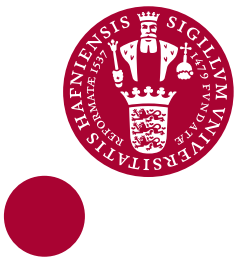

Joint identification of sex and sex-linked scaffolds in non-model organisms using low depth sequencing data

Nursyifa, Casia; Brüniche-Olsen, Anna; Garcia-Erill, Genis ; Heller, Rasmus; Albrechtsen, Anders

Published in:
Molecular Ecology Resources

DOI:
10.1111/1755-0998.13491

Publication date:
2021

Document version
Peer reviewed version

Joint identification of sex and sex-linked scaffolds in non-model organisms using low depth sequencing data

Casia Nursyifa[1*], Anna Brüniche-Olsen[1*], Genis Garcia Erill[1], Rasmus Heller[1$], Anders Albrechtsen[1$],

[1]*Section for Computational and RNA Biology, Department of Biology, University of Copenhagen, Denmark.*

[*] equal contribution

[$] corresponding authors

Anders Albrechtsen: albrecht@binf.ku.dk

Rasmus Heller: rheller@bio.ku.dk

Keywords

Autosomes, bioinformatics, resequencing, scaffold-level assembly

Abstract

Being able to assign sex to individuals and identify autosomal and sex-linked scaffolds are essential in most population genomic analyses. Non-model organisms often have genome assemblies at scaffold-level and lack characterization of sex-linked scaffolds. Previous methods to identify sex and sex-linked scaffolds have relied on synteny between the non-model organism and a closely related species or prior knowledge about the sex of the samples to identify sex-linked scaffolds. In the latter case, the difference in depth of coverage between the autosomes and the sex chromosomes are used. Here we present 'Sex Assignment Through Coverage' (SATC), a method

to assign sex to samples and identify sex-linked scaffolds from next generation sequencing (NGS) data. The method works for species with a homogametic/heterogametic sex determination system and only requires a scaffold-level reference assembly and sampling of both sexes with whole genome sequencing (WGS) data. We use the sequencing depth distribution across scaffolds to jointly identify: i) male and female individuals and ii) sex-linked scaffolds. This is achieved through projecting the scaffold depths into a low-dimensional space using principal component analysis (PCA) and subsequent Gaussian mixture clustering. We demonstrate the applicability of our method using data from five mammal species and a bird species complex. The method is freely available at https://github.com/popgenDK/SATC as R code and a graphical user interface (GUI).

Introduction

The increasing number of non-model organism scaffold-level genome assemblies provides new information on biodiversity and how evolutionary processes have shaped it (Ellegren 2014). An essential part of genome assembly and annotation is the identification of autosomes and sex chromosomes. Vertebrate species are generally diploid with the majority of their genome represented by a variable number of autosomes and two sex chromosomes (Graves 2008).

In mammals, the homogametic sex is the female (XX) and the heterogametic sex is the male (XY). This is opposite in birds, where males are homogametic (ZZ) and females heterogametic (ZW). Due to their inheritance, sex chromosomes differ from autosomes in several aspects of their population genetics and molecular evolution, e.g. by having a smaller effective population size (N_e) than autosomes (Ellegren 2009) and by having different patterns of population differentiation, especially under incipient or complete speciation (Presgraves 2018). Therefore, it is often preferable to separate them from autosomes in population genetic analyses.

Ideally, whole genome assemblies should be at chromosome-level and fully annotated, but due to the high cost and challenges associated with complete genome assembly this is often not prioritized for the first generation of a reference genome (Ellegren 2014). Consequently, several approaches to identify sex chromosomes in scaffold-level assemblies have been developed (for a review see Palmer *et al.*, (2019)). One of them is whole genome synteny alignment (Grabherr *et al.*2010), where the scaffold-level genome assembly is aligned to a chromosome-level assembly from a closely related species, and sex-linked scaffolds are identified based on sequence similarity to the reference. There are several obstacles to this approach, most importantly the availability of a chromosome-level assembly of a closely related species, but also the accelerated evolution of sex chromosomes in many lineages which causes a high degree of divergence even for closely related species (Charlesworth *et al.*2018; Irwin 2018; Meisel & Connallon 2013; Presgraves 2018), and computational time (Pennell *et al.*2018). Moreover, neo-sex chromosomes are likely to be missed due to their partial or complete synteny with autosomes in the closely related species genome (Graves 2008).

An alternative approach is to use genome depth of coverage (DoC) based methods. These are represented by two groups of methods, Y-linked and X-linked. Both methods require prior sexing information of individuals, which can be challenging to obtain for species with limited sexual dimorphism, for cryptic species and in non-invasive sampling situations. Y-linked scaffolds can be identified based on their low DoC compared to the autosomes, due to the expectation of ½×autosomal DoC in samples of the heterogametic sex (Hall *et al.*2013). X-linked scaffolds can be identified when sequencing reads from both sexes are mapped to a reference, with the expectation that the homogametic-linked scaffolds will have 1×autosomal DoC in the homogametic sex and ½×autosomal DoC in the heterogametic sex (Ellegren 2009, Graves 2008).

Due to noise in next generation sequencing (NGS) data, scaffold DoC distributions often deviate from this expectation. For example, autosomal and sex-linked scaffolds can have overlapping DoC distributions, making it challenging to clearly identify individual X and Y scaffolds (Malde *et al.*, 2019). The DoC approaches are furthermore highly sensitive to parameters used for the read mapping, pre- and post-mapping filtering steps and data quality, e.g. whether or not repetitive regions are removed, the average genome-wide DoC, etc. (Smeds *et al.*, 2015).

Here we present 'Sex Assignment Through Coverage' (SATC): a method and software to jointly identify sex-linked scaffolds and determine the sex of each sample mapped to a scaffold-level assembly. The method requires sequencing depth information from whole genome resequencing of male and female samples and applies principal component analysis (PCA) and Gaussian mixture clustering to group the dataset into males and females. Hence, the method harnesses the systematic—but noisy—difference in DoC of the heterogametic sex-linked scaffolds compared to the autosomal scaffolds. To illustrate how the method works we applied it to five mammal species and a bird species complex with different levels of DoC, assembly quality (N50) and sample sizes. Our method is very fast with computational time being less than a minute for 100 samples. We also evaluated the quality of input data required for SATC to work by subsampling our original data in various ways. We anticipate that it will be widely useful for inferring individual sex and identification of sex-linked scaffolds for non-model organisms.

Materials and methods

Method

The inputs for SATC are scaffold lengths and mapping statistics (e.g., number of reads mapped to each scaffold for each sample). These are quickly generated with the 'idxstats' command in

SAMTOOLS (Li *et al.*, 2009) from indexed bam files. The SATC method works by: i) normalizing the DoC of each scaffold within each sample, ii) reducing the dimensionality of the normalized DoC using PCA, iii) clustering the samples using Gaussian mixture clustering on the top PCs, and iv) identifying the sample sex and sex-linked scaffolds from the clustering and the DoC

In the first step, we calculate for each sample the average DoC per scaffold and normalize it by the mean DoC of the *M* longest scaffolds in the reference assembly. Suppose we have *s*= 1, 2, …, *S* scaffolds, *n*= 1,…, *N* samples and a matrix $R \in \mathbb{N}^{SxN}$containing the number of reads mapped to each scaffold for each sample. For simplicity we here assume that the scaffolds are ordered by length with scaffold 1 being the largest; however, this is not a requirement when using the method. From this we generate a matrix of normalized mean DoC for each scaffold and sample

$$D_{sn} = \frac{R_{sn}}{l_s} \times \frac{\sum_{i=1}^{M} l_i}{\sum_{i=1}^{M} R_{in}}$$

where R_{sn} is the number of mapped reads on scaffold *s,* sample *n,* l_s is the length in bases of scaffold *s,* R_{in} is the number of mapped reads on scaffold i and l_i is the length in bases of scaffold *i* , i=1, 2, ..., *M* longest scaffolds. In our default implementation we set *M* equal to 5 as we found it to give robust results, but this number can be changed by the user. We choose to use more than one scaffold to allow SATC to give warnings if one of the normalization scaffolds has a substantially different DoC than the others, which can alert the user that the scaffold should not be used in the normalization—it could for example be a sex-linked scaffold. Using multiple scaffolds for the normalization also provides the option to normalize by the median instead of the mean DoC if there is a depth difference between scaffolds. After this normalization, most scaffolds will have a normalized DoC close to 1. Following preliminary analyses, we found that filtering out scaffolds < 100kb in length and those with a mean normalized DoC outside the range of 0.3 – 2.0 improved

the performance of the method, but these thresholds can be changed by the user. We then center the matrix by subtracting the mean normalized DoC from that of each scaffold

$$\widetilde{D_{sn}} = D_{sn} - (\sum_{n=1}^{N} D_{sn} / N)$$

and perform PCA on the $\widetilde{\boldsymbol{D}}$ matrix. We proceed by identifying two clusters of samples that tentatively represent different sex groups (e.g. male and female), on the assumption that a large proportion of DoC variance is attributable to differences between males and females in the DoC of reads mapping to sex-linked scaffolds. For this we use the first two principal components as input for a clustering analysis using Gaussian finite mixture models in *mclust* (Scrucca *et al.*, 2016). Mclust has several models and the default option chooses the model that gives the lowest Bayesian Information Criterion (BIC). We chose a model with ellipsoidal distribution, equal volume and variable shape and distribution (EVV), that assumes equal variance in the two groups. Sometimes, this model fails to identify groups and in that case SATC will choose the model with the lowest BIC.

To identify the sex-linked scaffolds we use the two groups identified by the above clustering. We first apply a *t*-test assuming unequal variance for each scaffold to test for significant differences in mean DoC between the groups. We use a Bonferroni corrected *p*-value cut off of $0.05/S$ to identify the sex-linked scaffolds, and scaffolds identified with this test are broadly referred to as sex-linked in the following. In the *t*-test we assume normally distributed normalized DoC within each group, which might be violated for some datasets. We then calculate the average DoC for each scaffold in each group. The heterogametic (XY/ZW) and homogametic (XX/ZZ) sexes are expected to show a mean normalized DoC ratio of 0.5:1 between the two groups for scaffolds situated on the X/Z chromosome. Therefore, to identify scaffolds that are highly likely to be situated on the X/Z chromosome, we retain sex-linked scaffolds identified with the above

method for which the mean difference of normalized DoC between two sex groups is between 0.4 and 0.6. We refer to these scaffolds as X/Z-linked.

For the datasets analyzed in our study SATC performed well without any filtering of the mapped reads or scaffold contents. However, if the user wishes to add additional filters, for example to exclude repetitive regions of the genome or remove N's in the genome assembly, then the input file for the method is easily customizable. All analyses were done in R (R Core Team 2019) and are freely available at https://github.com/popgenDK/SATC as R code and a graphical user interface (GUI)..

Application to empirical datasets

To test our method we used six mammal and bird whole genome sequencing (WGS) datasets with low to medium DoC and mapped to scaffold-level assemblies. We used unpublished WGS for impala (*Aepyceros melampus*), muskox (*Ovibos moschatus*), waterbuck (*Kobus ellipsiprymnus*) and gray whale (*Eschrichtius robustus*). WGS for leopard (*Panthera pardus*) were from Pečnerová *et al.,* (2021) and we used only the 49 individuals that passed their QC. The scaffold assemblies for the waterbuck, impala and muskox were from the ruminant genome project (Chen *et al.,* 2019), the leopard were from Kim *et al.,* (2016) while the gray whale assembly is unpublished. WGS data from Darwin's finches species complex represents 15 species (Lamichhaney et al., 2015; Lamichhaney et al., 2016)—the mangrove finch (*Camarhynchus heliobates*), the woodpecker finch (*Camarhynchus pallidus*), the small tree finch (*Camarhynchus parvulus*), the medium tree finch (*Camarhynchus pauper*), the large tree finch (*Camarhynchus psittacula*), the gray warbler-finch (*Certhidea fusca*), the green warbler-finch (*Certhidea olivacea*), the Española cactus finch (*Geospiza conirostris*), the sharp-beaked ground finch (*Geospiza difficilis*), the medium ground

finch (*Geospiza fortis*), the small ground finch (*Geospiza fuliginosa*), the large ground finch (*Geospiza magnirostris*), the common cactus finch (Geospiza scandens), the Cocos finch (*Pinaroloxias inornate*), and the vegetarian finch (*Platyspiza crassirostris*). We also included two related tanager species; the black-faced grass-quit (*Tiaris bicolor*) and the lesser Antillean bullfinch (*Loxigilla noctis*). All finches species were mapped to the medium ground finch assembly (Zhang *et al* 2014).

The sequencing data from the six different datasets were generated as part of separate studies and were therefore filtered and pre-processed in different ways (for a detailed description see Supplementary Information Text S1). Hence, their heterogeneity allows us to assess whether our method is broadly applicable across a range of data treatment regimes, representing the variety of pipelines used in practice for non-model sequencing analysis.

Validation of sexing and sex-linked scaffolds

To evaluate the sensitivity of the sample sex assignment we mapped the read data from the five mammal species to a closely related, well-annotated chromosome-level reference genome that included an annotated X chromosome. We did not include the Y chromosome because it was not available in all reference genomes and is in general much harder to assemble. For the impala we mapped to the goat (ARS1), for the leopard to the domestic cat (Felis_catus_9.0), for the waterbuck we used the cow (bosTau8), for the muskox we used the sheep (Oar_rambouillet_v1.0), and for the gray whale we used the blue whale (mBalMus1.pri.v3). Based on this cross-species mapping we calculated the normalized DoC of X-mapping reads for each individual by the DoC of reads mapping to the five largest autosomal chromosomes of the same external reference genome as an external evaluation of the SATC accuracy in assigning the correct sample sex. As 16 of the 17

species in the Darwin's finches dataset were mapped to an outgroup (the medium ground finch) we did not explore mapping this dataset to further outgroups. To evaluate if evolutionary divergence among taxa would influence our method we ran SATC on the finch data with and without the more distantly related species, the black-faced grass-quit and the lesser Antillean bullfinch.

In addition, we also used an independent verification of sexing for the gray whale dataset where we had sex information from field observations (n = 37) and from single nucleotide polymorphism (SNP) genotyping of two markers in the zinc finger proteins of the *Zfx* and *Zfy* genes (n = 70) (DeWoody *et al.,* 2017; Brüniche-Olsen *et al.,* 2018).

Finally, we used LASTZ (Harris 2007) to explore the degree of synteny between the SATC inferred sex-linked scaffolds and sex chromosomes from the closest relative with a chromosome-level assembly. For this we focused on the gray whale and Darwin's finches datasets, as these were the taxa for which the closely related, chromosome-level assembly was heterogametic (blue whale (mBalMus1.pri.v3) and chicken (bGalGal1.mat.broiler.GRCg7b), respectively). We ran LASTZ for each putative sex-linked scaffold using a low-sensitivity alignment with the following settings --notransition --step=20 --nogapped.

Testing the limitations of SATC performance under extreme data conditions

Three main factors are likely to influence the performance of SATC: i) overall sequencing data DoC, ii) the number and sex distribution of samples, and iii) reference genome assembly quality. We therefore tested the robustness of the SATC results to factors i and ii, while the impact of factor iii was implicitly evaluated by using species with highly variable reference genome qualities (Table 1).

To evaluate the robustness of our method for low DoC data we downsampled the sequencing reads to a DoC of 1X, 0.5X, 0.1X and 0.01X. Downsampling for each sample in each scaffold was done by randomly subsampling reads to achieve the desired DoC directly from the idxstats files. We then compared the inferred sex from these lower DoC scenarios with their externally validated sex assignment found by using the full dataset. We also compared the inferred sex-linked scaffolds and their combined length to the sex-linked scaffolds identified using the full, non-downsampled dataset.

To assess the accuracy of sexing in scenarios with large sex ratio imbalances, we performed stratified downsampling for each sex group, reducing each group to 75%, 50%, 25% and 10% of their original sample size and also downsampled each data set to only 2 samples for each sex. To evaluate the inferred sex, we visualized the normalized DoC of sex-linked scaffolds for each sample using boxplots where the median should be a half for the heterogametic and one for the homogamatic. We also visualize the PCA plot with uncertainty profiles calculated from *mclust* in which uncertainty is measured as probability belonging to its group and the two sex groups clusters are visualized by two ellipsoids as scaled variances of inferred Gaussian mixture components.

Results

We analyzed whole-genome resequencing data from five mammal species and one bird species complex. The WGS data was mapped to the scaffold-level genome assembly from the same species. The datasets varied both in terms of quality of assembly and in DoC, which ranged from 3.13–13.76X. As shown in Table S1 the quality of the genome assembly varied between species, ranging from a scaffold N50 of 344kb in impala to 46.8Mb in muskox and each genome assembly contained 2,796–88,935 scaffolds. These assemblies are representative for many low to medium

quality draft genomes. Even after removing scaffolds <100kb, a high number of scaffolds remained for some species i.e., up to 7,717 for the impala (Table S1).

The normalized DoC was very noisy across scaffolds for most species (Figure S1). However, when we performed a PCA on the DoC matrix we observed a clear separation into two groups for all taxa (Figure 1, left column). All taxa—except the impala—separated in two distinct groups based on PC1; for impala this partitioning was on PC2. After applying Gaussian finite mixture models clustering on the two first axes of the PCA we could clearly group the samples from each taxa into two groups with characteristic and distinct normalized DoC patterns. We interpret these two groups as the homogametic and heterogametic sex, with the homogametic sex being the one with the highest DoC of the sex-linked scaffolds (Figure S2).

To identify sex-linked scaffolds we performed a *t*-test for each scaffold testing differences in mean DoC between the two sex groups identified above. This test identified 54-589 sex-linked scaffolds across the six taxa, with by far the highest number found in the two most challenging data sets from the impala and waterbuck (Table 1 and Figure S2). These sex-linked scaffolds might not be exclusively from a sex chromosome, and we therefore define them loosely as sex-linked. We furthermore identify X/Z-linked scaffolds by retaining only those sex-linked scaffolds that had a difference of 0.4–0.6 between the mean normalized DoC in the heterogamatic and homogametic individuals, respectively. This yielded between 11 and 113 X/Z-linked scaffolds in each species (Table 1 and Figure 1, right column). The total length of sex-linked scaffolds was 126–187 Mb across the five mammals and 79 Mb in the bird species complex, which is close to the length of the assembled X/Z chromosomes for the close relative of each species (Table 1).

Many of the sex-linked scaffolds that do not conform to the X/Z-linked criterion were scaffolds with DoC that are correlated with sex, but do not adhere to the 0.5:1 ratio between the

heterogametic and homogametic sexes. Some of these scaffolds show large and unexplained deviations from the autosomal DoC, while some show an expected Y/W ratio of 0.5:0 (Figure S2, muskox and Darwin's finches). Most of the taxa had a small number of scaffolds that were assigned as sex-linked but not X/Z-linked, however the two species with the most fragmented assemblies, impala and waterbuck, showed a much larger difference between the cumulative length of sex-linked and X-linked scaffolds (Table 1). For the impala, the cumulative inferred X-linked scaffold length was just 1.4% of the sex-linked scaffold length, and for the waterbuck this was 24.8%, whereas the corresponding numbers were 79.4-98.4% for the remaining species (Table 1). We therefore conclude that our sex-linked scaffold identification method likely identifies the majority of the X/Z-linked scaffolds while our stricter ratio-based criterion will not identify most of the X/Z-linked scaffolds when the reference genome quality and/or resequencing DoC is low.

Verification of SATC sex assignment

To validate the inferred sex for the samples, we mapped the five mammal species reads to a closely related reference genome containing an annotated chromosome-level assembly, and used the normalized DoC of reads mapping to the X chromosome to classify samples as heterogametic or homogametic, respectively. This validation showed 100% agreement with the SATC inferred sex in all cases and across all species (Figure S3). The Darwin's finches dataset represented 15 closely related species and two more distantly related species. We found that even with such a heterogeneous dataset, SATC was able to identify the sample sex and the sex-linked scaffolds regardless of whether we included the most distantly related species in the complex (Figure S4).

We further verified the SATC sample sex assignment using SNP-genotyping and field observations. The SNP-genotyping and field observations for the gray whale dataset showed high

concordance (Table S2). We found a 100% agreement between the SATC results and field observations. We identified one misclassification in the SNP-genotyping, a female identified by SATC and field observations, but misidentified as a male with genotyping, probably due to a sample mixup in the SNP-genotyping or sample contamination.

We evaluated the synteny of the inferred sex-linked scaffolds with known sex chromosomes by aligning the inferred sex-linked scaffolds to XY and ZW chromosomes from closely related species and visualizing the similarity across the sex chromosomes (Figure S5 and S6). For both the gray whale and Darwin''s finches we found a higher percentage of scaffold sequence aligning to the X/Z chromosomes (X mean aligned sequence % = 62.9 SD = 15.5 and Z mean aligned sequence % = 6.0; SD = 3.4) than the sex-limiting Y/W chromosomes (Y mean aligned sequence % = 1.2; SD = 0.6 and W mean aligned sequence % 1.9; SD = 2.9). Generally, the percentage of aligned sequence was higher in the gray whales than in Darwin's finches, as illustrated by the low amount of aligned sequence % in Darwin's finches (Figure S5+S6). However, as seen in Fig. S5 and Fig. S6 the majority of the X/Z chromosomes of the closely related reference genomes had clear synteny with one of the sex-linked scaffolds identified by SATC, confirming that we have succesfully identified the majority of the X/Z-linked scaffolds in our study taxa.

SATC performance in challenging data situations

We evaluated how the method performed for: i) low DoC data, ii) data with a low number of samples, and iii) data with imbalanced sex ratios. By downsampling sequencing reads we showed that SATC still correctly assigned sample sex for all taxa down to 0.1X. For five of the datasets, SATC also worked at a very low DoC of 0.01X. However, at this extremely low DoC the method failed to assign the correct sex for the impala dataset. (Figure S7). Moreover, we inferred the same set of sex-linked scaffolds at lower DoC for most datasets except the impala and the waterbuck, in

which SATC missed many of the sex-linked scaffolds at 0.1X and 0.01X. (Table S3). SATC also remains highly accurate in assigning sex with either a low total number of samples or with highly unbalanced sex ratios among the input samples. For the gray whales and leopard the method correctly identified the sexes both at very low samples sizes (n_{XX}=2, n_{XY}=2) and with extreme sex ratio imbalance (e.g. n_{XX}=47, n_{XY}=3). For the impala, Darwin's finches, muskox and waterbuck SATC performed well for a range of scenarios, but failed at sex ratio imbalances of 81:8, 5:126, 5:56, and 5:20 respectively (Figure S8). For all of the failed scenarios we detected issues with both sample DoC clustering and with the DoC distribution on the inferred sex-linked scaffolds, both of which are clear indications of poor performance (Figure S9). The PCA plots for these failed SATC analyses illustrate how the clustering can diagnose overlapping or low separation of the two ellipsoidals representing the two putative sex groups. We also observed that some individuals could not be assigned to a cluster with high certainty, which is illustrated in the diagnostic PCA plots by large point sizes. The corresponding boxplots of normalized DoC in all the SATC inferred sex-linked scaffolds clearly do not follow the expected normalized DoC of 0.5 for the heterogametic and 1.0 for the homogametic sex (Fig. S9). For comparison, the corresponding evaluation plots in the situations where SATC works well are shown in Figure S10.

We also evaluated the effect of using sex-linked scaffolds in the DoC normalization using the muskox data as an example. We did not see any effect of including the largest inferred sex-linked scaffold along with the four largest non-sex-linked scaffolds in the DoC normalization. We also tested the more extreme case of using the largest non-sex-linked scaffold and the 20 largest sex-linked scaffolds. This resulted in correct clustering and assignment of sex (Figure S11). However, the method now identified the autosomal scaffolds as sex-linked scaffolds, with most of them being categorized as having abnormal DoC ratios. We believe that this latter case is extremely

unlikely to occur in a real analysis and it can be avoided easily by inspecting the DoC plots given by the software. In addition the software gives several warnings to the user that the chosen normalizing scaffolds have very different DoCs and suggests to the user to change the normalization scaffolds or use the median DoC instead.

Discussion

The increasing amount of whole genome resequencing data presents new avenues for population genomic analyses. Here we add to the analytical toolset by introducing SATC, a method for joint individual sex assignment and identification of sex-linked scaffolds from NGS data in species with a homogametic/heterogametic sex-chromosome system. Our method is automated, computationally light, robust to pre-mapping filtering and has a high accuracy even with challenging data. We anticipate this will be particularly useful for non-model organisms and for samples collected in the field, where information on the sex of individuals and a chromosome-level assembly is often lacking.

The benefits of identifying sex-linked scaffolds when carrying out population genetic studies are many. First, sex chromosomal sites may be desirable for specific analyses, such as association (Lee *et al.*2017; Luciano *et al.*2019; Zuo *et al.*2013), gene expression (Grath & Parsch 2016), or any evolutionary genetic studies on X/Y or Z/W chromosomes (Gottipati *et al.*2011). Second, if sex-linked scaffolds are not flagged and treated separately they may bias analyses such as demographic history inference (Li & Durbin 2011), genome scans or genome-wide values of summary statistics, including F_{ST} (Lambert *et al.*2010), genetic diversity (Hammer *et al.*2010) and allele frequency distribution (Clayton 2008). Third, analyzing males and females separately can

elucidate patterns of sex-biased dispersal (Bidon *et al.*2014) or unequal contributions to offspring diversity (Pérez-González *et al.*2014).

We show here that PCA on normalized scaffold DoC is a robust approach to identify individual sex for a range of data situations, including having only low-depth resequencing data and a low quality draft assembly as a reference genome. Having assigned sample sex we can easily reverse the perspective and utilize this information to identify which scaffolds are sex-linked by exploiting the expected 0.5:1 ratio between heterogamatic and homogamatic DoC for each X/Z linked scaffold. Our SATC approach does not rely on prior knowledge of sample sex or sex-linked scaffolds, and is, to our knowledge, the only automated software that accomplishes this without any external information. The recommended usage of SATC would be to flag all scaffolds identified as sex-linked and remove them from further analyses that assume autosomal chromosome data. Conversely, if X or Z-linked sites are desired, we recommend to include only those that are flagged as X/Z-linked i.e., approximately follow the expected 0.5:1 ratio between DoC in the heterogamatic and homogamatic sex when compared between the two inferred groups of same-sex samples. Note that it is much harder to identify Y/W-linked scaffolds due to the smaller size of these chromosomes and their highly repetitive sequence content. This leads to a higher occurrence of misassembled Y/W scaffolds, which can distort the DoC. Despite this, some of the sex-linked scaffolds we identify show DoC patterns consistent with being Y/W-linked, having approximately 0.5 normalized DoC in males and very low (virtually zero) DoC in females (Figure S2).

We show that highly fragmented genome assemblies can be used in SATC, albeit with reduced performance of the sex-linked scaffold identification. The two examined species with the lowest reference genome scaffold N50s, impala and waterbuck, showed deviating patterns from

the rest by having very noisy scaffold DoC (Figure S1). The most obvious explanation for this is poor assembly quality, e.g. regions from both sex chromosomes and autosomes could be erroneously joined in the same scaffold. The impala had the lowest-quality genome assembly (N50 = 344kb), and for this species we found that the grouping of sexes occurred in PC2 rather than PC1 (Figure 1). Despite this, our clustering approach was still able to assign sample sex with 100% accuracy. In addition to lower assembly quality, biological factors could also influence the ability of SATC to correctly identify sex-linked scaffolds. For example, impalas are known to have segregating karyotypic polymorphisms (Pagacova *et al.*2011), which could potentially influence the depth across scaffolds and exacerbate the noise in DoC. The waterbuck had a higher than expected amount of sex-linked scaffold content with about 40Mb more content than the X chromosome of the cow reference genome. This could again be influenced by karyotypic polymorphisms, which are known to occur both within and between different subspecies of waterbuck (Kingswood *et al.*1998). Autosome-to-X translocations are known from several species of bovids (Effron *et al.*1976; Gallagher Jr & Womack 1992; Kumamoto *et al.*1996), and if such are segregating within our samples they would complicate the DoC-based identification of sex-linked scaffolds. We also observed a large difference between the total amount of sex-linked scaffold content identified by the DoC ratio for these two species, whereas this was much smaller for the other species (Table 1), confirming that excessive noise in scaffold DoC can challenge the use of hard thresholds for identifying sex-linked scaffolds.

Our SATC method also works well with heterogeneous datasets. Darwin's finches encompass around 18 species of passerine birds (Grant & Grant 2020). We analyzed 15 of these species, which diverged during the last 150,000-900,000 years but still have some degree of interspecies gene flow (Lamichhaney *et al.*2015). Despite the heterogeneous data we were able to

assign sex and to identify sex-linked scaffolds in the medium ground finch reference genome assembly when read data from across these species were analyzed in SATC together. We extended the Darwin's finches dataset with two more distantly related (>900ky) tanager species, the black-faced grassquit and lesser Antillean bullfinch (Lamichhaney *et al.*2015), and show that the PCA clustering method was still able to reliably assign sample sex as well as identify sex-linked scaffolds (Figure S4).

Finally, we found that there was a high degree of synteny between the identified sex-linked scaffolds and regions on the sex chromosomes in the closest chromosomal assembly for both mammals and birds (Figure S5 and S6). The synteny was apparent despite a relatively low mean sequence identity (< 6%) between the chicken sex chromosomes and the medium ground finch sex-linked scaffolds. This illustrates that in cases where there is high divergence (i.e., low sequence similarity, chromosome degradation), synteny based approaches alone might have challenges. SATC has an advantage over synteny-based approaches because it can differentiate between sex-linked scaffolds on actual sex chromosomes and on neo-sex chromosomes. The fusion between a sex chromosome and an autosome generates a neo-sex chromosome, and these can not be distinguished from the true sex chromosomes through synteny. Therefore, SATC is expected to perform better at identifying sex-linked scaffolds than synteny-based methods in taxa with more dynamic sex-chromosomes (i.e. non-homeotherms) where neo-sex chromosomes are fairly common.

Under extreme data conditions we showed that SATC still works well in both sample sex assignment and sex-linked scaffolds identification. Ultra-low DoC (0.01X) is not an issue for SATC when genome assembly quality is reasonably good, in our analyses represented by the muskox, leopard, gray whale, and Darwin's finches cases (Figure S7). However, SATC did fail in

some scenarios with large sex imbalance or with a very low number of individuals (Figure S8). The amount of sex imbalance at which SATC no longer performed well varied according to genome assembly quality. In the better-quality scaffold-level assemblies from the four taxa mentioned above, SATC performed well up to a sex imbalance of 1:10, while for the highly fragmented assemblies from impala and waterbuck SATC failed at 1:4. In failed scenarios the parameters in the model can be tweaked for better performance, as we have for now focused on identifying generic settings that work well across many cases. Importantly, the user can easily evaluate SATC performance using a PCA plot with uncertainty profile and individual DoC plot (Figure S9 and S10). If these diagnostic plots show unsatisfactory results, the user should assume that SATC has not performed well and the results should not be trusted. In this case, the user can try either to change the scaffolds used for normalization, normalization method (e.g. using median instead of mean) or change the clustering method where we also have an option to perform hierarchical clustering instead of Gaussian mixture model-based clustering.

We explored SATC's performance on a range of genomic datasets from mammals and birds. However, we did not test how well the method works for taxa with highly degenerated sex-chromosomes or taxa with more homomorphic sex chromosomes. We also did not perform extensive testing of the methods on reduced-representation-sequencing (RRS) datasets, but preliminary analyses showed that the noisy DoC distributions across individual RRS loci and possible allelic dropout (Heller et al., 2021) make such data challenging for SATC without further attention to pre-filtering, e.g. on missingness. We consider it outside the scope of this study to investigate how SATC could be improved to take the peculiarities of RRS data into account. We emphasize that our method works on WGS data without any need for sophisticated filtering. For example, we did not exclude repeat annotated regions or remove regions without mapped reads

prior to calculating the DoC in any of the species datasets. It is possible that additional filtering of the data could improve the identification of sex-linked scaffolds in some cases, but we focused instead on demonstrating the robustness of the method by showing that it works in challenging data situations. We found that a single set of settings for the different cutoff values—minimum scaffold length, maximum DoC, ratio of homogametic/heterogametic scaffold DoC—yielded usable results for all the species analyzed here. However, the SATC software allows the user to modify these settings if needed. We encourage users to try different cutoff settings to assess the sensitivity of the analyses.

Acknowledgements

ABO was supported by a Carlsberg Foundation Reintegration Fellowship (CF19-0427). RH, GGE and CN were supported by a DFF Sapere Aude research grant (DFF8049-00098B), and RH was furthermore supported by an ERC Starting Grant (No 853442). AA and GGE are supported by the Lundbeck foundation (R215-2015-4174). AA is supported by the Novo Nordisk Foundation (NNF20OC0061343). We thank Jonas Meisner for clustering input suggestions. We thank John W. Bickham and J. Andrew DeWoody for providing gray whale data, Peter van der Wolf for gray whales photograph, and Patrícia Chrzanová Pečnerová for providing muskox summary data. We also thank Kristian Ebbesen Hanghøj for helping with writing R code implementation and the PopGen group at University of Copenhagen for helpful comments on previous versions of the manuscript.

Data accession

The datasets analyzed in this study are available at the European Nucleotide Archive under the BioProject accession codes: leopard (PRJEB41230), waterbuck (PRJEB28089) and Darwin's Finches is on short read archive (PRJNA263122 and PRJNA301892). Idxstats files are available for all taxa (leopard, waterbuck, Darwin's finches, gray whale, impala and muskox) at https://github.com/popgenDK/SATC/tree/main/examples/idxstats. The genome assemblies used were downloaded from NCBI for goat (*Capra hircus*, ARS1, GCA_001704415.1), domestic cat (*Felis catus*, Felis_catus_9.0, GCA_000181335.4), cow (*Bos taurus*, bosTau8, GCA_000003055.4), sheep (*Ovis aries*, Oar_rambouillet_v1.0, GCA_002742125.1), blue whale (*Balaenoptera musculus*, mBalMus1.pri.v3, GenBank assembly accession GCA_009873245.2), the medium ground finch (*G. fortis*, GCF_000277835.1_GeoFor_1.0), impala (*Aepyceros melampus*,IMP GCA_006408695.1), leopard (*Panthera pardus*, PanPar1.0, GCA_001857705.1) and waterbuck (*Kobus ellipsiprymnus*, DFW, GCA_006410655.1). Rasmus Heller contributed the unpublished muskox and impala data. The unpublished gray whale data is contributed by Anna Brüniche-Olsen. The software framework is freely available at https://github.com/popgenDK/SATC.

Author contributions

The work was conceived by CN and AA. The research design was planned by CN, ABO, GGE, RH and AA. The data was analyzed by CN and GGE. CN, ABO and RH wrote the manuscript with input from GGE and AA. All authors read and approved the final version of the manuscript.

References

Bidon T *et al.,* (2014) Brown and polar bear Y chromosomes reveal extensive male-biased gene flow within brother lineages. *Molecular Biology and Evolution* **31**, 1353-1363.

Brüniche-Olsen A *et al.,* (2018) Genetic data reveal mixed-stock aggregations of gray whales in the North Pacific Ocean. Biol. Lett. 14: 20180399.

Charlesworth B, Campos JL, & Jackson BC (2018) Faster-X evolution: Theory and evidence from Drosophila. *Molecular Ecology* **27**, 3753-3771.

Chen L *et al.,* (2019) Large-scale ruminant genome sequencing provides insights into their evolution and distinct traits. *Science* 364.6446.

Clayton D (2008) Testing for association on the X chromosome. *Biostatistics* **9**, 593-600.

Dewoody JA *et al.,* (2017) "Characterization of the gray whale *Eschrichtius robustus* genome and a genotyping array based on single-nucleotide polymorphisms in candidate genes." *The Biological Bulletin* 232.3: 186-197.

Effron M, Bogart M, Kumamoto A, & Benirschke K (1976) Chromosome studies in the mammalian subfamily Antilopinae. *Genetica* **46**, 419-444.

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

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

Gallagher Jr D & Womack J (1992) Chromosome conservation in the Bovidae. *Journal of Heredity* **83**, 287-298.

Gottipati S, Arbiza L, Siepel A, Clark AG, & Keinan A (2011) Analyses of X-linked and autosomal genetic variation in population-scale whole genome sequencing. *Nature Genetics* **43**, 741.

Grabherr MG *et al.,* (2010) Genome-wide synteny through highly sensitive sequence alignment: Satsuma. *Bioinformatics* **26**, 1145-1151.

Grant PR & Grant BR (2020) *How and why species multiply* Princeton University Press.

Grath S & Parsch J (2016) Sex-biased gene expression. *Annual Review of Genetics* **50**, 29-44.

Graves JAM (2008) Weird animal genomes and the evolution of vertebrate sex and sex chromosomes. *Annual review of genetics* 42: 565-586.

Hall AB *et al.*(2013) Six novel Y chromosome genes in Anophelesmosquitoes discovered by independently sequencing males and females. *BMC Genomics* **14**, 273.

Hammer MF *et al.,* (2010) The ratio of human X chromosome to autosome diversity is positively correlated with genetic distance from genes. *Nature Genetics* **42**, 830.

Harris R.S. (2007) Improved pairwise alignment of genomic DNA. Ph.D. Thesis, The Pennsylvania State University.

Heller R *et al.,* (2021). A reference-free approach to analyse RADseq data using standard next generation sequencing toolkits. *Molecular Ecology Resources*, *21*(4), 1085–1097.

Irwin DE (2018) Sex chromosomes and speciation in birds and other ZW systems. *Molecular Ecology* **27**, 3831-3851.

Kim S *et al.,* (2016). Comparison of carnivore, omni- vore, and herbivore mammalian genomes with a new leopard assembly. Genome Biol. *17*, 211.

Kingswood S, Kumamoto A, Charter S, Aman R, & Ryder O (1998) Brief communication. Centric fusion polymorphisms in waterbuck (Kobus ellipsiprymnus). *Journal of Heredity* **89**, 96-100.

Kumamoto A, Charter S, Houck M, & Frahm M (1996) Chromosomes of Damaliscus (Artiodactyla, Bovidae): simple and complex centric fusion rearrangements. *Chromosome Research* **4**, 614-621.

Lambert CA *et al.,* (2010) Highly punctuated patterns of population structure on the X chromosome and implications for African evolutionary history. *The American Journal of Human Genetics* **86**, 34-44.

Lamichhaney S *et al.,* (2015) Evolution of Darwin's finches and their beaks revealed by genome sequencing. *Nature* **518**, 371-375.

Lamichhaney S *et al.*(2017) X chromosome-wide association study identifies a susceptibility locus for inflammatory bowel disease in Koreans. *Journal of Crohn's and Colitis* **11**, 820-830.

Li H & Durbin R (2011) Inference of human population history from individual whole-genome sequences. *Nature* **475**, 493-U484.

Li H *et al.,* (2009) The sequence alignment/map format and SAMtools. *Bioinformatics* **25**, 2078 - 2079.

548 Luciano M *et al.*(2019) The influence of X chromosome variants on trait neuroticism.
549 *Molecular Psychiatry*, 1-9.
550 Malde K, Skern R, & Glover KA (2019) Using sequencing coverage statistics to identify
551 sex chromosomes in minke whales. *arXiv preprint arXiv:1902.06654*.
552 Meisel RP & Connallon T (2013) The faster-X effect: integrating theory and data. *Trends*
553 *in Genetics* **29**, 537-544.
554 Pagacova E, Cernohorska H, Kubickova S, Vahala J, & Rubes J (2011) Centric fusion
555 polymorphism in captive animals of family Bovidae. *Conservation Genetics* **12**, 71-77.
556 Palmer DH, Rogers TF, Dean R, & Wright AE (2019) How to identify sex chromosomes
557 and their turnover. *Molecular Ecology* **28**, 4709-4724.
558 Pečnerová P *et al.*, (2021) High Genetic Diversity and Low Differentiation Reflect the
559 Ecological Versatility of the African Leopard. *Current Biology: CB* 31 (9): 1862–71.e5.
560 Pennell MW, Mank JE, & Peichel CL (2018) Transitions in sex determination and sex
561 chromosomes across vertebrate species. *Molecular Ecology* **27**, 3950-3963.
562 Pérez-González J *et al.*, (2014) Males and females contribute unequally to offspring genetic
563 diversity in the polygynandrous mating system of wild boar. *Plos One* **9**, e115394.
564 Presgraves DC (2018) Evaluating genomic signatures of "the large X-effect" during
565 complex speciation. *Molecular Ecology***27**, 3822-3830.
566 Scrucca L, Fop M, Murphy TB, & Raftery AE (2016) mclust 5: clustering, classification
567 and density estimation using Gaussian finite mixture models. *The R journal* **8**, 289.
568 Smeds L *et al.*, (2015) Evolutionary analysis of the female-specific avian W chromosome.
569 *Nature Communications* **6**.

570 Team RDC (2019) R: A language and environment for statistical computing, Vienna,
571 Austria.
572 Zhang G *et al.*, (2014) Comparative genomics reveals insights into avian genome evolution
573 and adaptation. Science 346(6215):1311-20.
574 Zuo L *et al.*, (2013) Sex chromosome-wide association analysis suggested male-specific
575 risk genes for alcohol dependence. *Psychiatric Genetics* **23**, 233.
576
577

Table 1. Basic properties of the species analyzed with our SATC method. For each species the number of samples (*N*), depth of coverage (DoC), total number of scaffolds (#scaffolds), number of inferred X or Z scaffolds based on mean difference of normalized DoC being between 0.4 and 0.6 (#scaffolds X/Z), total numbers of sex-linked scaffolds based on t-test (#scaffolds sex-linked), length of inferred X/Z scaffolds (X/Z (Mb), total length of all identified sex-linked scaffolds (Mb), and the sex ratio for the samples. Inferred sex was estimated based on Gaussian mixture clustering from the top 2 PCs inferred from closely related species with a chromosomal level assembly. DoC is the average number of reads on each position in the genome calculated by summing all mapped reads multiplied by average read length and divided by total length of scaffolds.

Species	***N***	**DoC**	**#scaffolds**	**#scaffolds X/Z**	**#scaffolds sex-linked**	**X/Z (Mb)**	**Sex-linked scaffolds (Mb)**	**sex ratio (hetero/homo)**
Leopard	49	4.2X	50,378	60	66	113	126	30/19
Impala	113	3.1X	24,159	11	589	2.1	143	81/32
Muskox	103	11.4X	7,072	47	54	126	128	47/56
Gray whales	73	5.8X	2,796	39	62	104	131	46/27

Waterbuck	40	3.4X	88,935	113	400	46.3	187	20/20
Darwin's finches	172	13.8X	27,240	42	62	72.8	78.5	122/50

587

588

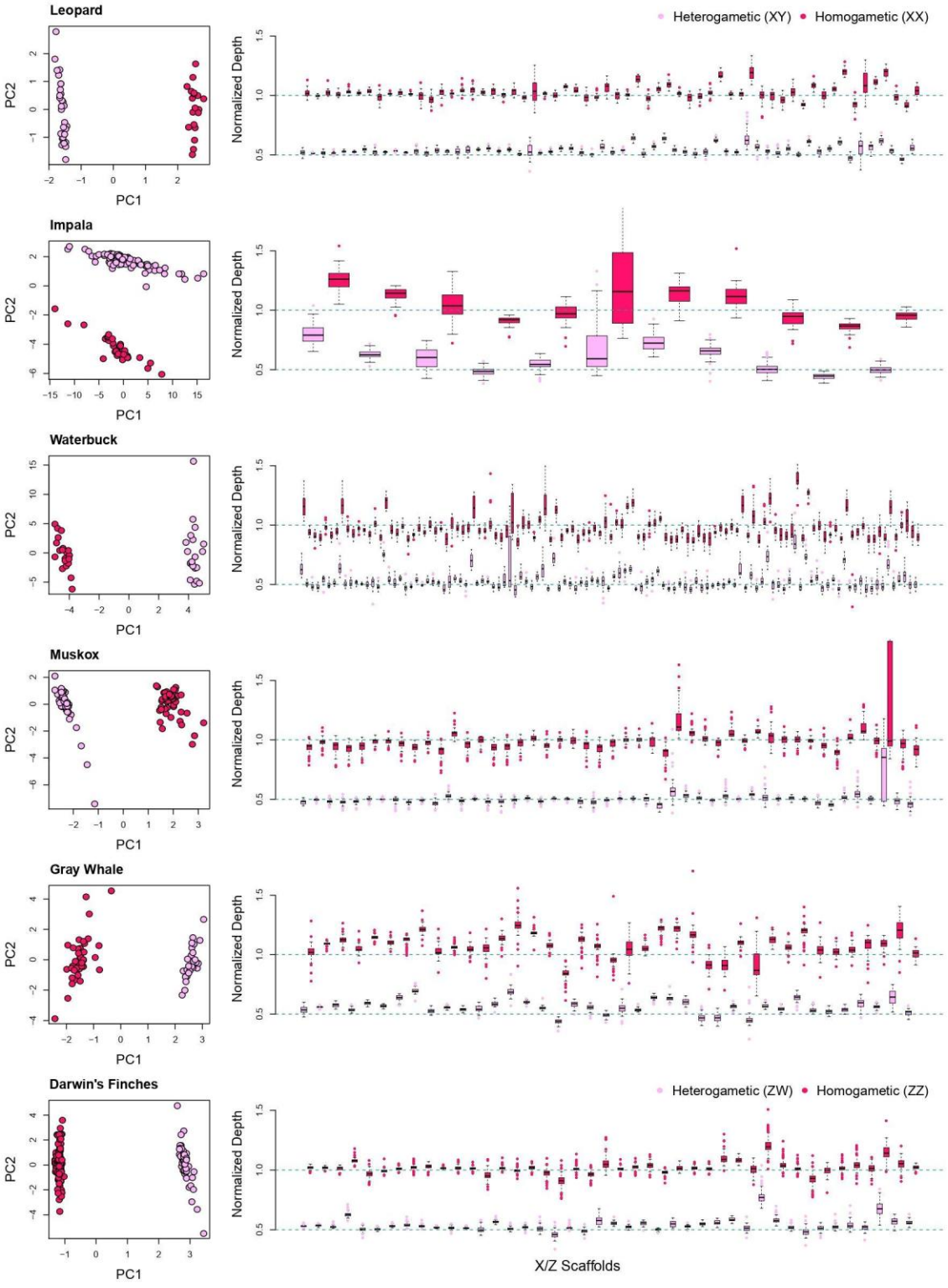
Leopard
Impala
Waterbuck
Muskox
Gray Whale
Darwin's Finches
PC1
PC2
Normalized Depth
Heterogametic (XY)
Homogametic (XX)
Heterogametic (ZW)
Homogametic (ZZ)
X/Z Scaffolds

Figure 1: PCA plots with Gaussian mixture clustering and boxplots.

Left column: PCA plots of normalized DoC across all scaffolds and samples from 5 mammalian species and the 15 species making up Darwin's finches species complex. Two clusters are inferred homogametic (dark pink) and heterogametic (light pink). Right column: Boxplot of normalized DoC from inferred X/Z scaffolds based on mean difference of two sex clusters within range of 0.4 and 0.6. Scaffolds are sorted based on their length (x-axis). Each scaffold is represented by two boxplots from homogametic and heterogametic groups. Expected median values for each group are shown by horizontal green dashed lines of 0.5 (heterogametic) and 1.0 (homogametic).

600