Evaluation of Four Methods to Identify the Homozygotic Sex Chromosome in Small Populations

Charles Christian Riis Hansen (ccr3@hi.is)
University of Iceland
Kristen M. Westfall
Fisheries and Oceans Canada, Pacific Biological Station
Snaebjörn Pálsson
University of Iceland

Research Article

Keywords: Homogametic sex chromosome, population genetics, non-model organisms, white-tailed eagle

Posted Date: September 22nd, 2021

DOI: https://doi.org/10.21203/rs.3.rs-892602/v1

Version of Record: A version of this preprint was published at BMC Genomics on February 24th, 2022. See the published version at https://doi.org/10.1186/s12864-022-08393-z.

Abstract

Background

Whole genomes are commonly assembled into a collection of scaffolds and often lack annotations of autosomes, sex chromosomes, and organelle genomes (i.e., mitochondrial and chloroplast). As these chromosome types differ in effective population size and can have highly disparate evolutionary histories, it is imperative to take this information into account when analysing genomic variation. Here we assessed the accuracy of four methods for identifying the homogametic sex chromosome in a small population using two whole genome sequences (WGS) and 133 RAD sequences of white-tailed eagles (*Haliaeetus albicilla*): i) difference in read depth per scaffold in a male and a female, ii) heterozygosity per scaffold in a male and a female, iii) mapping to a reference genome of a related species (chicken) with identified sex chromosomes, and iv) analysis of SNP-loadings from a principal components analysis (PCA), based on the low-depth RADseq data.

Results

The best performing approach was the reference mapping (method iii), which identified 98.12% of the expected homogametic sex chromosome (Z). The read depth per scaffold (method i) identified 86.41% of the homogametic sex chromosome with few false positives. The SNP-loading scores (method iv) found 78.6% of the Z-chromosome and had a false positive discovery rate of more than 10%. The heterozygosity per scaffold (method ii) 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). The evaluation of these methods also revealed 10 Mb of likely PAR and gametologous regions.

Conclusion

Identification of the homogametic sex chromosome in a small population is best accomplished by reference mapping or examining read depth differences between sexes.

Background

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 [1]. 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 use of a different reference genome from the study species is frequently done [2–4], and the use of a different species as a reference can be beneficial, as references of the same species have been shown to potentially introduce errors in population genetic analyses [5]. 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 [6], thus this small population is expected to lack heterozygosity overall. The golden eagle (*Aquila chrysaetos*) and the white-tailed eagle (*Haliaeetus albicilla*) are large raptors with a wide distribution in the northern hemisphere [7, 8]. Currently there are four genome assemblies available for the golden eagle, consisting of 142 (size: 1,233.7 Mb, N50: 46.9 Mb); 1,142 (size: 1,192.7 Mb, N50: 9.2 Mb); 35,366 (size: 1,196.0 Mb, N50: 0.11 Mb); and 42,881 (size: 1,548.4 Mb, N50: 1.7 Mb) scaffolds, where only the first has scaffolds assigned to chromosomes [9]. Currently only three fragmented genomes exist for the white-tailed eagle (consisting of 50,905 scaffolds with the size: 1,133.5 Mb, and N50: 0.05 Mb; 35,313 scaffolds with the size: 1,196.5 Mb and N50: 0.12 Mb; and 6,418 scaffolds with the size: 1,222.6 Mb and N50: 4.5 Mb), with no chromosomes identified [10]. The mitochondrial genomes of both the white-tailed and golden eagle have been identified [9, 11]. The Z-chromosome has been identified in golden eagle (88.2 Mb) and is large in comparison with many other bird species where the Z chromosome has been identified (however far from the largest), which range in size from 37.9-195.3 Mb, e.g., rock dove (*Columba livia*) 37.9 Mb [12], peregrine falcon (*Falco peregrinus*) 40 Mb [12], zebra finch (*Taeniopygia guttata*) 72.8 Mb [13], chicken (*Gallus gallus*) 82.5 Mb [14], and Eurasian skylark (*Alauda arvensis*) [15]. 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. We used four types of information separately to identify Z chromosome scaffolds in a small population of white-tailed eagles, which is expected to lack heterozygosity: 1) sequencing depth, and 2) patterns of heterozygosity in high-depth whole genome sequence data from one male and one female, 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 [16]. 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), in species with differentiated sex chromosomes, as in birds and mammals [17, 18], but not in species with little differentiation between sex chromosomes such as in several fish species [19, 20]. Thus, for example, identification of the Z (and X) chromosome through depth filtering has been successfully applied to flycatchers [21], and depth is also partly used programmatically for discovering the sex chromosomes [22–24].

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 [25], furthermore Xu and Zhou [17] 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 feature of 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 [26], LAST [27] and YASS [28]. 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 [29, 30], sometimes even in closely related species [31]. Thus, this method may be only applicable for taxa with relatively stable genomes such as mammals and birds, though some groups of birds have also recently been shown to have dynamic sex chromosomes [32].

In a PCA of genotypes from all scaffolds i.e., belonging to both autosomes and sex-chromosomes, it is possible that one or more principal components (PCs) split males and 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 on low-depth RAD-sequencing data from 133 White-tailed eagles (Figure S1) - to assess if they contribute to separation along a specific principal axis [33] 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 the PCA, and heterozygosity provide valuable additional information and shed light on some key challenges faced by researchers working with genomic data from species with partially assembled reference genomes.

Results

Depth. The overall modes of depth for 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 (Fig. 1, Figure S2, Figure S3). 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 (Fig. 1A and S2). As expected, this was not observed for the male (Fig. 1B). 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 (r_{mf}) of sequence depth is 1 for autosomal and 2 for Z scaffolds. Implementation of r_{mf} for the scaffolds revealed an even clearer split between the Z and the autosomes (Fig. 1C), 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 (Fig. 1D). Thereof 93 had $r_{mf} > 1.5$, consistent with the expected depth of Z scaffolds. Of these, 79 (76.2 Mb) identify as Z and 14 (0.09 Mb) as autosomal chromosomes in the golden eagle genome. We observed 525 scaffolds with $r_{mf} <= 1.5$, consistent with the expected depth of autosomes. Of these, 512 scaffolds (1,100.7 Mb) identify as autosomes and 13 (10.05 Mb) as Z in the golden eagle genome.

Heterozygosity. Only 32% of scaffolds (365 of 1,141), covering 97.5% of the genome, 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 Figs. 3 and 4), but a majority of the autosomal scaffolds lack heterozygous sites (67%, 1.1% in size). Furthermore, there are more autosomal scaffolds than 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 Fig. 3), but some show autosomal levels of heterozygosity in the female (separately marked in Fig. 2A). Four of these scaffolds also exhibited autosomal levels of depth in the female (Fig. 1) 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.

Table 1
Information about heterozygosity for a female and male. Heterozygosity for each of the male and female for scaffolds that map to the A and Z in the golden eagle genome with known chromosomes. Numbers of heterozygous (hets.) sites, scaffolds and windows of size 50,000 bases. Total number of scaffolds and 50k windows were 1,141 and 23,585 respectively.

	Female Z	Female A	Male Z	Male A
Proportion of heterozygous sites before filtering	0.00534	0.00067	0.00050	0.00065
Proportion of heterozygous sites after filtering	0.00010	0.00018	0.00007	0.00019
Scaffolds with no heterozygous sites	134 (80%)	720 (74%)	130 (77%)	712 (73%)
Size of scaffolds with no heterozygous sites (kb)	26,625 (31%)	55,018 (5%)	20,010 (23%)	65,907 (6%)
Scaffolds with heterozygous sites	34	254	38	262
Heterozygous sites per window (50kb) (median)	0	1	0	2
Standard deviation per window (50kb)	43.0	12.3	8.1	12.2
Coefficient of dispersion (CD)	360.6	16.1	20.2	15.6
Windows with no heterozygous site	1,398	10,264	1,267	9,857
Windows with heterozygous sites	304	11,619	435	12,026
	Female Z	Female A	Male Z	Male A
Proportion of heterozygous sites before filtering	0.00534	0.00067	0.00050	0.00065
Proportion of heterozygous sites after filtering	0.00010	0.00018	0.00007	0.00019
Scaffolds with no heterozygous sites	134 (80%)	720 (74%)	130 (77%)	712 (73%)
Size of scaffolds with no heterozygous sites (kb)	26,625 (31%)	55,018 (5%)	20,010 (23%)	65,907 (6%)
Scaffolds with heterozygous sites	34	254	38	262
Heterozygous sites per window (50kb) (median)	0	1	0	2
Standard deviation per window (50kb)	43.0	12.3	8.1	12.2
Coefficient of dispersion (CD)	360.6	16.1	20.2	15.6
Windows with no heterozygous site	1,398	10,264	1,267	9,857
Windows with heterozygous sites	304	11,619	435	12,026

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 S6). 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 (Fig. 2B). The filtering removed most of this excess heterozygosity in the female (Fig. 2C, 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 S4 and S5). 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) (Table 1).

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).

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 (Figs. 3A and 3B, 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.

Comparison of the four methods.

Table 2
Classification of scaffolds identified as Z or autosomal scaffolds. Classification for each of the approaches: depth, heterozygosity, LASTZ and SNP-loading analysis. The identification was found by comparison to the genome bAquChr1.2 (GCA_900496995.2) with known chromosomes. Results for the different methods are given in a) for total size of scaffolds (bp), and in b) for the number of scaffolds, missing is compared to the golden eagle scaffold assembly.

		Depth		Heterozygosity		LASTZ		SNP-loading	
a)		Z	A	Z*	A	Z	A	Z	A
	Z	76,239,124	10,056,095	-	60,214,856	86,569,008	270,522	69,355,267	13,642,226
	A	93,786	*1,100,765,118*	-	1,050,885,219	597,603	1,105,305,943	11,720,756	1,078,283,284
	Total		*1,187,154,123*		*1,159,757,217*		*1,192,725,744*		*1,173,001,533*
	Missing		*5,571,621*		*29,104,198*		*0*		*19,714,211*
b)	Z	79	13	-	34	110	58	28	12
	A	14	512	-	254	33	941	9	231
	Total		618		*365*		*1,141*		*280*
	# NA		*523*		*776*		*0*		*861*

*values not assigned due to lack of heterozygosity on the Z chromosome

Using chromosome assignments obtained by mapping the golden eagle scaffold assembly to the golden eagle genome with assigned chromosomes, the most successful method, finding 98.12% of the expected size, was mapping to the chicken genome (Table 2, Fig. 4). In second place was the depth analysis with 86.41% and, in third, the SNP-loading with 78,61%. Heterozygosity was poorly suited to find Z-chromosomal scaffolds 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 categorised as Z-chromosomal scaffolds (0.09, 0.59 and 11.72 Mb, respectively). All approaches resulted in false negatives i.e., Z-chromosomal scaffolds categorised as autosomal (Table 2), but least with mapping to the chicken (0.27 Mb), whereas depth, heterozygosity, and SNP-loading had 10.05, 60.21 and 13.64 Mb of false negatives, 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 golden eagle with assembled chromosomes revealed 98.42% of the whole known Z-chromosome (Table 2, Fig. 4). Though the goal of the study was to evaluate the approaches separately, a combined analysis (Fig. 4), where at least two of our three approaches, depth, mapping to the chicken, and SNP-loading, were compared, 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.

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 which had been mapped on the golden eagle scaffold assembly. The success of the methods varied as they may be affected differently by the small population size of the study species. The approaches applying heterozygosity and PCA, can be expected to be more affected in a small population, as they analyse diversity in the genome and population, whereas depth and especially mapping can be expected to be less or not at all affected by the low diversity in a small population.

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 [17] and in the case of mapping the golden eagle scaffold assembly to the chicken, with a split time > 80 million years [34], 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. Just as with all the other analysis, the SNP-loading approach also found false negatives (Table 2), but as the focus of this paper was the identification of the Z-chromosome, this was not studied further, but we feel this deserves further research.

Here the approach of looking at all scaffolds in a single PCA was used, however this could also be done in sliding windows [35], looking at signals different from the overall population signal, which could potentially optimize this method. However, this also require diversity on the homogametic sex chromosome in males compared to females, which may be lacking in small populations such as in the Icelandic white-tailed eagle.

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 [36] and thus being a Z/W. In birds, PAR vary greatly in size from just a few Mb to more than 60 Mb [25]. Alternatively, they could represent non-recombining homologous regions (gametologs) [17, 37] 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. Thus, we find that two of the scaffolds (NW_011950990.1 and NW_011951051.1), display a higher heterozygosity ratio in the female compared to the male (17, and 2.5 times higher, respectively), as expected for gametologous regions, whereas the two others (NW_011950951.1, and NW_011951047.1) may present PARs, as they display a ratio close to one between the sexes (1.08 and 0.78, respectively).

Inspection of the heterozygosity for all scaffolds revealed that it is difficult to distinguish between autosomal and Z-chromosomal scaffolds without any prior knowledge. However, there was 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 [6], have reduced 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. As shown in others species [38], 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.

Although depth analysis has shown to be a promising method to identify sex chromosomes [21, 39], 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 [17, 40, 41]. 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 and negatives. Further, it should be noted that the methods used here maybe more applicable in taxa with relative stable sex chromosomes, such as mammals and birds [17, 18], but less effective in taxa such as fish where the sex chromosomes can be less differentiated [19, 20].

Though the use of data from such divergent species as the assemble from golden eagle and the white-tailed eagle genomes can be problematic due to potential deviations in synteny, and dynamic nature of the Z-chromosome as shown for songbirds [32, 42], we feel this study bares value and relevance. Firstly, the approach using a different reference from the study species is a relevant scenario and not seldom used [2–4], secondly the use of a different reference species has been suggested, as a reference of the same species have been shown to potentially introduce errors in the analyses [5]. Finally, using our novel white-tailed sequences against the golden eagle assemblies, with one scaffold assembled, and one chromosome assembled, made it possible to evaluate the approaches precision to a greater extent. Thus, this study highlights potential problems when trying to identify the homogametic sex chromosome especially in small populations. In such cases, methods which are less affect by the population size should be preferred.

Even though all known eukaryote species may soon be sequenced [43], 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.

Conclusion

The best performing approach for identifying the homogametic sex chromosome in a small island population, of white-tailed eagle, was the reference mapping to a related but distant species. The second-best approach was analysis read depth per scaffold, thirdly SNP-loading in PCA. Identification using genomic diversity approaches; SNP-loading and heterozygotic differences between sexes are potentially affected by the small population size and a recent population bottleneck.

Evaluation of these methods are highly relevant as genomic regions vary in effective population size and can have different evolutionary histories. Furthermore, the use of a different reference genome to the study species is still a widely used approach, which has several upsides.

Methods

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 135 chicks were extracted using the ThermoFisher GeneJET Whole Blood Genomics DNA Purification Mini Kit following the standard protocol [44]. 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 of 135 extracts were double digest restriction-site associated DNA sequenced on Illumina HiSeq2500 (see supplementary text 1 for full description).

Two individuals of white-tailed eagle (the last two of 135), 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 [45].

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 [9, 46].

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 version 1.47 [47, 48]. Default settings were used for the RADseq data, applying the option “r” to rescue barcodes and RAD-tags.

After demultiplexing, FastQC[49] was run for quality control. For the RADseq data, an excess of specific sequences (kmers) were removed using AdapterRemoval v2 (version 2.2.2) [50]. The high depth shotgun sequenced individuals were tested in the same way but found no excess of kmers.

The Burrows-Wheeler Aligner (BWA) mem and SAMtools version 0.7.17-r1188 and 1.7, respectively [51, 52] 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) [9] using default settings in both instances.

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

Four different approaches were used to identify scaffolds in the white-tailed eagle genome belonging to the Z-chromosome, by comparison with 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 Fig. 5 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 [53]. 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 (corresponding to 523 scaffolds, and 0.47% of the genome). 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 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 [54, 55] with default settings. 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 homologous regions. As scaffolds vary in length and may include short variable regions, the variation was also analysed per 50kb window. Genotypes were filtered for quality using vcftools and bcftools version 0.1.15 and 1.7, respectively [56, 57] 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 the short reads from the white-tailed eagle to chromosomes, the 1142 scaffolds from the golden eagle scaffold assembly (which the white-tailed eagle genome had been mapped on) were mapped to the chicken genome, which has assigned chromosomes, using LASTZ [26]. 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 version 1.0 [58], an extension of ANGSD [59], 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 [58] 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 [59, 60]. 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 (Fig. 5 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).

Declarations

Ethics approval and consent to participate

Not applicable

Consent for publication

Not applicable

Availability of data and materials

The raw dataset supporting the conclusions of this article is available in the DRYAD data repository https://doi.org/10.5061/dryad.v9s4mw6vs.

Further, the analysed dataset supporting the conclusions of this article is included in the supplementary.

Competing interests

The authors declare no competing interests.

Funding

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.

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.

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.

Authors' information

Charles Christian Riis Hansen[1*], Kristen M. Westfall[1,2], and Snæbjörn Pálsson[1,3]

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

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

[3]Senior Author

*Correspondence: CCRH

References

1. Hill WG, Robertson A. The effect of linkage on limits to artificial selection. Genetical Research. 1966;8:269–94. doi:10.1017/S0016672300010156.
2. de Manuel M, Barnett R, Sandoval-Velasco M, Yamaguchi N, Garrett Vieira F, Zepeda Mendoza ML, et al. The evolutionary history of extinct and living lions. Proceedings of the National Academy of Sciences. 2020;117:10927–34. doi:10.1073/pnas.1919423117.
3. Pedersen CET, Albrechtsen A, Etter PD, Johnson EA, Orlando L, Chikhi L, et al. A southern African origin and cryptic structure in the highly mobile plains zebra. Nature Ecology and Evolution. 2018;2:491–8. doi:10.1038/s41559-017-0453-7.
4. Pečnerová P, Garcia-Erill G, Liu X, Nursyifa C, Waples RK, Santander CG, et al. High genetic diversity and low differentiation reflect the ecological versatility of the African leopard. Current Biology. 2021; February.
5. Gopalakrishnan S, Samaniego Castruita JA, Sinding MHS, Kuderna LFK, Räikkönen J, Petersen B, et al. The wolf reference genome sequence (Canis lupus lupus) and its implications for Canis spp. population genomics. BMC Genomics. 2017;18:1–11.
6. Skarphéðinsson KH. Haförninn. Reykjavik: Fuglavernd (Fuglaverndarfélag íslands); 2013.
7. BirdLife International. Aquila chrysaetos. The IUCN Red List of Threatened Species 2016. 2016;e.T2269606.
8. BirdLife International. Haliaeetus albicilla. The IUCN Red List of Threatened Species 2016. 2016;e.T2269513. doi:10.2305/IUCN.UK.2016-3.RLTS.T22695137A93491570.en.
9. Doyle JM, Katzner TE, Bloom PH, Ji Y, Wijayawardena BK, DeWoody JA. The genome sequence of a widespread apex predator, the golden eagle (Aquila chrysaetos). PLoS ONE. 2014;9:20–2.
10. Zhang G, Li C, Li Q, Li BB, Larkin DM, Lee C, et al. Comparative genomics reveals insights into avian genome evolution and adaptation. Science. 2014;346:1311–20.
11. Kim JA, Kang SG, Jeon HS, Jeon JH, Jang JH, Kim S, et al. Complete mitogenomes of two Accipitridae, Haliaeetus albicilla, and Pernis ptilorhynchus. Mitochondrial DNA Part B: Resources. 2019;4:392–3. doi:10.1080/23802359.2018.1547155.
12. Damas J, O'Connor R, Farré M, Lenis VPE, Martell HJ, Mandawala A, et al. Upgrading short-read animal genome assemblies to chromosome level using comparative genomics and a universal probe set. Genome Research. 2017;27:875–84.
13. Warren WC, Clayton DF, Ellegren H, Arnold AP, Hillier LW, Künstner A, et al. The genome of a songbird. Nature. 2010;464:757–62.
14. Bellott DW, Skaletsky H, Pyntikova T, Mardis ER, Graves T, Kremitzki C, et al. Convergent evolution of chicken Z and human X chromosomes by expansion and gene acquisition. Nature. 2010;466:612–6.
15. Sigeman H, Ponnikas S, Chauhan P, Dierickx E, Brooke M de L, Hansson B. Repeated sex chromosome evolution in vertebrates supported by expanded avian sex chromosomes. Proc R Soc B. 2019;:20192051. doi:10.1098/rspb.2019.2051.
16. Palmer DH, Rogers TF, Dean R, Wright AE. How to identify sex chromosomes and their turnover. Molecular Ecology. 2019; September:1–16.
17. Xu L, Zhou Q. The female-specific W chromosomes of birds have conserved gene contents but are not feminized. Genes. 2020;11:1–14.
18. Graves JAM. Sex chromosome specialization and degeneration in mammals. Cell. 2006;124:901–14.
19. Kitano J, Peichel CL. Turnover of sex chromosomes and speciation in fishes. Environmental Biology of Fishes. 2012;94:549–58.
20. Kikuchi K, Hamaguchi S. Novel sex-determining genes in fish and sex chromosome evolution. Developmental Dynamics. 2013;242:339–53.
21. Nadachowska-Brzyska K, Burri R, Ellegren H. Footprints of adaptive evolution revealed by whole Z chromosomes haplotypes in flycatchers. Molecular Ecology. 2019;:mec.15021.
22. Feron R, Pan Q, Wen M, Imarazene B, Jouanno E, Anderson J, et al. RADSex: A computational workflow to study sex determination using restriction site-associated DNA sequencing data. Molecular Ecology Resources. 2021;21:1715–31.
23. Rangavittal S, Stopa N, Tomaszkiewicz M, Sahlin K, Makova KD, Medvedev P. DiscoverY: a classifier for identifying Y chromosome sequences in male assemblies. BMC Genomics. 2019;20:641. doi:10.1186/s12864-019-5996-3.
24. Nursyifa C, Brüniche-Olsen A, Garcia Erill G, Heller R, Albrechtsen A. Joint identification of sex and sex-linked scaffolds in non-model organisms using low depth sequencing data. Molecular Ecology Resources. 2021;:0–3.
25. Zhou Q, Zhang J, Bachtrog D, An N, Huang Q, Jarvis ED, et al. Complex evolutionary trajectories of sex chromosomes across bird taxa. Science. 2014;346.
26. Harris RS. Improved pairwise alignmen of genomic DNA. Ph.D. Thesis. The Pennsylvania State University; 2007. 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.
27. Kiełbasa SM, Wan R, Sato K, Horton P, Frith MC. Adaptive seeds tame genomic sequence comparison. Genome Research. 2011;21:487–93.

28. Noé L, Kucherov G. YASS: Enhancing the sensitivity of DNA similarity search. Nucleic Acids Research. 2005;33 SUPPL. 2:540–3.
29. Sætre G-P, Ravinet M. Evolutionary Genetics. 1st edition. Oxford University Press; 2019.
30. Jobling M, Hollox E, Hurles M, Kivisild T, Tyler-Smith C. Human evolutionary genetics. 2nd edition. Garland Science, Taylor & Francis Group, LLC; 2014.
31. Hooper DM, Price TD. Chromosomal inversion differences correlate with range overlap in passerine birds. Nature Ecology and Evolution. 2017;1:1526–34.
32. Xu L, Auer G, Peona V, Suh A, Deng Y, Feng S, et al. Dynamic evolutionary history and gene content of sex chromosomes across diverse songbirds. Nature Ecology and Evolution. 2019;3:834–44.
33. Sim SC, van Deynze A, Stoffel K, Douches DS, Zarka D, Ganal MW, et al. High-Density SNP Genotyping of Tomato (Solanum lycopersicum L.) Reveals Patterns of Genetic Variation Due to Breeding. PLoS ONE. 2012;7:1–18.
34. Jarvis ED, Mirarab S, Aberer AJ, Li B, Houde P, Li C, et al. Phylogenomic analyses data of the avian phylogenomics project. GigaScience. 2015;4:1–9.
35. Li H, Ralph P. Local PCA Shows How the Effect of Population. Genetics. 2019;211 January:289–304. https://search.proquest.com/docview/2168065525/fulltextPDF/40FB8A65E6B34C81PQ/1?accountid=12598.
36. Otto SP, Pannell JR, Peichel CL, Ashman TL, Charlesworth D, Chippindale AK, et al. About PAR: The distinct evolutionary dynamics of the pseudoautosomal region. Trends in Genetics. 2011;27:358–67.
37. Smeds L, Warmuth V, Bolivar P, Uebbing S, Burri R, Suh A, et al. Evolutionary analysis of the female-specific avian W chromosome. Nature Communications. 2015;6.
38. Zhou R, Macaya-Sanz D, Carlson CH, Schmutz J, Jenkins JW, Kudrna D, et al. A willow sex chromosome reveals convergent evolution of complex palindromic repeats. Genome Biology. 2020;21:1–19.
39. Huylmans AK, Toups MA, MacOn A, Gammerdinger WJ, Vicoso B. Sex-biased gene expression and dosage compensation on the artemia franciscana Z-chromosome. Genome Biology and Evolution. 2019;11:1033–44.
40. Trukhina A V., Smirnov AF. Problems of Birds Sex Determination. Natural Science. 2014;06:1232–40.
41. Graves JAM. Avian sex, sex chromosomes, and dosage compensation in the age of genomics. Chromosome Research. 2014;22:45–57.
42. Sigeman H, Ponnikas S, Hansson B. Whole-genome analysis across 10 songbird families within Sylvioidea reveals a novel autosome–sex chromosome fusion. Biology Letters. 2020;16.
43. Lewin HA, Robinson GE, Kress WJ, Baker WJ, Coddington J, Crandall KA, et al. Earth BioGenome Project: Sequencing life for the future of life. Proceedings of the National Academy of Sciences. 2018;115:4325–33.
44. Thermo Fisher. Thermo Scientific GeneJET Genomic DNA Purification Kit #K0721, #K0722. 2016; October:1–8. 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=VXNlciBHdWlkZTogR2VuZUpF\ Accessed 14 Feb 2020.
45. Illumina. TruSeq® Nano DNA Library Prep. https://support.illumina.com/content/dam/illumina-support/documents/documentation/chemistry_documentation/samplepreps_truseq/truseqnanodna/truseq-nano-dna-library-prep-guide-15041110-d.pdf. 2015.
46. Hillier LW, Miller W, Birney E, Warren W, Hardison RC, Ponting CP, et al. Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. Nature. 2004;432:695–716.
47. Catchen J, Hohenlohe PA, Bassham S, Amores A, Cresko WA. Stacks: an analysis tool set for population genomics. Molecular ecology. 2013;22:3124–40.
48. Catchen JM, Amores A, Hohenlohe P, Cresko W, Postlethwait JH, De Koning D-J. Stacks: Building and genotyping loci de novo from short-read sequences. Genes|Genomes|Genetics. 2011;1:171–82.
49. Babraham Bioinformatics. FastQC. 2010. http://www.bioinformatics.babraham.ac.uk/projects/fastqc/. Accessed 16 Jun 2016.
50. Schubert M, Lindgreen S, Orlando L. AdapterRemoval v2: rapid adapter trimming, identification, and read merging. BMC Research Notes. 2016;9:88. doi:10.1186/s13104-016-1900-2.
51. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009;25:1754–60.
52. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009;25:2078–9. doi:10.1093/bioinformatics/btp352.
53. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010;26:841–2. doi:10.1093/bioinformatics/btq033.
54. Eggertsson HP, Jonsson H, Kristmundsdottir S, Hjartarson E, Kehr B, Masson G, et al. Graphtyper enables population-scale genotyping using pangenome graphs. Nature Genetics. 2017;49:1654–60. doi:10.1038/ng.3964.
55. Eggertsson HP, Kristmundsdottir S, Beyter D, Jonsson H, Skuladottir A, Hardarson MT, et al. GraphTyper2 enables population-scale genotyping of structural variation using pangenome graphs. Nature Communications. 2019;10:1–8. doi:10.1038/s41467-019-13341-9.
56. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, et al. The variant call format and VCFtools. Bioinformatics. 2011;27:2156–8. doi:10.1093/bioinformatics/btr330.
57. Li H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics. 2011;27:2987–93.
58. Meisner J, Albrechtsen A. Inferring Population Structure and Admixture Proportions in Low-Depth NGS Data. Genetics. 2018;210:719–31. doi:10.1534/genetics.118.301336.

59. Korneliussen TS, Albrechtsen A, Nielsen R. ANGSD: Analysis of next generation sequencing data. BMC bioinformatics. 2014;15:356.
60. da Fonseca RR, Albrechtsen A, Themudo GE, Ramos-Madrigal J, Sibbesen JA, Maretty L, et al. Next-generation biology: Sequencing and data analysis approaches for non-model organisms. Marine Genomics. 2016;30:3–13.
61. Elshire RJ, Glaubitz JC, Sun Q, Poland JA, Kawamoto K, Buckler ES, et al. A Robust, Simple Genotyping-by-Sequencing (GBS) Approach for High Diversity Species. PLoS ONE. 2011;6(5):e19379.
62. Peterson BK, Weber JN, Kay EH, Fisher HS, Hoekstra HE. Double Digest RADseq: An Inexpensive Method for De Novo SNP Discovery and Genotyping in Model and Non-Model Species. PLoS ONE. 2012;7:e37135.

Figures

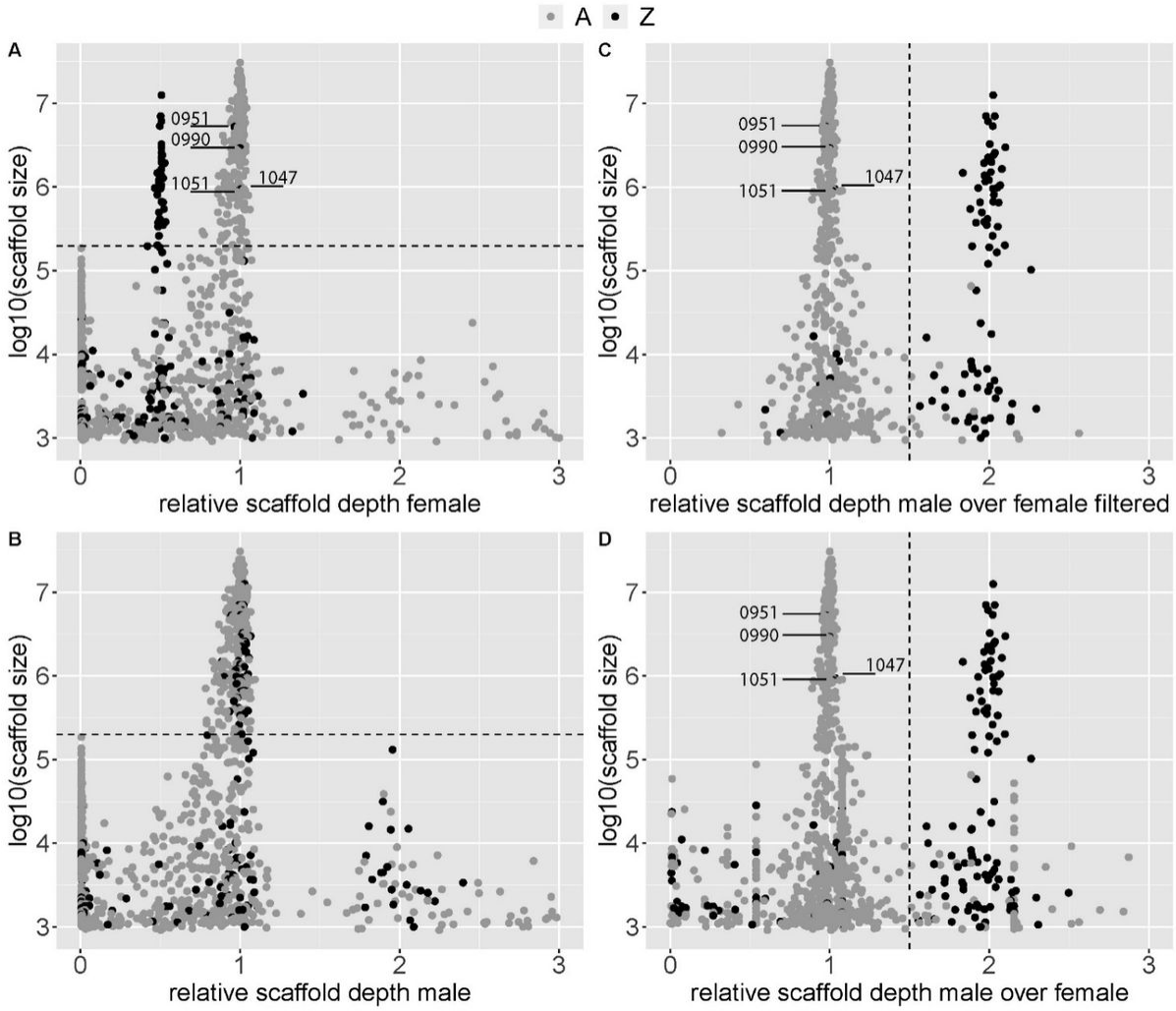

Figure 1

Relative sequencing depth of scaffolds in a male and a female white-tailed eagle. Relative scaffold depth was estimated as mode of scaffold depth / overall genomic depth, which was 195 for the female and 181 for the male. The dots, representing scaffolds, are shaded by whether they map to the Z or autosomal (A) chromosomes in the golden eagle genome with known chromosomes. A) Relative depth in the female. B) Relative depth in the male. C) The male to female ratio (rmf) of relative scaffold depth after filtering (removing scaffolds with relative depth outside the range of 0.25-1.5 in either the male or female). D) The male to female ratio (rmf) of relative depth for all scaffolds. In A and B the dashed line represents the scaffold size threshold value of 198,789 bases (log10 5.29). In A and B, points lower than the threshold value of 198,789 bases displayed high variation for relative depth (Figure S2). Scaffolds below the threshold in A and B make up 1.1% of data, only 0.0071% is below the threshold and above a relative depth of 3. Dashed line in C and D is 1.5, which is right between expectation for autosomal (1) and Z chromosomes scaffolds (2). “0951”, “0990”, “1047”, and “1051”, in A, C, and D, refers to the scaffolds NW_011950951.1, NW_011950990.1, NW_011951047.1 and NW_011951051.1.

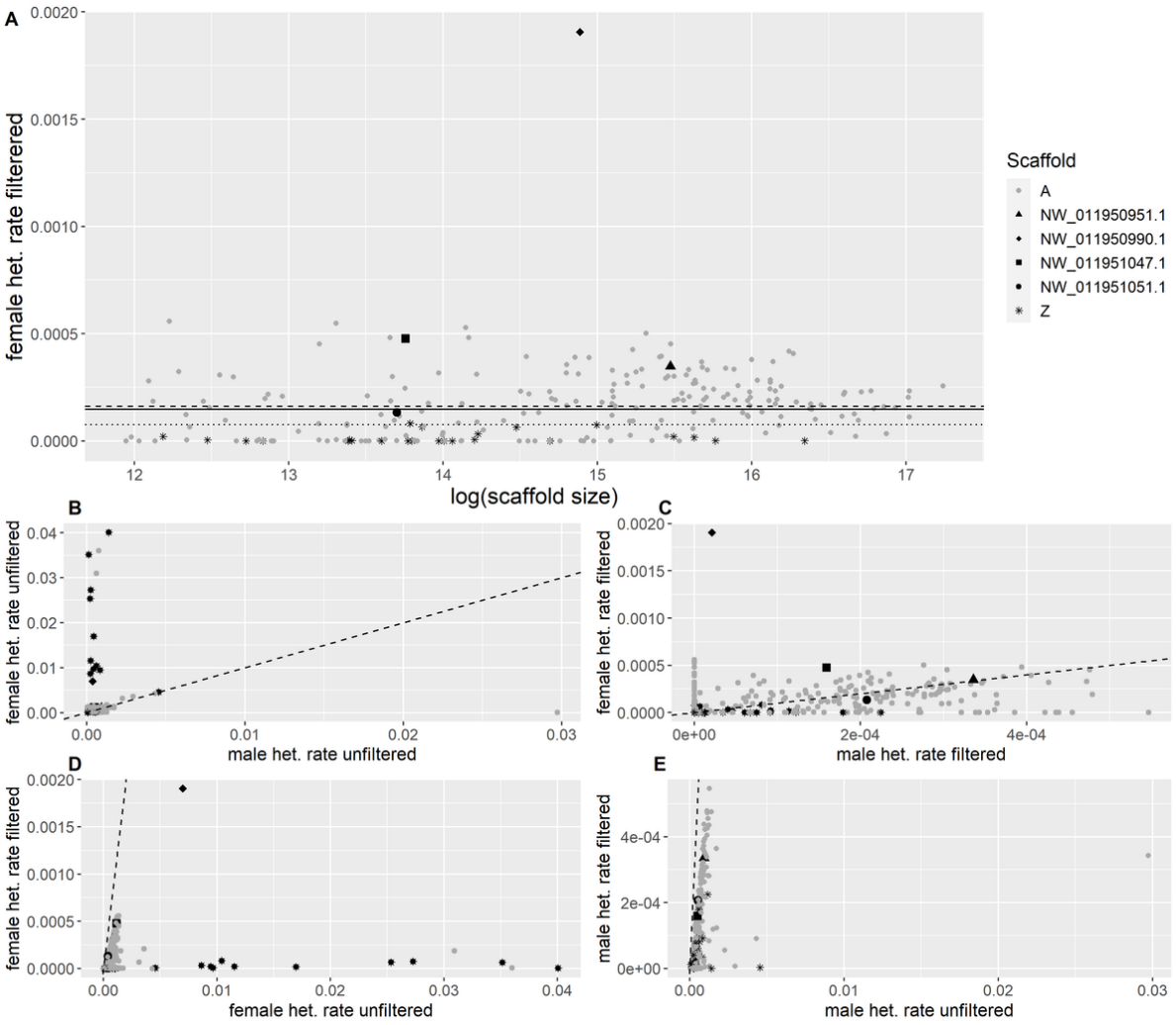

Figure 2

Heterozygosity rate of scaffolds mapped to autosomes and the Z-chromosome in white tailed eagle. Heterozygosity rate as number of heterozygous positions divided by length. A) filtered heterozygosity rate of the female. The dashed and dotted lines represent the mean filtered heterozygosity rate for autosomal and Z scaffolds, respectively. The full line is for all scaffolds. B) unfiltered heterozygosity rate for the female plotted against the male. C) filtered heterozygosity rate for the female plotted against the male. D) filtered versus unfiltered heterozygosity rate in the female. E) filtered versus unfiltered heterozygosity rate in the male. In all plots shape and color reflect scaffold type; grey dot is autosomal; black star Z-chromosomal; triangle, diamond, square and black dot are scaffolds NW_011950951.1, NW_011950990.1, NW_011951047.1, NW_011951051.1, respectively, which all show high heterozygosity in the female and have a relative depth as being autosomal. In B through E, dashed lines represent the identity line (slope=1).

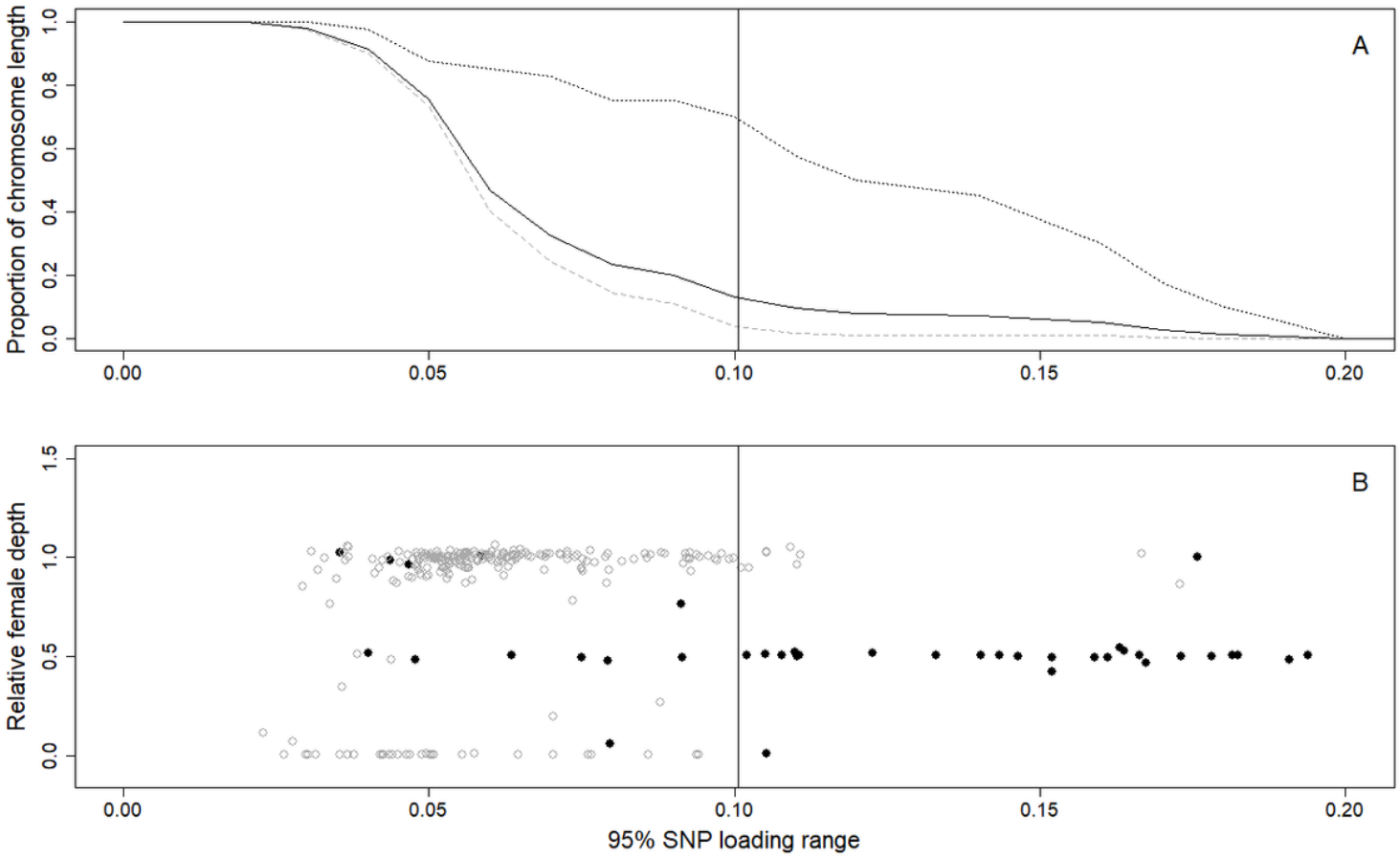

Figure 3

Proportion of chromosome length and female relative depth compared with SNP-loading range. A) Proportion of genome against 95% range of SNP-loading values for PC1. Dotted black line is Z-chromosomal, dashed grey line is autosomal and full black line is all scaffolds pooled. B) Relative scaffold depth the female sequenced to a high depth against 95% SNP-loading range. Open grey circles are autosomal scaffolds and full black dots are Z-chromosomal. The vertical line in both plots represents 3 SDs above the mean. For legibility, panel B was limited on the y axis to 0-1.5. Two scaffolds had relative depth greater than 1.5 (both >15), an autosomal scaffold around 0.025 SNP loading range, and a Z-chromosomal scaffold around 0.15.

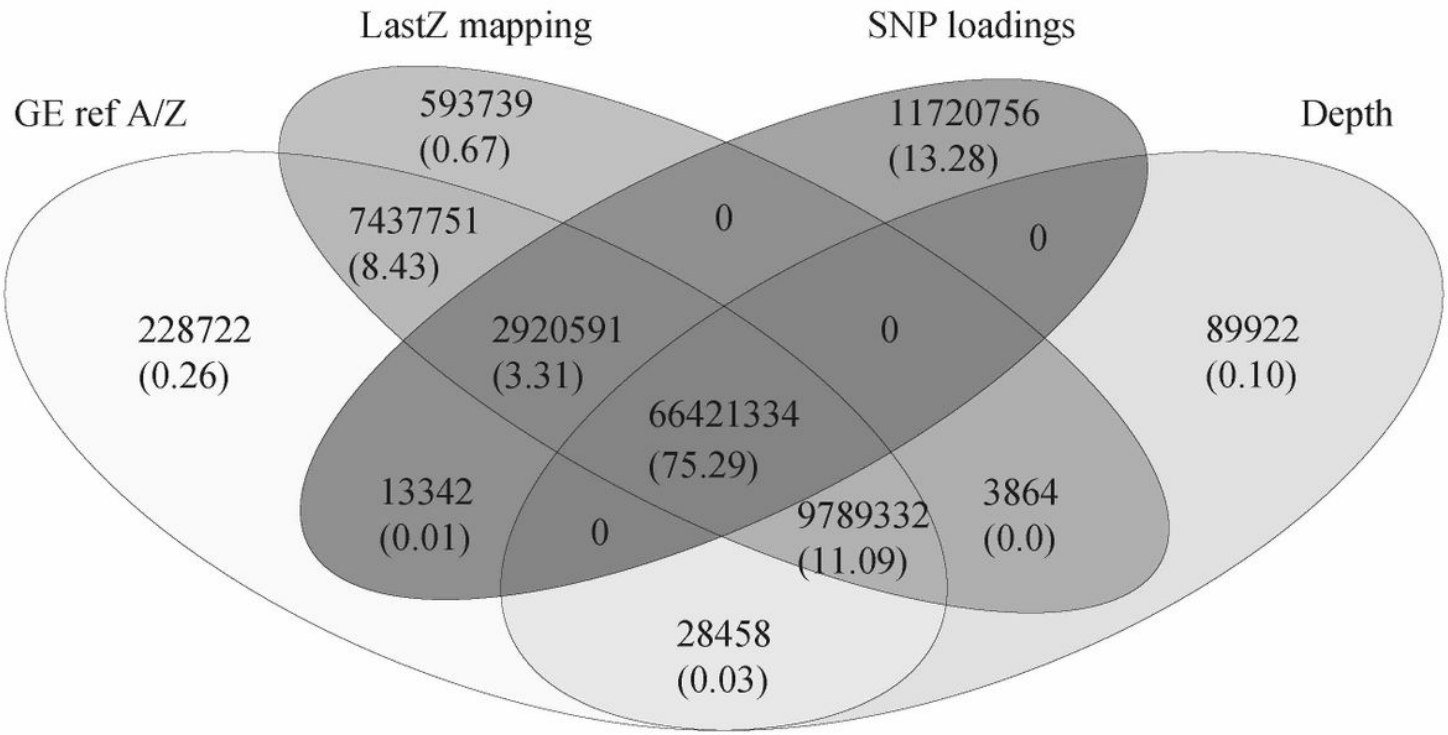

Figure 4

Venn diagram summarizing the size of scaffolds in bases identified as Z-chromosome with the three different analyses: mapping, depth and SNP-loadings. The Z-chromosomal scaffolds were assigned by mapping the genome with scaffolds to the genome with known chromosomes. Values in parentheses represent percentage size compared to the size of the known Z-chromosome. Notice that the percentage found by mapping the golden eagle scaffold assembly to the golden eagle genome is only 98.42%.

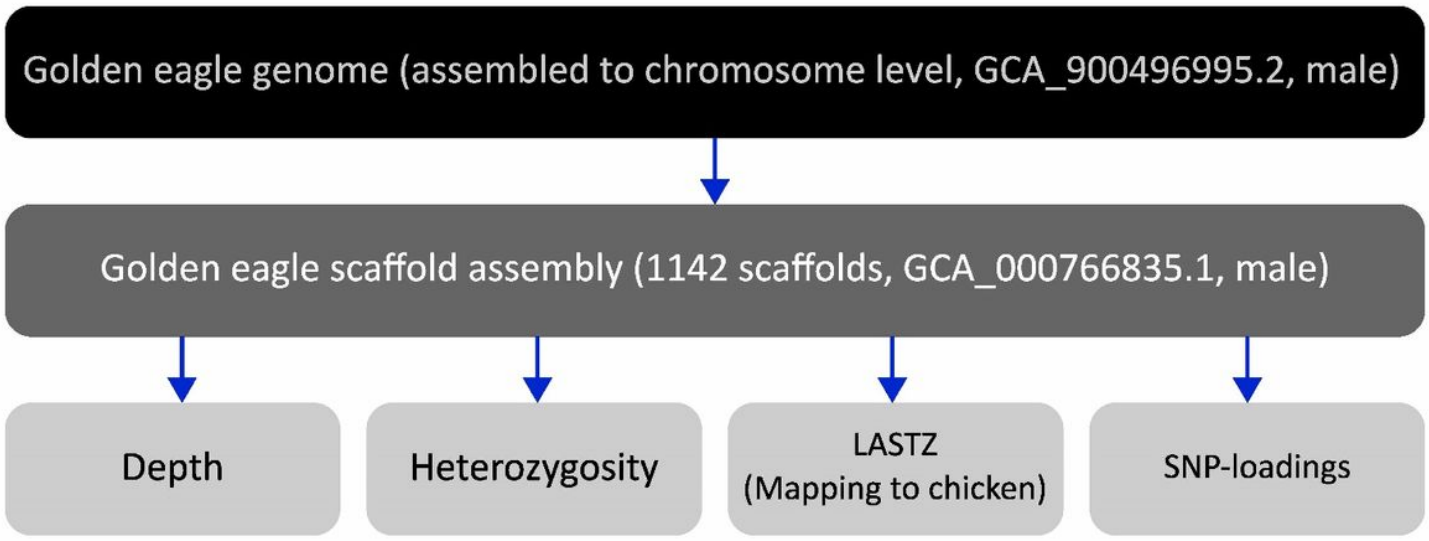

Figure 5

Schematic overview of the methods used to identify the Z-chromosome in a scaffold assembled genome. The golden eagle genome referred to in the dark grey box represents the reference in which we are attempting to identify scaffolds belonging to the Z-chromosome. The golden eagle genome in the black bar is the genome with known chromosomes, used to identify which scaffolds in the dark grey boxed genome probably belong to Z-chromosome (and

autosomes) – to use as a reference. The light grey boxes are the four approaches we tested to find the scaffolds belonging to the Z-chromosome: 1) Depth: analysis of difference in sequencing depth between scaffolds in a high depth whole genome sequenced white-tailed eagle female. 2) Heterozygosity: analysis of the difference in heterozygosity per scaffold a high depth whole genome sequenced white-tailed eagle male and female. 3) LASTZ: mapping of the golden eagle reference genome to the chicken genome using LASTZ. 4) SNP-loadings: analysis of SNP-loadings for principal components splitting the sexes, in 133 RADseq white-tailed eagle individuals.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- 9.DiscoveringtheZ100921SupplementaryTable1.xlsx
- 10.DiscoveringtheZ100921SupplementaryInformation.docx