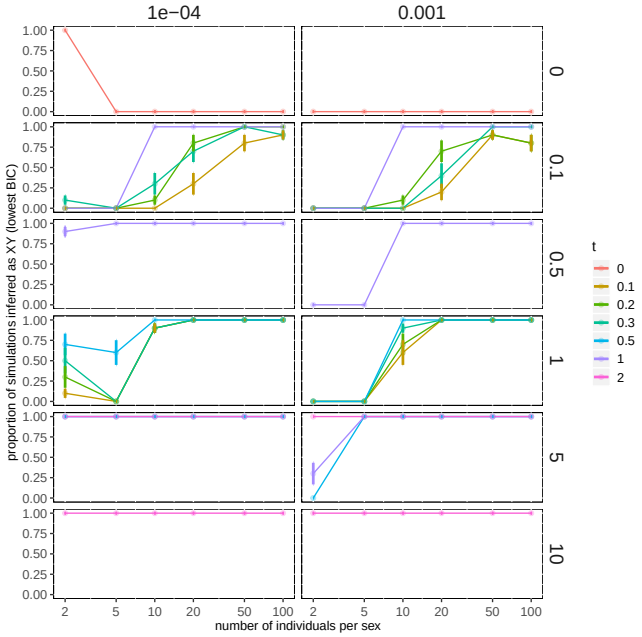
where $\tau_k \geq T_k$ is the total length of the contig k, including monoallelic sites. The divergence is

$$D^k_{XY} = \frac{1}{\tau_k}\sum_t \left(\widehat{f}^{kt}_X\left(1 - \widehat{f}^{kt}_Y\right) + \widehat{f}^{kt}_Y\left(1 - \widehat{f}^{kt}_X\right)\right)$$

in the XY case; extension to ZW chromosomes is trivial.

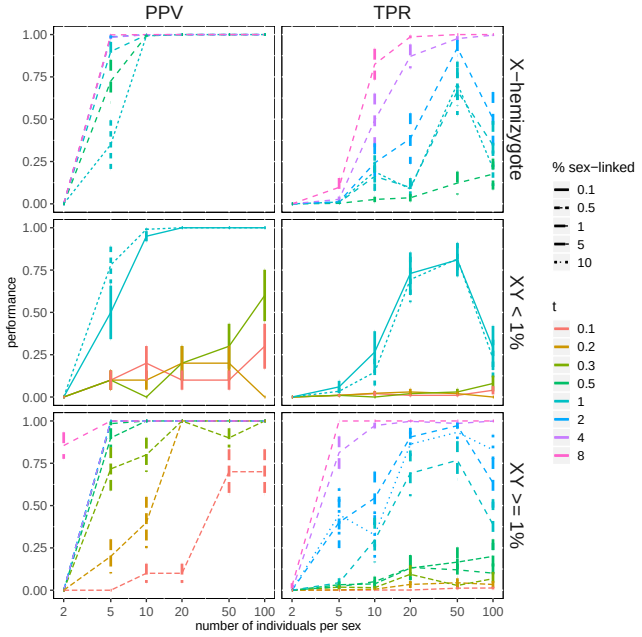
bioRxiv preprint doi: https://doi.org/10.1101/2020.01.31.928291; this version posted July 22, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Supplementary figures

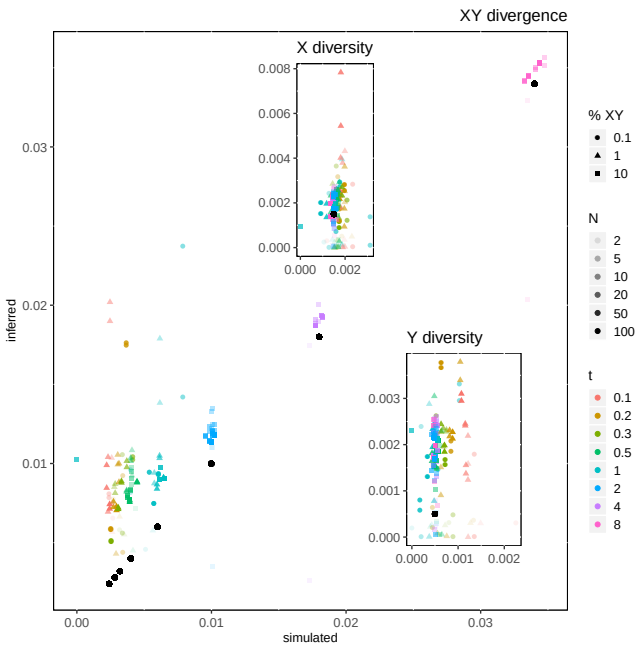


Supplementary Figure 1: Model choice by SDpop on simulated data. The proportion of simulations for which the XY model had the lowest BIC is indicated; each combination of simulation parameter values was repeated 10 times from a random seed. Left column: simulated error rate $e = 0.0001$, right column: $e = 0.001$; the percentage sex-linked sequences varies from none (top row) to 10% (bottom).

bioRxiv preprint doi: https://doi.org/10.1101/2020.01.31.928291; this version posted July 22, 2020. 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-NC-ND 4.0 International license.

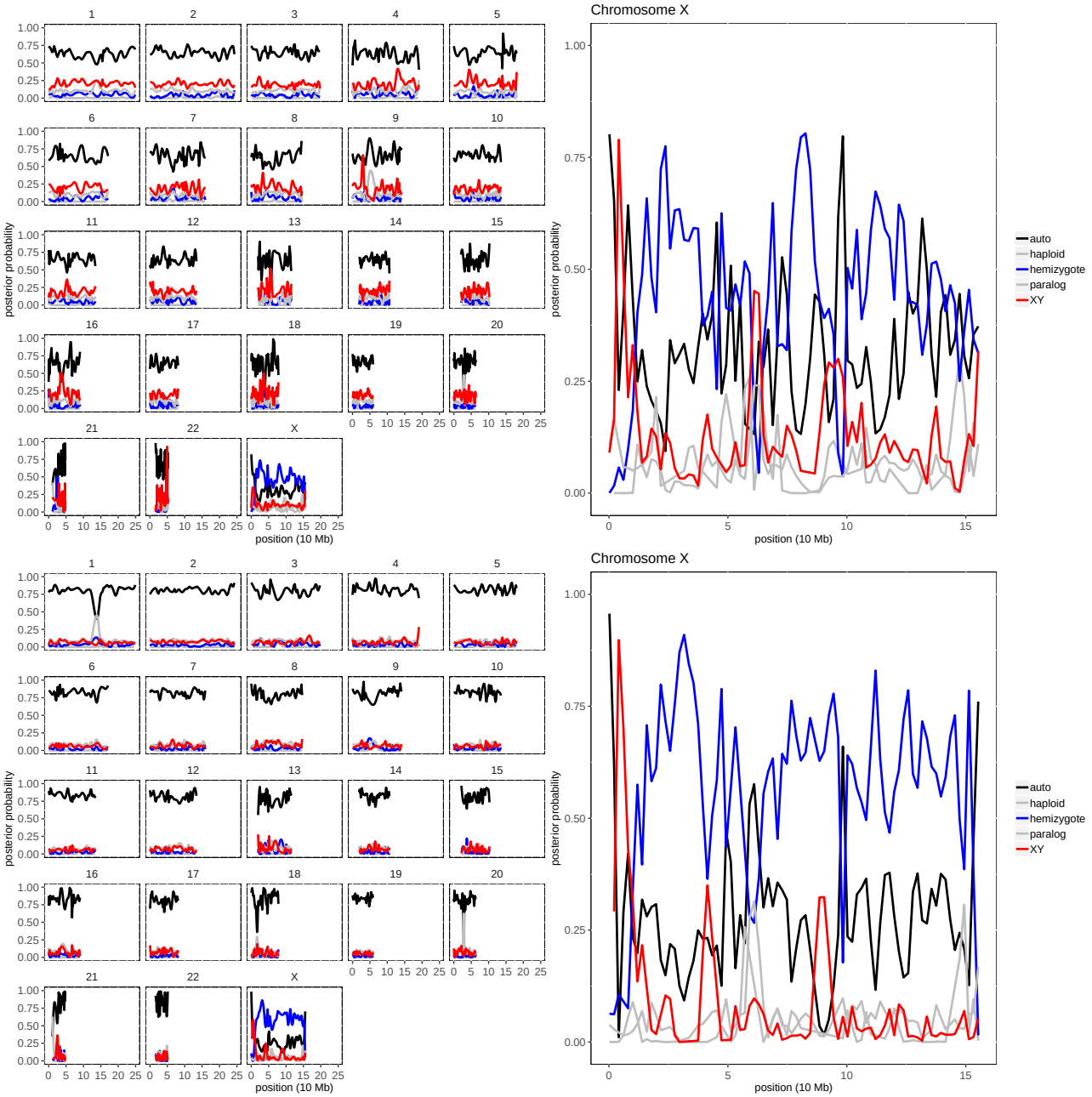


Supplementary Figure 2: Precision (Positive Predictive Value, left) and Power (True Positive Rate, right) of the detection of sex-linked contigs in simulated data, using a threshold for the posterior probability of 0.8.. Top graphs: X-hemizygous genes. Middle graphs: XY gametologs, simulations with a small fraction of gametologs in the genome ($< 1\%$); bottom graphs: XY gametologs, simulations with a larger fraction of gametologs in the genome ($\geq 1\%$). Each point is the average of 100 simulations, with the bars representing the standard error. For all cases shown here, the simulated error rate was 0.001; for $e = 0.0001$, see Fig. 2.



Supplementary Figure 3: Population genetic inferences of SDpop at higher error rates than Fig. 3; here, $e = 0.001$. The values of nucleotide diversity and divergence calculated directly from the simulation results are compared to the values inferred from SDpop's output. Comparisons are based on SDpop's assignment of the genes (i.e. all genes with a posterior probability > 0.8 were used). Main graph: gametolog divergence (D_{XY}); left inset: X nucleotide diversity (π_X); right inset: Y nucleotide diversity (π_Y). N is the number of individuals per sex, t the time since recombination suppression, and %XY the simulated proportion of gametologs. The black points indicate the theoretical values (D_{XY}: one for each t; π_X and π_Y: one value for all runs).

bioRxiv preprint doi: https://doi.org/10.1101/2020.01.31.928291; this version posted July 22, 2020. 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-NC-ND 4.0 International license.



Supplementary Figure 4: Test of SDpop's performance on the human exome-targeted sequencing data from the 1000 genome project, using 5 individuals per sex (top graphs) or 20 (bottom graphs). Smoothed gene-level posterior probabilities for autosomal (black), X-hemizygote (blue) and XY (red) segregation are shown; haploid and paralogous posterior probabilities are indicated in gray. The right panels show the results on the X chromosome: the extremities corresponding to the pseudo-autosomal regions are predicted to be autosomal by SDpop, while most XY gametologous genes are found close to the pseudo-autosomal region on the left arm, which represents the youngest stratum where Y copies have not yet been lost. The rest of the chromosome consists of X-hemizygous genes, that lack a Y copy.