# Identification of the homozygotic sex chromosome of non-model organisms

Charles Christian Riis Hansen1\*, Kristen M. Westfall1,2, and Snæbjörn Pálsson1,3

1Department of Life and Environmental Sciences, University of Iceland, Reykjavik, Iceland.

2Current: Fisheries and Oceans Canada, Pacific Biological Station, Nanaimo, BC, Canada.

3Senior Author

\*Correspondence: CCRH, ccr3@hi.is

## Abstract

Whole genomes are commonly assembled into a collection of scaffolds and often lack annotations of autosomes, sex chromosomes and, and organelle genomes (i.e., mitochondrial and chloroplast). As these chromosome types can have highly disparate evolutionary histories, it is imperative to take this information into account when analyzing genomic variation. Here we assessed the accuracy of four methods for identifying the homogametic sex chromosome using two whole genome sequenced (WGS) and 133 RAD sequenced white-tailed eagles (*Haliaeetus albicilla*): i) difference in read depth per scaffold, ii) heterozygosity per scaffold in a male and female bird, iii) mapping to a reference genome of a related species (chicken) with identified sex chromosomes, and iv) an analysis of SNP-loadings from a principal components analysis (PCA), based on low-depth RADseq data from 133 individuals. In i and ii, the WGS were mapped to a reference genome consisting of 1142 assembled scaffolds from the golden eagle (Aquila chrysaetos) with no identified chromosomes. The read depth per scaffold identified 86.41% of the homogametic sex chromosome (Z) with few false positives. The SNP-loading scores found 78.6% of the Z-chromosome but had a false positive discovery rate of more than 10%. The heterozygosity per scaffold did not provide clear results due to a lack of diversity in both the Z and autosomal chromosomes, and potential interference from the heterogametic sex chromosome (W).

## 1. Introduction

Inferences about genetic variation, effective population size and population structure from genomic data are dependent on the correct identification of different genomic regions, i.e., autosomes, sex chromosomes and the plastid genomes. As these different genomic regions typically have different ploidy numbers, substitution rates and recombination rates, it follows that they will also be affected differently by genetic drift and selection (Hill & Robertson, 1966). Knowledge about genomic regions can be obtained either from a high-quality reference genome of the same species, a closely related species, or from the more computationally intensive and time-consuming method of de novo assembly.

Here, we use genomic data from white-tailed eagles in Iceland, mapped to a golden eagle reference genome, to determine which scaffolds belong to the Z and autosomal chromosomes. The geographically isolated population of white-tailed eagles in Iceland currently consists of 80 breeding pairs and is known to have gone through a severe bottleneck in population size during the 20th century, when the number of breeding pairs declined to about 20 for more than 50 years (Skarphéðinsson, 2013). The golden eagle (*Aquila chrysaetos*) and the white-tailed eagle (*Haliaeetus albicilla*) are large raptors with a wide distribution in the northern hemisphere (BirdLife International, 2016a, 2016b). Currently there are four genome assemblies available for the golden eagle, consisting of 142, 1,142, 35,366 and 42,881 scaffolds, where only the first has scaffolds assigned to chromosomes (Doyle et al., 2014). Currently only three fragmented genomes exist for the white-tailed eagle (consisting of 50,905, 35,313 and 6,418 scaffolds, respectively), with no chromosomes identified (Zhang et al., 2014). The Z-chromosome has been identified in golden eagle (88.2 Mb) and is large in comparison with other bird species, which range in size from 37.9-82.5 Mb, e.g., rock dove (*Columba livia*) 37.9 Mb (Damas et al., 2017), peregrine falcon (*Falco peregrinus*) 40 Mb (Damas et al., 2017), zebra finch (*Taeniopygia guttata*) 72.8 Mb (Warren et al., 2010) and chicken (*Gallus gallus*) 82.5 Mb (Bellott et al., 2010). Resolving the chromosomal composition of the white-tailed eagle genome will facilitate research on the genetics and history of the species and for other eagle species. Furthermore, assessing the accuracy of methods for identifying the homozygotic sex chromosome facilitates annotation for other species genome assemblies for downstream analyses. The mitochondrial genomes of both species have been identified (Doyle et al., 2014; Kim et al., 2019). We used four types of information to help identify Z chromosome scaffolds in the white-tailed eagles: 1) sequencing depth and 2) patterns of heterozygosity in high-depth whole genome sequence data from a male and female bird, 3) mapping the golden eagle reference genome to that of the chicken, and 4) a PCA of genotypes from low-depth RAD-sequencing data from 133 white-tailed eagles.

A recent review describes various methods for identifying sex chromosomes (Palmer, Rogers, Dean, & Wright, 2019). When template DNA molecules from a genome are sequenced randomly, it is expected that equivalent chromosomal classes will have similar average sequencing depths, and thus the depth can be used to identify different parts of the genome. For example, mitochondrial DNA is expected to have relatively high read depth, due to a greater per-cell copy number than the nuclear chromosomes (this also applies to repeated regions). In addition, the sex chromosome found in the homogametic sex (ZZ or XX) is expected to have double the sequencing depth obtained from the heterogametic sex (ZW or XY). Thus, for example, identification of the Z (and X) chromosome through depth filtering has been successfully applied to flycatchers (Nadachowska‐Brzyska, Burri, & Ellegren, 2019).

Sex differences in heterozygosity can also be used to assess which scaffolds belong to the homogametic sex chromosome. Thus, for any given set of individuals from the same population, the Z-chromosome is expected to have fewer heterozygous positions in females (ZW) than in males (ZZ), whereas autosomal scaffolds are expected to have a similar number of heterozygous positions in both sexes. However, several factors can limit the discriminatory power of heterozygosity to identify Z scaffolds when comparing males and females. First, the difference between the sexes will be reduced for scaffolds containing pseudoautosomal (PAR) and gametologous regions (conserved but non-recombining homologous regions). A study on PAR-regions in birds have shown large variation in the size and divergence of W- and Z-chromosomes across species (Q. Zhou et al., 2014), furthermore Xu and Zhou (Xu & Zhou, 2020) showed that the W-chromosome has retained its gene function in birds better than the Y-chromosome in mammals and that the proportion of gametologs can be high. Moreover, long runs of homozygosity affecting Z scaffolds in males and autosomal scaffolds in both sexes, due to inbreeding or small population size, can mask the expected pattern of sex differences in heterozygosity. This is expected to be a marked problem in the white-tailed eagles analysed in this study.

Another approach is to map scaffolds from an incompletely assembled reference genome to a more fully annotated genome from a “closely” related species. Such mapping can be done with several available programs e.g. LASTZ (Harris, 2007), LAST (Kiełbasa, Wan, Sato, Horton, & Frith, 2011) and YASS (Noé & Kucherov, 2005). The accuracy of chromosomal locations of scaffolds obtained from this approach depends on the evolutionary distance between the two reference genomes, which can differ due to chromosomal translocations, transposed regions and repetitive regions (Jobling, Hollox, Hurles, Kivisild, & Tyler-Smith, 2014; Sætre & Ravinet, 2019), sometimes even in closely related species (Hooper & Price, 2017).

In a PCA of genotypes from scaffolds belonging to both autosomes and sex-chromosomes, it is possible that one or more principal components (PCs) split males from females, due to sex differences at markers from the sex chromosomes. It therefore follows that a PCA could be used to identify scaffolds belonging to sex chromosomes, much in the same way as for population or group differentiation. We tested this by examining the loadings of SNPs from a PCA based in low-depth RAD-sequencing data from 133 White-tailed eagles (Figure S1) - to assess if they contribute to separation along specific principal axis (Sim et al., 2012) by sex.

We show that sex differences in sequencing depth and mapping to a more complete reference genome from a related species provide the most effective means to identify Z chromosome scaffolds in the white-tailed eagles. However, the approaches based on PCA and heterozygosity provide valuable additional information and shed light on some key problems faced by researchers working with genomic data from species with partially assembled reference genomes. (BirdLife International, 2016a, 2016b)(BirdLife International, 2016a, 2016b)(BirdLife International, 2016a, 2016b)(BirdLife International, 2016a, 2016b)

## 2. Methods

### 2.1 Sample collection, laboratory work and sequencing

Blood samples were collected from white-tailed eagle chicks as a part of an ongoing monitoring program in Iceland since 2001 by the Natural History Institute of Iceland. The sex of the chicks was determined in the field based on morphology. Three to ten mL of blood was extracted from each chick. The blood was stored in EDTA buffer at -20 degrees until DNA extraction.

DNA from blood samples from 133 chicks were extracted using the ThermoFisher GeneJET Whole Blood Genomics DNA Purification Mini Kit following the standard protocol (Thermo Fisher, 2016). DNA concentration was estimated using the NanoDrop 1000 and run on 0.7% agarose gels to evaluate the fragment size. Samples with concentration higher than 60 ng/µl were selected for library preparation and sequencing.

The 133 samples were prepared for double digest restriction-site associated DNA sequencing (ddRADseq) using modified protocols from Elshire et al. (2011) and Peterson et al. (2012). Total genomic DNA (100-500 ng) was sequentially digested using the restriction endonucleases Sau3AI (1U) and ApeKI (2U), respectively, each for four hours at manufacturer (NEB) recommended temperatures in NEB Buffer 4. Digested DNA (100 ng) was ligated to adapters (sequences in Elshire et al. (2011)) containing unique combinatorial barcodes (16 unique 5 bp barcodes for ApeKI adapters and five unique 6 bp barcodes for Sau3AI adapters) for each individual (barcode and adapter sequences in Supplementary Information S1) using T4 DNA ligase (NEB) in supplied buffer at 21°C for four hours. Ligation reactions contained a 6:1 molar excess of adapter to fragmented DNA, calculated using the mean fragment size determined from an agarose gel. Ligated DNA was pooled and purified using magnetic beads (Macherey-Nagel NGS clean-up and size selection) following the manufacturers protocol. Size selection of ligated DNA fragments was performed on a Pippin Prep (Sage Science) with 2% ethidium-free agarose gels and external size standard. The narrow range setting included a mean fragment size of 350 bp ± 18 bp. The eluate was split among eight PCR reactions and amplified using the primers and PCR conditions as in Elshire et al. (2011). Each PCR reaction had a total volume of 25 μL containing; 1x OneTaq Master Mix with Standard Buffer (NEB), 0.5 mM each primer, and 8 μL template DNA. PCR products were pooled and purified using magnetic beads before quantification using a SYBR Gold fluorometric assay (protocol in Supplementary Information S2). The library was prepared for sequencing following manufacturer’s instructions with a final concentration of 38 nM. The library was sequenced on an Illumina HiSeq2500 using the Illumina TruSeq kit (2x125bp) following the manufacturer’s instructions. The sequencing was done on one lane and obtained 303 million unambiguous PE reads.

Two individuals of white-tailed eagle, a male and a female, were selected for high-depth whole genome shotgun sequencing with two lanes each on an Illumina HiSeqX. Library preparation and sequencing was done at deCODE genetics, using the TruSeq Nano sample preparation method.

Two reference assemblies from male golden eagles (ZZ), one in 1142 scaffolds and one assembled to chromosome level (GenBank Assembly Accession numbers: GCA\_000766835.1 and GCA\_900496995.2, respectively) and female chicken (ZW) (GenBank Assembly Accession: GCA\_000002315.3) were downloaded from NCBI and used in the analysis (Doyle et al., 2014; Hillier et al., 2004).

### 2.2 Sequence cleaning and mapping

The white-tailed eagle RADseq data was demultiplexed, sorting sequence reads into individual files, both for forward and reverse sequences using the command process\_radtags in Stacks (J. Catchen, Hohenlohe, Bassham, Amores, & Cresko, 2013; J. M. Catchen et al., 2011). Standard setting was used for the RADseq data, applying the option “r”, to rescue barcodes and RAD-Tags.

After demultiplexing, FastQC (Babraham Bioinformatics, 2010) was run for quality control. For the RADseq data, we found an excess of specific sequences (kmers) which were removed using AdapterRemoval v2 (version 2.2.2) (Schubert, Lindgreen, & Orlando, 2016). The high depth shotgun sequenced individuals were tested in the same way but found no excess of kmers.

The Burrows-Wheeler Aligner (BWA) and SAMtools (Li & Durbin, 2009; Li et al., 2009) were used to process RADseq and high depth shotgun data and map reads to the golden eagle scaffold assembly of 1142 scaffolds with no identified chromosomes (GCA\_000766835.1) (Doyle et al., 2014).

### 2.3 Four different approaches to find the Z-chromosome - Depth, Heterozygosity, Mapping and SNP-loadings

Four different approaches were used to identify scaffolds belonging to the Z-chromosome in the golden eagle scaffold assembly with no chromosomes (GCA\_000766835.1). An assembly consisting of 1,141 assembled scaffolds, excluding mtDNA, and a total of 1,192,725,744 bp, ranging in size from 913 to 30,727,332 bp with a median of 5,587 bp, and average length of 1,045,334 bp (SD 3,203,066 bp). An overview of the methods is presented in Figure 1 and the data used in each analysis is available in supplementary Table S1.

**Depth.** For the high-depth white-tailed eagle sequencing data, the average autosomal sequencing depth was estimated for the male and female separately, as the mode of the number of mapped reads per position across all scaffolds, based on results from the command “bedtools coverage” from Bedtools v2.18.2 (Quinlan & Hall, 2010). Using these averages, 195 for the female and 181 for the male, the relative sequencing depth was calculated for each position in each scaffold for both individuals. The per-scaffold relative sequence depth was then estimated for the female and male, separately, as the mode across positions. Positions in autosomal scaffolds are expected to have a relative depth of 1 in both sexes, whereas Z-chromosomal scaffolds are expected to have a relative depth of 0.5 in females and 1 in males. As the estimate of relative depth may be less reliable for smaller scaffolds, the dependency of the relative mode depth due to scaffold size was analysed by calculating the variance in the depths per interval of scaffold sizes, transformed to a log scale. The distribution of the proportions of scaffolds at each interval was summarized with a cumulative percentage curve. In addition, the depth per scaffold was evaluated by comparing the per-scaffold relative sequencing depth between the two individuals: male over female. Scaffolds with a relative sequencing depth below 0.25 and above 1.5 were removed. This ratio is expected to be around two for Z-chromosomal scaffolds and one for the autosomal scaffolds, as the male has two copies of Z and the female one. Thus, a cut-off between the two was set at 1.5.

**Heterozygosity.** Sex differences in heterozygosity were assessed by comparing numbers of heterozygous sites per scaffold based on genotypes of the high-depth white-tailed eagle male and female, called using Graphtyper (Eggertsson et al., 2017, 2019). The variation on the Z-chromosome is expected to be ¾ of the autosomes and it should be restricted to the male, except for the PAR and non-recombining homologus regions. As scaffolds vary in length and they may include short variable regions, the variation was analysed also per windows of 50 kb. Genotypes were filtered for quality using vcftools and bcftools (Danecek et al., 2011; Li, 2011) before counting, using minimum GQ score 20, minimum Q score 1000, missingness 1 (both individuals had to have a valid genotype at the site), mapping quality equal or above 60 (MQ), and only biallelic sites. Two additional criteria were applied to remove sites with likely spurious heterozygous genotypes. First, heterozygous genotypes where the number of mapped reads deviated significantly from the mode depth of the scaffold, based on a two-sided Poisson test (P < 0.01) were excluded. Second, we used a binomial test to assess whether the proportion of reads in heterozygous genotypes, either in the male or the female, deviated from the 50/50 expectation, using P < 0.05 as the exclusion threshold.

**Mapping.** In order to assign each of the 1142 scaffolds from the golden eagle scaffold assembly to chromosomes, LASTZ (Harris, 2007) was used to map (Harris, 2007)(Harris, 2007)(Harris, 2007)(Harris, 2007)them to the chicken genome, which has assigned chromosomes. Standard settings were used with the following modifications: ambiguous=iupac, gfextend, chain, gapped. Scaffolds in the golden eagle which mapped better to the Z-chromosome than any other chromosome, measured as most bases mapped, were deemed to belong to the golden eagle Z-chromosome.

**SNP-loadings.** A PCA analysis of 133 low-depth RAD sequenced white-tailed eagle individuals was constructed using PCangsd (Meisner & Albrechtsen, 2018), an extension of ANGSD (Korneliussen, Albrechtsen, & Nielsen, 2014), as described below. A clear split between males and females was observed along the first principal component (PC) (Figure S1). Loadings obtained with PCangsd were used to identify which parts of the scaffolds induced the split, with the “-selection” option (Meisner & Albrechtsen, 2018) and with sites passing the following filters: a minimum 25% of individuals had to have valid genotypes, only unique mapping sites, base quality minimum 20, mapping quality minimum 30, SNP p-value 1e-6. ANGSD uses genotype likelihoods to tackle the restrictions of low depth (da Fonseca et al., 2016; Korneliussen et al., 2014). To assess which scaffolds contributed to the split on the first axis (PC1), a 95% range of loading values for all SNPs per scaffold was calculated using R and compared between scaffolds with more than 50 SNPs. The distributions of the range of loading values were summarized with accumulation curves, combined for all scaffolds and separately based on the results obtained by the mapping on the autosomes and Z chromosome. Scaffolds were assigned to the Z-chromosome or autosomes depending on whether the range-values were above or below a threshold of three standard deviations from the mean (covering ~99% of a normally distributed variable).

**Comparison of the four methods.** To evaluate how well the four approaches performed, the golden eagle scaffold assembly (GCA\_000766835.1) was mapped to a golden eagle genome with known chromosomes (GCA\_900496995.2) using LASTZ with the same settings and cutoff as described previously. In what follows, the outcome of this mapping was used as the true chromosome identity of the 1141 scaffolds that was used to assess the accuracy of our four different approaches to identify Z chromosome scaffolds (Figure 1 and Table 2). A total of 168 scaffolds were assigned to the Z-chromosome, with a total length of 86,839,530 bp (mean = 516,902, sd = 1,509,132, and median = 5,236), which is slightly smaller than the Z-chromosome in the newly released genome of 88,216,475 bp (GenBank Assembly Accession: GCA\_900496995.2). The autosomal loci mapped to 973 scaffolds of a size of 1,105,886,214 bp (mean = 1,136,574, sd = 3,403,676, and median = 5,674).

## 3. Results

**3.1. Depth***.* The overall modes of depth for the two high-depth shotgun sequenced female and male were 195 and 181, respectively, which were used to estimate the relative sequence depth for each position on each scaffold. Some variation was observed in mode of relative sequence depth across scaffolds, but this was mainly due to smaller scaffolds (Figure 2, Figure S2). For the per individual scaffold comparison, discarding the shortest scaffolds (<198,789 bases, log10 < 5.29) resulted in a clear bimodal distribution and a good prediction of the Z-chromosome (0.5) and the autosomes (1) for the female (Figure 2A and S2). As expected, this was not observed for the male (Figure 2B). After removing the short scaffolds, 257 scaffolds out of the 1141 scaffolds remained, but covering 98.9 % of the full genome. In the female, 36 scaffolds, comprising ~75.2 Mb, had a relative depth close to 0.5 (from 0.466 to 0.533), all from the Z-chromosome. In comparison, 211 scaffolds (1.0947 Gb) had a relative depth around 1 (from 0.764 to 1.062), whereof 207 were autosomal. The remaining four scaffolds (NW\_011950951.1, NW\_011950990.1, NW\_011951047.1 and NW\_011951051.1) map to the Z chromosome, comprising ~10 Mb or 0.91% of the scaffolds identified as autosomes (see Table 2 and Table S1 for all numbers).

The expected male to female ratio (*rmf*) of sequence depth is 1 for autosomal and 2 for Z scaffolds. Implementation of *rmf* for the scaffolds revealed an even clearer split between the Z and the autosomes (Figure 2C), particularly after removing the primarily small scaffolds with relative depth outside the credible range of 0.25-1.5 in either the male or female. This left 618 scaffolds that account for 99.53% of the total sequence (Figure 2D). Thereof 93 had *rmf* > 1.5, consistent with the expected depth of Z scaffolds. Of these, 79 (76.2 Mb) map to the Z and 14 (0.09 Mb) map to autosomal chromosomes in the golden eagle genome. We observed 525 scaffolds with *rmf* <= 1.5, consistent with the expected depth of autosomes. Of these, 512 scaffolds (1,100.7 Mb) map autosomes and 13 (10.05 Mb) map to Z in the golden eagle genome.

**3.2 Heterozygosity.** Only 32% of scaffolds (365 of 1,141), but covering 97.5% of the genome size, had at least one heterozygous genotype after filtering in either of the two individuals, with slightly fewer in the female (288) than in the male (300). The majority of the scaffolds with no heterozygous sites mapped to the Z (80% in the female, corresponding to 30% of the Z chromosome; 77% in the male, covering 23% of Z). The Z has generally lower number of heterozygous sites after filtering (Table 1, Supplement Figure 3 and 4), a majority of the autosomal scaffolds lack heterozygous sites (67%, 1.1% in size). Furthermore, the autosomal scaffolds are in much larger frequency than the Z’s. Seventy-seven scaffolds (52.5 Mb, ranging from 1.5-5,565 kb) had no heterozygous genotypes in the female but a minimum of one heterozygous genotype in the male and ten of those scaffolds (10.1 Mb) map to the Z-chromosome in the golden eagle genome. Aside the larger fraction of the Z scaffolds which have no variation on Z, about 62% of the Z-chromosome in the female has also considerably fewer heterozygous sites than the male (supplement Figure 3), but some show autosomal levels of heterozygosity in the female (separately marked in Figure 3A). Four of these scaffolds also exhibited autosomal levels of depth in the female (Figure 2) and two of those scaffolds ("NW\_011950951.1" "NW\_011950990.1") in the female had the highest number of heterozygous sites (1823, 5568), followed by NW\_011951047.1 which had 450 sites.

The four Z chromosomal scaffolds that had a male-like pattern of autosomal depth and heterozygosity in the female were further analysed in windows of 50Kb, as heterozygous sites can be restricted to small parts of the scaffold (Figure S5). An examination of the number of filtered heterozygous sites per 50Kb window in these four scaffolds in the female showed that NW\_011950951.1, NW\_011950990.1 consisted of either 1 or 2 continuous regions, whereas the other two were more fragmented.

The average heterozygosity per scaffold, prior to filtering, was >10-fold higher in the female than the male for the Z-chromosome (Table 1), and several scaffolds were even higher (Figure 3B). The filtering removed most of this excess heterozygosity in the female (Figure 3C, D and E). As the pattern of excess heterozygosity in the female was primarily seen in Z rather than autosomal scaffolds, we postulate that these instances might represent the mapping of diverged homologous reads from the W chromosome.

Overall, the distributions of heterozygous sites per window was similar for the male and the female and almost half of the windows had no heterozygosity (49% in the female and 47% in the male). When the windows were grouped by Z and autosomes, a difference between the sexes is observed for the Z-chromosome (Table 1 and Figure S3 and S4). As expected, there is a higher proportion of windows on Z with no heterozygous sites in the female (82%) than in the male (74%) (P = 6.111\*10-8, Fishers exact test). However, the 10 most variable 50kb windows in the female, with rate of heterozygous sites ranging from 0.17-1.73% all come from the scaffold NW\_011950990.1 which map to Z. The window in male with largest rate of heterozygous sites has 0.15%. This difference in the distribution of heterozygosity per 50 kb windows on the Z chromosome per sex, is also reflected in the average number and standard deviation of heterozygous genotypes per window, which is larger in the female Z (5.1 and 43) than in the male Z (3.2 and 8.1), whereas no differences are observed in these descriptive statistics for the autosomes. This means that the distribution of heterozygous genotypes is more clumped for Z in the female (Coefficient of dispersion, CD=360.5) than in the male (20.2) and the autosomes of both sexes (~16).

**3.3 Mapping**

Mapping the 1141 scaffolds from the golden eagle scaffold assembly to the chicken genome, using LASTZ, resulted in 110 scaffolds (86.5 Mb) correctly assigned to the Z-chromosome, and 940 scaffolds correctly assigned to autosomes, according to the golden eagle chromosome-based genome. On the other hand, 33 scaffolds (0.59 Mb, amounting to 0.69% of the total length of scaffolds) were wrongly assigned to the Z-chromosome, and 58 scaffolds (0.27 Mb, 0.024%) were wrongly assigned to autosomes (Table 2).

**3.4 PCA**

The analysis of the loadings of 164,952 SNPs from the PCA analysis (Figure S1), based on 133 RADseq individuals with an average sequencing depth per site of 2.25 per individual, was limited to the 280 scaffolds (40 Z and 240 autosomal) that had more than 50 SNPs (accounting for 98.3% of the genome). We calculated the 95% range of SNP-loadings for PC1 in our attempt to identify scaffolds belonging to the Z, using a threshold (0.1006) that corresponds to 3 standard deviations above the mean 95% range across scaffolds (Figures 4A and 4B, Table 2). Of the scaffolds included in this analysis, 28 (78%) scaffolds from the Z-chromosome were above this threshold, accounting for 69.3 Mb (83.6% of the total length of Z scaffolds used in this analysis). In contrast, only 9 (3.75%) of the autosomal scaffolds were above the threshold, amounting to 11.7 Mb (1.1% of the total length of autosomal scaffolds used in this analysis). Thus, the range of PC1 loadings provides some discriminatory power to distinguish Z from autosomal scaffolds.

**3.5 Comparison of the four methods.**

Using chromosome assignments obtained by mapping the golden eagle scaffold assembly to the golden eagle genome with assigned chromosomes, we find that the most successful of the four methods tested, finding 98.12% of the expected size, was mapping to the chicken genome (Table 2, Figure 5). In second place was the depth analysis with 86.41% and, in third, the SNP-loading with 78,61%. Heterozygosity turned out to be poorly suited to find Z-chromosomal scaffolds, based on data from only a single male and female, as a large fraction of scaffolds had no variation, and some Z-chromosomal scaffolds were found to be highly variable in the female (likely due to the mapping of reads that belong to the W chromosome). Depth, mapping to the chicken and SNP-loading all found false positives, i.e., autosomal scaffolds that were categorized as Z-chromosomal scaffolds (0.09, 0.59 and 11.72 Mb, respectively). Forty-five very short Z-chromosomal scaffolds (with a total length of 0.22 Mb) were not found by any of the analysis but were only found when the golden eagle scaffold assembly was mapped to the golden eagle with known chromosomes. Mapping of the golden eagle scaffold assembly to the one with assembled chromosomes revealed 98.42% of the whole known Z-chromosome (Table 2, Figure 5). A combined analysis, where at least two of our three approaches, depth, mapping to the chicken and SNP-loading, detected between 75.29-86.29% of the size of the Z-chromosome of the golden eagle genome, and of these only the approach combining depth and mapping to the chicken found false positives, which was less than <0.01% of the size of the golden eagle Z-chromosome.

## 4. Discussion

Three of the four methods evaluated in this study, the relative depth, mapping to chicken and SNP-loadings were able to detect a high fraction of the Z-chromosome of the white-tailed eagle, in the golden eagle scaffold assembly.

The mapping of contigs to genome sequences from a distantly related species such as golden eagle to chicken can be problematic due to chromosomal changes such as translocations and inversions. Minor mismatches e.g., transposable elements and mutations may further impact the success of finding the Z-chromosome. In birds, sex chromosomes may however be well preserved e.g., Xu and Zhou (2020) and in the case of mapping the golden eagle scaffold assembly to the chicken, with a split time >80 million years (Jarvis et al., 2015), the effect seems to be minimal.

The Z scaffolds that were not detected using the SNP-loading approach are likely due to parts of the Z-chromosome which lack variation, or which share homologous regions in the distinct sex chromosomes and do thus not contribute to the difference between the sexes in the PCA-plot. The PCA approach found few false positives, possibly due to the lack of a precise distinction between the range of loadings observed for the autosomal and Z-chromosomal scaffolds. Considering the information from the mapping it is though clear that the Z-scaffolds have higher impact, as most false positives were just above the threshold of three SDs (0.10 95% SNP loading range), and only two autosomal scaffolds were larger than ~0.11 comprising only a total size of 1.73 Mb, or 14% of the false positives.

The relative depth analysis revealed 86.41% of the expected size of the Z-chromosome and found few false positives. Four scaffolds were especially noted and were false negatives in one of the two depth analysis. These four scaffolds (NW\_011950951.1, NW\_011950990.1, NW\_011951047.1, and NW\_011951051.1) make up about 10 Mb and also show the highest heterozygosity of all Z-chromosomal scaffolds after filtering; their levels are comparable or even higher to that of the autosomal scaffolds. Three of the four scaffolds also showed low 95% SNP-loading ranges (all around 0.05), unlike the scaffolds contributing to the separation of the sexes. One scaffold (NW\_011950990.1) had both a very high 95% SNP-loading range, and a very high heterozygosity. The signal in these four Z scaffolds may be because they belong to the pseudo-autosomal regions (PAR) known to exist between the W and Z-chromosomes (Otto et al., 2011) and thus being a Z/W. In birds, PAR vary greatly in size from just a few Mb to more than 60 Mb (Q. Zhou et al., 2014). Alternatively, they could represent non-recombining homologous regions (gametologs) (Smeds et al., 2015; Xu & Zhou, 2020) which can be expected to have even higher heterozygosity in females than within the recombining Z-chromosomes in the homogametic males or the autosomes, as such regions may have evolved independently for millions of years.

Inspection of the heterozygosity for all scaffolds revealed that it is difficult to distinguish between autosomal and Z-chromosomal scaffolds without any prior knowledge. There were however a difference in average heterozygosity per scaffold between autosomal and Z-chromosomal scaffolds, and difference between the autosomal and Z-chromosomal scaffolds especially in the female. Low effective population sizes such as for the white-tailed eagle in Iceland (Skarphéðinsson, 2013) has reduced the heterozygosity and long runs of homozygosity were observed on the Z-chromosome and the autosomes, making it more difficult to distinguish among the chromosomal types. Further, there is a clear overlap in scaffolds with some heterozygosity which might belong to PAR and non-homologus regions on the Z- and W-chromosomes. PAR and the nonrecombining homologous regions, could explain some deviations in the prediction of the Z-chromosome in the SNP-loading analysis but these regions are probably small, and thus they don’t display the signal of an autosome in the depth analysis. Changes in synteny of the sex chromosomes and the autosomes between golden eagle and the white-tailed eagle might have happened during their evolutionary divergence but this remains to be explored further (R. Zhou et al., 2020).

Although depth analysis has shown to be a promising method to identify sex chromosomes (Huylmans, Toups, MacOn, Gammerdinger, & Vicoso, 2019; Nadachowska‐Brzyska et al., 2019), it is not error free. Scaffolds belonging to the Z-chromosome can have as high depth as autosomes, as variance in depth can be large in small scaffolds which may be poorly sampled due to low variation, or the scaffolds include regions from both Z- and W-chromosomes i.e., gametologs and the PAR regions. Here the best approach for identifying the homogametic sex chromosome was mapping to a reference with known homogametic sex chromosome, and in birds even a very distant species could be used (Graves, 2014; Trukhina & Smirnov, 2014; Xu & Zhou, 2020). To identify the Z-chromosome, a combination of the mapping with at least one other analysis is recommended as it may result in fewer potential false positives.

Even though all known eukaryote species may soon be sequenced (Lewin et al., 2018), it will still be a long time before all parts of their chromosomes have been identified. Thus, it is important to further explore these different methods and how they depend on sequence variation and scaffold sizes, as variation in the different chromosomes will differ due to different effective population sizes and evolutionary histories.

## Acknowledgement

Thanks to Gunnar T. Hallgrimsson from Department of Life and Environmental Sciences, University of Iceland, Menja von Schmalensee and Robert A. Stefansson from West-Iceland Centre of Natural History, and Kristinn Haukur Skarphédinsson from the Icelandic Institute of Natural History for sample collection.

Thanks to Jonas Meisner from section for Computational and RNA Biology, University of Copenhagen, for great help with PCAngsd.

A special thanks to Agnar Helgason from deCODE genetics for guidance in analysis, writing, and access to facilities at deCODE genetics. Further, thanks to deCODE genetics for sequencing of the two white-tailed genomes.

The study was supported by Research Grant nr 185280-052 from The Icelandic Research Council, the Doctoral student fund of the University of Iceland and The University of Iceland Research fund.

## Author contribution

CCRH and SP designed the study; KMW prepared the RADseq libraries; CCRH and SP analyzed the data; CCRH, KMW and SP wrote the paper.

## Competing interests

The authors declare no competing financial interests.

## References

Babraham Bioinformatics. (2010). FastQC. Retrieved June 16, 2016, from http://www.bioinformatics.babraham.ac.uk/projects/fastqc/

Bellott, D. W., Skaletsky, H., Pyntikova, T., Mardis, E. R., Graves, T., Kremitzki, C., … Page, D. C. (2010). Convergent evolution of chicken Z and human X chromosomes by expansion and gene acquisition. *Nature*, *466*(7306), 612–616. doi: 10.1038/nature09172

BirdLife International. (2016a). Aquila chrysaetos. *The IUCN Red List of Threatened Species 2016*, *e.T2269606*. doi: 10.2305/IUCN.UK.2016-3.RLTS.T22696060A93541662.en

BirdLife International. (2016b). Haliaeetus albicilla. *The IUCN Red List of Threatened Species 2016*, *e.T2269513*. doi: 10.2305/IUCN.UK.2016-3.RLTS.T22695137A93491570.en

Catchen, J., Hohenlohe, P. A., Bassham, S., Amores, A., & Cresko, W. A. (2013). Stacks: an analysis tool set for population genomics. *Molecular Ecology*, *22*(11), 3124–3140. doi: 10.1111/mec.12354

Catchen, J. M., Amores, A., Hohenlohe, P., Cresko, W., Postlethwait, J. H., & De Koning, D.-J. (2011). Stacks: Building and genotyping loci de novo from short-read sequences. *Genes|Genomes|Genetics*, *1*(3), 171–182. doi: 10.1534/g3.111.000240

da Fonseca, R. R., Albrechtsen, A., Themudo, G. E., Ramos-Madrigal, J., Sibbesen, J. A., Maretty, L., … Pereira, R. J. (2016). Next-generation biology: Sequencing and data analysis approaches for non-model organisms. *Marine Genomics*, *30*, 3–13. doi: 10.1016/j.margen.2016.04.012

Damas, J., O’Connor, R., Farré, M., Lenis, V. P. E., Martell, H. J., Mandawala, A., … Larkin, D. M. (2017). Upgrading short-read animal genome assemblies to chromosome level using comparative genomics and a universal probe set. *Genome Research*, *27*(5), 875–884. doi: 10.1101/gr.213660.116

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., … Durbin, R. (2011). The variant call format and VCFtools. *Bioinformatics*, *27*(15), 2156–2158. doi: 10.1093/bioinformatics/btr330

Doyle, J. M., Katzner, T. E., Bloom, P. H., Ji, Y., Wijayawardena, B. K., & DeWoody, J. A. (2014). The genome sequence of a widespread apex predator, the golden eagle (Aquila chrysaetos). *PLoS ONE*, *9*(4), 20–22. doi: 10.1371/journal.pone.0095599

Eggertsson, H. P., Jonsson, H., Kristmundsdottir, S., Hjartarson, E., Kehr, B., Masson, G., … Halldorsson, B. V. (2017). Graphtyper enables population-scale genotyping using pangenome graphs. *Nature Genetics*, *49*(11), 1654–1660. doi: 10.1038/ng.3964

Eggertsson, H. P., Kristmundsdottir, S., Beyter, D., Jonsson, H., Skuladottir, A., Hardarson, M. T., … Melsted, P. (2019). GraphTyper2 enables population-scale genotyping of structural variation using pangenome graphs. *Nature Communications*, *10*(1), 1–8. doi: 10.1038/s41467-019-13341-9

Elshire, R. J., Glaubitz, J. C., Sun, Q., Poland, J. A., Kawamoto, K., Buckler, E. S., & Mitchell, S. E. (2011). A Robust, Simple Genotyping-by-Sequencing (GBS) Approach for High Diversity Species. *PLoS ONE*, *6(5)*(5), e19379. doi: 10.1371/journal.pone.0019379

Graves, J. A. M. (2014). Avian sex, sex chromosomes, and dosage compensation in the age of genomics. *Chromosome Research*, *22*(1), 45–57. doi: 10.1007/s10577-014-9409-9

Harris, R. S. (2007). *Improved pairwise alignmen of genomic DNA. Ph.D. Thesis* (The Pennsylvania State University). Retrieved from http://scholar.google.com/scholar?q=related:J\_7kcJuUwCoJ:scholar.google.com/&hl=en&num=20&as\_sdt=0,5%5Cnhttp://www.bx.psu.edu/~rsharris/rsharris\_phd\_thesis\_2007.pdf

Hill, W. G., & Robertson, A. (1966). The effect of linkage on limits to artificial selection. *Genetical Research*, *8*(3), 269–294. doi: 10.1017/S0016672300010156

Hillier, L. W., Miller, W., Birney, E., Warren, W., Hardison, R. C., Ponting, C. P., … Sequencing, G. (2004). Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. *Nature*, *432*(7018), 695–716. doi: 10.1038/nature03154

Hooper, D. M., & Price, T. D. (2017). Chromosomal inversion differences correlate with range overlap in passerine birds. *Nature Ecology and Evolution*, *1*(10), 1526–1534. doi: 10.1038/s41559-017-0284-6

Huylmans, A. K., Toups, M. A., MacOn, A., Gammerdinger, W. J., & Vicoso, B. (2019). Sex-biased gene expression and dosage compensation on the artemia franciscana Z-chromosome. *Genome Biology and Evolution*, *11*(4), 1033–1044. doi: 10.1093/gbe/evz053

Jarvis, E. D., Mirarab, S., Aberer, A. J., Li, B., Houde, P., Li, C., … Zhang, G. (2015). Phylogenomic analyses data of the avian phylogenomics project. *GigaScience*, *4*(1), 1–9. doi: 10.1186/s13742-014-0038-1

Jobling, M., Hollox, E., Hurles, M., Kivisild, T., & Tyler-Smith, C. (2014). *Human evolutionary genetics* (2nd ed.; E. Owen, D. Borrowdale, & G. Lucas, Eds.). Garland Science, Taylor & Francis Group, LLC.

Kiełbasa, S. M., Wan, R., Sato, K., Horton, P., & Frith, M. C. (2011). Adaptive seeds tame genomic sequence comparison. *Genome Research*, *21*(3), 487–493. doi: 10.1101/gr.113985.110

Kim, J. A., Kang, S. G., Jeon, H. S., Jeon, J. H., Jang, J. H., Kim, S., & An, J. (2019). Complete mitogenomes of two Accipitridae, Haliaeetus albicilla, and Pernis ptilorhynchus. *Mitochondrial DNA Part B: Resources*, *4*(1), 392–393. doi: 10.1080/23802359.2018.1547155

Korneliussen, T. S., Albrechtsen, A., & Nielsen, R. (2014). ANGSD: Analysis of next generation sequencing data. *BMC Bioinformatics*, *15*(1), 356. doi: 10.1186/s12859-014-0356-4

Lewin, H. A., Robinson, G. E., Kress, W. J., Baker, W. J., Coddington, J., Crandall, K. A., … Zhang, G. (2018). Earth BioGenome Project: Sequencing life for the future of life. *Proceedings of the National Academy of Sciences*, *115*(17), 4325–4333. doi: 10.1073/pnas.1720115115

Li, H. (2011). A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics*, *27*(21), 2987–2993. doi: 10.1093/bioinformatics/btr509

Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, *25*(14), 1754–1760. doi: 10.1093/bioinformatics/btp324

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., … Durbin, R. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, *25*(16), 2078–2079. doi: 10.1093/bioinformatics/btp352

Meisner, J., & Albrechtsen, A. (2018). Inferring Population Structure and Admixture Proportions in Low-Depth NGS Data. *Genetics*, *210*(2), 719–731. doi: 10.1534/genetics.118.301336

Nadachowska‐Brzyska, K., Burri, R., & Ellegren, H. (2019). Footprints of adaptive evolution revealed by whole Z chromosomes haplotypes in flycatchers. *Molecular Ecology*, mec.15021. doi: 10.1111/mec.15021

Noé, L., & Kucherov, G. (2005). YASS: Enhancing the sensitivity of DNA similarity search. *Nucleic Acids Research*, *33*(SUPPL. 2), 540–543. doi: 10.1093/nar/gki478

Otto, S. P., Pannell, J. R., Peichel, C. L., Ashman, T. L., Charlesworth, D., Chippindale, A. K., … McAllister, B. F. (2011). About PAR: The distinct evolutionary dynamics of the pseudoautosomal region. *Trends in Genetics*, *27*(9), 358–367. doi: 10.1016/j.tig.2011.05.001

Palmer, D. H., Rogers, T. F., Dean, R., & Wright, A. E. (2019). How to identify sex chromosomes and their turnover. *Molecular Ecology*, (September), 1–16. doi: 10.1111/mec.15245

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. doi: 10.1371/journal.pone.0037135

Quinlan, A. R., & Hall, I. M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, *26*(6), 841–842. doi: 10.1093/bioinformatics/btq033

Sætre, G.-P., & Ravinet, M. (2019). *Evolutionary Genetics* (1st ed.). doi: 10.1093/oso/9780198830917.001.0001

Schubert, M., Lindgreen, S., & Orlando, L. (2016). AdapterRemoval v2: rapid adapter trimming, identification, and read merging. *BMC Research Notes*, *9*(1), 88. doi: 10.1186/s13104-016-1900-2

Sim, S. C., van Deynze, A., Stoffel, K., Douches, D. S., Zarka, D., Ganal, M. W., … Francis, D. M. (2012). High-Density SNP Genotyping of Tomato (Solanum lycopersicum L.) Reveals Patterns of Genetic Variation Due to Breeding. *PLoS ONE*, *7*(9), 1–18. doi: 10.1371/journal.pone.0045520

Skarphéðinsson, K. H. (2013). *Haförninn*. Reykjavik: Fuglavernd (Fuglaverndarfélag íslands).

Smeds, L., Warmuth, V., Bolivar, P., Uebbing, S., Burri, R., Suh, A., … Ellegren, H. (2015). Evolutionary analysis of the female-specific avian W chromosome. *Nature Communications*, *6*. doi: 10.1038/ncomms8330

Thermo Fisher. (2016). Thermo Scientific GeneJET Genomic DNA Purification Kit #K0721, #K0722. Retrieved February 14, 2020, from https://www.thermofisher.com/document-connect/document-connect.html?url=https%3A%2F%2Fassets.thermofisher.com%2FTFS-Assets%2FLSG%2Fmanuals%2FMAN0012667\_GeneJET\_Whole\_Blood\_Genomic\_DNA\_Purification\_Mini\_Kit\_UG.pdf&title=VXNlciBHdWlkZTogR2VuZUpFVCBXaG9sZSBC

Trukhina, A. V., & Smirnov, A. F. (2014). Problems of Birds Sex Determination. *Natural Science*, *06*(15), 1232–1240. doi: 10.4236/ns.2014.615111

Warren, W. C., Clayton, D. F., Ellegren, H., Arnold, A. P., Hillier, L. W., Künstner, A., … Wilson, R. K. (2010). The genome of a songbird. *Nature*, *464*(7289), 757–762. doi: 10.1038/nature08819

Xu, L., & Zhou, Q. (2020). The female-specific W chromosomes of birds have conserved gene contents but are not feminized. *Genes*, *11*(10), 1–14. doi: 10.3390/genes11101126

Zhang, G., Li, C., Li, Q., Li, B. B., Larkin, D. M., Lee, C., … Froman, D. P. (2014). Comparative genomics reveals insights into avian genome evolution and adaptation. *Science*, *346*(6215), 1311–1320. doi: 10.1126/science.1251385

Zhou, Q., Zhang, J., Bachtrog, D., An, N., Huang, Q., Jarvis, E. D., … Zhang, G. (2014). Complex evolutionary trajectories of sex chromosomes across bird taxa. *Science*, *346*(6215). doi: 10.1126/science.1246338

Zhou, R., Macaya-Sanz, D., Carlson, C. H., Schmutz, J., Jenkins, J. W., Kudrna, D., … Difazio, S. P. (2020). A willow sex chromosome reveals convergent evolution of complex palindromic repeats. *Genome Biology*, *21*(1), 1–19. doi: 10.1186/s13059-020-1952-4

## Data accessibility

Individual data has deposited for DRYAD data depository and will be released when the manuscript is published at https://doi.org/10.5061/dryad.v9s4mw6vs.