Easy-to-use *R* functions to separate reduced-representation genomic datasets into sex-linked and autosomal loci, and conduct sex-assignment

Diana Robledo-Ruiz[1], Lana Austin[2], J Amos[2], Jesús Castrejón-Figureoa[2], Michael Magrath[3], Paul Sunnucks[1], and Alexandra Pavlova[1]

[1]Monash University
[2]Monash University School of Biological Sciences
[3]Zoos Victoria

December 2, 2022

Abstract

Identifying sex-linked markers in genomic datasets is important, because their analyses can reveal sex-specific biology, and their presence in supposedly neutral autosomal datasets can result in incorrect estimates of genetic diversity, population structure and parentage. But detecting sex-linked loci can be challenging, and available scripts neglect some categories of sex-linked variation. Here, we present new R functions to (1) identify and separate sex-linked loci in ZW and XY sex determination systems and (2) infer the genetic sex of individuals based on these loci. Two additional functions are presented, to (3) remove loci with artefactually high heterozygosity, and (4) produce input files for parentage analysis. We test these functions on genomic data for two sexually-monomorphic bird species, including one with a neo-sex chromosome system, by comparing biological inferences made before and after removing sex-linked loci using our function. We found that standard filters, such as low read depth and call rate, failed to remove up to 28.7% of sex-linked loci. This led to (i) overestimation of population FIS by [?] 9%, and the number of private alleles by [?] 8%; (ii) wrongly inferring significant sex-differences in heterozygosity, (iii) obscuring genetic population structure, and (iv) inferring ~11% fewer correct parentages. We discuss how failure to remove sex-linked markers can lead to incorrect biological inferences (e.g., sex-biased dispersal and cryptic population structure) and misleading management recommendations. For reduced-representation datasets with at least 15 known-sex individuals of each sex, our functions offer convenient, easy-to-use resources to avoid this, and to sex the remaining individuals.

Easy-to-use *R* functions to separate reduced-representation genomic datasets into sex-linked and autosomal loci, and conduct sex-assignment

Authors: Diana A. Robledo-Ruiz[1+], Lana Austin[1], J. Nevil Amos[1,2], Jesús Castrejón-Figueroa[1], Michael J. L. Magrath[3,4], Paul Sunnucks[1], Alexandra Pavlova[1]

[1]School of Biological Sciences, Monash University, Clayton, Vic. 3800, Australia

[2]Arthur Rylah Institute for Environmental Research, Department of Environment, Land, Water and Planning, Heidelberg, Vic. 3084, Australia

[3]Department of Wildlife Conservation and Science, Zoos Victoria, Parkville, Vic. 3052, Australia

[4]School of BioSciences, University of Melbourne, Parkville, Vic. 3010, Australia

[+]Correspondence author: diana.robledoruiz1@monash.edu

ABSTRACT

Identifying sex-linked markers in genomic datasets is important, because their analyses can reveal sex-specific biology, and their presence in supposedly neutral autosomal datasets can result in incorrect estimates of genetic diversity, population structure and parentage. But detecting sex-linked loci can be challenging, and available scripts neglect some categories of sex-linked variation. Here, we present new *R*functions to (1) identify and separate sex-linked loci in ZW and XY sex determination systems and (2) infer the genetic sex of individuals based on these loci. Two additional functions are presented, to (3) remove loci with artefactually high heterozygosity, and (4) produce input files for parentage analysis. We test these functions on genomic data for two sexually-monomorphic bird species, including one with a neo-sex chromosome system, by comparing biological inferences made before and after removing sex-linked loci using our function. We found that standard filters, such as low read depth and call rate, failed to remove up to 28.7% of sex-linked loci. This led to (i) overestimation of population F_{IS} by [?] 9%, and the number of private alleles by [?] 8%; (ii) wrongly inferring significant sex-differences in heterozygosity, (iii) obscuring genetic population structure, and (iv) inferring ~11% fewer correct parentages. We discuss how failure to remove sex-linked markers can lead to incorrect biological inferences (e.g., sex-biased dispersal and cryptic population structure) and misleading management recommendations. For reduced-representation datasets with at least 15 known-sex individuals of each sex, our functions offer convenient, easy-to-use resources to avoid this, and to sex the remaining individuals.

INTRODUCTION

Population genetic datasets are a rich source of information for wildlife managers (Hoffmann et al. 2015; Hohenlohe et al. 2021). They provide data on genetic structure, adaptation and evolutionary trajectories of species and populations (e.g., local adaptation, hybridization, population dynamics, evolutionary potential; Willi et al. 2021). They can reveal biological and ecological processes that could not otherwise be studied (e.g., mating systems, sex-specific dispersal and gene flow; Ellegren 2014; Amos et al. 2014). In addition, they help to identify genetic problems in small populations—notably loss of genetic diversity, inbreeding, inbreeding depression—and develop simple and cost-effective management solutions towards their conservation (e.g., genetic augmentation, genetic rescue; Frankham et al. 2017; Harrisson et al. 2019; Kardos 2021).

With the massive amount of genomic data that can be generated, the level of expertise in bioinformatics required for analysing genomic datasets has increased (MacMahon et al. 2014; Holderegger et al. 2019; Hohenlohe et al. 2020). Conservation geneticists spend a great part of their time learning the use of new software, which reduces their availability to engage in other important activities needed to bridge the gap between research and conservation practice (e.g., facilitating communication with wildlife managers, building relationships with primary industry, informing and shaping policy; Galla et al. 2016; Taylor et al. 2017; Britt et al. 2018). Accordingly, there is much interest in creating easy-to-use resources to automate and streamline dataset filtering and genomic analyses. This has included the development of packages for*R,* which tends to be a more welcoming environment for biologists than does command-line software (e.g., *dartR* : Gruber et al. 2018, Mijangos et al. 2022; *SambaR* : de Jong et al. 2021; *snpR* : Hemstrom & Jones 2022; SNPfiltR: DeRaad 2022; Hogg et al. 2022).

Most population genetic analyses assume autosomal loci; thus best-practice filtering includes removal of sex-linked loci from SNP datasets. If sex-linked loci are not removed, estimates of population genetic diversity such as heterozygosity, Wright's fixation indices including F_{IS}, polymorphism, and allelic richness may be biased depending on the sex ratio of the sample and the sex-chromosome-to-autosome diversity ratio (Ellegren 2009; Allendorf et al. 2022; Frankham et al. 2017). Assessment of population genetic structure also benefits from the removal of sex-linked loci because they can mask genetic structure that is due to evolutionary processes (e.g., gene flow, natural selection, genetic drift; Pritchard et el. 2000; Radosavljević et al. 2015; Benestan et al. 2016). Similarly, parentage analyses assume autosomal Mendelian inheritance and so their accuracy can be affected by the presence of sex-linked loci because they create apparent genetic mismatches between true parent-offspring pairs (Jones & Wang 2010). On the other hand, focusing on sex-linked markers can help assign sex to individuals of sexually-monomorphic species, as well as reveal interesting patterns of sex-specific ecology and evolution (e.g., natural selection, philopatry; Castella et al. 2001; Pavlova et al.

2013; Arnold & Wilkinson 2015). Thus, correct identification of sex-linked loci is important for making appropriate management recommendations.

In animal species, the two most common chromosomal sex-determination systems are XY and ZW. In an XY system, typical for mammals and some insects, males are the heterogametic sex with one X and one Y chromosome, and females are the homogametic sex with two X chromosomes. In contrast, in the ZW system, typical for birds, and some reptiles and insects, females are heterogametic (ZW) and males homogametic (ZZ) (Beaukeboom & Perrin 2014). The SNP markers on sex chromosomes can be classified into three types with different inheritance and characteristics (Figure 1):

1. Those present only on the W or Y chromosome (hereafter 'W-linked/Y-linked'; Figure 1 in yellow). In SNP datasets, such markers are called only in the heterogametic sex and are missing in the homogametic sex.
2. Those present only on the Z or X chromosome (hereafter 'Z-linked/X-linked'; Figure 1 in orange). In SNP datasets, the heterogametic sex possesses only one allele (i.e., they are*hemizygous*) and individuals appear homozygous when genotyped. The homogametic sex, which possesses two alleles, can be heterozygous or homozygous as for an autosomal locus.
3. Those present in homologous regions of both sex chromosomes, Z*and* W or X *and* Y, and similar enough to be considered alleles of the same locus (hereafter 'gametologs', Figure 1 in green). In some cases, gametologous loci have one allele that is found exclusively on one sex chromosome while the other allele appears exclusively on the other. As a result, all members of the heterogametic sex appear heterozygous, and the homogametic sex homozygous. These loci are known as 'fixed' gametologs and are typical of old sex chromosomes. In other cases (i.e., in recently evolved [neo-] sex chromosomes), the 'Z-allele' (or 'X-allele') is still found on some versions of the W (or Y) chromosome, and thus, some individuals of the heterogametic sex are homozygous. In these cases, the gametologs are 'non-fixed'.

The simplest way to distinguish sex-linked loci from autosomal ones is to identify those found in reads that mapped to the sex chromosomes of the reference genome. However, this is not possible when (i) a reference genome is not available–as is the case for most wildlife species–and*de novo* genotyping is required, (ii) there is little conserved synteny between the studied genome and the reference, or (iii) the W/Y chromosome of the reference genome is fragmented into numerous unmapped scaffolds, as is common in many genome projects (Carvalho & Clark 2013).

Some methods to identify sex-linked SNPs have been developed. MendelChecker, for example, uses the deviation from Mendelian inheritance to calculate the probability that a specific SNP is sex-linked, with the disadvantage that it requires genotype probabilities and pedigree information for analysis (Chen et al. 2014). Other methods use a set of individuals of known sex to test whether the allele frequencies of a given locus differ between the sexes. For instance, RADSEX is a command-line software that uses identical raw reads as non-polymorphic markers and uses their presence or absence in males and females to identify those significantly associated with sex (Feron et al. 2021). Some other studies have identified sex-linked markers by testing for differentiation between the sexes usingF_{ST}, but this approach can be used only for Z-linked/X-linked and gametologous loci (Benestan et al. 2017; Drinan et al. 2018; Trenkel et al. 2020). Function gl.report.sexlinked from*dartR* package (v2; Mijangos et al. 2022) uses arbitrary heterozygosity thresholds as default parameters to identify fixed gametologs, and can be used to identify non-fixed gametologs and Z-linked/X-linked loci by fine-tuning parameters (Pavlova et al. 2022). Nevertheless, this approach has the disadvantages that there are no clear instructions on how to tune parameters, the user has to manually adjust thresholds on a trial-and-error basis for each genomic dataset, and its precision declines with heterozygosity, risking either the erroneous removal of autosomal loci with rare alleles or the failure to remove sex-linked loci with low heterozygosity. Overall, these methods could be improved upon by developing an intuitive statistical approach that systematically identifies and distinguishes among types of SNPs (autosomal, W-linked/Y-linked, Z-linked/X-linked, and gametologs) that is automated in a ready-to-use *R* function with little user intervention needed.

In the same way that it is possible to use a set of known-sex individuals to identify sex-linked loci, the opposite

is also possible: use a set of known sex-linked loci to identify the sex of an individual. Sex-assignment is usually done utilizing a handful of sex-linked loci of only one type (Trenkel et al. 2020). For example, if using non-fixed ZW-gametologs (for which heterozygous individuals are never male), an individual is declared female if it is heterozygous for at least one locus, yet by chance, depending on allelic frequencies and the number of evaluated gametologs, some females may not be heterozygous for any of the loci. Similarly, depending on genotyping error rates, some males may appear heterozygous for some loci. To our knowledge, despite the rich information that the three types of sex-linked loci contain to improve sex-assignment, the comparison of their information is rarely, if ever, done. Thus, the sexing of individuals using large SNP datasets can benefit from a methodical procedure that uses the information from all available sex-linked loci and that can be integrated as a standard step in bioinformatics pipelines.

Another best-practice during filtering autosomal and sex-linked datasets is minimizing the presence of 'multilocus' SNPs (also known as multilocus contigs, multicopy loci or homeologs; Hohenlohe et al. 2011; Willis et al. 2017; O'Leary et al. 2018). These artefactual SNPs arise during bioinformatic processing of raw reads and are product of erroneously fusing two physically separate loci that are very similar, either because they are paralogs, repetitive elements or otherwise very much alike. Because a multilocus SNP is actually two loci, multilocus SNPs tend to present abnormally high read depths. This characteristic allows their removal by setting a maximum read depth threshold during filtering (usually twice the mode or mean; Willis et al. 2017). In some cases, there are fixed or near-fixed differences between the artificially fused loci, which makes multilocus SNPs exhibit heterozygosity well-above the expectation of 0.5 for biallelic markers at Hardy-Weinberg proportions. As a consequence, these SNPs can inflate estimates of heterozygosity (O'Leary et al. 2018). A common practice to identify these artefactual loci using heterozygosity is to set an arbitrary maximum threshold (e.g., heterozygosity [?] 0.6). It has been found that using more than one approach to identify multilocus SNPs–and removing those that are flagged by any method—constitutes the best strategy (Willis et al. 2017).

Lastly, parentage analyses and sibship reconstruction using molecular markers have great relevance in wildlife conservation. Resolving unknown parent-offspring relationships gives insights into the behaviour, ecology and evolution of plant and animal populations (e.g., extra-pair mating, inbreeding avoidance, dispersal, natural selection, effective population size; Flanagan & Jones 2018). Their application extends into very practical instances such as monitoring the success of translocations and genetic rescue, and spotting illegal trade of wild individuals (Fitzpatrick et al. 2020; Van Rossum 2022; Mucci et al. 2020). Moreover, captive breeding programs also benefit from parentage analyses that allow them to estimate founder relationships (typically assumed unrelated), and validate pedigrees and correct errors (Moran et al. 2021; Overbeek et al. 2020; Galla et al. 2021). Among the variety of parentage analysis software in existence, one of the most popular is COLONY, which simultaneously infers sibship and parentage, and can handle thousands of SNPs (Jones & Wang 2010). However, handling large amounts of genetic data in order to format it into the specific input file for COLONY requires some degree of bioinformatics expertise (Flanagan & Jones 2018). Often, researchers need to create different input files because several runs are usually required to maximize detection of true relationships. This can be a time-consuming task worth automating.

In this study, we aim to create four *R* functions to assist researchers analyzing reduced-representation genomic datasets. The study consists of two parts. First, we describe four *R* functions that we designed to automate common tasks in conservation genomic studies: (1) identify and remove sex-linked loci (function*filter.sex.linked*), (2) use sex-linked loci to identify the genetic sex of individuals (function *infer.sex*), (3) filter out excessively heterozygous loci that are likely to be genotyping errors (function *filter.excess.het*), and (4) create input files for parentage analyses in COLONY (function *gl2colony*). Second, we use the four new functions to process genetic datasets for two bird species and show how incomplete removal of sex-linked loci affects downstream analyses of (i) population genetic diversity, (ii) individual heterozygosity, (iii) population genetic structure, and (iv) parentage.

METHODS

1. Design of functions

The following four *R* functions were designed for SNP datasets, such as those produced by reduced-representation technologies (DArT, RAD or ddRAD; Kilian et al. 2012; Baird et al. 2008; Davey & Blaxter 2010; Peterson et al. 2012). The functions require the data to be imported to *R* as a genlight object (*adegenet* ; Jombart & Ahmed 2011) so that individual genotypes for each locus are scored as '0' (homozygous reference), '1' (heterozygous), and '2' (homozygous alternate). The functions make use of the information stored in genlight object's 'ind.metrics' and are available at github.com/drobledoruiz/conservation_genomics.

1.1 Function *filter.sex.linked*

Purpose: Detecting and filtering out sex-linked loci.

Input: One genlight object with at least 30 individuals of known sex (15 of each sex; see Results section 3), and a user-specified parameter declaring the sex-determination system of the species ('zw' or 'xy'). Known sex is provided in 'ind.metrics' with a column named 'sex' and individuals assigned 'F' (females) or 'M' (males). Individuals with unknown sex (i.e., assigned anything other than 'F' or 'M') are ignored by the function.

How it works: The rationale behind this function is that the scoring rate and heterozygosity of autosomal loci should not differ between the sexes, but they do differ for sex-linked loci. Based on this, the function works in two phases:

Phase I. Use locus call rate to identify W-linked/Y-linked loci and other loci with sex-biased call rates. The function counts, for each locus, the number of known females and the number of known males with NA (i.e., missing data) and with a called genotype (i.e., '0', '1' or '2'). These four counts are used to build a 2×2 contingency table per locus on which a Fisher's exact test is performed in order to test for the independence of call rate and sex ($\alpha = 0.05$). The logic is that autosomal loci should present roughly the same call rate for males and females (Figure 2a, diagonal cloud in gray), and therefore, a locus in which one sex has significantly more missing data than the other is likely to be sex-linked. The p-values of all loci are adjusted for False Discovery Rate with *R* function p.adjust (Benjamini & Hochberg, 1995). Of the loci with adjusted $p < 0.05$, those whose male call rate is [?] 0.1 are assigned as W-linked (because males lack a W chromosome; Figure 2a, in yellow), or as Y-linked if female call rate is [?] 0.1 (because females lack a Y chromosome). Remaining loci with adjusted $p < 0.05$ are identified as 'sex-biased' (Figure 2a, in blue).*Phase II.* Use locus heterozygosity to identify Z-linked/X-linked loci and gametologs. The function counts, for each locus, the number of known females and the number of known males that are heterozygous (i.e., '1'), and homozygous (i.e., '0' or '2'). In the same way as for*Phase I*, these four counts are used to build a 2×2 contingency table per locus and to perform a Fisher's exact test to test for the independence of heterozygosity and sex ($\alpha = 0.05$). Under the logic that autosomal loci should present no difference in proportion of heterozygous individuals between sexes (Figure 2b, diagonal cloud in dark gray), a locus in which one sex has significantly more heterozygous individuals than the other is likely to be sex-linked. P-values are adjusted for False Discovery Rate with *R* function p.adjust (Benjamini & Hochberg, 1995). Of the loci with adjusted $p < 0.05$, those whose proportion of heterozygous males is greater than the proportion of heterozygous females are identified as Z-linked (because females have only one Z chromosome, and should be mainly scored as homozygous; Figure 2b, in orange). On the other hand, loci whose proportion of heterozygous females is larger than the proportion of heterozygous males are identified as gametologs (because males have two Z chromosomes, and thus should present only the Z-associated allele and be scored as homozygous; Figure 2b, in green). The same logic, with reversed expectations for sexes, is applied to XY-sex determination system (X-linked: proportion of heterozygous females > proportion of heterozygous males; gametologs: proportion of heterozygous males > proportion of heterozygous females).

The loci that are not identified as belonging to any category of sex-linkage are inferred autosomal. The function finishes by splitting each category of loci into its own genlight object.

Output:

A list containing six elements: one table–with per-locus counts, Fisher's exact test estimates, p-values and

true/false columns for each type of sex-linked loci–and five genlight objects: one with autosomal loci, and one with each type of sex-linked loci.

Two sets of 'before' and 'after' plots: one set with female call rate plotted against male call rate with each data point representing one locus (one plot before and one plot after removing sex-linked loci identified by call rate). The other set with proportion of heterozygous females plotted against proportion of heterozygous males, and each point representing one locus (one plot before and one after removing sex-linked loci identified by heterozygosity).

Recommended use: In order to minimize the number of loci analysed by the function to speed computation time, it is advantageous to use the *filter.sex.linked* function after removing of secondary loci (i.e., those in the same sequenced fragment). This may not be needed when computation time is not a concern or the number of loci is small (i.e., ~50,000 SNPs), which will help identify sex-linked markers in species with short or little-differentiated sex chromosomes. Additionally, it is strongly recommended to use this function before other quality filters in order to ensure that (i) variation in call rate has not been truncated, and (ii) downstream filtering is done on autosomal loci only. When known-sex individuals are scarce, we recommend using 15 known-sex individuals of each sex to identify as many sex-linked loci as possible (even if few), then use those sex-linked loci to sex all individuals with function*infer.sex* , and then use the new sex assignments to identify the remaining sex-linked loci with function *filter.sex.linked* (see Results section 3).

1.2 Function *infer.sex*

Purpose: Identify the genetic sex of individuals.

Input: The output of function *sex.linked.filter* (list of six elements), a user-specified parameter that declares the sex-determination system of the species ('zw' or 'xy'), and a seed number.

How it works: This function uses the types of loci available in the input (W-linked/Y-linked, Z-linked/X-linked and gametologous loci) to assign one preliminary sex for each type of sex-linked loci:

W-linked/Y-linked loci. For a ZW-system, it preliminarily assigns 'M' (male) to an individual if it presents more loci with NA (i.e., missing data) than loci with called genotype (i.e., '0', '1' or '2'), and 'F' (female) otherwise. For a XY-system, the assignment is the opposite.*Z-linked/X-linked loci.* It uses the matrix of genotypes for all individuals to perform k-means clustering with two centers (using the provided seed number). The rationale is that individuals would form two distinctive clusters, one per sex. As a result, individuals are assigned to one of two sex clusters. The individual with the most loci scored as heterozygous is used to identify the sex of its cluster ('M' for ZW-system, and 'F' for XY-system), while the other cluster is identified as the opposite sex.*Gametologs.* It follows the same method as Z-linked/X-linked loci: performs k-means clustering in which individuals are assigned to one of two sex clusters. It also uses the individual with the most loci scored as heterozygous to identify the sex of its cluster ('F' for ZW-system, and 'M' for XY-system).

If a type of sex-linked locus was not available (e.g., zero gametologs), it assigns NA to that preliminary assignment. The function uses the preliminary assignments to output a final sex assignment: 'F' or 'M' if all preliminary assignments match, '*F' or '*M' if they do not.

Output: a table with the three preliminary, and final sex assignments per individual. The Table 1lso includes the raw data on which the preliminary assignments were based on: number of W-linked/Y-linked loci with missing/called genotype, number of Z-linked/X-linked loci scored as homozygous/heterozygous, and number of gametologs scored as homozygous/heterozygous

Recommended use: We created this function with the explicit intent that a person inspects the final sex assignments for which not all three preliminary assignments agree (denoted as '*M' or '*F'). Some individuals may have ambiguous genotypes for one type of sex-linked loci, and given the nature of k-means clustering, they may be assigned the wrong preliminary sex. It is recommended that the user checks the output table to

make a definite final assignment. We recommend it being used straight after using function *filter.sex.linked* .

1.3 Function *filter.excess.het*

Purpose: Remove loci with excessively high heterozygosity that are suspected to be bioinformatic artefacts (i.e., multilocus SNPs).

Input: A genlight object in which 'ind.metrics' contains a column named 'pop' and each individual is assigned to one population.

How it works: This function considers a locus to be 'excessively heterozygous' if its heterozygosity $> 0.5 and$ it significantly exceeds 0.5 assuming Hardy-Weinberg (HW) proportions. The rationale is that applying an absolute heterozygosity cut-off (e.g., 0.5 or 0.6) may remove some loci that conform to HW proportions but exceed the threshold due to sampling error. The function starts by dividing the genlight object by population, and identifying loci whose heterozygosity > 0.5. It then performs a χ^2 test to detect heterozygote excess significantly beyond that from sampling variance assuming HW proportions in a given population ($\alpha = 0.05$), and adjusts the p-values for False Discovery Rate with *R* function p.adjust (Benjamini & Hochberg, 1995). Loci whose adjusted p-values [?] 0.5 in any population are considered excessively heterozygous and are removed from the input genlight object.

Output:

1. A table with information on each excessively heterozygous locus, including its number of observed genotypes, number of expected genotypes, χ^2 statistic, and p-values.
2. A genlight object without excessively heterozygous loci.
3. A vector with the names of the removed loci (i.e., excessively heterozygous ones).
4. Two plots: one 'before' plot with the heterozygosity of the loci present in the *input* genlight, and one 'after' plot with the heterozygosity of the loci present in the *output* genlight (i.e., without excessively heterozygous loci).

Recommended use: We recommend caution when using this function, because it has the potential to remove loci that reflect population processes. For example, some loci may exhibit excessive heterozygosity due to (i) recent admixture between previously isolated populations (i.e., Wahlund-breaking), (ii) inbreeding avoidance, (iii) balancing selection, such as heterozygous advantage. Therefore, the use of this filter is best suited when there is previous understanding of the system, and for studies assuming neutral loci.

1.4 Function *gl2colony*

Purpose: Automate the creation of a COLONY input file from a genlight object.

Input: A genlight object in which 'ind.metrics' contains three columns named 'offspring', 'mother' and 'father' taking values 'yes' or 'no' to indicate if an individual should be considered a candidate offspring, mother and/or father. The desired name of the exported file.

Output: a ready-to-analyse COLONY file with the specified name.

Recommended use: We recommend using this function after all filtering is finished.

2. Use of the functions on biological datasets

We applied the designed functions (available at github.com/drobledoruiz/conservation genomics) on the *de novo* -scored DArT SNP datasets of two species: eastern yellow robin (EYR, *Eopsaltria australis*) and yellow-tufted honeyeater (YTH, *Lichenostomus melanops*). In order to assess the impact of removing sex-linked loci, we compared the biological inferences drawn from analyses of population genetic diversity, individual heterozygosity, population structure and parentage analyses before and after using the function *filter.sex.linked*(hereafter referred to as 'before' and 'after'). For that, we applied two SNP filtering regimes to the EYR and YTH datasets: a) 'Standard', which included only standard filtering steps, and b) 'Removing

sex-linked loci', which included standard filtering plus the use of*filter.sex.linked* function (Table 1). Functions *infer.sex* ,*filter.excess.het* and *gl2colony* were used in regime 'Removing sex-linked loci' where appropriate (Table 1). We also test what is the minimum number of known-sex individuals required by the combination of *filter.sex.linked* and *infer.sex* functions to be able to identify all sex-linked loci.

2.1 Empirical SNP datasets

DNA samples from two species of common eastern Australian passerine birds were genotyped commercially with Diversity Arrays Technology Pty. Ltd. (Canberra, Australia; Kilian et al. 2012). Briefly, DArTseq starts with DNA digestion, adapter ligation, and amplification of adaptor-ligated fragments. Amplification products are pooled and sequenced (single-read) on the Illumina HiSeq 2500 in batches of 94 samples per sequence lane, with 25% random technical replicates to enable assessment of loci scoring repeatability. Sequencing reads are processed using DArT proprietary analytical pipelines (for details see Harrisson et al. 2019). The end product is a spreadsheet with locus information and individual genotypes for each locus scored as '0' (homozygous reference), '1' (heterozygous), and '2' (homozygous alternate; Gruber et al. 2019). Both species are sexually monomorphic, with most individuals sexed using PCR-based methods (Pavlova et al. 2013).

Eastern yellow robin (EYR). The eastern yellow robin (*Eopsaltria australis*) is an avian model system for climate adaptation through mitonuclear interactions, with two diverged mitochondrial lineages occurring roughly east and west of the Great Dividing Range, and corresponding differentiation on neo-sex chromosomes enriched with mitonuclear genes (Morales et al. 2019, Gan et al 2021, Pavlova et al. 2013). In this study, we used data for 782 individuals sampled between 2016 and 2021 in four locations in Central Victoria (Crusoe, Muckleford, Timor and Wombat), in the zone of contact between the mitochondrial lineages (Austin et al. *unpublished manuscript*). Blood samples were collected under DELWP permit 10007910 under the Wildlife Act 1975 and the National Parks Act 1975, and NW11047F under section 52 of the Forest Act 1958, Australian Bird and Bat Banding Scheme permit, and approval 24225 of Monash University animal ethics committee. DArTseq yielded 53,324 binary SNPs for 238 Crusoe, 421 Muckleford, 52 Timor and 71 Wombat individuals.

Yellow-tufted honeyeater (YTH). The yellow-tufted honeyeater (*Lichenostomus melanops*) is a bird comprising four subspecies (*'cassidix', 'gippslandicus', 'melanops' and 'meltoni'* , Pavlova et al. 2014). Of these, *cassidix* (helmeted honeyeater) is Critically Endangered (Environment Protection and Biodiversity Conservation Act 1999, Advisory List of Threatened Vertebrate Fauna in Victoria 2013), restricted to a single small population, and supplemented by a captive breeding program (Harrisson et al. 2016). We used existing DArT SNP data of 641 YTH individuals used in a previous study (Harrisson et al. 2019). Of these, 540 were*cassidix* , 48 *gippslandicus* , 12 *melanops* , 33*meltoni* , 4 *cassidix* × *gippslandicus* crosses (hereafter 'hybrids'), 2 presumed hybrids, 1 presumed*gippslandicus* and 1 presumed *gippslandicus* ×*melanops* F1 individual. The initial DArTseq dataset consisted of 118,732 binary SNPs for 641 individuals.

2.2 SNP filtering regimes

The genetic datasets were imported into *R* as genlight objects and filtered using *dartR* package v2.0.4 (Mijangos et al. 2022) and our designed functions in *R* v4.2.1 (R Core Team 2022). The tally of filtering steps and remaining loci and individuals is presented in Table 2.

The first step in both filtering regimes controlled for very close physical linkage by keeping only one randomly-selected SNP per sequenced fragment (i.e., remove secondaries; method = 'random'). The resultant datasets served as the starting point for both regimes:

a) 'Standard' regime. The next filtering step removed SNPs with exceptionally low (< 5) and twice the average read depth, followed by the removal of SNPs with large amounts of missing data ($> 70^{\text{th}}$ percentile). At this point, individuals with $> 20\%$ missing data were dropped from the datasets, as were loci that became monomorphic as a result.

b) 'Removing sex-linked loci' regime. We started by removing sex-linked loci with function *filter.sex.linked* .

For EYR, all but one individual in the input genlight were of known sex (352 females and 429 males), while for YTH, 646 out of 641 individuals had known sex (289 females and 347 males). The output was used to infer the genetic sex of all individuals with function *infer.sex* , which led to a number of more-accurate sex assignments (one *de novo* sex assignment and one re-assignment in EYR, and five *de novo* sex assignments in YTH). We continued by removing highly heterozygous SNPs with function*filter.excess.het* . The rest of the steps (i.e., filtering for read depth, missing data and monomorphic loci) were done using the same parameters as for the 'Standard' regime.

2.3 Impact of filtering sex-linked loci on genetic diversity, individual heterozygosity, genetic structure and parentage analyses

Population genetic diversity. Six measures of population genetic diversity were calculated for 'before' and 'after' datasets: observed (Ho) and expected heterozygosity (He), Wright's fixation index (F_{IS}), polymorphism (P), number of private alleles not present in any other population (PA), and allelic richness (AR). Ho, He, F_{IS} and PA were calculated with*dartR* package v2.0.4 (function gl.report.heterozygosity method = 'pop', and function gl.report.pa method = 'one2rest'). AR was calculated using *hierfstat* package v0.5-11 (function allelic.richness; Goudet 2004). P was calculated as the proportion of loci that were polymorphic in a given population.

Individual observed heterozygosity (Ho). Individual Ho was calculated with *dartR* function gl.report.heterozygosity (method = 'ind'). In order to measure whether individual Ho changed when sex-linked loci were removed, we compared 'before' and 'after' individual Ho with a paired t-test ($\alpha = 0.05$) per sex. We also tested for significant differences in individual Ho between males and females (independent sample t-test), with 'before' and 'after' datasets. Cohen's d was used to measure effect sizes.

Genetic structure. Genetic structure between populations was qualitatively assessed with Pearson Principal Component Analyses (PCA,*dartR* function gl.pcoa). In order to reduce computation time, loci whose Minor Allele Count (MAC) was below 3 were removed from all datasets (*dartR* function gl.filter.maf, threshold = 3). We report results for the first two PCs, but the six major PCs were explored.

Parentage analyses. Given the potential for sex-linked chromosomes to affect the inference of parentage relationships, we performed separate parentage analyses using 'before' and 'after' datasets. We analysed 677 EYR individuals, and 527 YTH individuals (*cassidix* only). In both cases, MAC = 3 was applied to keep only loci shared between at least two individuals in order to reduce computation time. The genetic datasets for EYR consisted of 13,685 and 12,618 SNPs for the 'before and 'after' datasets, respectively. For*cassidix* , the 'before' dataset comprised 11,477 SNPs, and the 'after' dataset, 10,848 SNPs (Table 2).

Parentage analyses were run in COLONY v2.0.6.8 (Jones & Wang 2010). The function *gl2colony* was used to transform the genetic datasets to a COLONY input file. We assigned all individuals as candidate offspring, all females as candidate mothers (EYR: n = 308, *cassidix* : n = 255), and all males as candidate fathers (EYR: n = 369, cassidix: n = 272). In the case of EYR, candidate parents for 203 offspring were excluded based on year of birth, year of death (when known) and excessive geographical distance (Austin et al. *unpublished manuscript*). For both species, we used a full-likelihood approach ('likelihood = 1') with medium runs ('length_run = 2') at medium precision ('precision_fl = 1'). We assumed polygamy ('polygamy_male = 0', 'polygamy_female = 0') and a prior probability that the true parent is present in the sample of 0.5 ('probability_mother', 'probability_father'). Allele frequencies were not updated in order to minimize computational time ('update_allele_freq = 0'). For*cassidix* , we indicated the presence of inbreeding ('inbreed = 1') and set genotyping error to 0.05 ('other_typ_err = 0.05@') after Robledo-Ruiz et al. (2022). Genotyping error for EYR was set to empirically-determined 0.03, following Austin et al. (*unpublished manuscript*). Due to the stochasticity of the method implemented in COLONY (Jones & Wang 2010), we performed five independent runs per dataset (each with a different seed) to better explore the space of potential pedigree configurations.

Parentage assignments per run were compared to a set of known parentage relationships: 119 social EYR mothers observed consistently attending the nest and incubating (Austin et al. *unpublished manuscript*),

and 45 YTH known parent-offspring relationships from *cassidix*captive breeding (Robledo-Ruiz et al. 2022). The accuracy of parentage assignments was measured in two ways: (i) by counting how many runs out of five correctly identified a parent per known parentage relationship, and comparing before and after averages using a paired t-test, and (ii) by assigning as final parents those that were identified in at least three out of five runs (following Robledo-Ruiz et al. 2022) and testing whether the number of correct final assignments was positively associated with the removal of sex-linked loci with a χ^2-test.

Minimum number of known-sex individuals for*filter.sex.linked* function

We used both EYR and YTH datasets to estimate the number of sex-linked loci that are identified with subsets of known-sex individuals of variable size. We created eight subsets: 20, 24, 30, 40, 50, 100, 200 and 400 individuals chosen at random, all with 1:1 sex ratio, and applied function *filter.sex.linked* to each. We then identified the smallest subset of known-sex individuals with which it was still possible to identify sex-linked loci, and tested whether those loci were useful to sex the rest of the individuals and in turn, use the new sex assignments to identify all sex-linked loci. For this, we created five random subsets of known-sex individuals of the smallest size (24 and 30 known-sex individuals for EYR and YTH, respectively; see Results 3), applied function *filter.sex.linked* followed by function*infer.sex* , and used the new sex assignments to re-run*filter.sex.linked* .

RESULTS

Identification and removal of sex-linked loci

The function *filter.sex.linked* identified and removed 3,807 sex-linked loci in EYR (10.7% of the total 35,663 loci tested; Table 3). Of these, 69.3% were identified based on differential call rate between the sexes (i.e., W-linked and sex-biased; Figure 3a, b) and 30.7% based on differential heterozygosity between the sexes (i.e., Z-linked and gametologs; Figure 3c, d). For YTH, the function identified 3,414 sex-linked loci (4.6% of the total 74,470 loci tested; Table 3) of which 65% were identified by call rate, and 35% by heterozygosity (Figure S1).

Comparison of 'before' and 'after' datasets revealed that, when the function *filter.sex.linked* was not used, 28.7% (n = 1,093) and 19.0% (n = 650) of the sex-linked loci remained in the final SNP datasets of EYR and YTH, respectively. Standard locus-filters had variable efficiency in removing different types of sex-linked loci (Figure 4): together, read depth and loci missing data filters were capable of removing all W-linked loci, and 90% and 99% of sex-biased loci from EYR and YTH datasets, respectively. However, they were unable to remove 75% and 57% of Z-linked loci (EYR: n = 620 were not removed; YTH: n = 652), and 71% and 37% of gametologs (EYR: n = 241; YTH: n = 21). Other filtering steps such as removing individual missing data and applying a minor allele count (MAC) had little effect on removing additional sex-linked loci (Figure 4). This inefficiency translated in 7.8% and 5.7% of the final dataset SNPs being sex-linked in EYR and YTH, respectively.

Impact of removing sex-linked loci on population genetic diversity, individual heterozygosity, genetic structure and parentage analyses

Population genetic diversity. In general, removal of sex-linked loci produced a decrease in estimates of population genetic diversity (Figure S2 and S3). However, the magnitude of this change varied with different measures of genetic diversity and, importantly, magnitude and direction of the change ranged across populations (Figure 5): the largest impact was on F_{IS}, which ranged from 9.3% decrease to 2% increase, and private alleles (PA), which ranged from 8% decrease to 0.5% increase. Expected heterozygosity (He) experienced decreases ranging from 0.7% to 2.4%. The direction and magnitude of the change did not correspond to the F:M ratios of samples (EYR: Crusoe = 0.87, Muckleford = 0.93, Timor = 0.79, Wombat = 0.39; YTH: Cassidix = 0.94, Gippslandicus = 0.55, Melanops = 1.0, Meltoni = 0.1).

Individual observed heterozygosity (Ho). The removal of sex-linked loci produced a statistically significant change in individual Ho whose magnitude and direction varied between sexes and species (Table 4). For EYR, the decrease in female and male Ho was significant but small (F: 0.2% decrease, Cohen's D = 0.35;

M: 0.3% decrease, Cohen's D = 0.23). For YTH, the change was an order of magnitude larger and went in opposite directions between the sexes: female Ho increased 3.8% (p-value < 0.001, Cohen's D = -8.7) and male Ho decreased 2.9% (p-value < 0.001, Cohen's D = 1.9). This opposite effect in male and female Ho translated into the disappearance of the significant (but misleading) difference between male and female Ho (p-value < 0.001) after the removal of sex-linked loci from the YTH dataset (p-value = 0.1; Table 5). There were no significant differences in Ho between the sexes in EYR before or after removing sex-linked loci.

Genetic structure. Before the removal of sex-linked loci, PC1 explained 2.4% of the genetic variation in EYR, and divided the individuals into two groups (Crusoe-Timor and Muckleford-Wombat; Figure 6a). PC2, on the other hand, explained 1.6% of variation and captured genetic structure due to the presence of sex-linked loci: it divided the individuals into males and females (Figure 6b). This division between male and females disappeared from PC2 after removing sex-linked loci (Figure 6c, d). For YTH, none of PC1, PC2, PC3 or PC4 showed sex genetic structure, before or after using function *filter.sex.linked*(Figure S4).

Accuracy of parentage analyses. For EYR, before removing sex-linked loci, an average of 3.83 runs out of five identified the correct parent. After removing sex-linked loci, the average increased significantly to 4.26 (p-value = 0.003; Table 6). We also found a significant association between the removal of sex-linked loci and the number of correct final parentage assignments (χ^2 = 4.8, df = 1, p-value = 0.03): before removing sex-linked loci, 91 out of 119 (76.5%) final assignments were correct, compared to 104 (87.4%) correct final assignments after removing sex-linked loci. For YTH (*cassidix*), we found that removing sex-linked loci did not significantly change the average number of runs that correctly identified parents, which started with the high average of 4.9 runs (Table 6).

Minimum number of known-sex individuals for*filter.sex.linked* function

For EYR, 24 known-sex individuals (12 females and 12 males) were the minimum with which it was still possible to identify sex-linked loci:*filter.sex.linked* identified 267 loci which represented 7% of the total sex-linked loci (Figure 7a, Table S1). For YTH, 30 known-sex individuals (15 females and 15 males) were the minimum:*filter.sex.linked* identified 61 loci which represented 1.8% of the total sex-linked loci in the full dataset (Figure 7b, Table S1). With fewer known-sex individuals the function was unable to identify any sex-linked loci.

For EYR, *filter.sex.linked* function identified, on average, only 7.2% (range = 6.6-7.9%) of all sex-linked loci for the five subsets of 24 known-sex individuals (91.5% of all W-linked loci, 0% of all sex-biased, 0.1% of all Z-linked and 32.6% of all gametologs). For YTH, *filter.sex.linked* function identified only 1.9% (range = 1.8-2.0%) of all sex-linked loci for the five subsets of 30 known-sex individuals (99.3% of all W-linked loci, 0.1% of all sex-biased, 0% of all Z-linked and 8.6% of all gametologs). These retrieved sex-linked loci allowed *infer.sex* to correctly identify the sexes of all individuals which it assigned as 'M' or 'F' (cf. marked as '*M' or '*F'; 587 EYR and 519 YTH; the same individuals for the five sets). Using the new 587 EYR and 519 YTH assignments to re-run *filter.sex.linked*identified 100% of all sex-linked loci for both EYR and YTH (3,807 and 3,414 sex-linked loci, respectively). It is likely that function*filter.sex.linked* was able to identify sex-linked loci with fewer known-sex EYR individuals than YTH individuals because EYR has larger sex chromosomes (i.e., it has neo-sex chromosomes in which a portion of chromosome 1A got fused to the Z chromosome while the other portion got fused to the W chromosome; Gan et al. 2019). We recommend the use of at least 15 males and 15 females to allow the identification of all sex-linked loci, although a larger number might be needed for species with shorter, less differentiated or less variable sex chromosomes.

DISCUSSION

In this study, we designed and developed four *R* functions that automate tasks commonly needed in conservation genomic analyses: (1)*filter.sex.linked* to identify and remove sex-linked loci, (2)*infer.sex* to infer the genetic sex of individuals using sex-linked loci, (3) *filter.excess.het* to remove loci with abnormally high heterozygosity, and (4) *gl2colony* to produce input files for parentage analysis software. Use of these functions on genomic data for two bird species revealed that standard filters, such as low read depth and call rate, are inefficient at removing sex-linked loci, removing fewer than half of Z-linked loci and only 29-63%

of gametologs. In the two studied species, the failure to comprehensively remove sex-linked loci led to one or more of: (i) overestimation of up to 9% of population F_{IS}, and up to 8% of the number of private alleles (ii) incorrectly inferring sex differences in individual heterozygosity, (iii) capturing sex genomic differences instead of population structure, and (iv) inferring ˜11% fewer parent-offspring relationships in parentage analyses. We also found that our functions were capable of identifying all sex-linked loci using as few as 15 known males and 15 known females, through a preliminary run of *filter.sex.linked* , followed by running *infer.sex* and then re-running *filter.sex.linked* .

Appropriate filtering is a challenging part of population genomic analyses. It is widely acknowledged that filtering can significantly affect the inferences drawn from different analyses, ranging from 'simple' standard measures like heterozygosity, all the way to GEA (e.g., Fu 2014; Linck & Battey 2019; Graham et al. 2020; William et al. 2022; Ahrens et al. 2021). Given this awareness, there is surprisingly little mention of best-practices for filtering out sex-linked loci from SNP datasets in population genomics research (but see Benestan et al. 2017 and Trenkel et al. 2020). Unless using per-marker F_{ST} or *dartR* 's gl.report.sexlinked function to explicitly identify sex-linked markers, studies rarely address them, and seem to rely mainly on read depth and loci missing data filters to remove sex-linked loci from large SNP datasets. We have demonstrated that this untargeted approach fails to remove ˜19-29% of all sex-linked loci. Filtering sex-linked markers based only on assumed synteny with the chromosome location of a heterospecific reference genome can also result in failing to account for neo-sex chromosomes in evolutionary studies (Morales et al. 2018). Recent discoveries of neo-sex chromosome systems in Sylvioidea (Sigeman et al. 2020; Sigeman et al. 2022), Australian robins (Gan et al. 2019), insects (Wang et al. 2022) and other systems highlight dangers of assuming synteny with reference genomes of other species while detecting sex-linked loci. Thus, we propose that use our *filter.sex.linked*function to remove sex-linked loci *before* applying SNP quality filters can comprise best-practice that will ensure that downstream filters are in fact evaluating the quality of autosomal loci.

We showed that the failure to remove sex-linked loci meant that a considerable proportion—7.8% and 5.7%—of the SNPs in the final datasets were not autosomal, and therefore, yielded incorrect estimates of population diversity. Interestingly, the effect of sex-linked loci on genetic diversity biases varied among populations unpredictably, and was not influenced by the within-population sex-ratio (Figure 5). This is likely because there are many factors intervening in addition to sample sex-bias, such as the proportions of different types of sex-linked loci, their different allelic frequencies in the populations, the total amount of sex-linked versus autosomal loci, the sex-chromosome-to-autosome diversity ratio, and the level of recombination between sex chromosomes. This highlights the necessity of searching for and carefully filtering out sex-linked loci, because it would be hard to control for their presence in other ways (e.g., by introducing sample sex ratio in statistical models).

Despite the relatively small impact of the presence of sex-linked loci on *population* Ho, there was a significant impact on*individual* Ho that was large enough to erroneously indicate that YTH females were 5% less heterozygous than males (Table 5). This spurious significant difference could have mistakenly suggested that females are philopatric (which is not true in *cassidix* ; Smales 2004) or that they experience less inbreeding depression for survival (the reverse is true in *cassidix* ; Harrisson et al. 2019). If these hypotheses were not known in advance to be incorrect, they might have been accepted or at least further investigated; thus, poor filtering of sex-linked loci can lead to incorrect ecological and evolutionary inferences and wasted resources.

Our results also illustrated how the presence of sex-linked SNPs can obscure population structure. The first PC on EYR data showed population structure due to geographically separated groups. The second PC, however, simply captured the genetic differences between sexes when sex-linked markers were not removed, obscuring the fact that in reality, the second largest source of genetic variation comes from within the Muckleford population (Figure 6). This masking of population structure has also been observed in the Discriminant Analysis of Principal Components (DAPC) of two species of lobsters due to the presence of a few sex-linked loci (Benestan et al. 2017). If not properly checked against sex, the PC2 split in two could

have been interpreted as, for instance, the presence of two cryptic sympatric species. Researchers studying populations with little genetic variation should be particularly careful, because this effect is expected to be more pronounced for populations with low genetic differentiation.

Importantly, we found that failing to remove sex-linked loci led to ˜11% fewer correct parentage assignments (Table 6). Such a substantial loss of correct assignments could have repercussions for the management of endangered species. For example, releases of captive-bred individuals or translocations/introductions are usually done avoiding the release of close relatives in the same group in order to maximize genetic diversity and discourage inbreeding (e.g.,*cassidix* , Harrisson et al. 2016; Frankham et al. 2017). Removing sex-linked loci will be even more crucial in the absence of a set of known parentages with which to calibrate parentage analyses as is likely to apply to many species of conservation concern such as (i) those whose breeding season cannot be monitored because it occurs in inaccessible locations or because of lack of resources, (ii) polygamous and cooperative-breeding species, (iii) those with external fertilisation like amphibian and fish species (Nakamura 2009). Accounting for sex-linked loci is also likely to have the largest impact on species with large sex chromosomes (including neo-sex chromosomes, which have been discovered in many taxa including EYR) because sex-linked loci will represent a large proportion of the potential genomic markers for parentage analysis (Sigeman et al. 2022; Beukeboom & Perrin 2014; Gan et al. 201).

The functions we propose were created with the needs of conservation genomicists and wildlife managers in mind. Sexing individuals is especially important for species without sex dimorphism, or for sexually-dimorphic species whose youngs' sex is undistinguishable. With the combination of the functions *filter.sex.linked* and*infer.sex* we offer a formal statistical framework that systematically identifies and uses sex-linked loci to make sex assignments with as few as 15 known-sex individuals of each sex. Unlike current practices, *infer.sex* was designed to use the complementary information contained in all types of sex-linked loci available, which makes the sex-assignments more robust. The use of all types of sex-linked loci will be advantageous for low-density marker datasets because it uses information that would otherwise be neglected, and it facilitates development of SNP panels that include sex-specific loci (Blåhed et al. 2018; Willis et al. 2020). It also allows for error-checking and confirming congruence between genetic and phenotypic sex of individuals, which may assist in detecting cases of environmental sex-reversal (Stelkens & Wedekind 2010). The separation of sex-linked loci can be used to validate the assembly of W and Y chromosomes, and to study sex-specific processes (e.g., natural selection, philopatry). Furthermore, it reduces the cost in time, genetic material and resources of using other sexing methods (e.g., PCR amplification of CHD1-Z and CHD1-W genes; Fridolfsson & Ellegren 1999).

The function *filter.excess.het* provides a statistically-backed method to identify artefactual multilocus SNPs that show abnormally high heterozygosity. The function circumvents the problem of choosing an arbitrary heterozygosity threshold by, instead, testing loci whose heterozygosity [?] 0.5 and also have significant excess of heterozygotes beyond sampling error. This has the advantage of taking into account random sampling and genotyping errors that affect loci differently. In fact, this approach is available in *VCFtools* but not yet in *dartR* , *snpR* or*SNPfiltR* (Hohenlohe et al. 2011; Denecek et al. 2011; Mijangos et al. 2022; Hemstrom & Jones 2022; DeRaad 2022). Nonetheless, we would like to emphasize that this is *not* a Hardy-Weinberg equilibrium filter (which requires critical thinking to be correctly applied and interpreted; Waples 2015), and should be used only when looking to obtain neutral autosomal loci (cf. looking for signatures of selection).

In conclusion, we demonstrated how incomplete removal of sex-linked loci can bias conservation genomic inferences. We argue that comprehensively removing sex-linked loci should be best practice when handling genomic data, and we offer convenient easy-to-use resources to automate this and other bioinformatic steps. The functions presented here can be integrated into bioinformatic pipelines and widely used *R*packages such as *dartR* , *sambaR* , *SNPfiltR* and*snpR* . By developing functions that can be easily adopted by conservation biologists and incorporated in wildlife management workflows, this study will contribute to a better understanding of the processes occurring in threatened species, such as inbreeding, inbreeding depression, population structure.

ACKNOWLEDGEMENTS

This work was funded by the Australian Research Council though Linkage Grant LP160100482 with Partner Organizations Department of Environment, Land, Water and Planning (DELWP, Victoria), Diversity Arrays Technology, Zoos Victoria, Environment, Planning & Sustainable Development Directorate (ACT Government), and Department of Biodiversity, Conservation and Attractions (Western Australia), and Discovery Project Grants DP180102359 and DP210102275. Additional support was provided by Zoos Victoria, the Faculty of Science (Monash University), and Holsworth Wildlife Research Endowment (Ecological Society of Australia). Alexandra Pavlova was supported by the Catalyst Science Fund from Revive & Restore. Lana Austin was supported by an Australian Government Research Training Program (RTP) Scholarship. We thank Bruce Quin, Friends of the Helmeted Honeyeater, Jessica Zhou, Pete Collins, Thomas Richard, Anna Polesskiy, Alice Sunnucks and numerous volunteers for assistance collecting and processing genetic samples and field data. We also thank Blair Venn and Bendigo City Council for access to Crusoe Reservoir. Computational resources were provided by the Monash eResearch Centre (MeRC) and Monash eSolutions-Research Support Services through the use of the MonARCH and MASSIVE HPC Clusters. Special thanks to Gabriel W. Low for discussion on the rationale of the functions.

AUTHOR CONTRIBUTIONS

Diana A. Robledo-Ruiz, Alexandra Pavlova and Paul Sunnucks led initial project conceptualization and design. Lana Austin and J. Nevil Amos collected EYR field samples and genotyping data, and conducted initial analyses of sex-linked loci. Diana A. Robledo-Ruiz and Jesus Castrejon-Figueroa wrote the *R* functions and analysed the data with guidance from Paul Sunnucks and Alexandra Pavlova. Alexandra Pavlova and Diana A. Robledo-Ruiz wrote the first draft of the manuscript and all authors contributed to editing and writing. All authors approved the final version of this manuscript for publication. Paul Sunnucks, Alexandra Pavlova, Michael J. L. Magrath and Lana Austin secured funding for the project.

DATA ACCESSIBILITY STATEMENT

The *R* functions are available at github.com/drobledoruiz/conservation_genomics. All data and scripts used in this manuscript have been archived in Bridges Monash University research repository (Robledo Ruiz et al. 2022).

BENEFIT-SHARING STATEMENT

The contributions of all individuals to the research, including volunteers, are described in the Acknowledgements. The results of this research have been shared with stakeholders and the broader scientific community. Benefits from this research accrue from the sharing of our data and results on public databases as described above.

REFERENCES

Allendorf, F. W., Funk, W. C., Aitken, S. N., Byrne, M., & Luikart, G. (2022). *Conservation and the genomics of populations* . Oxford University Press.

Amos, J. N., Harrisson, K. A., Radford, J. Q., White, M., Newell, G., Mac Nally, R., Sunnucks, P., Pavlova, A. (2014) Species- and sex-specific connectivity effects of habitat fragmentation in a suite of woodland birds. *Ecology* **95** , 1556–1568.

Arnold, B. D., & Wilkinson, G. S. (2015). Female natal philopatry and gene flow between divergent clades of pallid bats (*Antrozous pallidus*). *Journal of Mammalogy* , *96* (3), 531-540.

Baird, N. A., Etter, P. D., Atwood, T. S., Currey, M. C., Shiver, A. L., Lewis, Z. A., ... & Johnson, E. A. (2008). Rapid SNP discovery and genetic mapping using sequenced RAD markers. *PloS one* ,*3* (10), e3376.

Beukeboom, L. W., & Perrin, N. (2014). *The evolution of sex determination* . Oxford University Press, USA.

Benestan, L., Moore, J. S., Sutherland, B. J., Le Luyer, J., Maaroufi, H., Rougeux, C., ... & Bernatchez, L. (2017). Sex matters in massive parallel sequencing: Evidence for biases in genetic parameter estimation

and investigation of sex determination systems. *Molecular Ecology* , *26* (24), 6767-6783.

Benjamini, Y., and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing.*Journal of the Royal Statistical Society Series B, 57* , 289–300.

Blahed, I. M., Konigsson, H., Ericsson, G., & Spong, G. (2018). Discovery of SNPs for individual identification by reduced representation sequencing of moose (*Alces alces*). *PloS one* , *13* (5), e0197364.

Britt, M., Haworth, S. E., Johnson, J. B., Martchenko, D., & Shafer, A. B. (2018). The importance of non-academic coauthors in bridging the conservation genetics gap. *Biological Conservation* , *218* , 118-123.

Carvalho, A. B., & Clark, A. G. (2013). Efficient identification of Y chromosome sequences in the human and Drosophila genomes. *Genome Research* , *23* (11), 1894-1907.

Castella, V., Ruedi, M., & Excoffier, L. (2001). Contrasted patterns of mitochondrial and nuclear structure among nursery colonies of the bat*Myotis myotis* . *Journal of Evolutionary Biology* ,*14* (5), 708-720.

Chen, N., Van Hout, C. V., Gottipati, S., & Clark, A. G. (2014). Using Mendelian inheritance to improve high-throughput SNP discovery.*Genetics* , *198* (3), 847-857.

Davey, J. W., & Blaxter, M. L. (2010). RADSeq: next-generation population genetics. *Briefings in functional genomics* ,*9* (5-6), 416-423.

DeRaad, D. A. (2022). SNPfiltR: an R package for interactive and reproducible SNP filtering. *Molecular Ecology Resources* .

Ellegren, H. (2009). The different levels of genetic diversity in sex chromosomes and autosomes. *Trends in Genetics* , *25* (6), 278-284.

Ellegren, H. (2014). Genome sequencing and population genomics in non-model organisms. *Trends in ecology & evolution* ,*29* (1), 51-63.

Fitzpatrick, S. W., Bradburd, G. S., Kremer, C. T., Salerno, P. E., Angeloni, L. M., & Funk, W. C. (2020). Genomic and fitness consequences of genetic rescue in wild populations. *Current Biology* ,*30* (3), 517-522.

Flanagan, S. P., *&* Jones, A. G. *(* 2019*).* The future of parentage analysis: From microsatellites to SNPs and beyond. *Molecular Ecology, 28(3),* 544*–567.*

Frankham, R., Ballou, J. D., Ralls, K., Eldridge, M., Dudash, M. R., Fenster, C. B., ... & Sunnucks, P. (2017). *Genetic management of fragmented animal and plant populations* . Oxford University Press.

Fridolfsson, A. K., & Ellegren, H. (1999). A simple and universal method for molecular sexing of non-ratite birds. *Journal of avian biology* , 116-121.

Fu, Y. B. (2014). Genetic diversity analysis of highly incomplete SNP genotype data with imputations: an empirical assessment. *G3: Genes, Genomes, Genetics* , *4* (5), 891-900.

Galla, S. J., Buckley, T. R., Elshire, R., Hale, M. L., Knapp, M., McCallum, J., ... & Steeves, T. E. (2016). Building strong relationships between conservation genetics and primary industry leads to mutually beneficial genomic advances.

Galla, S. J., Brown, L., Couch-Lewis, Y., Cubrinovska, I., Eason, D., Gooley, R. M., ... & Steeves, T. E. (2022). The relevance of pedigrees in the conservation genomics era. *Molecular Ecology Resources, 31(* 1*),* 41*–54.*

Gan, H. M., Falk, S., Morales, H. E., Austin, C. M., Sunnucks, P., & Pavlova, A. (2019). Genomic evidence of neo-sex chromosomes in the eastern yellow robin. *GigaScience* , *8* (9), giz111.

Goudet, J. (2005). Hierfstat, a package for R to compute and test hierarchical F-statistics. *Molecular ecology notes* , *5* (1), 184-186.

Graham, C. F., Boreham, D. R., Manzon, R. G.,Stott, W., Wilson, J. Y., & Somers, C. M.(2020). How "simple" methodological decisions affect interpretation of population structure based on reduced representation library DNA sequencing: A case study using the lake whitefish. *PLoS One, 15(1), e0226608.*

Gruber, B., Unmack, P., Berry, O., & Georges, A. (2019). Introduction to dartR. *User Manual* , 51.

Harrisson, K. A., Pavlova, A., Goncalves da Silva, A., Rose, R., Bull, J. K., Lancaster, M. L., ... & Sunnucks, P. (2016). Scope for genetic rescue of an endangered subspecies though re-establishing natural gene flow with another subspecies. *Molecular Ecology* , *25* (6), 1242-1258.

Harrisson, K. A., Magrath, M. J., Yen, J. D., Pavlova, A., Murray, N., Quin, B., Menkhorst, P., Miller, K. A., Cartwright, K., Sunnucks, P. (2019) Lifetime fitness costs of inbreeding and being inbred in a critically endangered bird. *Current Biology 29* , 2711–2717.

Hoffmann, A., Griffin, P., Dillon, S., Catullo, R., Rane, R., Byrne, M., ... & Sgro, C. (2015). A framework for incorporating evolutionary genomics into biodiversity conservation and management. *Climate Change Responses* , *2* (1), 1-24.

Hogg, C. J., Ottewell, K., Latch, P., Rossetto, M., Biggs, J., Gilbert, A., ... & Belov, K. (2022). Threatened Species Initiative: Empowering conservation action using genomic resources. *Proceedings of the National Academy of Sciences* , *119* (4), e2115643118.

Hohenlohe, P. A., Amish, S. J., Catchen, J. M., Allendorf, F. W., & Luikart, G. (2011). Next-generation RAD sequencing identifies thousands of SNPs for assessing hybridization between rainbow and westslope cutthroat trout. *Molecular ecology resources* , *11* , 117-122.

Hohenlohe, P. A., Funk, W. C., & Rajora, O. P. (2021). Population genomics for wildlife conservation and management. *Molecular Ecology* , *30* (1), 62-82.

Holderegger, R., Balkenhol, N., Bolliger, J., Engler, J. O., Gugerli, F., Hochkirch, A., Nowak, C., Segelbacher, G., Widmer, A., & Zachos, F. E. (2019). Conservation genetics: Linking science with practice.*Molecular Ecology, 28* (17), 3848-3856.

Jombart, T., & Ahmed, I. (2011). adegenet 1.3-1: new tools for the analysis of genome-wide SNP data. *Bioinformatics* , *27* (21), 3070-3071.

Jones, O. R., & Wang, J. (2010). COLONY: a program for parentage and sibship inference from multilocus genotype data. *Molecular ecology resources, 10* (3), 551-555.

Kardos, M. (2021). Conservation genetics. *Current Biology, 31* (19), R1185-R1190.

Kilian, A., Wenzl, P., Huttner, E., Carling, J., Xia, L., Blois, H., ... & Uszynski, G. (2012). Diversity arrays technology: a generic genome profiling technology on open platforms. In *Data production and analysis in population genomics* (pp. 67-89). Humana Press, Totowa, NJ.

Linck, E., & Battey, C. J. (2019). Minor allele frequency thresholds strongly affect population structure inference with genomic data sets. *Molecular Ecology Resources, 19(* 3*)*,639– 647.

Mijangos, J. L., Gruber, B., Berry, O., Pacioni, C., & Georges, A. (2022). dartR v2: An accessible genetic analysis platform for conservation, ecology and agriculture. *Methods in Ecology and Evolution* .

Morales, H. E., Pavlova, A., Amos, N., Major, R., Killian, A., Greening, C., Sunnucks, P. (2018) Concordant divergence of mitogenomes and a mitonuclear gene cluster in bird lineages inhabiting different climates.*Nature Ecology and Evolution 2* , 1258–1267.

Moran, B. M., Thomas, S. M., Judson, J. M., Navarro, A., Davis, H., Sidak-Loftis, L., ... & Steiner, C. C. (2021). Correcting parentage relationships in the endangered California Condor: Improving mean kinship estimates for conservation management. *The Condor* ,*123* (3), duab017.

Mucci, N., Giangregorio, P., Cirasella, L., Isani, G., & Mengoni, C. (2020). A new STR panel for parentage analysis in endangered tortoises.*Conservation genetics resources* , *12* (1), 67-75.

Nakamura, M. (2009, May). Sex determination in amphibians. In*Seminars in cell & developmental biology* (Vol. 20, No. 3, pp. 271-282). Academic Press.

O'Leary, S. J., Puritz, J. B., Willis, S. C., Hollenbeck, C. M., & Portnoy, D. S. (2018). These aren't the loci you'e looking for: Principles of effective SNP filtering for molecular ecologists.

Overbeek, A., Galla, S., Brown, L., Cleland, S., Thyne, C., Maloney, R., & Steeves, T.(2020). Pedigree validation using genetic markers in an intensively-managed taonga species, the critically endangered kaki (*Himantopus novaezelandiae*). *Notornis, 67(4),*709–*716.*

Pavlova, A., Amos, J. N., Joseph, L., Loynes, K., Austin, J. J., Keogh, J. S., Stone, G., Nicholls, J. A., Sunnucks, P. (2013) Perched at the mito-nuclear crossroads: divergent mitochondrial lineages correlate with environment in the face of ongoing nuclear gene flow in an Australian bird. *Evolution* **67** , 3412–3428.

Pavlova, A., Selwood, P., Harrisson, K. A., Murray, N., Quin, B., Menkhorst, P., Smales, I., Sunnucks, P. (2014) Integrating phylogeography and morphometrics to assess conservation merits and inform conservation strategies for an endangered subspecies of a common bird species. *Biological Conservation* **174** , 136–146.

Pavlova, A., Harrisson, K. A., Turakulov, R., Lee, Y. P., Ingram, B. A., Gilligan, D., ... & Gan, H. M. (2022). Labile sex chromosomes in the Australian freshwater fish family Percichthyidae. *Molecular Ecology Resources* , *22* (4), 1639-1655.

Pearman, W. S., Urban, L., & Alexander, A. (2022). Commonly used Hardy–Weinberg equilibrium filtering schemes impact population structure inferences using RADseq data. *Molecular ecology resources* , *22* (7), 2599-2613.

Peterson, B. K., Weber, J. N., Kay, E. H., Fisher, H. S., & Hoekstra, H. E. (2012). Double digest RADseq: an inexpensive method for de novo SNP discovery and genotyping in model and non-model species. *PloS one* , *7* (5), e37135.

Pritchard, J. K., Stephens, M., & Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics* ,*155* (2), 945-959.

R Core Team. (2022). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/

Radosavljević, I., Satovic, Z., & Liber, Z. (2015). Causes and consequences of contrasting genetic structure in sympatrically growing and closely related species. *AoB Plants* , 7.

Robledo Ruiz, D., Pavlova, A., & Sunnucks, P., (2022) Supporting data for Robledo-Ruiz et al. (submitted) Monash University. Dataset. https://doi.org/10.26180/21608028

Robledo-Ruiz, D. A., Pavlova, A., Clarke, R. H., Magrath, M. J., Quin, B., Harrisson, K. A., ... & Sunnucks, P. (2022). A novel framework for evaluating in situ breeding management strategies in endangered populations. *Molecular Ecology Resources* , *22* (1), 239-253.

Shafer, A. B. A., Peart, C. R., Tusso, S., Maayan, I., Brelsford, A., Wheat, C. W., & Wolf, J. B. W. (2017). Bioinformatic processing of RAD-seq data dramatically impacts downstream population genetic inference. *Methods in Ecology and Evolution, 8(* 8*),* 907– 917.

Sigeman, H., Ponnikas, S., Hansson, B. (2020) Whole-genome analysis across 10 songbird families within Sylvioidea reveals a novel autosome–sex chromosome fusion. *Biology Letters* **16** , 20200082.

Sigeman, H., Zhang, H., Ali Abed, S., & Hansson, B. (2022). A novel neo-sex chromosome in Sylvietta brachyura (Macrosphenidae) adds to the extraordinary avian sex chromosome diversity among Sylvioidea songbirds.*Journal of Evolutionary Biology* .

Smales, I. J. (2004). *Population ecology of the Helmeted Honeyeater Lichenostomus melanops cassidix: long-term investigations of a threatened bird* (Doctoral dissertation, University of Melbourne, School of Botany).

Stelkens, R. B., & Wedekind, C. (2010). Environmental sex reversal, Trojan sex genes, and sex ratio adjustment: conditions and population consequences. *Molecular Ecology* , *19* (4), 627-646.

Taylor, H. R., Dussex, N., & van Heezik, Y. (2017). Bridging the conservation genetics gap by identifying barriers to implementation for conservation practitioners. *Global Ecology and Conservation* ,*10* , 231-242.

Trenkel, V. M., Boudry, P., Verrez-Bagnis, V., & Lorance, P. (2020). Methods for identifying and interpreting sex-linked SNP markers and carrying out sex assignment: application to thornback ray (Raja clavata). *Molecular ecology resources* , *20* (6), 1610-1619.

Van Rossum, F. Sibship and parentage reconstruction as a genetic tool for designing and monitoring plant translocations. *Restoration Ecology* , e13726.

Wang, S., Nalley, M. J., Chatla, K., Aldaimalani, R., MacPherson, A., Wei, K. H. C., ... & Bachtrog, D. (2022). Neo-sex chromosome evolution shapes sex-dependent asymmetrical introgression barrier.*Proceedings of the National Academy of Sciences* , *119* (19), e2119382119.

Waples, R. S. (2015). Testing for Hardy–Weinberg proportions: have we lost the plot?. *Journal of heredity* , *106* (1), 1-19.

Willi, Y., Kristensen, T. N., Sgrò, C. M., Weeks, A. R., Ørsted, M., & Hoffmann, A. A. (2022). Conservation genetics as a management tool: The five best-supported paradigms to assist the management of threatened species. *Proceedings of the National Academy of Sciences, 119* (1) e2105076119.

Willis, S. C., Hess, J. E., Fryer, J. K., Whiteaker, J. M., Brun, C., Gerstenberger, R., & Narum, S. R. (2020). Steelhead (Oncorhynchus mykiss) lineages and sexes show variable patterns of association of adult migration timing and age-at-maturity traits with two genomic regions. *Evolutionary applications* , *13* (10), 2836-2856.

Willis, S. C., Hollenbeck, C. M., Puritz, J. B., Gold, J. R., & Portnoy, D. S. (2017). Haplotyping RAD loci: an efficient method to filter paralogs and account for physical linkage. *Molecular Ecology Resources* , *17* (5), 955-965.

FIGURE LEGENDS

Figure 1. Schematic of the distribution patterns of three types of sex-linked loci in the ZW sex-determination system: W-linked loci are found only in the W chromosome (yellow); Z-linked loci are found only in the Z chromosome (orange); gametolog loci are present in both chromosomes (green). The same principles apply to the XY sex-determination system but males are heterogametic (XY) and females homogametic (XX).

Figure 2. Distribution of some diagnostic parameters for autosomal and sex-linked loci. **a)** Autosomal loci (grey) are expected to present roughly the same call rate for males and females. W-linked loci (yellow) are expected to be called in females but absent in males because males lack a W chromosome. We refer to other loci whose call rate is biased by sex as 'sex-biased' (blue, drawn here for male-bias in call rate). **b)** Autosomal loci (grey) are expected to present roughly the same proportion of heterozygous males and females. For Z-linked loci (orange), females are expected to be homozygous because they have only one Z chromosome. For gametologous loci (green), males are expected to be homozygous because they have two Z chromosomes each with the same Z-associated allele.

Figure 3. Plots produced by function *filter.sex.linked*after being used to identify and remove sex-linked loci from eastern yellow robin (EYR) genetic data. Top panels: plots of female call rate against male call rate in which each point represents a locus, before (**a**) and after (**b**) removing 2,639 sex-linked loci with differential call rate between the sexes. Bottom panels: plots of the proportion of heterozygous females against the proportion of heterozygous males with each point representing a locus, before (**c**) and after (**d**) removing 1,168 sex-linked loci with differential heterozygosity between the sexes.

Figure 4. Progression of four types of sex-linked loci after different SNP filtering steps were applied to eastern yellow robin (EYR) and yellow-tufted honeyeater (YTH) datasets. Arrows to the right indicate the percentage of sex-linked loci (out of the initial 100%) that were removed. Down arrows indicate the percentage of sex-linked loci (out of the initial 100%) that remain in the dataset.

Figure 5. Percentage change of six measures of population genetic diversity after removing sex-linked loci (Ho: observed heterozygosity, He: expected heterozygosity, FIS: Wright's F_{IS}, P: polymorphism, PA: private alleles, and AR: allelic richness). Estimates are given per population of eastern yellow robin (EYR) and yellow-tufted honeyeater (YTH).

Figure 6. Principal Component Analyses (PCA) of the genomic dataset of eastern yellow robin, EYR, before (top panels) and after (bottom panels) removing sex-linked loci. On (**a**) and (**c**), individuals are coloured according to their population. On (**b**) and (**d**), individuals are coloured by sex.

Figure 7. Proportion of sex-linked loci that function*filter.sex.linked* was able to identify with variable number of known-sex individuals for EYR (**a**) and YTH (**b**) datasets. The sex ratio of known sex-individuals was 1:1, except for 'all' which included the whole set of known-sex individuals (EYR: 352 females and 429 males, YTH: 289 females and 347 males).

TABLES

Table 1 . Ordered steps of two SNP filtering regimes applied to the genetic datasets of eastern-yellow robin (EYR) and yellow-tufted honeyeater (YTH). '*' indicates that a filtering step was used in the filtering regime.

	Standard	**Removing sex-linked loci**
Secondaries	*	*
filter.sex.linked + infer.sex		*
filter.excess.het		*
Read depth	*	*
Locus missing data	*	*
Individual missing data	*	*
Minor allele count	*	*
gl2colony	*	*

Table 2 . Count of remaining loci and individuals after each step of two filtering regimes ('Standard' and 'Removing sex-linked loci') applied to the genetic datasets of eastern-yellow robin (EYR) and yellow-tufted honeyeater (YTH). Minor Allele Count filter was used as an extra filtering step before performing PCA and parentage analyses.

	EYR	**EYR**	**EYR**	**YTH**	**YTH**
		Standard	**Removing sex-linked loci**		**Standard**
Filter	**# individuals**	**# loci**	**# loci**	**# individuals**	**# loci**
No filtering	782	53,324	53,324	641	118,732
Secondaries	782	35,663	35,663	641	74,470
Sex-linked			31,856		
Excessively heterozygous			31,832		
Read depth	782	21,577	19,531	641	53,179
Locus missing data	782	13,972	12,899	641	16,481
Individual missing data	753	13,925	12,853	628	16,421
Minor Allele Count	753	13,693	12,626	628	14,908

Table 3. Number of sex-linked loci found in the genetic datasets of eastern yellow robin (EYR; 35,663 loci tested) and yellow-tufted honeyeater (YTH; 74,470 loci tested). In brackets is the proportion that loci represent with respect to the total number of loci tested.

	W-linked	Sex-biased	Z-linked	Gametologs	Total
EYR	146 (0.4%)	2,493 (7.0%)	827 (2.3%)	341 (1.0%)	3,807 (10.7%)
YTH	59 (0.1%)	2,220 (3.0%)	1079 (1.4%)	56 (0.1%)	3,414 (4.6%)

Table 4. Paired t-tests measuring the difference in individual observed heterozygosity (Ho) before and after removing sex-linked loci, per sex, in each species. Results are presented for eastern yellow robin (EYR) and yellow-tufted honeyeater (YTH). Significant p-values are signalled in bold letters.

Sex	Mean before	Mean after	% change	Μεαν Δ	Δ SD	t statistic	df	p-value	Co D
EYR	**EYR**	**EYR**	**EYR**	**EYR**	**EYR**	**EYR**	**EYR**	**EYR**	**EY**
F	0.188	0.188	-0.2%	-0.0003	0.0014	4.3	340	< **0.001**	0.2
M	0.187	0.187	-0.3%	-0.0005	0.0013	7.6	411	< **0.001**	0.3
YTH	**YTH**	**YTH**	**YTH**	**YTH**	**YTH**	**YTH**	**YTH**	**YTH**	**YT**
F	0.156	0.162	3.8%	0.0060	0.0007	-145.6	280	< **0.001**	-8.
M	0.164	0.159	-2.9%	-0.0047	0.0024	35.8	338	< **0.001**	1.9

Table 5. t-tests measuring the difference in individual observed heterozygosity (Ho) between females and males. Tests were done before and after removing sex-linked loci, for eastern yellow robin (EYR) and yellow-tufted honeyeater (YTH). Significant p-values are signalled in bold letters.

	Mean Females	Mean Males	s.e. Females	s.e. Males	t statistic	df	p-value	Cohen D
EYR	**EYR**	**EYR**	**EYR**	**EYR**	**EYR**	**EYR**	**EYR**	**EYR**
Before	0.188	0.187	0.001	0.001	0.8	748.6	0.4	0.05
After	0.188	0.187	0.001	0.001	0.9	749.9	0.4	0.06
YTH	**YTH**	**YTH**	**YTH**	**YTH**	**YTH**	**YTH**	**YTH**	**YTH**
Before	0.156	0.164	0.001	0.001	-6.4	606.4	< **0.001**	-0.5
After	0.162	0.159	0.001	0.001	1.5	613.7	0.1	0.1

Table 6. Paired t-tests measuring the difference in the average number of COLONY runs (out of five) that identified the correct parent of an offspring before and after removing sex-linked loci. Results are presented for eastern yellow robin (EYR) and subspecies *cassidix* . Significant p-value is in bold.

Mean before	Mean after	% change	Mean Δ	Δ SD	t statistic	df	p-value	Cohen
EYR	**EYR**	**EYR**	**EYR**	**EYR**	**EYR**	**EYR**	**EYR**	**EYR**

Mean before	Mean after	% change	Mean Δ	Δ SD	t statistic	df	p-value	Cohen
3.83	4.26	11.2%	0.43	1.5	-3.029	118	**0.003**	-0.28
YTH (*cas-sidix*)	**YTH (*cas-sidix*)**	**YTH (*cas-sidix*)**	**YTH (*cas-sidix*)**	**YTH (*cas-sidix*)**	**YTH (*cas-sidix*)**	**YTH (*cas-sidix*)**	**YTH (*cas-sidix*)**	**YTH (*cas-sidix*)**
4.90	4.85	-1.0%	-0.05	0.4	0.81	39	0.421	0.13

Hosted file

`Figure_1.pptx` available at https://authorea.com/users/562169/articles/609421-easy-to-use-r-functions-to-separate-reduced-representation-genomic-datasets-into-sex-linked-and-autosomal-loci-and-conduct-sex-assignment