PhaseWY: A bioinformatic pipeline for phasing and retrieving Y and W sequences from population genomic data

SIMON JACOBSEN ELLERSTRAND & BENGT HANSSON | DEPARTMENT OF BIOLOGY, LUND UNIVERSITY

Introduction

In species with homologous sex chromosomes, retrieving sequence data from Y and W is challenging. This is unfortunate as such data would facilitate understanding interesting aspects of sex chromosome evolution such as:

- Neo-sex chromosome formation.
- Degeneration of heterogametic Y and W chromosomes.
- Dosage compensation on the X and Z chromosomes in response to Y and W degeneration and loss of function.
- The accumulation of sexually antagonistic mutations on X and Z.

Here we present PhaseWY, a bioinformatic pipeline that aims to identify and extracts Y and W sequences by phasing and clustering haplotypes from multiple females and males.

The PhaseWY pipeline

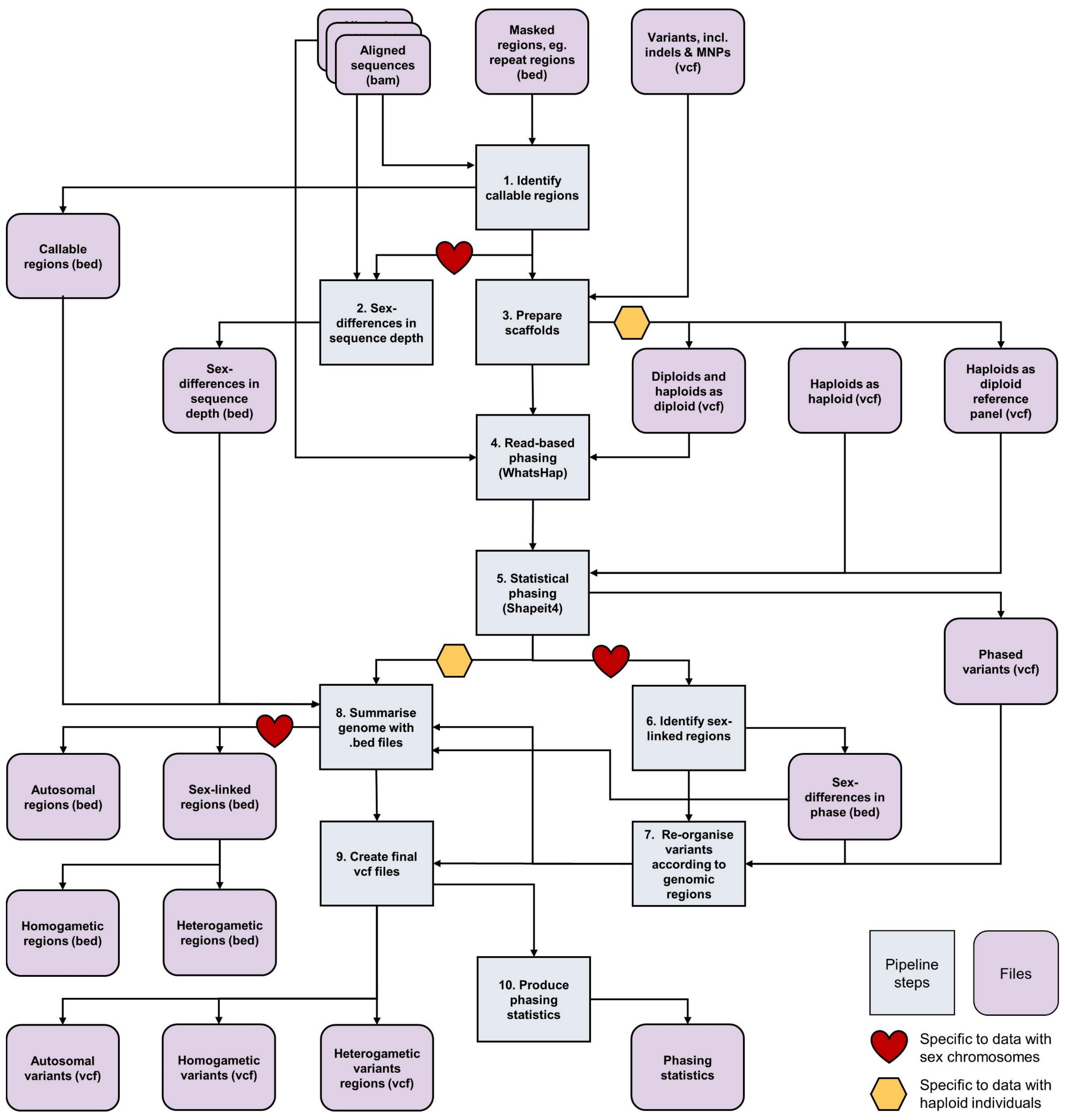
Aligned sequences (bam)
Masked regions, eg. repeat regions (bed)
Variants, incl. indels & MNPs (vcf)
1. Identify callable regions
Callable regions (bed)
2. Sex-differences in sequence depth
3. Prepare scaffolds
Sex-differences in sequence depth (bed)
Diploids and haploids as diploid (vcf)
Haploids as haploid (vcf)
Haploids as diploid reference panel (vcf)
4. Read-based phasing (WhatsHap)
5. Statistical phasing (Shapeit4)
Phased variants (vcf)
8. Summarise genome with .bed files
6. Identify sex-linked regions
Sex-differences in phase (bed)
Autosomal regions (bed)
Sex-linked regions (bed)
9. Create final vcf files
7. Re-organise variants according to genomic regions
Homogametic regions (bed)
Heterogametic regions (bed)
10. Produce phasing statistics
Autosomal variants (vcf)
Homogametic variants (vcf)
Heterogametic variants regions (vcf)
Phasing statistics
Pipeline steps
Files
Specific to data with sex chromosomes
Specific to data with haploid individuals

Figure 1. Flowchart of the PhaseWY pipeline.

To run the pipeline, only knowledge of the sex determination system and the sex of each sequenced individual is required. The pipeline runs independently per scaffold and goes through several key steps. (1) It identifies callable regions of the genome, and (2) sex-linked regions based on sex differences in sequencing depth. Then, it (3) subsets variants per scaffold, it (4) performs read-based phasing with WhatsHap, and (5) statistical phasing with Shapeit4. It (6) identifies sex-linked regions based on clustering of phased haplotypes. Finally, it (7) re-genotypes the heterogametic sex to haploids for both the homogametic and the heterogametic sex chromosomes. The pipeline summarises the genomic regions in bed format as autosomal, homogametic, and heterogametic. It further outputs phased variants in vcf format for the corresponding regions. The pipeline is not limited to phasing sex chromosomes and performs phasing on strictly autosomal data. If the dataset contains haploid and diploid individuals (e.g., haplodiploid Hymenoptera), haploid individuals can be used as a haplotype reference panel for statistical phasing of the diploid individuals.

Study system

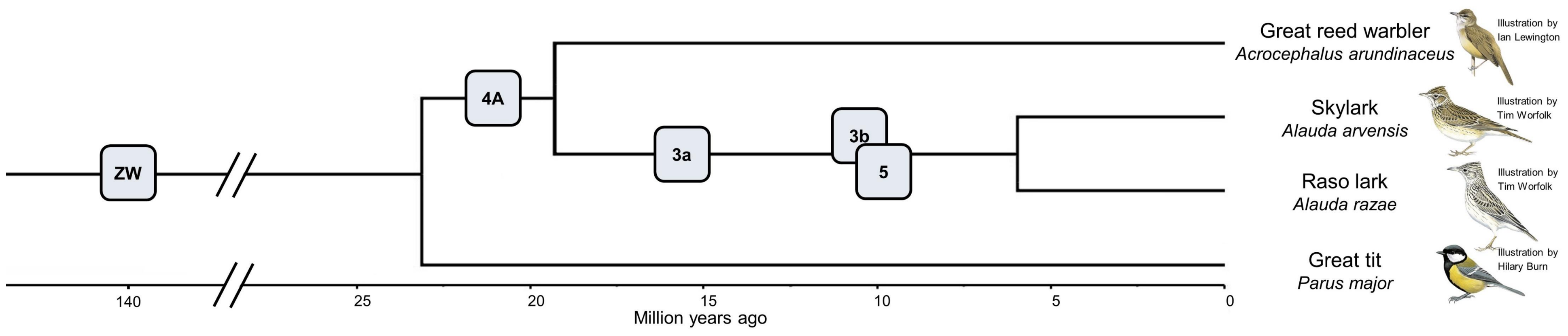
ZW
4A
3a
3b
5
Great reed warbler
Acrocephalus arundinaceus
Illustration by Ian Lewington
Skylark
Alauda arvensis
Illustration by Tim Worfolk
Raso lark
Alauda razae
Illustration by Tim Worfolk
Great tit
Parus major
Illustration by Hilary Burn
140
25
20
15
10
5
0
Million years ago

Figure 2. A phylogeny of study species with neo-sex chromosome strata at the time of recombination cessation. All avian species share the ancestral ZW sex chromosomes formed ~140 million years ago. Later, 9.6 Mbp of chromosome 4A fused to the ancestral ZW after the superfamily Sylvioidea diverged from other passerines. The largest known avian sex chromosomes are found in the lark genus *Alauda*, which has further acquired 76.5 Mbp of chromosome 3 (3a and 3b) and 36.3 Mbp of chromosome 5.

Sex chromosome identification

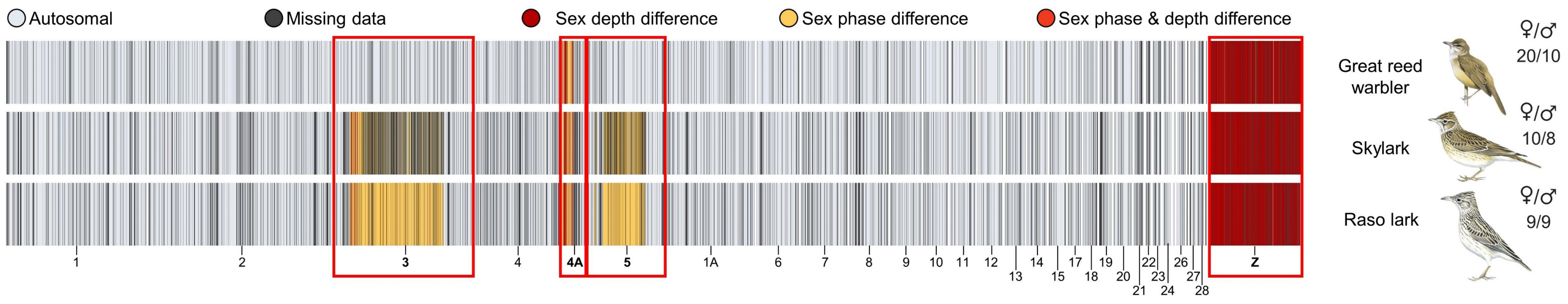

Figure 3. PhaseWY classifies the genome into different categories. Callable regions are based on specified depth and missingness. Older sex-linked regions are identified by sex differences in sequencing depth. Younger sex-linked regions are identified in sliding windows through Euclidean distances and k-means clustering of haplotypes according to the expected frequency of heterogametic and homogametic sex chromosomes in the dataset. Here, chromosomal regions are illustrated with the Great tit as synteny species. Missing data represents uncallable regions, regions with unknown classification, and missing alignments to the Great tit. Genome alignments produced with Satsuma2 v.2016-12-07 and lift-over of genome classification produced with kraken v.2017-07-06.

Phased W- and Z-haplotypes

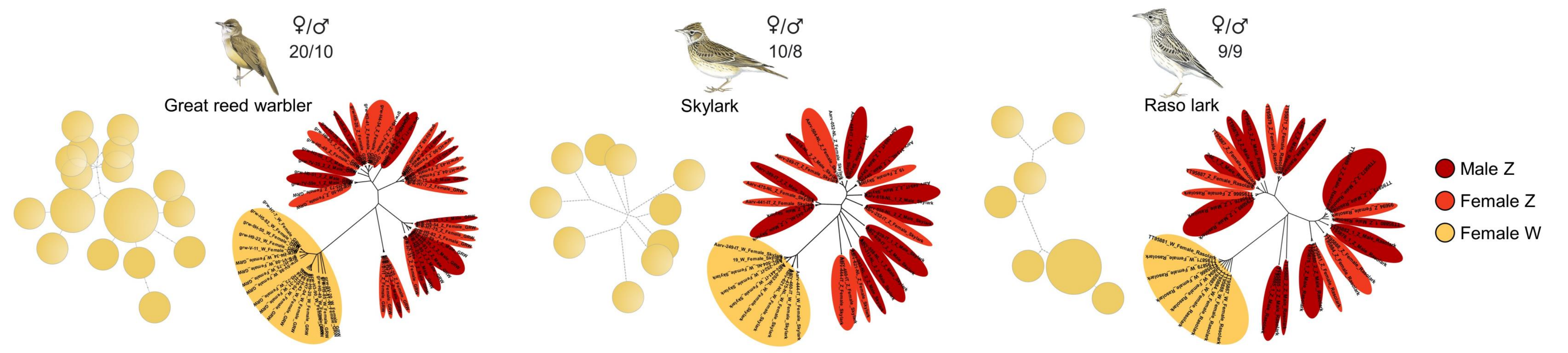

Figure 4. PhaseWY re-genotypes the heterogametic sex in regions classified as sex-linked and outputs haploid genotypes as heterogametic and homogametic. Here, resulting haplotype genealogies of W are shown. Further, ML-trees of Z and W sequences show no obvious methodogical artifact resulting from sex biased clustering of Z. Haplotype genealogies produced with Fitchi 1.1.4 and phylogenies produced with IQ-TREE v.2.2.2.2 with 1 000 ultrafast bootstraps. Analyses based on CDS sequences.

Haplotype bifurcation at the PAR border

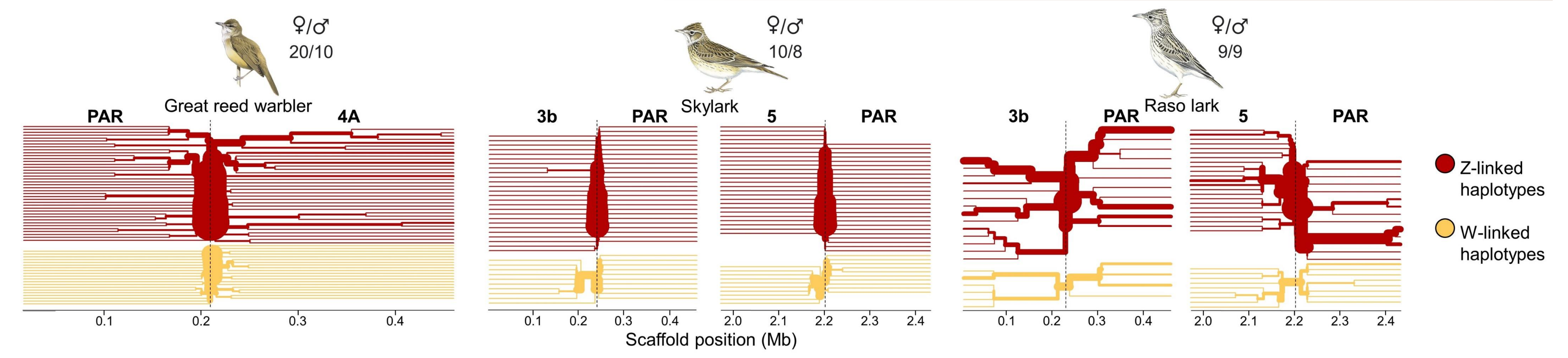

Figure 5. During meiosis in the heterogametic sex, cross-over is often concentrated to the pseudoautosomal region (PAR), which therefore may experience an average rate of linkage decay higher than other regions of the genome. Here, haplotype bifurcation plots visualise the breakdown of haplotypes at the PAR border. Haplotype bifurcations produced with rehh v. 3.2.2. Minor alleles observed at a frequency below 0.05 not considered.

Sexual antagonism on Z

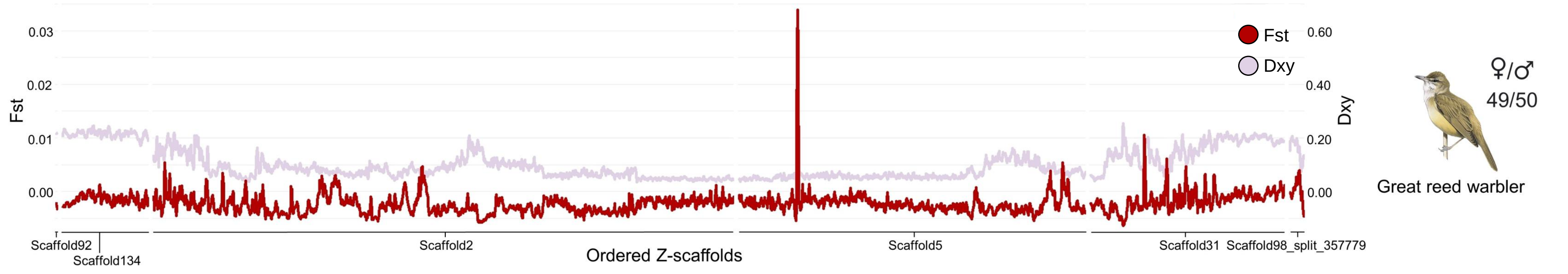

Figure 5. The study of intralocus sexually antagonistic mutations on the homogametic sex chromosome can be extremely difficult since homologous sequences from the heterogametic sex chromosome may bias the results toward false positives. It is therefore necessary to reduce such biases by seperating out these sequences in the heterogametic sex. Here, a hundred individuals of female and male Great reed warblers that survived a minimum of three years are used to identify candidate genes for sexual antagonism on the Z chromosome. An Fst scan between the sexes finds one outlying region on the ancestral Z that could still be a methodological articat, or warrant a closer investigation. Genome scans produced by popgenWindows.py. Scans performed in 100 000 bp sliding windows with 25 000 bp steps.

Simon Jacobsen Ellerstrand
PhD student
Department of Biology
Lund University
simon.jacobsen_ellerstrand@biol.lu.se

Swedish Research Council

Follow for updates:

sjellerstrand