A BIRD'S WHITE-EYE VIEW ON NEO-SEX CHROMOSOME EVOLUTION

Thibault Leroy[1], Yoann Anselmetti[1], Marie-Ka Tilak[1], Sèverine Bérard[1], Laura Csukonyi[1,2], Maëva Gabrielli[2], Céline Scornavacca[1], Borja Milá[3], Christophe Thébaud[2], Benoit Nabholz[1]*

[1] ISEM, Univ Montpellier, CNRS, IRD, EPHE, Montpellier, France
[2] Laboratoire Evolution et Diversité Biologique (EDB), UMR 5174 CNRS – Université Paul Sabatier – IRD, Toulouse, France
[3] National Museum of Natural Sciences (MNCN), Spanish National Research Council (CSIC), Madrid, Spain

** Corresponding author: benoit.nabholz@umontpellier.fr*

ABSTRACT

Chromosomal organization is relatively stable among avian species, especially with regards to sex chromosomes. Members of the large Sylvioidea clade however have a pair of neo-sex chromosomes which is unique to this clade and originate from a parallel translocation of a region of the ancestral 4A chromosome on both W and Z chromosomes. Here, we took advantage of this unusual event to study the early stages of sex chromosome evolution. To do so, we sequenced a female (ZW) of two Sylvioidea species, a *Zosterops borbonicus and a Z. pallidus*. Then, we organized the *Z. borbonicus* scaffolds along chromosomes and annotated genes. Molecular phylogenetic dating under various methods and calibration sets confidently confirmed the recent diversification of the genus *Zosterops* (1-3.5 million years ago), thus representing one of the most exceptional rates of diversification among vertebrates. We then combined genomic coverage comparisons of five males and seven females, and homology with the zebra finch genome (*Taeniopygia guttata*) to identify sex chromosome scaffolds, as well as the candidate chromosome breakpoints for the two translocation events. We observed reduced levels of within-species diversity in both translocated regions and, as expected, even more so on the neoW chromosome. In order to compare the rates of molecular evolution in genomic regions of the autosomal-to-sex transitions, we then estimated the ratios of non-synonymous to synonymous polymorphisms (π_N/π_S) and substitutions (d_N/d_S). Based on both ratios, no or little contrast between autosomal and Z genes was observed, thus representing a very different outcome than the higher ratios observed at the neoW genes. In addition, we report significant changes in base composition content for translocated regions on the W and Z chromosomes and a large accumulation of transposable elements (TE) on the newly W region. Our results revealed contrasted signals of molecular evolution changes associated to these autosome-to-sex chromosome transitions, with congruent signals of a W chromosome degeneration yet a surprisingly weak support for a fast-Z effect.

Key-words : Sex chromosome, molecular evolution, molecular dating, bird diversification, Sylvioidae, *Zosterops*.

INTRODUCTION

Spontaneous autosomal rearrangements are common across higher metazoan lineages (Choghlan et al. 2005). By contrast, sex chromosome architectures are much more conserved, even across distant lineages (Murphy et al. 1999; Raudsepp et al. 2004; Fraïsse et al. 2017). A growing number of studies have recently observed departures from this pattern of evolutionary conservation by detecting changes in the genomic architecture of sex chromosomes in some particular lineages – so-called neo-sex chromosomes – that are mainly generated by fusion or translocation events of at least one sex chromosome with an autosome (*e.g.* Kitano & Peichel, 2012; Zhou & Bachtrog, 2012). Considering the long-term conservation of sex chromosome synteny, neo-sex chromosomes provide opportunities to investigate the processes at work during the early stages of sex chromosome evolution (Charlesworth et al. 2005). Previous detailed studies have for example investigated its important role for divergence and speciation (*e.g.* Kitano et al. 2009; Weingartner & Delph, 2014; Yoshida et al. 2014; Bracewell et al. 2017).

From a molecular evolution point of view, an important consequence of the transition from autosomal to sex chromosome is the reduction in effective population size (N_e). Assuming a 1:1 sex ratio, N_e of sex-linked regions on the Y (or W) and X (or Z) chromosomes are expected to decrease by three-fourths and one-fourths, respectively (see Ellegren, 2009 for details). According to the neutral theory, nucleotide diversity is then expected to be reduced proportionally to the reduction in N_e due to increased drift effects (Vicoso & Charlesworth, 2006; Pool & Nielsen, 2007). N_e reduction also induces a change in the balance between selection and drift, with drift playing a greater role after the translocation, thus reducing the efficacy of natural selection to purge deleterious mutations from populations. Mutations - including deleterious ones - may also drift to fixation at a faster rate in sex-chromosomes than in autosomes, thus generating expectations for faster evolution at X and Y chromosomes (so-called fast-X or fast-Y effects) (Mank et al. 2007; Rousselle et al. 2016). In addition, given that mutations are on average recessive, positive and negative selection are expected to be more efficient in the heterogametic sex (Hvilsom et al. 2012; Nam et al. 2015). Suppression of recombination is also expected to initiate a degenerative process on the Y chromosome, that may result in the accumulation of nonsynonymous deleterious substitutions owing to a series of processes acting simultaneously: Muller's ratchet, the Hill-Robertson effect, and linked selection (Charlesworth and Charlesworth, 2000). For the same reason, transposable elements (TEs) are also expected to accumulate soon after the cessation of recombination in the Y chromosomes (Charlesworth 1991; Charlesworth et al. 1994).

Except for few reported examples (de Oliveira et al. 2005; Nanda et al. 2006; Kapusta & Suh, 2017; O'Connor et al. 2018), most birds share a high degree of synteny conservation across autosomal chromosomes (Griffin et al. 2007; Nanda et al. 2008; Ellegren, 2010; Völker et al. 2010; Warren et al. 2010; Ellegren, 2013) and an even higher one at the Z chromosome (Nanda et al. 2008). A notable exception is the neo-sex chromosome of Sylvioidea species, a superfamily of passerine birds in which a translocation of a large part of the Zebra finch 4A chromosome onto both the W and the Z chromosomes, as characterized by genetic mapping (hereafter translocations of the neoW-4A and neoZ-4A on ancestral W and Z sex chromosomes; Pala et al. 2012a). Terminologically, these original sex chromosomes (W or Z) are considered as specific regions of the neoW and neoZ chromosomes (hereafter neoW-W and neoZ-

Z). All along the manuscript, we have used this terminology to emphasize the fact that these translocations also induce substantial evolutionary shifts on original sex chromosomes. These two autosome-to-sex chromosome transitions are present in reed warblers (Acrocephalidae), old-world warblers (Sylviidae) and larks (Alaudidae) (see also Brooke et al. 2010), and therefore likely occurred in the common ancestor of all present-day Sylvioidea (Pala et al. 2012a,b; Sigeman et al. 2018), between 15 and 30 million years ago (Myrs) (Ericson et al. 2014; Prum et al. 2015; Nabholz et al. 2016). These two sex chromosome translocations provide unique opportunities to investigate the early stages of the W and Z chromosome evolution.

Based on phylogenetic trees calibrated using geological events (Moyle et al. 2009), *Zosterops* species of the family Zosteropidae (more commonly referred to as white-eyes) are considered to have emerged around the Miocene/Pliocene boundary. Considering both this recent emergence and the remarkable high diversity currently observed in this genus (more than 80 species), this group appears to have one of the highest diversification rates reported to date for vertebrates and is therefore considered as one of the 'great speciator' examples (Diamond et al. 1976; Moyle et al. 2009). White-eye species are typical examples of taxa spanning the entire "grey zone" of speciation (Roux et al. 2016). As a consequence of these different degrees of reproductive isolation between taxa, white-eyes have long been used as models to study bird speciation (*e.g.* Clegg and Philimore 2010; Melo et al. 2011; Oatley et al. 2012; Oatley et al. 2017). Among all white-eyes species, the Reunion grey white-eye *Zosterops borbonicus* received considerable attention over the last 50 years. This species is endemic from the volcanic island of Reunion and shows an interesting pattern of microgeographical variation, with five distinct colour variants distributed over four specific regions across the 2,500 km² of island surface. Both plumage color differentiation data (*e.g.* Gill et al. 1973; Milá et al. 2010; Cornuault et al. 2015) and genetic data (*e.g.* Milá et al. 2010; Delahaie et al. 2017) support this extensive within-island diversification. Despite its important role in the understanding of the diversification of *Zosterops* species, no genome sequence is currently available for this species. More broadly, only one *Zosterops* species has been sequenced to date (the silvereye *Z. lateralis,* Cornetti et al. 2015).

Here, we obtained detailed genome data from *Z. borbonicus*, a member of the Zosteropidae family and then arranged scaffolds into pseudochromosomes to provide insights into the evolutionary processes that may have contributed to the early stages of the sex chromosome evolution. We also generated a more fragmented genome sequence for *Z. pallidus* for molecular dating and molecular evolution analyses. We found similar breakpoint locations for both translocated regions suggesting evolution from the same initial gene sets, and studied the molecular evolution of the two newly sex-associated regions. By comparing levels of within-species nucleotide diversity at autosomal and sex chromosomes, we found support for a substantial loss of diversity on both translocated regions, largely consistent with expectations under neutral theory. We then compared patterns of polymorphisms and divergence at neo-sex chromosome genes and found support for a considerable fast-W effect, but surprisingly weak support for a fast-Z effect. Investigations of candidate changes in base composition led to the identification of specific signatures associated with abrupt changes in recombination rates (reduction or cessation) of the two neo-sex regions. Finally, we reported higher transposable elements (TE) content on the newly W than on the newly Z regions, suggesting ongoing neoW chromosome degeneration.

RESULTS

ZOSTEROPS BORBONICUS GENOME ASSEMBLY

Using a strategy combining long-read sequencing with PacBio and short-read Illumina sequencing with both mate-pairs and paired-end reads, we generated a high-quality reference genome for a female Reunion grey white-eye captured during a field trip to Reunion (Mascarene archipelago, southwestern Indian Ocean). The 1.22 gigabase genome sequence comprises 97,503 scaffolds (only 3,757 scaffolds after excluding scaffolds smaller than 10 kbp), with a scaffold N50 of 1.76 Mb (Fig. 1, Table S1). The completeness of the assembly is very high based on the BUSCO statistic (93.0%). Among all investigated avian species, the 'GRCg6a' chicken genome assembly is the only one exceeding this value (93.3%) (Fig. S1). Compared to the other species, our *Z. borbonicus* reference assembly also exhibits the lowest proportions of 'missing' (2.5%, a value only observed for the reference chicken genome) and 'fragmented' genes (4.3%, a value which is 0.1% higher than for the reference chicken assembly).

Using the PASA pipeline combined with EVidenceModeler (Haas et al. 2008; Haas et al. 2011) and several *in silico* tools trained by the RNAseq data, a total of 22,558 gene models were predicted. Among these 22,558 genes, half of the "hidden" avian genes identified by Botero-Castro et al. (2017), *i.e.* a large fraction of avian genes in GC-rich regions missing from most avian genome assemblies, were recovered (1,072 out of 2,132).

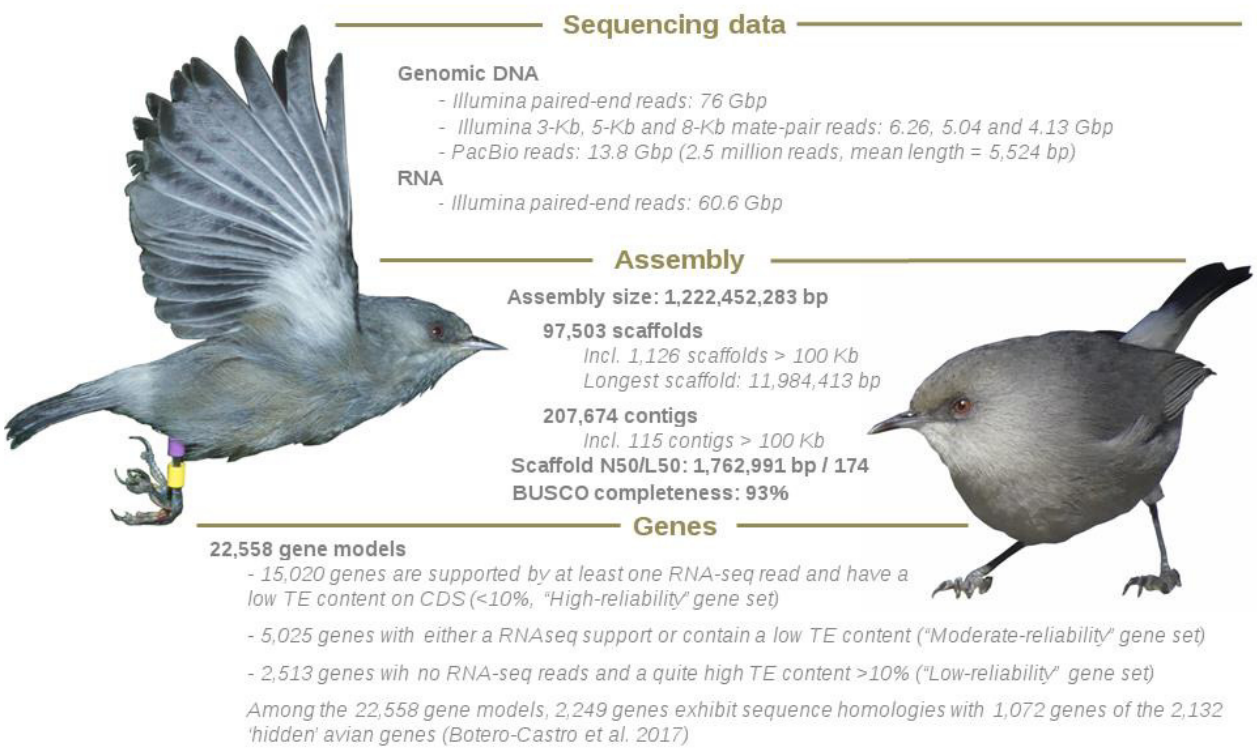


Figure 1*: Z. borbonicus *genome assembly and gene model summary statistics. *The reported numbers of reads and the total numbers of bases sequenced correspond to the trimmed reads. The three different gene sets are indicated as specified in the gff file. Photo credits: Maëva Gabrielli & Laurent Brillard (http://laneb.re)*

The vast majority of the 22,558 CDS (83.3%) are supported by at least one RNA-seq read (18,793 CDS), including 17,474 with a FPKM above 0.01. Among the 22,558 gene models, TE content is globally low (8.5%), but 4,786 genes exhibit a TE content in coding regions greater than 0.25, including 1,512 genes predicted to be TE over the total length of the CDS (Fig. S2). Still more broadly, the level of expression is strongly negatively correlated with the within-gene TE content (r^2=0.038, $p<2.2x10^{-16}$, after excluding 709 genes with FPKM>100).

Additionally, we generated a genome assembly of a female Orange River white-eye (*Z. pallidus*) using 10X mate-pair and 72X paired-end reads after cleaning. This assembly is much more fragmented (170,557 scaffolds and scaffold N50 = 375 Kb, Table S1) than the *Z. borbonicus*. Both *Z. borbonicus* and *Z. pallidus* assemblies are available from Figshare repository URL: https://figshare.com/s/122efbec2e3632188674 (see also the data availability section).

Reference-assisted genome organization

We then anchored scaffolds using the v3.2.4 reference genome of the zebra finch (Warren et al. 2010) assuming synteny. We used the zebra finch as a pivotal reference since this reference sequence is of high-quality, with 1.0 of 1.2 gigabases physically assigned to 33 chromosomes including the Z chromosome, plus three additional linkage groups based on genetic linkage and BAC fingerprint maps (Warren et al. 2010). We anchored 928 among the longest scaffolds to the zebra finch chromosomal-scale sequences, thus representing a total of 1.01 Gb among the 1.22 Gb of the *Z. borbonicus* assembly (82.8%).

In parallel to the zebra finch-oriented approach, we used DeCoSTAR (Duchemin et al. 2017), a tool that improves the assembly of several fragmented genomes by proposing evolutionary-induced adjacencies between scaffolding fragments containing genes. To perform this analysis, we used the reference sequences of 27 different avian species (Table S2) with associated gene tree phylogeny of 7,596 single copy orthologs (trees are available from the following Figshare repository, URL: https://figshare.com/s/122efbec2e3632188674). Among the 97,503 scaffolds (800 containing at least one orthologous gene), DeCoSTAR organized 653 scaffolds into 188 super-scaffolds for a total of 0.837 Gb (68.5% of the *Z. borbonicus* assembly), thus representing a 2.59-fold improvement of the scaffold N50 statistic (4.56 Mb). Interestingly, among the 465 scaffold junctions, 212 are not only supported by gene adjacencies within the other species, but are also supported by at least one *Z. borbonicus* paired-end read. From a more global point of view, DeCoSTAR not only improved the *Z. borbonicus* genome, but also those of 25 other species (mean gain in scaffold N50 over 11 Mb, representing 3.30-fold improvement on average, range: 1.0-6.02). The only exception is the already well-assembled chicken genome reference. For all these species, the proposed genome organizations ("agp files") were made available at the following URL: https://figshare.com/s/122efbec2e3632188674.

We then combined zebra finch-oriented and DeCoSTAR approaches for the *Z. borbonicus* genome, by guiding DeCoSTAR using the a priori information of the zebra finch-oriented approach to get beyond two limitations. First, DeCoSTAR is a gene-oriented strategy, and thus cannot anchor scaffolds without genes that have orthologous analogues in the other species, which is generally the case for short scaffolds. Second, the zebra-finch oriented approach assumes a perfect synteny and collinearity between *T. guttata* and *Z. borbonicus*, which is unlikely. By combining both approaches, we were able to anchor 1,082 scaffolds, including 1,045 scaffolds assigned to chromosomes representing a total 1.047 Gb (85.7% of the *Z. borbonicus* assembly). In addition, DeCoSTAR

helped propose more reliable *Z. borbonicus* chromosomal organizations for these 1,045 scaffolds by excluding some *T. guttata*-specific intra-chromosomal inversions.

Assigning scaffolds to W and Z chromosomes

To identify sex chromosome scaffolds, we first mapped trimmed reads from males and females *Z. borbonicus* individuals which were previously sequenced by Bourgeois et al. (2017) and then computed median per-site coverage over each scaffold for males and females (Fig. 2). After taking into account differences in coverage between males and females, we then identified scaffolds that significantly deviated from 1:1 and identified neoW and neoZ scaffolds (see methods section). This strategy led to the identification of 218 neoW (7.1 Mb) and 360 neoZ scaffolds (91.8 Mb) among the 3,443 scaffolds longer than 10 kb (Fig. 2). Among the 360 neoZ scaffolds assigned by coverage, 339 scaffolds were already anchored to the neoZ chromosome based on the synteny-oriented approach, thus confirming the accuracy of our previous assignation and suggesting that we have generated a nearly complete Z chromosome sequence The list of scaffolds identified on the two neo-sex chromosomes was made available : https://figshare.com/s/5ad54809ed89dba83db7.

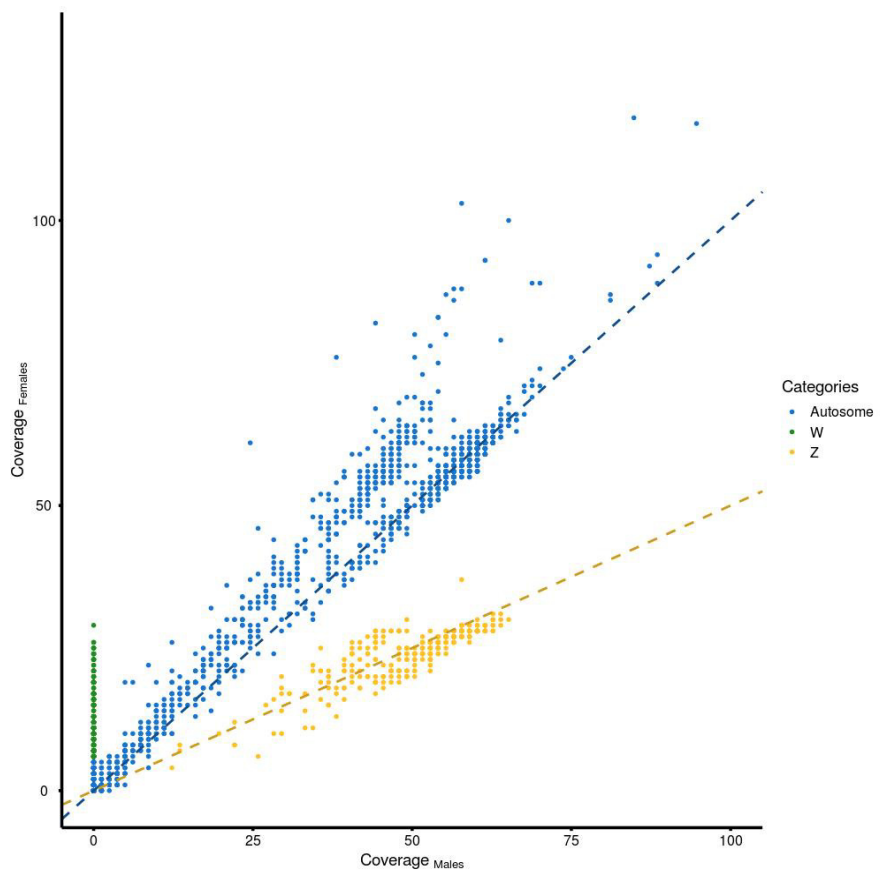


Figure 2: Identification of Z and W scaffolds using coverages of male and female individuals. *Normalized sum of median per-site coverage of trimmed reads over the five males (x-axis) and seven females (y-axis). Values were only calculated for scaffolds longer than 10-Kb. Linear regressions for slope=1 (blue) and slope=0.5 (yellow) are shown. Scaffolds were assigned following the decision rule described in the material and methods section.*

Due to the absence of a chromosomal-scale W sequence for a passerine bird species, we are unable to provide a chromosomal structure for the 218 neoW scaffolds. We however assigned 174 among the 218 neoW scaffolds to the neoW-4A region. These scaffolds representing a total length of 5.48 Mb exhibited high levels of homology with the zebra finch 4A chromosome. We found support for neoW-4A scaffolds aligning between positions 21,992 and 9,603,625 of the *T. guttata* 4A chromosome (thus representing 57% of the corresponding 9.6 Mb 4A region). Leaving aside the difficulty of sequencing and assembling the neoW chromosome (Tomaszkiewicz et al. 2017), a large part of the difference between the total assembled size and the corresponding region in *T. guttata* is likely due to a 1.75 Mb chromosomal deletion on the neoW-4A. Indeed, we found no neoW scaffold with a homology to the large 4A *T. guttata* region between positions 1,756,080 and 3,509,582. Based on the *T. guttata* reference genome, this region was initially gene-poor, since only two *T. guttata* genes were found in this large genomic region (*i.e.* 1.1 genes/Mb), as compared to the 142 genes observed on the whole translocated region (*i.e.* 14.8 genes/Mb). Remaining neoW scaffolds, *i.e.* those having no homology with the 4A *T. guttata* region, were considered as sequences belonging to the ancestral W chromosome (hereafter neoW-W region), except for six scaffolds showing reliable hits but only at specific locations of the scaffold, and for which the accuracy of any assignation was considered too low.

Molecular dating

The molecular phylogenetic analyses were aimed at estimating the divergence time of the *Zosterops* genus. Indeed, even if our focal dataset is composed of only three species, the divergence between *Z. pallidus / Z. borbonicus* and *Z. lateralis* represents the first split within the genus except *Z. wallacei*, *i.e.* the origin of clade B in Moyle et al. (2009). As a consequence, our phylogeny (Fig. 3, S3) is expected to provide an accurate estimate for the onset of the diversification of the *Zosterops* lineage. For this molecular dating, we added three species to the 27 species used in the previous analyses and generated gene sequence alignments. Given that these newly added species - namely the silvereye, Orange River white-eye and willow warbler - have no gene models available yet, we used the AGILE pipeline (Hughes & Teeling 2018, see method) to obtain orthologous sequences. Due to the inherent computational burden of Bayesian molecular dating analyses, we randomly selected 100 alignments among the least GC rich single-copy orthologs and performed ten replicated analyses (chronograms available at FigShare URL: https://figshare.com/s/122efbec2e3632188674). Indeed, GC poor genes are known to be slowly and more clock-like evolving genes as compared to the other genes (Jarvis et al. 2014).

We used several combinations of fossil calibrations and substitution models (Table S3). For the radiation of Neoaves, all analyses led to molecular dating consistent with Jarvis et al. (2014) and Prum et al. (2015) with estimates around 67-70 Myrs, except for the calibration set 4 (82 Myrs), albeit with large 95% confidence intervals (CI = 64 - 115 Myrs) (Table 1). Calibration set 4 is very conservative with no maximum calibration bound except for the Paleognathae / Neognathae set to 140 Myrs. In contrast, calibration set 3 is the more constrained with the Suboscines / Oscines split bounded between 28 - 34 Myrs. Unsurprisingly, this calibration led to the youngest estimates, dating the origins of passerines at 59 Myrs (56 - 63 Myrs) and the Paleognathae / Neognathae at 65 Myrs (63 - 69 Myrs).

In all runs, our estimates of the origin of *Zosterops* have lower limits of CI including 2.5 Myrs and mean estimates are also often considerably lower than this value (Table 1). Using the calibration set 1, the ten replicated datasets gave a mean age estimate for the origin of *Zosterops* around 2 million years ago (Table 1).

Table 1: Molecular dating analyses. *Mean dates are indicated and 95% confidence intervals (CI) are provided within parentheses*

Datasets*	Calibration set**	Crown *Zosterops*	*Zosterops* / *Phylloscopus*	*Corvus* / Passerida	Passerines	Crown Neoaves
1	1	2.8 (1.7, 4.2)	17.2 (14.2, 20.4)	25.8 (22.6, 28.9)	62 (58.3, 65.5)	67.1 (63.2, 71.2)
1	2	3.1 (1.8, 4.6)	18.4 (15.2, 22.2)	27.3 (23.7, 31.5)	64.8 (58.5, 71.9)	70.1 (63.5, 77.5)
1	3	1.9 (1.3, 2.8)	12.7 (10.9, 14.7)	19.5 (17.5, 21.4)	59.1 (56.5, 62.9)	65 (63, 69.1)
1	4	3.5 (1.9, 6.1)	21.4 (15.4, 31.6)	31.9 (23.9, 45.2)	75.9 (59.4, 106.4)	82.1 (64.2, 114.9)
2	1	0.9 (0.5, 1.4)	7.8 (5.1, 10.3)	13.4 (9.7, 16.5)	62.9 (58.9, 66.3)	68 (63.7, 71.9)
3	1	2.0 (1.3, 3.2)	12.4 (9.7, 16.2)	19.5 (15.9, 23.5)	62 (58.5, 65.6)	67.1 (63.1, 71.1)
4	1	1.8 (1.2, 2.9)	11.9 (8.6, 15.4)	18.7 (14.1, 22.8)	61.1 (57.5, 65.3)	68.2 (64.2, 72.6)
5	1	1.7 (1, 2.6)	11 (8, 14.7)	17.3 (13.1, 21.8)	61.6 (57.7, 65.4)	68.1 (64, 72.3)
6	1	3.2 (1.9, 4.9)	16.8 (13.9, 19.9)	23.8 (20.8, 26.5)	62 (58.3, 65.4)	67.2 (63.1, 71.3)
7	1	2.2 (1.2, 3.5)	13.3 (9.8, 16.8)	20.6 (16.5, 24.3)	64.7 (62.9, 65.7)	71.8 (69.7, 73.9)
8	1	2.1 (1.3, 3.5)	13.2 (9.9, 17.2)	21.2 (16.9, 25.3)	64.6 (62.4, 65.7)	69.7 (67.3, 71.6)
9	1	1.5 (0.9, 2.5)	9.5 (6.6, 12.6)	16.9 (12.8, 20.9)	65.3 (63.5, 66.7)	72.3 (69.9, 74.5)
10	1	1.9 (1.3, 2.8)	12.4 (9.6, 14.9)	19.4 (15.6, 22.3)	61.7 (57.4, 65.5)	68.5 (63.8, 73.1)

** Different numbers indicate independent replicated datasets of 100 randomly selected orthologs*

*** Number corresponds to the different calibration sets used and presented in Table S3*

415

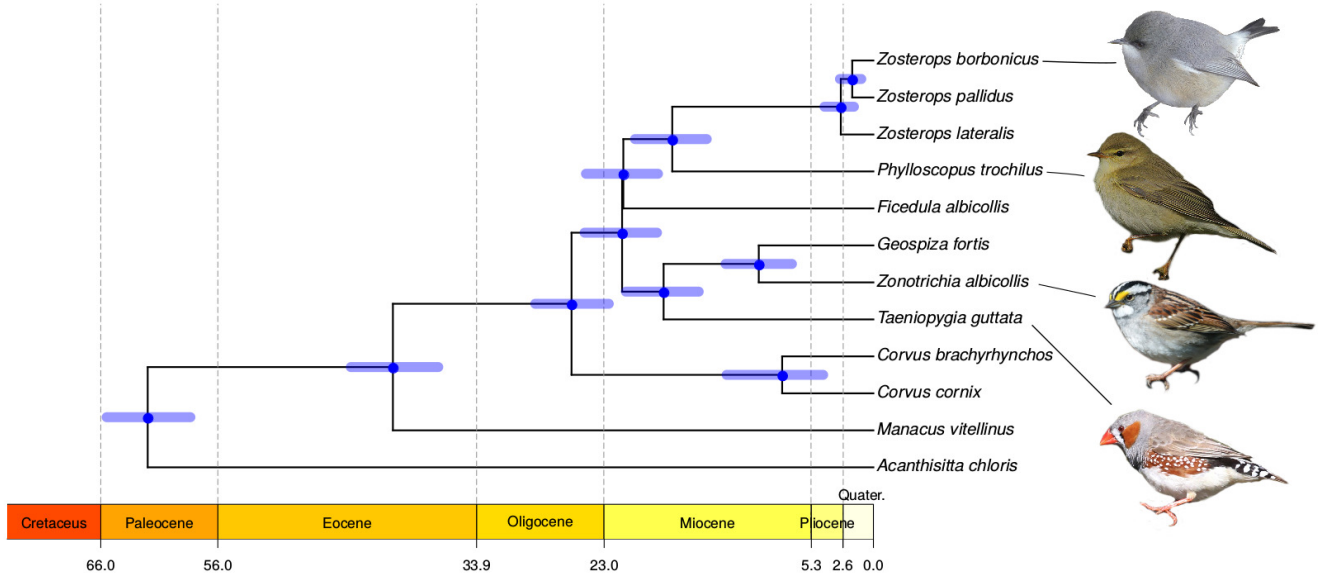


Figure 3: Molecular phylogeny and dating of the investigated passerine birds. *Estimates of molecular dating are based on dataset 1 and calibration set 1 (Table S3). For greater clarity, the cladogram focuses on passerine birds only, but see Fig S3 for a cladogram of the inferred phylogeny for the whole set of investigated avian species.*

More broadly, our analysis is consistent with a recent origin of *Zosterops*, with a crown clade age of less than 5 Myrs and probably between 1 and 3.5 Myrs (Table 1).

Additionally, we performed a completely independent molecular dating by applying the regression method proposed in Nabholz et al. (2016). Based on nine full *Zosterops* mitochondrial genomes, we calculated molecular divergence. The divergence between *Z. lateralis* and the other *Zosterops* has a median of 0.205 subst./site (max = 0.220, min = 0.186) for the third codon position. Assuming a median body mass for the genus of 10.7 g (Dunning, 2007), we estimated a divergence date between 2.3 and 6.2 Myrs. These estimates are in line with our previous dating based on nuclear data and with molecular dating based on fossil calibration confirming the extremely rapid diversification rate of white-eyes.

CHROMOSOMAL BREAKPOINTS

Our synteny-oriented approach using pairwise whole-genome alignments of *Z. borbonicus* and *T. guttata* sequences helped us in identifying the chromosomal breakpoint of the neoZ sex chromosome. We reported a scaffold (scaffold329) with long sequence alignments with both the 4A and the Z chromosomes (Fig. 4). Considering the intervals between the last LASTZ hit on the 4A and the first one on the Z chromosome, we estimated that this breakpoint occurred between positions 9,605,374 and 9,606,437 of the 4A zebra finch chromosome (genome version : v.3.2.4), which is fairly close to the estimate of 10 Mb previously reported by Pala et al. (2012a). Based on the soft-masked version 3.2.4 of the zebra finch genome, this 1-Kb region is well assembled (no ambiguous “N” bases) and shows no peculiarities in TE or GC content (7.4%

and 39.0%, respectively) as compared to the rest of the zebra finch 4A chromosome (18.7% and 43.7%).

To discard the potential confounding factor of a sequence artifact due to a chimerism in the *Z. borbonicus* assembly, we used the procedure for the sequence assembly of *Z. lateralis* (Cornetti et al. 2015) and identified a scaffold (LAII01000008.1) supporting the same chromosomal breakpoint. Using the *Z. lateralis* sequence, we estimated that this breakpoint occurred between positions 9,605,524 and 9,606,431.

We then investigated the chromosomal breakpoint for the neoW (Fig. 4). We identified a candidate scaffold, *scaffold1768_size33910*, with alignment hits on both the 4A *T. guttata* and the W of *F. albicollis* (Smeds et al. 2015). Among all W scaffolds, this scaffold aligns at the latest positions of the 4A *T. guttata* chromosome (from 9,590,067 to 9,603,625), which is remarkably close to our estimation for the region translocated on the neoZ chromosome. Considering also alignments against contigs of the *F. albicollis* W chromosome, we estimated that the chromosomal breakpoint probably occurred between positions 9,603,625 and 9,605,621.

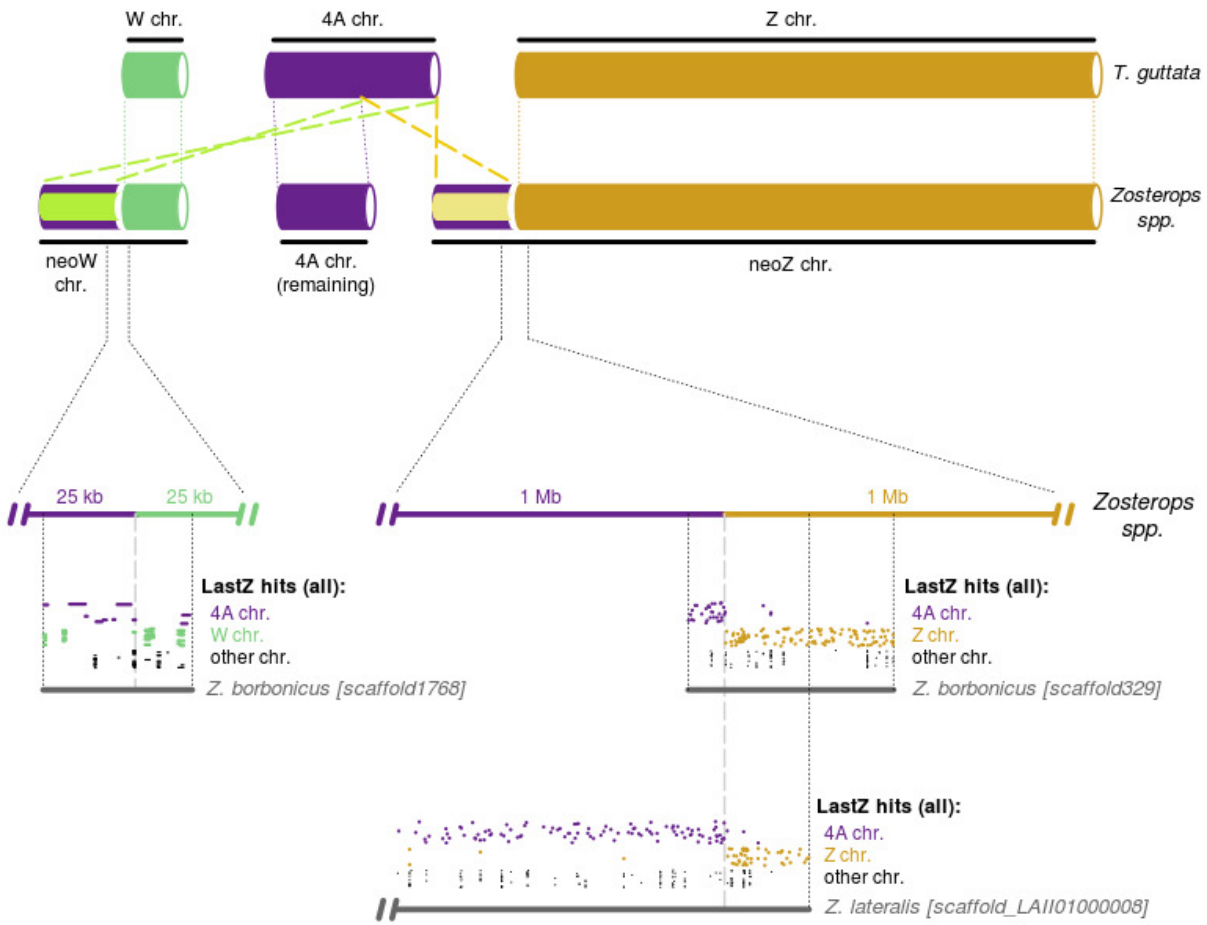


Figure 4: Fine-scale genome architecture of the Sylvioidea neo-sex chromosomes. *Original W and Z and autosomal 4A chromosomes are shown in green, yellow and purple, respectively. Translocated regions of the ancestral 4A chromosome (neoW-4A and neoZ-4A) are shown with a purple outline. LASTZ hits on scaffold supporting the chromosomal breakpoints are shown using the same color code. For ease of illustration, coordinates on the Z and 4A chromosomes are reversed (i.e. the white disk represents the beginning of the* T. guttata *chromosome sequence).*

Chromosomal-scale estimates of nucleotide diversity

We used sequencing data from six *Z. borbonicus* individuals (three males and three females) sequenced by Bourgeois et al. (2017) to explore the genomic landscape of within-species diversity. Over all autosomal 10-Kb windows, mean nucleotide diversity (Tajima's π; Tajima 1983) estimates were roughly similar in males and females (π_{males}=1.82e-3 and $\pi_{females}$=1.81e-3, respectively). The nucleotide diversity landscape greatly varies within and between chromosomes (A in Fig 5, Fig S4 & S5). In addition, we identified some series of 10-Kb windows with very low level of nucleotide diversity (red bars, Fig. 5A). Interestingly, some of these regions also exhibit the highest negative values of Tajima's D (red bars, Fig. 5B & S6; *e.g.* the end of the chromosome 2). Small interchromosomal differences in the distribution of nucleotide diversity values were detected for both female- and male-based estimates, suggesting that both datasets give similar results at the chromosomal level, except for the neoZ chromosome for which substantial differences were observed between the two datasets (Fig. S4). Even considering this source of variability, the neoZ chromosome still shows significant deviation from the mean autosomal diversity for both datasets ($\pi_{females}$=1.04e-3 and π_{males}=1.34e-3, respectively; t-tests, p<2e-16 for both datasets), thus representing 57.6% and 73.5% of the mean autosomal diversity. This reduced level of nucleotide diversity was similarly detected for the neoZ-Z and the neoZ-4A regions of the neo-Z chromosome (C, Fig. 5). Based on both datasets, a lower nucleotide diversity was observed on neoZ regions as compared to the autosomal chromosomes. Mean π_{males} was estimated to 1.21e-3 for the neoZ-4A region and 1.35e-3 for the neoZ-Z region, corresponding to 66.6% and 74.2% of the autosomal diversity. Mean $\pi_{females}$ values are roughly similar with 1.37e-3 and 1.00e-3, thus representing 76.2% and 55.5% of the autosomal diversity, respectively.

Similarly, data from females only were used to estimate the level of within-species diversity variation on the neoW chromosome (C, Fig. 5). We similarly observed a reduced level of diversity as compared to the autosomal chromosomes (mean $\pi_{females}$=5.87e-4; t-test, $p<2e^{-16}$), with only one-third (32.5%) of the mean nucleotide diversity estimated for the autosomes. Differences in Tajima's $\pi_{females}$ were observed on the neoW-W region of the neoW chromosome and on the neoW-4A translocated region, albeit non-significant (t-test, p=0.102), with higher diversity on the neoW-W (π=1.08e-3) as compared to the neoW-4A (π=4.65e-4). Median values are however much more consistent between the two regions (π=1.25e-4 and π=1.16e-4, respectively), suggesting that few neoW-W windows greatly contributed to this discrepancy (C, Fig. 5).

Ratios of non-synonymous to synonymous polymorphisms (Π_N/Π_S) and substitutions (D_N/D_S)

We computed π_N/π_S ratios among all genes in autosomal, neoZ and neoW chromosomes. Estimated π_N/π_S of chromosome 4A genes is slightly lower as compared to the rest of autosomal chromosomes (π_N/π_S= 0.212 vs. π_N/π_S = 0.170). NeoZ-Z exhibits slightly higher values than both autosomal sets (π_N/π_S=0.282), however no or very little difference has been observed between neoZ-4A (π_N/π_S = 0.181) and autosomal chromosomes. On the contrary, π_N/π_S ratios on genes of the neoW chromosome are very high, with π_N/π_S=0.418 for the neoW-4A and π_N/π_S= 0.780 for the neoW-W regions.

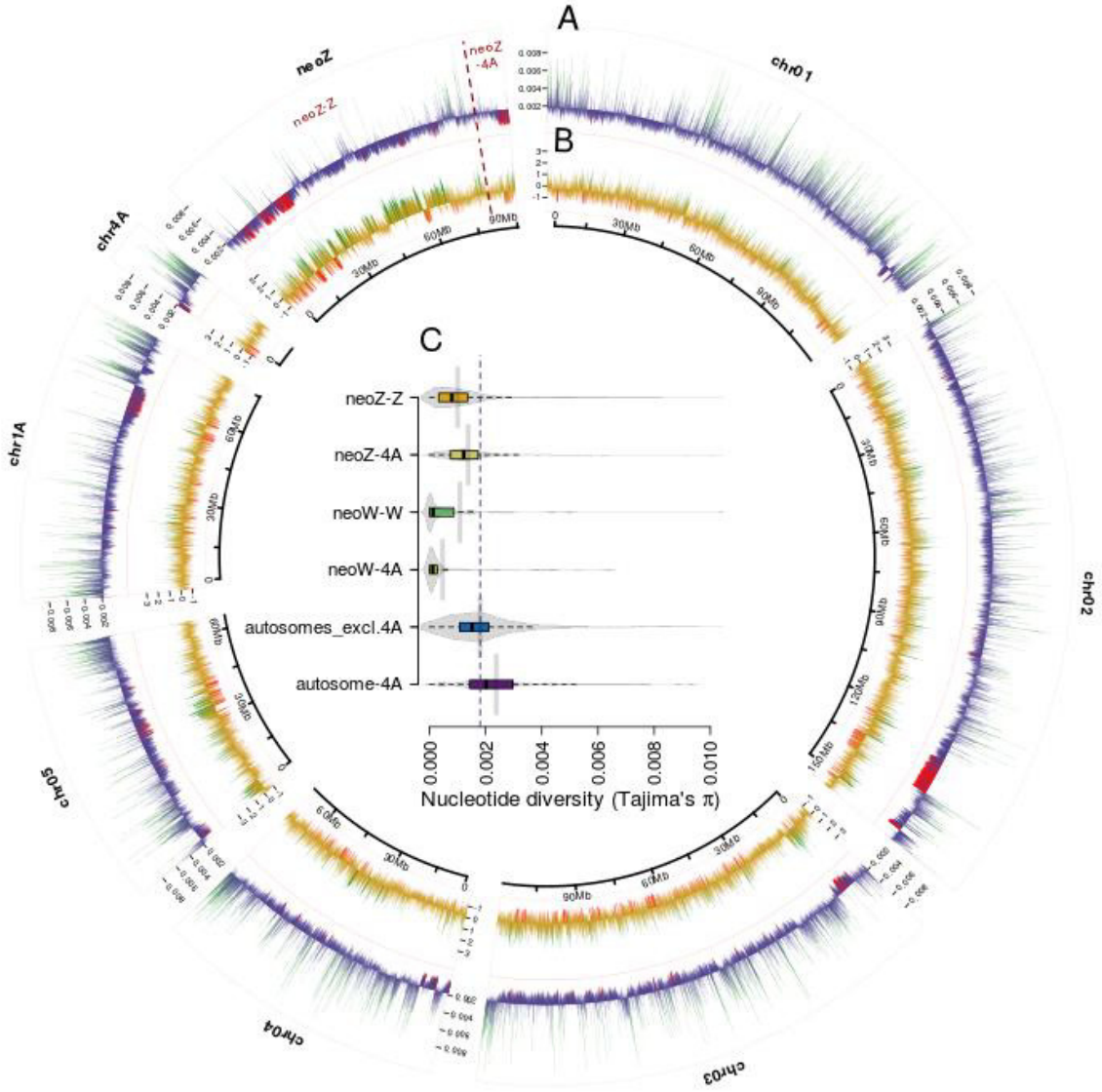


Figure 5: Intra- and inter-chromosomal variations in nucleotide diversity. *Variations of Tajima's* π_{males} *(A) and* D_{males} *(B) estimates along the six Z. borbonicus macrochromosomes, the autosomal 4A and the neo-Z chromosome. The two metrics were calculated in non-overlapping 10-Kb sliding windows. Top 2.5% and bottom 2.5% windows are shown in green and red, respectively. For both Tajima's* π_{males} *and Tajima's* D_{males}*, each bar shows the deviation from the mean genomic value over the whole genome (Tajima's* π_{males} *and Tajima's* D_{males} *baselines: 1.82e-3 and -0.26, respectively). C) Interchromosomal differences in Tajima's* $\pi_{females}$ *between autosomes and sex chromosomes. Fig. S5 and S6 show Tajima's pi and D along most* Z. borbonicus *chromosomes.*

We estimated d_N/d_S ratios for *Z. borbonicus* and ten additional passerine species (all passerines except *A. chloris* in Fig. 1) for a total of 6,339 alignments of single-copy orthologs, corresponding to 6,073 autosomal genes, 66 on the remaining autosomal region of the 4A chromosome (hereafter autosome-4A), 164 on the neoZ-Z, 54 on the neoZ-4A and 17 on the neoW-4A. For the neoZ-4A and neoW-4A, we made a special effort to identify gametologs (homologous sequences between neoZ-4A and neoW-4A genes, which were previously excluded during the filtering of 1:1 orthologs).

We then compared the d_N/d_S of neoZ-4A and neoZ-Z genes. To do that, we randomly subsampled the data to match the number of genes in neoZ-4A region (54) before concatenation and then computed d_N/d_S ratios. The variability in d_N/d_S is evaluated by bootstrapping genes and creating repeated concatenation of 54 genes for each genomic region.

For all species, the d_N/d_S ratio reaches significantly higher values for neoZ-Z genes than for autosomal genes (Fig. 6). Surprisingly, we found no evidence for an increase in d_N/d_S ratios in the neoZ-4A when compared to autosomes (Fig. 6, species in the red frame). For both the willow warbler and *Zosterops* species, d_N/d_S ratios are lower in neoZ-4A genes than in autosomal regions (willow warbler: neoZ-4A d_N/d_S = 0.09 (95% CI=0.07-0.11) *vs.* autosomal d_N/d_S = 0.11 (95% CI=0.08-0.15); for white-eyes: neoZ-4A d_N/d_S = 0.10 (95% CI = 0.08-0.13) *vs.* autosomal d_N/d_S = 0.14 (95% CI=0.09-0.20)). This also holds true when we compare genes of the ancestral 4A chromosome translocated on the neoZ chromosome (neoZ-4A, "chromosome 4A:0-9.6 Mb" in Fig. 6) and genes of the 4A chromosome ("chromosome 4A:9.6-20.7 Mb"). Even if we report slightly higher d_N/d_S for the translocated region as compared to the rest of the 4A chromosome, such a difference between the two ancestral 4A regions is also observed in several other species that do not have the translocation (e.g., *M. vitellinus*, *T. gutatta* or *Z. albicollis;* Fig. 6), including a much bigger difference for *T. guttata* (Fig. 6). As a consequence, our d_N/d_S analysis did not provide any support for a higher d_N/d_S ratio associated to the autosomal-to-Z translocation.

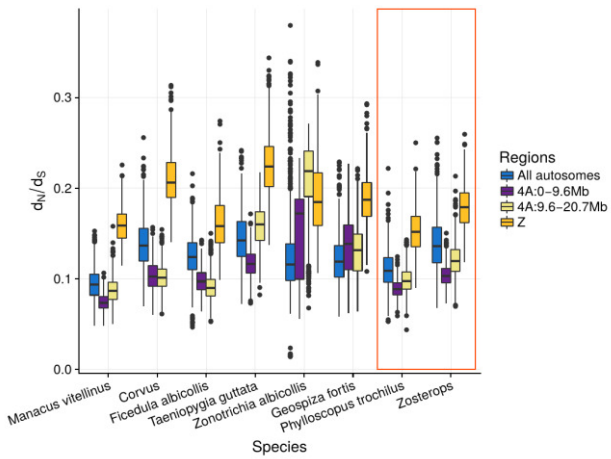


Figure 6: Variation in d_N/d_S ratios among chromosomal regions for the 11 investigated species. *Estimates are performed at the genus level, i.e. on branches before the split of the two* Corvus *(*C. cornix *and* C. brachyrhynchos*) and the three* Zosterops *species (*Z. lateralis, Z. pallidus *and* Z. borbonicus*). The red box shows species with the translocated neoZ-4A region For these species, the 4A:0-9.6 Mb region corresponds to the neoZ-4A and the Z corresponds to the neoZ-Z.*

Next, we computed the d_N/d_S of the branch leading to the neoZ-4A and neoW-4A copies in the *Zosterops borbonicus* genome. In this case, the d_N/d_S of the neoW-4A genes were significantly higher than the d_N/d_S of the neoZ-4A copies (mean d_N/d_S = 0.531, sd = 0.417 for neoW-4A genes; mean d_N/d_S = 0.194, sd = 0.234 for neoZ-4A genes; Wilcoxon signed rank test, p-value =7.7e-5, Fig. 7). The increase in d_N/d_S is particularly strong, including some neoW-4A genes with d_N/d_S close or slightly higher than 1.

For the three genes with a d_N/d_S >1, we performed a likelihood-ratio test comparing a model with a fixed d_N/d_S value equaling 1 (null model) to a model with a dN/dS value free to vary. All observed values were not significantly different from the null model. Based on these tests, these results are therefore more consistent with ongoing pseudogenization than positive selection. However, it should be outlined that we detected no premature stop-codon or frameshift mutation in the six neoW-W genes with the highest dN/dS values (i.e. d_N/d_S reaching at least 0.5).

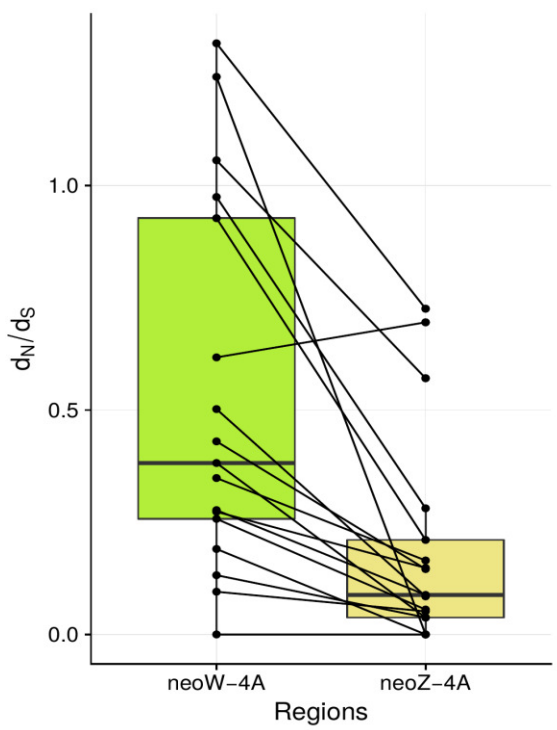


Figure 7: The d_N/d_S ratios of the neoW-4A and neoZ-4A copies (gametologs) of the* Zosterops borbonicus *genomes. *For each gene, horizontal lines linked values of the two gametologs.*

GC, GC* and Transposable Elements (TE) contents

Next, we investigated the potential change in base composition after the translocation events, due to changes in recombination rates and more precisely, the recombination-associated effect of GC-biased gene conversion (gBGC, Duret & Galtier 2009, Nabholz et al. 2011, Weber et al. 2014). For this purpose, we computed the GC content over 10-Kb sliding windows (Fig. S7) and the GC content at equilibrium for orthologous sequences (Fig. S8). As expected, differences in GC contents are observed between chromosomes, with a higher GC content in short chromosomes. Among all chromosomes, the neoZ chromosome showed the second lowest median GC rate. Interestingly, GC content is lower in the neoZ-4A regions as compared to the autosomal 4A chromosome (Student's t-test, p=2.2e-4) or as compared to the neoW-4A, although this difference is only marginally significant (t-test, p=0.059). Slight differences in GC content are observed between neoW-4A (t-test, p=0.066) and neoW-W (t-test, p=0.096) as compared to the autosomal-4A chromosome suggesting that the GC content of the neoW-4A likely decreased after the translocation, but to a lesser extent than for the neoZ-4A (Fig. 8). We also evaluated variation in GC content at equilibrium (GC*) for the new gametologs, *i.e.* homologous genes in the neoW-4A and neoZ-4A regions (Fig. 9A). For these 17 genes, all Passerida species without the neo-sex regions exhibited a GC* between 0.7 and 0.8 (Fig. 9A). Passerida species with these neo-sex chromosomes showed a slightly reduced GC* at neoZ-4A genes (mean GC* = 0.57 and 0.68 for the willow warblers and the white-eyes respectively, and a strongly reduced GC* content at neoW-4A genes (mean=0.4, 95% CI = 0.30-0.51, Fig. 9A). In contrast, the GC* of the non-translocated region of the ancestral chromosome 4A (position 9.6-20.7 Mb) apparently remains unchanged between the Sylvioidea and the other Passerida (GC* between 0.71 and 0.86, Fig. 9B).

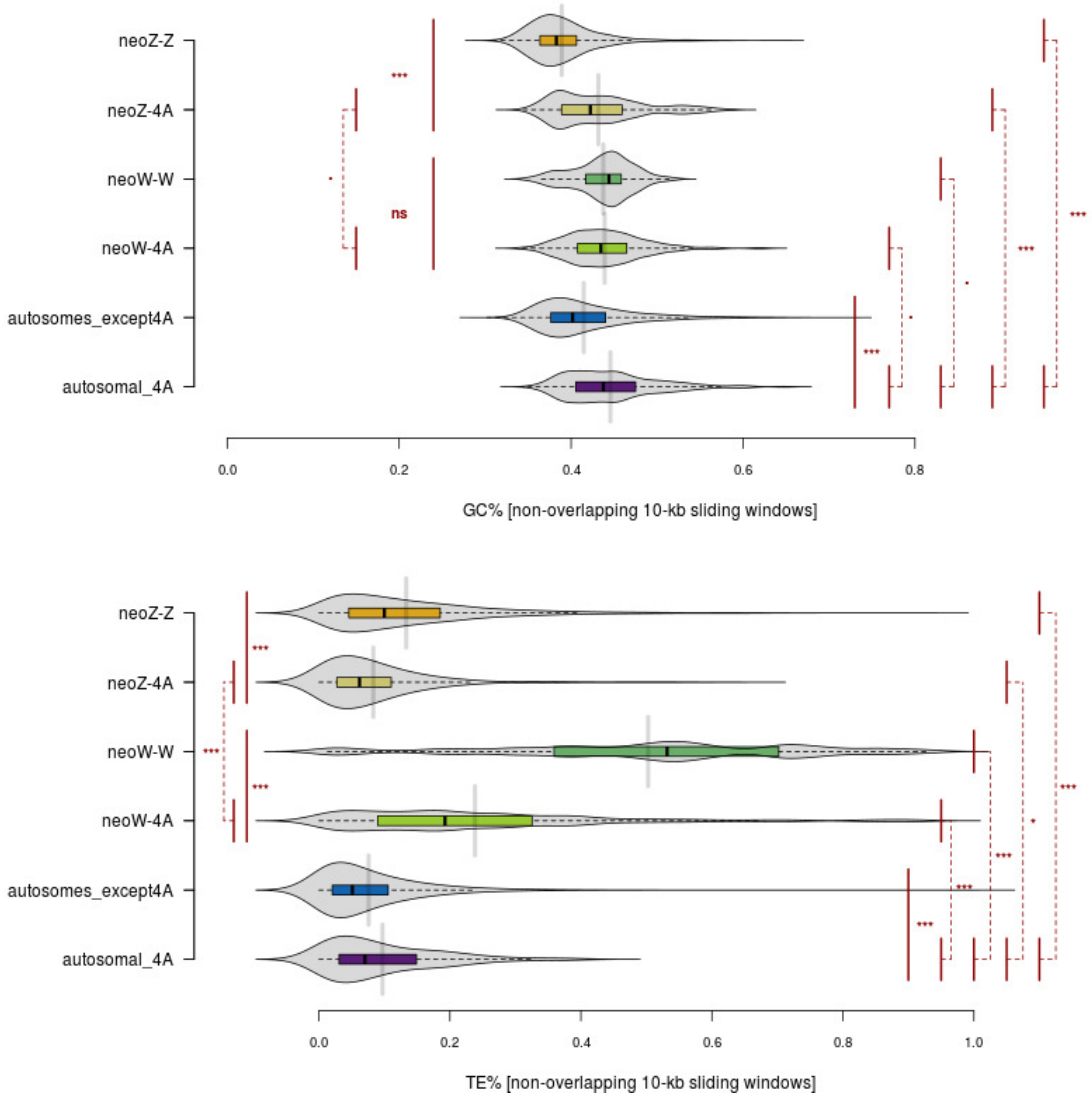


Figure 8: Variation in GC (top) and TE (below) contents between chromosomes. *GC and TE contents were estimated along the whole genome using non-overlapping 10-kb sliding windows. Means were compared using t-tests (pairwise comparisons between chromosome sets, without corrections for multiple testing). Thresholds: . < 0.1, * < 0.05, ** < 0.01, *** < 0.001*

695

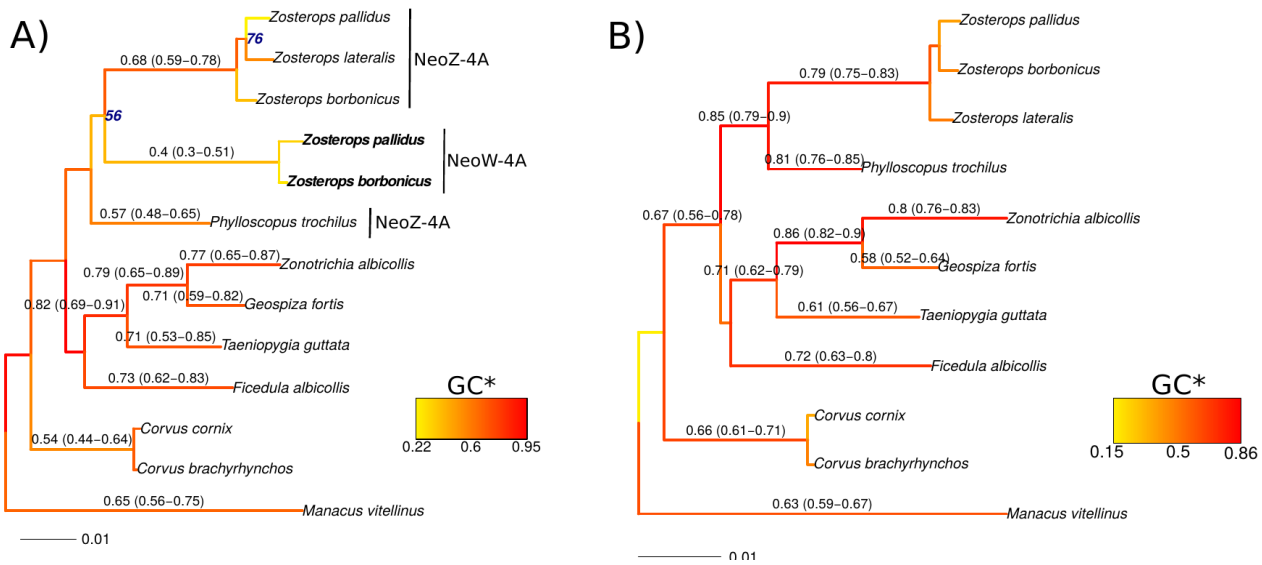


***Figure 9*: Phylogenetic relationships and estimation of GC equilibrium (GC*).** *A) Estimation using the 17 genes coded by the translocated region of the ancestral chromosome 4A (0-9.6 Mb). Bold font indicates NeoW-4A sequences. B) Estimation using the 79 genes coded by the non-translocated region of the ancestral chromosome 4A (9.6-20.7 Mb). Branch color and numbers above branches indicated GC*, mean and 95% confidence interval obtained by bootstrap are in parenthesis. Number in blue in front of the node indicates ultra-fast bootstrap values lower than 100%.*

We also found support for a higher abundance of transposable elements (TE) in the W chromosome as compared to all other chromosomes (Fig. 8 & S9). Overall, 45.1% of the cumulative size of the scaffolds assigned to the neoW-W is composed of transposable elements, which represents a 3.31- to 9.95-fold higher content than on the autosomal chromosomes. Interestingly, the TE content over 10-kb windows (Fig. 8) is also much higher on the neoW-4A than on the autosomal 4A chromosome, the other autosomes or, even more interesting, the neoZ-4A region (p<2e-16 for all these comparisons). Even if RepeatModeler was unable to classify most *Z. borbonicus*-specific TE ("no category" in Fig S9), Class I LTR elements seem to have greatly contributed to this higher content in the neoW-W, as well as in the neoW-4A chromosome (Fig S9).

Discussion

A new high-quality reference genome for Sylvioidea and *Zosterops*

Using a combination of short Illumina and long PacBio reads, we have generated a high-quality bird assembly of *Z. borbonicus* with a scaffold N50 exceeding one megabase, which is comparable to the best passerine reference genomes available to date (Kapusta & Suh, 2017; Peona et al. 2018). Similar conclusions can be drawn by comparing BUSCO analyses between this assembly and a set of other 26 extant avian genome assemblies (Fig. S1). This assembly is of equivalent quality to the well-assembled congeneric species *Z. lateralis* (Cornetti et al.

2015), thus jointly representing important genomic resources for *Zosterops*, a bird lineage exhibiting one of the fastest rates of species diversification among vertebrates (Moyle et al. 2009).

Avian GC-rich regions are known to be underrepresented in sequencing data because Illumina library construction protocols are biased toward intermediate GC-content (Botero-Castro et al. 2017; Tilak et al. 2018; Peona et al. 2018). Combining moderate coverage (10x) PacBio and Illumina sequencing technologies, we have generated gene models for half of the "hidden" genes of Botero-Castro et al. (2017). The use of long PacBio reads seems therefore a promising solution to partially address the under-representation of GC-rich genes in avian genomes. In the future, it will be interesting to combine long read technologies such as PacBio or Oxford Nanopore with the Illumina library preparation proposed by Tilak et al. (2018).

To further improve the contiguities of the *Z. borbonicus* sequence, we used 26 avian species as pivotal resources to chromosomally organized scaffolds of the *Z. borbonicus* assembly. This strategy has led to a 1.047 Gb chromosome-scale genome sequence for *Z. borbonicus* (Table S1). We were able to obtain assembly statistics comparable to some assemblies using high coverage long-read data (Weissensteiner et al. 2017) or a pedigree linkage map (Kawagami et al. 2014). Importantly, the application of the DeCoSTAR strategy has not only improved the *Z. borbonicus* genome, but also those of 25 other species. Indeed, at the notable exception of the chicken reference genome, all genome assemblies have been improved using DeCoSTAR (4.19-fold improvement of the median scaffold N50), thus demonstrating the utility of the inclusion of other genome assemblies, even fragmented, to polish the genome assembly of a species of interest (Duchemin et al. 2017; Anselmetti et al. 2018).

CONFIRMING *ZOSTEROPS* AS GREAT SPECIATORS

The availability of the *Z. borbonicus* genome sequence is also an important step to study the evolution of the *Zosterops* lineage, which was described as the "Great Speciator" (Moyle et al. 2009; Cornetti et al. 2015). Indeed, the availability of sequence data for two new species (*Z. borbonicus* and *Z. pallidus*), in addition with the sequence of *Z. lateralis* (Cornetti et al. 2015) helped us to validate the recent origin of this taxon. Moyle et al. (2009) estimated that the white-eyes genus (except *Z. wallacei*) originated in the early pleistocene (~2.5 Myrs). With more than 80 species, this clade exhibits an exceptional rate of diversification compared to other vertebrates (net rate of diversification without extinction (r) : 1-2.5 species per Myr, Magallon and Sanderson, 2001 cited in Harmon, 2018).

The divergence time estimated by Moyle et al. (2009) was based on a biogeographic calibration obtained from the ages of the Solomon Islands. These biogeographic calibrations should however be interpreted with care as previous phylogenetic studies found evidence for older lineages than the emergence ages of the islands for which they are endemic (Heads, 2005; Heads, 2011). This could be the consequence of extinction of mainland relatives leading to long branches of some island species (Heads, 2005; Heads, 2011). More recently, Cai et al. (2019) have obtained similar dates using a larger phylogeny but, again, have applied a biogeographic calibration. In this study, we took the opportunity to combine genetic information of these three *Zosterops* with the other investigated bird species to obtain new estimates using independent fossil calibrations to reassess the conclusions of Moyle et al. (2009) regarding the recent and extensive diversification of *Zosterops*. Using two independent datasets (mitochondrial and nuclear) and methods, we

confirmed the recent origin of the genus. Our analyses are consistent with a diversification of white-eyes over less than 5 Myrs and that could be as young as 1 Myr. With the exception of the African Great Lakes cichlids (Genner et al. 2007), the genus *Zosterops* represents one of the most exceptional diversification rates among vertebrates (Lagomarsino et al. 2016). As an example, it is more than ten times higher than the average diversification rate estimated across all bird species by Jetz et al. (2009). Even the large and relatively recent radiation of the Furnariidae (ovenbirds and woodcreepers) has a net rate of diversification markedly lower than the white-eyes (r = 0.16; Derryberry et al. 2010).

NO EVIDENCE FOR A FAST-NEOZ EFFECT

Taken all together, our analyses are surprisingly consistent with a pattern of substantial reduction of nucleotide diversity, but a low impact of the autosome-to-Z translocation on the molecular evolution of *Z. borbonicus*.

First, using pairwise genome alignments of the zebra finch genome with either the *Z. borbonicus* or the *Z. lateralis* genomes, we found support for a narrow candidate region of 1 kb around position 9.606 Mb of the v.3.2.4 zebra finch 4A chromosome, in which the chromosomal breakpoint likely occurred. This result is consistent and fairly close to the estimate of 10 Mb suggested by Pala and collaborators (2012a) who first demonstrated the translocation of approximately a half of the zebra finch 4A on the Z chromosome using an extended pedigree of the great reed warbler, a Sylvioidea species. A recent article reported a similar estimate (9.6 Mb) in another Sylvioidea species, the common whitethroat (Sigeman et al. 2018). To get this estimate, these latter authors used a very similar approach to ours (H. Sigeman, personal communication).

Second, relative estimates of within-species diversity on both sides of this chromosomal breakpoint (*i.e.* neoZ-4A and neoZ-Z regions) were obtained. As compared to all autosomes, neoZ-4A and neoZ-Z regions of the neo-Z chromosome exhibit reduced levels of within-species diversity in both the ancestral Z chromosome (*i.e.* neoZ-Z:autosomes = 0.605) as well as in the newly translocated region (neoZ-4A:autosomes=0.782), consistent with a substantial loss of diversity associated with this autosome-to-sex transition, following the expected effects of changes in effective population sizes. The neoZ-4A:autosomal-4A nucleotide diversity reported is slightly higher than 0.75 but is in line with a previous report for the common whitethroat (0.82, Pala et al. 2012b). Strongest deviations have however been reported in two other Sylvioidea species, namely the great reed warbler and the skylark (0.15 and 0.42, respectively) but some of the variation might be explained by the moderate number of loci analyzed (Pala et al. 2012b). Sex ratio imbalance or selection are known to contribute to strong deviations from neutral equilibrium expectations of three-fourths (reviewed in Ellegren 2009 and Wilson Sayres, 2018).

Third, we found no support for a fast-Z evolution in the neoZ-4A region, *i.e.* neither an elevated ratio of non-synonymous to synonymous polymorphisms (π_N/π_S) nor an elevated ratio of non-synonymous to synonymous substitutions (d_N/d_S) at neoZ-4A genes when compared with autosomal-4A sequences. This result is intriguing as the decrease of nucleotide diversity observed on the neoZ-4A is expected to reflect a decrease in N_e and, therefore, a decrease in the efficacy of natural selection. This should result in an increase of the frequency of slightly deleterious mutations (Ohta 1992; Lanfear et al. 2014). The increase in d_N/d_S of avian Z-linked genes compared to autosomes - a pattern that we recovered well in our analyses - has often been interpreted in that

way (Mank et al. 2010; Wright et al. 2015). We propose several potential hypotheses to explain the absence of fast-Z on the neoZ-4A regions. First, the hemizygous status of neoZ-4A regions could help to purge the recessive deleterious mutation and, therefore, limit the increase of π_N/π_S and d_N/d_S as reported in Satyrinae butterflies (Rousselle et al. 2016). Second, the intensity of purifying selection is not only determined by N_e but also by gene expression (Drummond & Wilke 2008; Nguyen et al. 2015) and recombination rate (Hill & Roberston 1966). It is therefore possible that the expression pattern of neoZ-4A genes has changed after the translocation. The change in recombination rate, however, seems to go in the opposite direction as GC and GC* decrease in the neoZ-4A region, suggesting a decrease in recombination rate and, therefore, a decrease in the efficacy of natural selection.

Finally, base composition has changed after the translocation to the Z chromosome. We indeed found evidence for lower GC content and GC* in the neoZ-4A region than in the remaining autosomal region of the 4A chromosome. In birds, as in many other organisms, chromosome size and recombination rate are negatively correlated (Backström et al. 2010), probably because one recombination event occurs per chromosome arm per meiosis. Given that GC content strongly negatively correlates with chromosome size (Eyre-Walker, 1993; Pessia et al. 2012), the observed difference in GC content is a likely consequence of changes in the intensity of the recombination-associated effect of gBGC (Duret & Arndt, 2008; Duret & Galtier, 2009), resulting from the instantaneous changes in chromosome sizes due to the translocation. From the translocated region of the ancestral 4A chromosome point of view, the chromosomal context has drastically changed from an ancestral ~21 Mb 4A chromosome to a ~90 Mb neo-Z chromosome probably with a lower recombination rate. Similarly, from the remaining 4A chromosome point of view, the chromosomal context has drastically changed too, from a ~21-Mb to a ~11-Mb chromosome. As a consequence, we can expect that base composition has evolved in the opposite direction with an increase in GC. However, this hypothesis must be qualified since we were unable to find any change in GC* at autosomal 4A genes, suggesting that the chromosome-scale recombination rate might not have been impacted by the translocation. This could be possible if the translocation occurred close to the centromere (*i.e.* a nearly whole-arm translocation). In this case, the overall recombination rate is not expected to change.

Convergent chromosomal footprints of a neoW degeneration

Degeneration of the non-recombining chromosome, *i.e.* the chromosome only present in the heterogametic sex, has been explored in depth in a variety of species (*e.g.* Bachtrog & Charlesworth, 2002; Papadopulos et al. 2015). Long-term gradual degeneration through the accumulation of deleterious mutations, partial loss of adaptive potential and gene losses are expected to start soon after species cease to recombine due to a series of factors including Muller's ratchet, linked selection and the Hill-Robertson effect (Charlesworth & Charlesworth, 2000; Bachtrog, 2005; Sun & Heitman 2012).

To investigate this degeneration, we have first identified 218 scaffolds with a female-specific pattern in read coverage, for a total of 7.1 Mb. Among these scaffolds, we have assigned 174 scaffolds to the neoW-4A region, because of a high level of homology with the zebra finch 4A chromosome, thus representing a neoW-4A of a total length of 5.48 Mb. The absence of any neoW scaffold homologous to the 4A *T. guttata* chromosome between positions 1,756,080 and 3,509,582 suggests a large chromosomal

deletion. The total length of the neoW-W region we have assigned is low, only representing 1.23 Mb of sequence, which is far from the 6.94 Mb sequence that Smeds and collaborators (2015) identified in the collared flycatcher genome or the 6.81 Mb sequence reported in the reference Chicken genome (GRCg6a version, International Chicken Genome Sequencing Consortium 2004; Warren et al. 2017). It is however important to specify that our objective was not to be exhaustive, but rather to focus on the longest scaffolds for which both estimates of the median coverage and alignments against the zebra finch chromosome 4A were considered reliable enough to be confident in their assignment to the W chromosome, particularly in a context of the intense activity of transposable elements. Given the known difficulty to sequence and assemble the W chromosome (*e.g.* Tomaszkiewicz et al. 2017), such a reduced-representation of the W chromosome was expected. Our intent was to get sufficient information to study the molecular evolution of neoW-W specific genes too. Obtaining a high-quality sequence of the neoW chromosome for the *Z. borbonicus* species, while possible, would require a considerable additional sequencing effort to be achieved.

Then, by aligning *Z. borbonicus* assembly against the zebra finch 4A chromosome, we found support for a candidate scaffold supporting the chromosomal breakpoint. Alignments on both ends of this scaffold suggest a potential chromosomal breakpoint occurring around positions 9.603-9.605 Mb of the v.3.2.4 zebra finch 4A chromosome, which is remarkably close to our estimate for the translocation on the neoZ-4A. Importantly, such an observation therefore supports an evolution of neoW-4A and neoZ-4A regions from initially identical gene sets. Interestingly, we have also identified a large chromosomal deletion on the W chromosome, which represents another expected early signature of the W degeneration (Charlesworth, 1991).

We have found support for a highly reduced level of nucleotide diversity in the neoW chromosome as compared to autosomes. This also holds true for the neoW-4A region (mean neoW-4A:autosomal nucleotide diversity = 0.36), which is in broad agreement with the hypothesis of a three-fourth reduction in effective population size associated to an autosomal-to-W or autosomal-to-Y translocation. Our overall result of low within-species diversity on the W chromosome is however not as drastic as compared with the dramatically reduced diversity observed by Smeds and collaborators (2015) on the W chromosome of several flycatcher species, with a W:autosomal diversity ranging from 0.96% to 2.16% depending on populations and species. Non-recombinant W chromosome of *Z. borbonicus* also exhibits elevated π_N/π_S in the neoW-W (0.78) as well as in the neoW-4A regions (0.42), representing 3.68-fold and 1.97-fold higher ratios than for the autosomal genes, respectively (4.59-fold and 2.46-fold higher ratios when compared with autosomal-4A genes only). Higher d_N/d_S at neoW-4A genes as compared to their neoZ-4A gametologs were also observed. This higher d_N/d_S ratio is in agreement with the pattern observed in another Sylvioidea species, the common whitethroat (Sigeman et al. 2018) and, more broadly, with other young sex chromosome systems (*e.g.* Marais et al. 2008). Sigeman et al. (2018) also reported an association between amino acid and gene expression divergences for the neoW-4A. Altogether, these results are consistent with an accumulation of deleterious mutations associated with the strong reduction of the net efficacy of natural selection to purge deleterious mutations (Charlesworth & Charlesworth, 1997).

TE accumulation on W or Y chromosome is suspected to play a particularly important role in the first phases of the evolution of chromosome differentiation (Bachtrog, 2003). To investigate this, we have *de novo* identified *Z. borbonicus*-specific TE and have analyzed

distribution and abundance of TEs. This has led to the identification of a high TE load in the ancestral W chromosome (45.1%), which is the same order of magnitude as the reported value for the W chromosome of *Ficedula albicollis* (48.5%, Smeds et al. 2015) or *Zonotrichia albicollis* (51.1%, Davis et al. 2010). Interestingly, we found support for a high TE load in the translocated region too (21.6%), which is approximately twice the observed TE content of any other autosomal chromosome, including the autosomal 4A chromosome. TE classification, albeit incomplete, supports an important contribution of class I LTR elements to the overall TE load. LTR elements seem to be particularly active in the zebra finch (Kapusta and Suh, 2017) or in the collared flycatcher genomes (Suh et al. 2018) suggesting that the recent burst of LTR elements on autosomes may have facilitated the accumulation of TEs on the neoW-4A chromosomes. The most probable hypothesis is that LTR elements are particularly retained in low-recombination regions. Following this hypothesis, the non-recombining neoW-4A region might therefore be viewed as an extreme case in terms of the retention of these TE insertions. Although much more data and work will be needed in the future to analyze in greater depth this accumulation of TEs, and particularly the determinants of this accumulation, our results suggest that LTR elements may have virtually played a major role in altering gene content, expression and/or chromosome organization of the newly translocated region of the Sylvioidea neo-W chromosome.

CONCLUSIONS

In this study, we generated a high quality reference genome for *Z. borbonicus* that has provided us unique opportunities to investigate the molecular evolution of neo-sex chromosomes, and more broadly to improve our understanding of avian sex chromosome evolution. Since this species belong to the Sylvioidea, one of the three major clades of passerines, comprising close to 1,300 species, we can reasonably anticipate that this chromosomal-scale assembly will serve as a reference for a large diversity of genome-wide analyses in the Sylvioidea lineage itself, and more generally in passerine birds. Interestingly, Sylvioidea is becoming an important animal taxon for the study of sex chromosomes (Dierickx et al. 2019; Sigeman et al. 2019).

Through detailed analyses of the evolution of the newly sex chromosome-associated regions, we found evolutionary patterns that were largely consistent with the classic expectations for the evolution of translocated regions on sex chromosomes, including evidence for reduction of diversity, ongoing neoW chromosome degeneration and base composition changes. A notable exception was the neoZ region for which no fast-neoZ effect was identified. Although most of the analyses are congruent, our report is based on a limited number of individuals and from only one population of *Z. borbonicus*. Further investigations based on a complementary and extensive dataset will probably help to fine-tune the conclusions, especially regarding the lack of any fast-Z effect or the drastic increase of TE on the neoW-4A. Lastly, given the huge difference in diversity between autosomes and sex chromosomes, we emphasize the importance of taking into account sex chromosomes for local adaptation studies, at least by scanning autosomes and sex chromosomes separately (see Bourgeois et al. 2018 for an example).

Materials & Methods

DNA and RNA extraction, sequencing

We extracted DNA from fresh tissues collected on a *Z. borbonicus* female individual (field code: 15-179), which died accidentally during fieldwork in May 2015 at Pas de Bellecombe (Gîte du volcan, Réunion; coordinates: S: -21.2174, E: 55.6872; elevation: 2246m above sea level). Sampling was conducted under a research permit (#602) issued to Christophe Thébaud by the Centre de Recherches sur la Biologie des Populations d'Oiseaux (CRBPO) – Muséum National d'Histoire Naturelle (Paris).

We also extracted DNA of a *Zosterops pallidus* female collected in February 2015 in South Africa, Free state province, Sandymount Park, 10 kms from Fauresmith (coordinates: S: -29.75508, E: 25.17733). The voucher is stored at the Museum National d'Histoire Naturelle (MNHN), Paris, France, under the code MNHN ZO 2015-572 and a tissue duplicate is deposited in the National Museum Bloemfontein (South Africa).

For both samples, 9µg of total genomic DNA were extracted from liver and/or muscle, using DNeasy Blood and Tissue kit (QIAGEN) following the manufacturer instructions. Each of these samples was sequenced as followed: one paired-end library with insert sizes of 300bp and three mate-pair libraries (3 kb, 5 kb and 8 kb) using Nextera kit. Libraries and sequencing were performed by platform "INRA plateformes Génomes et transcriptomes (GeT-PlaGe)", Toulouse, France. Illumina sequencing was also performed at the platform GeT-PlaGe using Illumina HiSeq 3000 technology.

To improve genome assembly of Z. *borbonicus*, an additional sequencing effort was made by generating 11X coverage of PacBio long reads data. 20µg of high molecular weight DNA were extracted from muscle using MagAttract HMW DNA kit (QIAGEN) following manufacturer instructions. The PacBio librairies and sequencing were performed at Genome Québec (Centre d'innovation Génome Québec et Université McGill, Montréal, Quebec, Canada) using a PacBio RS platform.

After the accidental death of the *Z. borbonicus*, the brain of the freshly dead bird was extracted and then stabilized using RNAlater (Sigma). Total RNA of *Z. borbonicus* individual was extracted from the dissected tissue sampleusing RNeasy Plus Mini Kit (Qiagen) following manufacturer's instructions (RNA integrity number: 7.9). Both the RNAseq library preparation and the Hiseq2500 sequencing (1 lane) were performed at the genomic platform MGX-Montpellier GenomiX.

Genome assembly

The paired-end reads were filtered using Trimmomatic (v0-33; Bolger et al., 2014) using the following parameters: "ILLUMINACLIP:TruSeq3-PE-2.fa:2:30:10 LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:50". The mate-pair reads were cleaned using NextClip (v1.3.1; Leggett et al. 2013) using the following parameters : "--min_length 20 --trim_ends 0 --remove_duplicates".

Paired-end and mate-pair reads were assembled using SOAPdenovo (v2.04; Luo et al 2012) with parameters "-d 1 -D 2". Several k-mers (from 27 to 37-mers) were tested and we chose the assembly maximizing the N50 scaffold length criteria. Next, we applied Gapcloser v1.10 (a companion program of SOAPdenovo) to fill the gap in the assemblies.

Given that the PacBio technology produces long reads but with quite high sequencing error rates, we used LoRDEC (v0.6; Salmela & Rivals 2014) to correct the PacBio

reads of the *Z. borbonicus* individual, using the following parameters: "-k 19 -s 3". In brief, LoRDEC corrects PacBio reads (both insert/deletions and base call errors) by the use of Illumina paired-end reads, a technology producing short reads only, but with a much higher base call accuracy and depth of coverage. The corrected PacBio reads were then used to scaffolds the SOAPdenovo assembly using SSPACE-LongRead (v1.1; Boetzer & Pirovano 2014). For *Z. borbonicus*, we also used MaSuRCA (v3.2.4; Zimin et al. 2017) to perform a hybrid assembly with a mixture of short and long reads. MaSuRCA produced an assembly with similar quality but slightly shorter than SOAPdenovo+SSPACE-LongRead (Table S1).

Several statistics were computed using assemblathon_stats.pl script (Bradnam et al. 2013; https://github.com/ucdavis-bioinformatics/assemblathon2-analysis) to evaluate the different genome assemblies. These statistics include genome size, number of scaffolds, scaffold N50, scaffold N90 and the proportion of missing data (N%). Additionally, we used BUSCO (v3.0.2; Waterhouse et al. 2017), a commonly used tool for evaluating the genome completeness based on the content in highly conserved orthologous genes (options: "-m genome -e 0.001 -l aves_odb9 -sp chicken").

The mitochondrial genome was assembled using mitobim (v1.8; Hahn et al. 2013) with the mitochondrial genome of *Z. lateralis* (accession : KC545407) used as reference (so-called "bait"). The mitochondrial genome was automatically annotated using the web server mitos2 (http://mitos2.bioinf.uni-leipzig.de; Bernt et al. 2013). This annotation was manually inspected and corrected using alignment with the other *Zosterops* mitochondrial genomes available in genbank.

Genome-guided De novo Transcriptome Assembly.

RNA-seq reads were used to generate a transcript catalogue to train the gene prediction software. First, RNA-seq reads were filtered with trimmomatic (v0-33; Bolger et al., 2014) using the following parameters : "ILLUMINACLIP:TruSeq3-PE.fa:2:30:10 SLIDINGWINDOW:4:5 LEADING:5 TRAILING:5 MINLEN:25". Second, the filtered RNA-seq reads were mapped onto the reference genome using HISAT2 (Kim et al. 2015). HISAT2 performed a splice alignment of RNA-Seq reads, outperforming the spliced aligner algorithm implemented in TopHat (Kim et al. 2015). Finally, two methods were used to assemble the transcripts from the HISAT2 output bam file: i) Cufflinks (Trapnell et al. 2010) using the following parameters: "-q -p 10 -m 300 -s 100" and ii) Trinity (v2.5.0; Haas et al. 2013) using the following parameters:

"--genome_guided_max_intron 100000".

Protein-coding Gene Annotation

Gene annotation was performed using the PASA pipeline combined with EVidenceModeler (Haas et al. 2008; Haas et al. 2011; https://github.com/PASApipeline/PASApipeline/wiki). The complete annotation pipeline involved the following steps:

1. *ab initio* gene finding with Augustus (Stanke and Waack 2003 ; http://bioinf.uni-greifswald.de/augustus/) using the parameter : "--species=chicken".
2. Protein homology detection and intron resolution using genBlastG (She et al. 2011; http://genome.sfu.ca/genblast/download.html). Protein sequences of several passerines were used as references, namely zebra finch (*Taeniopygia guttata*; assembly taeGut3.2.4; Warren et al. 2010), collared flycatcher (*Ficedula albicollis*; assembly FicAlb_1.4; Ellegren et al.

2012), medium ground-finch (*Geospiza fortis*; assembly GeoFor_1.0; (Zhang et al. 2012) and hooded crow (*Corvus cornix*; accession number JPSR00000000.1; Poelstra et al. 2014). Genblast parameters were : "-p genblastg -c 0.8 -r 3.0 -gff -pro -cdna -e 1e-10".

3. PASA alignment assemblies based on overlapping transcript from Trinity genome-guided *de novo* transcriptome assembly (see above) (Haas et al. 2003). This step involved the so-called PASAPipeline (v2.2.0; https://github.com/PASApipeline/PASApipeline/) used with the following parameters : "-C -r -R --ALIGNERS blat,gmap".

4. Next, EVidenceModeler (v1.1.1; Haas et al. 2008) was used to compute weighted consensus gene structure annotations based on the previous steps (1 to 3). We used the following parameters : "--segmentSize 500000 --overlapSize 200000" and an arbitrary weight file following the guidelines provided at http://evidencemodeler.github.io/ .

5. Finally, we used the script "pasa_gff3_validator.pl" of PASA add UTR annotations and models for alternatively spliced isoforms.

Finally, StringTie (Pertea et al. 2015) was also used to estimate the proportion of annotated CDS with RNA-seq information support.

We defined three sets of genes depending on their reliability, namely the "high reliability" set corresponding to genes with a low TE content in coding regions (<10%, Fig. S2) and with at least a transcript support ("high"), the "moderate reliability" set corresponding to genes with either a low TE content (<10%) or with at least a transcript support ("moderate") and a "low reliability" set containing the remaining genes.

Orthology detection

We used the available passerine genomes (namely, zebra finch, collared flycatcher, white-throated sparrow and hooded crow) plus the high-coverage genomes (>100x) of Zhang et al. (2014) (Table S2). Orthology detection was performed using OrthoFinder (v2.2.6; Emms & Kelly 2015). Single copy (one-to-one) orthologs were extracted from OrthoFinder results to perform multi-species alignment with TranslatorX (v1.1; Abascal et al. 2010) using MAFFT (Katoh et al. 2002) to build the alignment. Alignments were inspected by HMMCleaner (v1.8; Amemiya et al. 2013; Philippe et al. 2017) to exclude badly aligned sites. Next, dubious, highly divergent, sequences were excluded using trimAl (Capella-Gutiérrez et al. 2009) option "-resoverlap 0.60 -seqoverlap 80". We also used genome assemblies of three passerine species for which no gene annotation sets were publicly available, namely the silvereye, *Z. lateralis* (Cornetti et al. 2015), the Orange River white-eye *Z. pallidus* (this study) and the willow warbler *Phylloscopus trochilus* (Lundberg et al. 2017). For these genomes, gene orthology detection was conducted using AGILE (Hughes & Teeling 2018). AGILE is a pipeline for gene mining in fragmented assembly overcoming the difficulty that genes could be located in several scaffolds. We applied AGILE using *Z. borbonicus* single copy orthologs as query genes.

Filtering scaffolds originating from autosomes, W and Z chromosomes

Given that we have sequenced a female genome, we likely assembled contigs from W, Z and autosomal chromosomes. To identify scaffolds originating from sex chromosomes or autosomes, we mapped seven females and five males reads from *Z. borbonicus* birds published in Bourgeois et al. (2017) onto our female genome assembly. We first mapped all raw reads against the *Z. borbonicus* reference genome using BWA

mem v. 0.7.5a (Li, 2013), we then removed duplicates with Picard 1.112 (http://broadinstitute.github.io/picard). We then used Mosdepth (Pedersen & Quinlan, 2018) to compute median per-site coverage for each scaffold, following the same strategy than in Smeds et al. (2015). For each *i* scaffold, total coverage for males and females were computed as the sum of coverage of all individuals. Then, we compute a normalized coverage per scaffolds for male as :

$$Normalized(Cov.\ Scaffold_i(Males)) = Cov.\ Scaffold_i(Males) * \frac{MeanCov.(Females)}{MeanCov.(Males)}$$

Where "*Mean Cov. (Females)*" and "*Mean Cov. (Males)*" corresponds to the median coverages for males and females across all scaffolds longer than 1 Mb. This normalization is intended to take into account the different number of male and female individuals.

Next, we used the normalized median per-site coverage to detect W-linked scaffolds (zero and above 5X in males and females, respectively), Z-linked (female with less than 0.75 the male coverage and male with > 9X coverage) and autosomes (all the remaining scaffolds corresponding female and male with a roughly similar median coverage). To decrease the probability of false identification due to the mapping in repeat-rich regions, this approach has only been performed using coverage data from scaffolds longer than 10 kb (3,443 / 97,503 scaffolds, >96% of the assembly size).

Pseudo-chromosome assembly

We first aligned soft-masked *Z. borbonicus* scaffolds on the soft-masked Zebra Finch genome (*Taeniopygia guttata*) using LASTZ v. 1.04.00 (Schwartz et al. 2003). For *Z. borbonicus*, *de novo* transposable elements prediction were performed using Repeat Modeler v. 1.0.11 and genome assembly soft-masking using Repeat Masker v. 4.0.3 (Smit et al. 2013-2015). For *T. guttata*, we used the soft-masked genome v. 3.2.4 (taeGut3.2.4) made available by the Zebra Finch genome consortium. This soft-masking procedure was put in place in order to exclude these regions for the LASTZ's seeding stage, and thereby to avoid finding alignments in these highly repeated regions. We then filtered LASTZ alignments hits in order to only keep reliable hits (5% longest hits and with sequence identities greater than the median over all alignments). For each scaffold, we then defined syntenic blocks as adjacencies of several reliable hits. A syntenic block represents a homologous region between zebra finch and *Z. borbonicus* starting at the first reliable hit and ending at the last one. Syntenic blocks covering at least 80% of a *Z. borbonicus* scaffold size and 80-120% the corresponding homologous region in the Zebra Finch genome were automatically anchored to its *T. guttata* chromosome position, assuming complete synteny between the zebra finch and *Z. borbonicus*. Unanchored scaffolds were then manually identified by visually inspecting the summary statistics and positions of all raw alignments. In case of chimeric scaffolds, these scaffolds were cut into two or several new scaffolds assuming that this chimerism is due to a contigging or a scaffolding artifact.

Second, we used DeCoSTAR, a computational method tracking gene order evolution from phylogenetic signal by inferring gene adjacencies evolutionary histories, in order to not only improve the genome assembly of *Z. borbonicus*, but also 26 additional extant avian assemblies including the high coverage (>100X) of Zhang et al. 2014 and the other passerine genomes (Table S2). The gene trees were obtained for phylogenies of single-copy orthologs (see above) estimated using IQ-TREE model GTR+G4 and 1000 so-called "ultra-fast bootstrap" (v1.6.2; Nguyen et al. 2014). A snakemake pipeline (https://github.com/YoannAnselmetti/DeCoSTAR

_pipeline) were used to apply DeCoSTAR on gene orders and phylogenies.

Finally, scaffold orders given by LASTZ and DeCoSTAR were manually reconciled to determine the consensus scaffold orders along the *Z.borbonicus* chromosomes.

MOLECULAR DATING

We performed two independent molecular dating analyses. First, we used relaxed molecular clock analysis based on nuclear sequences and fossil calibrations applied on the full bird phylogeny. Second, we applied the approach of Nabholz et al. (2016) using white-eyes body mass to estimate their mitochondrial substitution rate.

In the first approach, molecular dating analyses were performed using the 27 species selected for the DeCoSTAR analysis plus the silvereye, the Orange River white-eye and the willow warbler leading to a total of 30 species. We restricted our analyses to single-copy orthologs with low GC content. These genes are known to evolve slowing and more clock-like than GC rich genes (Nabholz et al. 2011; Jarvis et al. 2014). To do that, we randomly selected 100 single-copy orthologs coded by chromosome 1 and 2 excluding genes at the beginning and at the end of the chromosomes (minimum distance from each chromosome tip based on the *T. guttata* genome: 10 Mbp). This step was replicated ten times to evaluate the robustness of the inferences among different sets of genes.

We used several fossil calibrations sets summarized in Table S3. Our different calibration sets reflect the current uncertainty surrounding bird diversification dates. We used a conservative set with only one maximum bound at 140 Myrs for the origin of Neornithes (set 4). At the other end, we used another set with a narrow constraint between 28 and 34 Myrs for the Suboscines / Oscines divergence (set 3). This constraint could turn out to be incorrect with future advances in the bird fossil records. For all the calibration sets, the Neognathae / Palaeognathae divergence minimal age was set to 66 Myrs using *Vegavis iaai* fossil (Benton et al. 2009; Mayr, 2013; Ksepka & Clarke, 2015). The maximum age of this node is much more difficult to select. We have opted for two maximal ages. The first one at 86.5 Myrs following the rationale of Prum et al. (2015) based of the upper bound age estimate of the Niobrara Formation (set 1, 2 and 3). We used another very conservative maximum bound at 140 Myrs (set 4). This is the maximum age estimated for the origin of Neornithes by Lee et al. (2014) using an extensive morphological clock analysis. Additionally, in sets 1 and 3, we constrained the divergence between Passeriformes / Psittaciformes to be between 53.5 and 65.5 Myrs, assuming the complete absence of passerine bird species during the Cretaceous. In sets 2 and 4, we only used a minimum bound on the first fossil occurrence of passerines in the Eocene (Boles, 1997) and on the stem Psittaciformes fossil *Pulchrapollia gracilis* (Dyke & Cooper, 2000). In calibration set 3, we constrained the Oscines/Suboscines split between 28 and 34 Myrs, following the rationale of Mayr (2013) assuming crown Oscines and Suboscines originated in the early Oligocene (28 Myrs, Mayr & Manegold 2006) and based on the absence of Eocene fossil records discovered so far (Eocene/Oligocene limit is at 34 Myrs). All the other minimum-bound calibrations we used followed the suggestions of Ksepka & Clarke (2015) and are presented in Table S3.

Molecular dating was performed using Phylobayes (v4.1; Lartillot et al. 2009) using a CAT-GTR substitution model. For the relaxed clock model, we used the log-nomal auto-correlated rates (ln) model. We also tested the uncorrelated gamma multipliers (ugam) model that gave similar results than the ln model (results not shown). We used uniform prior on

divergence times, combined with soft calibrations (Rannala and Yang, 2007; Yang and Rannala, 2006). The MCMC were run for at least 8,000 cycles. MCMC convergence were diagnosed by running two independents MCMC and by visually checking the evolution of the likelihood and other parameters along the Markov chain (in ".trace" files).

Additionally, we applied an independent molecular dating using the method proposed by Nabholz et al. (2016). Seven complete mitochondrial *Zosterops* genomes were downloaded from Genbank, including *Z. erythropleuros* (KT194322), Z. japonicus (KT601061), *Z. lateralis* (KC545407), *Z. poliogastrus* (KX181886), *Z. senegalensis* (KX181887), *Z. senegalensis* (KX181888). We also included the sequences of the Réunion grey white-eye and the Orange River white-eye assembled in the present study. The mitochondrial sequences of *Yuhina diademata* (KT783535) and *Zoothera dauma* (KT340629) were used as outgroups. Body mass for all *Zosterops* species were obtained from Dunning (2007) and the median of these body masses was computed. Phylogenetic relationship and branch length were estimated using IQ-TREE with a HKY+G4 substitution model using third codon positions only. Then, we applied the formula of Nabholz et al. 2016 to derive the substitution rate (substitution per site per Myr) as $\frac{10^{-0.141*log10(BodyMass)+0.367}}{100}$ and $\frac{10^{-0.243*log10(BodyMass)+0.905}}{100}$ for the minimal and maximal rate where "*Body Mass*" is the median body mass in grams (logarithm of base 10). Next, we computed the median divergence time between the silver-eye (*Z. lateralis)* and all the other white-eye species to obtain an estimate of the crown *Zosterops* clade age. Finally, we divided this divergence time by the rate obtained with the formula above to obtain divergence dates in Myrs.

GENOME-WIDE ESTIMATES OF NUCLEOTIDE DIVERSITY

Based on the previously generated BAM files (see '*Filtering scaffolds originating from autosomes, W and Z chromosomes*' section*)*, we then use GATK to generate the gvcf from the six parents of the three progenies described in Bourgeois et al. 2017 (three males and three females). Joint genotyping, as well as all subsequent analyses, was then performed separately for the three males and the three females. We followed all the GATK best practices under the GATK suite, except for variant filtration which was performed using a custom script to speed up computations. This step however followed the same procedures than under GATK, assuming the following thresholds: QD>2.0, FS<60.0, MQ>40.0, MQRANKSUM>-2.0, READPOSRANKSUM>-2.0 and RAW_MQ > 45000.

For each individual, we then reconstructed two genomic sequences. At each position of the genome, the position (reference or alternate if any) was added to the sequence if the coverage at the position was between 3 and 50. In any other case, a "N" character was added in order to keep the sequence length the same. We then computed the Tajima's π estimator of nucleotides diversity and the GC content over non-overlapping 10 kb windows. To be highly conservative in the analysis of the neoW chromosome, all windows associated to W scaffolds and found covered in the male (ZZ) dataset were excluded from the analysis of the female (ZW) dataset, even if this non-zero coverage was observed at a single base over the genomic window. Such a non-zero coverage is expected caused by read mismapping, particularly in TE regions.

Estimates of non-synonymous and synonymous divergences and GC equilibrium

For the neoZ-4A and neoW-4A, we made a specific effort to retrieve the paralogous sequences. Assuming that the two translocated regions evolved from the same gene sets (see results), each gene located on the neoZ-4A are expected to have a paralog on the neoW-4A (so called "gametologs", Pala et al. 2012a,b). As a consequence, most gametologs have been eliminated during our selection of the single-copy orthologs. We visually inspected all the alignments containing a neoZ-4A gene and then tried to identify the corresponding copy in a scaffold assigned to the neoW chromosome.

Given the drastically different number of genes between autosomes, neoZ-Z and neoZ-4A regions, we subsampled the data to match the category with the lowest number of genes (i.e., the neoZ-4A region). Then, we concatenated genes and computed d_N and d_S as the sum of non-synonymous and synonymous branch lengths respectively. This will decrease the variance of the estimated d_N/d_S ratios and limit problems associated to low d_S values (Wolf et al. 2009). Finally, the variability in d_N/d_S was evaluated by bootstrapping genes 1000 times within each genomic region. For GC equilibrium (GC*), we used the nonhomogeneous model T92 (Galtier and Gouy 1998) implemented in the BPPSUITE package (Dutheil and Boussau 2008, http://biopp.univ-montp2.fr/wiki/index.php/BppSuite) based on the Bio++ library (Guéguen et al. 2013) to infer GC* at third codon position for each branch. We used a similar bootstrapping strategy (1000 times within each genomic region) to evaluate the variation in GC*.

For the comparison between neoZ-4A and neoW-4A, we used the d_N/d_S and GC* of neoZ-4A and neoW-4A genes for the branch leading to Z. borbonicus. When the neoW-4A sequences of Z. borbonicus was very closely related to a sequence of Z. pallidus, we assumed that the Z. pallidus sequence also came from the neoW-4A regions and we computed the d_N/d_S of the ancestral branch of these two species. We evaluated the different in d_N/d_S using a paired-samples Wilcoxon test (also known as Wilcoxon signed-rank test). The d_N/d_S was estimated using CODEML (Yang 2007) using a free-ratios model (model = 2). We also checked for the presence of frameshift and premature stop codon within the neoW-W sequenced using macse (v2, Ranwez et al. 2007).

GC and TE contents

We used the automated approach implemented in RepeatModeler Open (v.1.0.11, Smit & Hubley, 2008) to *de novo* detect *Z.borbonicus*-specific TE consensus. The generated list of *de novo* TE sequences was merged to the chicken repetitive sequences publicly available in Repbase (Jurka et al. 2005). We then used this set of sequences as a custom library for RepeatMasker (v.open-4.0.3, Smit et al. 2013) to generate a softmasked version of the *Z. borbonicus* genome assembly. Then, we used a non-overlapping 10-kbp sliding windows approach to calculate the GC and TE contents along the whole genome. All genomic windows composed of more than 50% of Ns were excluded to ensure accurate estimates of local TE and GC contents.

All statistical analyses were performed using R v. 3.4.4 (R core Team, 2018). Some analyses and graphics were performed using several additional R packages: APE (Paradis & Strimmer, 2004), beanplot (Kampstra, 2008), circlize (Gu et al. 2014), cowplot (Wilke, 2016),

ggplot2 (Wickham, 2016), phytools (Revell, 2012) and plotrix (Lemon, 2006).

DATA AVAILABILITY

Nuclear and mitochondrial genome sequences, scripts and programs used are available at the following FigShare repository: https://figshare.com/s/122efbec2e3632188674. Raw reads and genome sequences are available on SRA and GenBank (Bioproject: PRJNA530916). Mitochondrial genome sequences are available on GenBank (accession numbers: MK524996, MK529728).

ACKNOWLEDGMENTS

This research was funded by the French ANR (BirdIslandGenomic project, ANR-14-CE02-0002). The analyses benefited from the Montpellier Bioinformatics Biodiversity (MBB) platform services and the genotoul bioinformatics platform Toulouse Midi-Pyrenees (Bioinfo Genotoul). We are grateful to Jérôme Fuchs for providing the mitochondrial alignments, to Fabien Condamine for constructive discussions on molecular dating and Anna-Sophie Fiston-Lavier for providing advice on the analysis of transposable elements. We thank the Reunion National Park for granting us permission to conduct fieldwork in Pas de Bellecombe, Reunion. We are grateful for the logistic support provided by the field station of Marelongue, funded by the P.O.E., Reunion National Park and OSU Reunion. We are grateful to the provincial authorities in the Free State (South Africa) for granting permission to collect samples and specimens (permit 01-24158) and to Dawie de Swardt (National Museum Bloemfontein) for help with organizing field work. We are also grateful to the PCI recommender Kateryna Makova as well as three reviewers (Melissa Wilson, Gabriel Marais and an anonymous reviewer) for providing excellent reviews based on a previous version of the manuscript.

CONFLICT OF INTEREST DISCLOSURE

The authors of this preprint declare that they have no financial conflict of interest with the content of this article. Benoit Nabholz and Céline Scornavacca are also PCI Evol Biol recommenders.

REFERENCES

Abascal F, Zardoya R, Telford MJ. **2010**. TranslatorX: multiple alignment of nucleotide sequences guided by amino acid translations. *Nucleic Acids Research* **38**: W7–W13.

Amemiya CT, Alföldi J, Lee AP, Fan S, Philippe H, MacCallum I, Braasch I, Manousaki T, Schneider I, Rohner N, *et al.* 2013. The African coelacanth genome provides insights into tetrapod evolution. *Nature* **496**: 311.

Anselmetti Y, Duchemin W, Tannier E, Chauve C, Bérard S. **2018**. Phylogenetic signal from rearrangements in 18 Anopheles species by joint scaffolding extant and ancestral genomes. *BMC Genomics* **19**: 96.

Bachtrog D. **2005**. Sex chromosome evolution: molecular aspects of Y-chromosome degeneration in Drosophila. *Genome research* **15**: 1393–1401.

Bachtrog D, Charlesworth B. **2002**. Reduced adaptation of a non-recombining neo-Y chromosome. *Nature* **416**: 323.

Backström N, Forstmeier W, Schielzeth H, Mellenius H, Nam K, Bolund E, Webster MT, Ost T, Schneider M, Kempenaers B, *et al.* 2010. The recombination landscape of the zebra finch Taeniopygia guttata genome. *Genome research* **20**: 485–495.

Benton M, Donoghue P, Asher R. **2009**. Calibrating and constraining molecular clocks. In: The timetree of Life. Oxford University Press, 35–86.

Bernt M, Donath A, Jühling F, Externbrink F, Florentz C, Fritzsch G, Pütz J, Middendorf M, Stadler PF. **2013**. MITOS: Improved de novo metazoan mitochondrial genome annotation. *Mitogenomics and Metazoan Evolution* **69**: 313–319.

Boetzer M, Pirovano W. **2014**. SSPACE-LongRead: scaffolding bacterial draft genomes using long read sequence information. *BMC Bioinformatics* **15**: 211.

Boles W. **1997**. *Fossil Songbirds (Passeriformes) from the Early Eocene of Australia*.

Bolger AM, Lohse M, Usadel B. **2014**. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**: 2114–2120.

Botero-Castro F, Figuet E, Tilak M-K, Nabholz B, Galtier N. **2017**. Avian Genomes Revisited: Hidden Genes Uncovered and the Rates versus Traits Paradox in Birds. *Molecular Biology and Evolution* **34**: 3123–3131.

Bourgeois YXC, Delahaie B, Gautier M, Lhuillier E, Malé P-JG, Bertrand JAM, Cornuault J, Wakamatsu K, Bouchez O, Mould C, *et al.* 2017. A novel locus on chromosome 1 underlies the evolution of a melanic plumage polymorphism in a wild songbird. *Royal Society open science* **4**: 160805–160805.

Bourgeois Y, Ruggiero R, Manthey J, Boissinot S. **2018**. Recent secondary contacts, background selection and variable recombination rates shape genomic diversity in the model species Anolis carolinensis. *bioRxiv*.

Bracewell RR, Bentz BJ, Sullivan BT, Good JM. **2017**. Rapid neo-sex chromosome evolution and incipient speciation in a major forest pest. *Nature Communications* **8**: 1593

Bradnam KR, Fass JN, Alexandrov A, Baranay P, Bechner M, Birol I, Boisvert S, Chapman JA, Chapuis G, Chikhi R, *et al.* 2013. Assemblathon 2: evaluating de novo methods of genome assembly in three vertebrate species. *GigaScience* **2**: 2047-217X-2–10.

Brooke M de L, Welbergen JA, Mainwaring MC, van der Velde M, Harts AMF, Komdeur J, Amos W. **2010**. Widespread Translocation from Autosomes to Sex Chromosomes Preserves Genetic Variability in an Endangered Lark. *Journal of Molecular Evolution* **70**: 242–246.

Cai T, Cibois A, Alström P, Moyle RG, Kennedy JD, Shao S, Zhang R, Irestedt M, Ericson PGP, Gelang M, *et al.* 2019. Near-complete phylogeny and taxonomic revision of the world's babblers (Aves: Passeriformes). *Molecular Phylogenetics and Evolution* **130**: 346–356.

Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T. **2009**. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **25**: 1972–1973.

Charlesworth B. **1991**. The evolution of sex chromosomes. *Science* **251**: 1030.

Charlesworth B, Charlesworth D. **1997**. Rapid fixation of deleterious alleles can be caused by Muller's ratchet. *Genetics Research* **70**: 63–73.

Charlesworth B, Charlesworth D. **2000**. The degeneration of Y chromosomes. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences* **355**: 1563–1572.

Charlesworth B, Sniegowski P, Stephan W. **1994**. The evolutionary dynamics of repetitive DNA in eukaryotes. *Nature* **371**: 215.

Clegg SM, Phillimore AB. **2010**. The influence of gene flow and drift on genetic and phenotypic divergence in two species of Zosterops in Vanuatu. *Philosophical transactions of the Royal*

Society of London. Series B, Biological sciences **365**: 1077–1092.

Coghlan A, Eichler EE, Oliver SG, Paterson AH, Stein L. **2005**. Chromosome evolution in eukaryotes: a multi-kingdom perspective. *Trends in Genetics* **21**: 673–682.

Cornetti L, Valente LM, Dunning LT, Quan X, Black RA, Hébert O, Savolainen V. **2015**. The Genome of the "Great Speciator" Provides Insights into Bird Diversification. *Genome Biology and Evolution* **7**: 2680–2691.

Cornuault J, Delahaie B, Bertrand JAM, Bourgeois YXC, Milá B, Heeb P, Thébaud C. **2014**. Morphological and plumage colour variation in the Réunion grey white-eye (Aves: *Zosterops borbonicus*): assessing the role of selection. *Biological Journal of the Linnean Society* **114**: 459–473.

Davis JK, Thomas PJ, Thomas JW, NISC Comparative Sequencing Program. 2010. A W-linked palindrome and gene conversion in New World sparrows and blackbirds. *Chromosome Research* **18**: 543–553.

Delahaie B, Cornuault J, Masson C, Bertrand JAM, Bourgeois YXC, Milá B, Thébaud C. **2017**. Narrow hybrid zones in spite of very low population differentiation in neutral markers in an island bird species complex. *Journal of Evolutionary Biology* **30**: 2132–2145.

Derryberry EP, Claramunt S, Derryberry G, Chesser RT, Cracraft J, Aleixo A, Pérez-Emán J, Remsen J JV, Brumfield RT. **2011**. Lineage diversification and morphological evolution in a large-scale continental radiation: the neotropical ovenbirds and woodcreepers (Aves: Furnariidae). *Evolution* **65**: 2973–2986.

Diamond JM, Gilpin ME, Mayr E. **1976**. Species-distance relation for birds of the Solomon Archipelago, and the paradox of the great speciators. *Proceedings of the National Academy of Sciences of the United States of America* **73**: 2160–2164.

Dierickx EG, Sin SYW, van Veelen P, Brooke ML, Liu Y, Edwards SV, Martin SH. 2019. Neo-sex chromosomes and demography shape genetic diversity in the Critically Endangered Raso lark. *bioRxiv.* 10.1101/617563

Drummond DA, Wilke CO. **2008**. Mistranslation-Induced Protein Misfolding as a Dominant Constraint on Coding-Sequence Evolution. *Cell* **134**: 341–352.

Duchemin W, Anselmetti Y, Patterson M, Ponty Y, Bérard S, Chauve C, Scornavacca C, Daubin V, Tannier E. **2017**. DeCoSTAR: Reconstructing the Ancestral Organization of Genes or Genomes Using Reconciled Phylogenies. *Genome Biology and Evolution* **9**: 1312–1319.

Dunning J. **2007**. *CRC Handbook of Avian Body Masses*. Boca Raton.

Duret L, Arndt PF. **2008**. The Impact of Recombination on Nucleotide Substitutions in the Human Genome. *PLOS Genetics* **4**: e1000071.

Duret L, Galtier N. **2009**. Biased Gene Conversion and the Evolution of Mammalian Genomic Landscapes. *Annual Review of Genomics and Human Genetics* **10**: 285–311.

Dutheil J, Boussau B. **2008**. Non-homogeneous models of sequence evolution in the Bio++ suite of libraries and programs. *BMC evolutionary biology* **8**: 255–255.

Dyke GJ, Cooper JH. **2008**. A new psittaciform bird from the London Clay (Lower Eocene) of England. *Palaeontology* **43**: 271–285.

Ellegren H. **2009**. The different levels of genetic diversity in sex chromosomes and autosomes. *Trends in Genetics* **25**: 278–284.

Ellegren H. **2010**. Evolutionary stasis: the stable chromosomes of birds. *Trends in Ecology &*

Evolution **25**: 283–291.
Ellegren H. **2013**. The Evolutionary Genomics of Birds. *Annual Review of Ecology, Evolution, and Systematics* **44**: 239–259.
Ellegren H, Smeds L, Burri R, Olason PI, Backström N, Kawakami T, Künstner A, Mäkinen H, Nadachowska-Brzyska K, Qvarnström A, *et al.* 2012. The genomic landscape of species divergence in Ficedula flycatchers. *Nature* **491**: 756.
Emms DM, Kelly S. **2015**. OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome Biology* **16**: 157.
Ericson PG, Klopfstein S, Irestedt M, Nguyen JM, Nylander JA. **2014**. Dating the diversification of the major lineages of Passeriformes (Aves). *BMC Evolutionary Biology* **14**: 8.
Eyre-Walker A. **1993**. Recombination and Mammalian Genome Evolution. *Proceedings: Biological Sciences* **252**: 237–243.
Fraïsse C, Picard MAL, Vicoso B. **2017**. The deep conservation of the Lepidoptera Z chromosome suggests a non-canonical origin of the W. *Nature Communications* **8**: 1486.
Galtier N, Gouy M. **1998**. Inferring pattern and process: maximum-likelihood implementation of a nonhomogeneous model of DNA sequence evolution for phylogenetic analysis. *Molecular Biology and Evolution* **15**: 871–879.
Genner MJ, Seehausen O, Lunt DH, Joyce DA, Shaw PW, Carvalho GR, Turner GF. **2007**. Age of Cichlids: New Dates for Ancient Lake Fish Radiations. *Molecular Biology and Evolution* **24**: 1269–1282.
Gill FB. **1973**. Intra-Island Variation in the Mascarene White-Eye *Zosterops borbonica*. *Ornithological Monographs*: iii–66.
Griffin DK, Robertson LBW, Tempest HG, Skinner BM. **2007**. The evolution of the avian genome as revealed by comparative molecular cytogenetics. *Cytogenetic and Genome Research* **117**: 64–77.
Gu Z, Gu L, Eils R, Schlesner M, Brors B. **2014**. circlize implements and enhances circular visualization in R. *Bioinformatics* **30**: 2811–2812.
Guéguen L, Gaillard S, Boussau B, Gouy M, Groussin M, Rochette NC, Bigot T, Fournier D, Pouyet F, Cahais V, *et al.* 2013. Bio++: Efficient Extensible Libraries and Tools for Computational Molecular Evolution. *Molecular Biology and Evolution* **30**: 1745–1750.
Haas BJ, Delcher AL, Mount SM, Wortman JR, Smith RK Jr, Hannick LI, Maiti R, Ronning CM, Rusch DB, Town CD, *et al.* 2003. Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies. *Nucleic acids research* **31**: 5654–5666.
Haas BJ, Papanicolaou A, Yassour M, Grabherr M, Blood PD, Bowden J, Couger MB, Eccles D, Li B, Lieber M, *et al.* 2013. De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nature Protocols* **8**: 1494.
Haas BJ, Salzberg SL, Zhu W, Pertea M, Allen JE, Orvis J, White O, Buell CR, Wortman JR. **2008**. Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. *Genome Biology* **9**: R7.
Haas BJ, Zeng Q, Pearson MD, Cuomo CA, Wortman JR. **2011**. Approaches to Fungal Genome Annotation. *Mycology* **2**: 118–141.
Hahn C, Bachmann L, Chevreux B. **2013**. Reconstructing mitochondrial genomes directly from genomic next-generation sequencing reads—a baiting and iterative mapping approach. *Nucleic Acids Research* **41**: e129–e129.
Harmon LJ. **2018**. *Phylogenetic Comparative*

Methods: Learning from Trees. CreateSpace Independent Publishing Platform.
Heads M. **2005**. Dating nodes on molecular phylogenies: a critique of molecular biogeography. *Cladistics* **21**: 62–78.
Heads M. **2011**. Old Taxa on Young Islands: A Critique of the Use of Island Age to Date Island-Endemic Clades and Calibrate Phylogenies. *Systematic Biology* **60**: 204–218.
Hill WG, Robertson A. **1966**. The effect of linkage on limits to artificial selection. *Genetics Research* **8**: 269–294.
Hughes GM, Teeling EC. **2018**. AGILE: an assembled genome mining pipeline. *Bioinformatics*: bty781–bty781.
Hvilsom C, Qian Y, Bataillon T, Li Y, Mailund T, Sallé B, Carlsen F, Li R, Zheng H, Jiang T, *et al.* 2012. Extensive X-linked adaptive evolution in central chimpanzees. *Proceedings of the National Academy of Sciences* **109**: 2054.
International Chicken Genome Sequencing Consortium, Hillier LW, Miller W, Birney E, Warren W, Hardison RC, Ponting CP, Bork P, Burt DW, Groenen MAM, *et al.* 2004. Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. *Nature* **432**: 695.
Jarvis ED, Mirarab S, Aberer AJ, Li B, Houde P, Li C, Ho SYW, Faircloth BC, Nabholz B, Howard JT, *et al.* 2014. Whole-genome analyses resolve early branches in the tree of life of modern birds. *Science* **346**: 1320.
Jetz W, Thomas GH, Joy JB, Hartmann K, Mooers AO. **2012**. The global diversity of birds in space and time. *Nature* **491**: 444.
Jurka J, Kapitonov VV, Pavlicek A, Klonowski P, Kohany O, Walichiewicz J. **2005**. Repbase Update, a database of eukaryotic repetitive elements. *Cytogenetic and Genome Research* **110**: 462–467.
Kampstra P. **2008**. Beanplot: A Boxplot Alternative for Visual Comparison of Distributions. *Journal of Statistical Software, Code Snippets* **28**: 1–9.
Kapusta A, Suh A. **2017**. Evolution of bird genomes—a transposon's-eye view. *Annals of the New York Academy of Sciences* **1389**: 164–185.
Katoh K, Misawa K, Kuma K, Miyata T. **2002**. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic acids research* **30**: 3059–3066.
Kawakami T, Smeds L, Backström N, Husby A, Qvarnström A, Mugal CF, Olason P, Ellegren H. **2014**. A high-density linkage map enables a second-generation collared flycatcher genome assembly and reveals the patterns of avian recombination rate variation and chromosomal evolution. *Molecular Ecology* **23**: 4035–4058.
Kim D, Langmead B, Salzberg SL. **2015**. HISAT: a fast spliced aligner with low memory requirements. *Nature Methods* **12**: 357.
Kitano J, Peichel CL. **2012**. Turnover of sex chromosomes and speciation in fishes. *Environmental Biology of Fishes* **94**: 549–558.
Kitano J, Ross JA, Mori S, Kume M, Jones FC, Chan YF, Absher DM, Grimwood J, Schmutz J, Myers RM, *et al.* 2009. A role for a neo-sex chromosome in stickleback speciation. *Nature* **461**: 1079.
Ksepka D, Clarke J. **2015**. *Phylogenetically vetted and stratigraphically constrained fossil calibrations within Aves*.
Lagomarsino LP, Condamine FL, Antonelli A, Mulch A, Davis CC. **2016**. The abiotic and biotic drivers of rapid diversification in Andean bellflowers (Campanulaceae). *New Phytologist* **210**: 1430–1442.
Lanfear R, Kokko H, Eyre-Walker A. **2014**.

Population size and the rate of evolution. *Trends in Ecology & Evolution* **29**: 33–41.
Lartillot N, Lepage T, Blanquart S. **2009**. PhyloBayes 3: a Bayesian software package for phylogenetic reconstruction and molecular dating. *Bioinformatics* **25**: 2286–2288.
Lee MSY, Cau A, Naish D, Dyke GJ. **2014**. Morphological Clocks in Paleontology, and a Mid-Cretaceous Origin of Crown Aves. *Systematic Biology* **63**: 442–449.
Leggett RM, Clavijo BJ, Clissold L, Clark MD, Caccamo M. **2014**. NextClip: an analysis and read preparation tool for Nextera Long Mate Pair libraries. *Bioinformatics* **30**: 566–568.
Lemon J. **2006**. Plotrix: A package in the red light district of R. *R-News* **6**: 8–12.
Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv*.
Lundberg M, Liedvogel M, Larson K, Sigeman H, Grahn M, Wright A, Åkesson S, Bensch S. **2017**. Genetic differences between willow warbler migratory phenotypes are few and cluster in large haplotype blocks. *Evolution Letters* **1**: 155–168.
Luo R, Liu B, Xie Y, Li Z, Huang W, Yuan J, He G, Chen Y, Pan Q, Liu Y, *et al.* 2012. SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. *GigaScience* **1**: 2047-217X-1–18.
Magallon S, Sanderson MJ. **2001**. Absolute diversification rates in angiosperm clades. *Evolution* **55**: 1762–1780.
Mank JE, Axelsson E, Ellegren H. **2007**. Fast-X on the Z: Rapid evolution of sex-linked genes in birds. *Genome Research* **17**: 618–624.
Mank JE, Nam K, Ellegren H. **2010**. Faster-Z Evolution Is Predominantly Due to Genetic Drift. *Molecular Biology and Evolution* **27**: 661–670.
Marais GAB, Nicolas M, Bergero R, Chambrier P, Kejnovsky E, Monéger F, Hobza R, Widmer A, Charlesworth D. **2008**. Evidence for Degeneration of the Y Chromosome in the Dioecious Plant Silene latifolia. *Current Biology* **18**: 545–549.
Mayr G. **2009**. *A small suboscine-like passeriform bird from the early Oligocene of France*.
Mayr G. **2013**. The age of the crown group of passerine birds and its evolutionary significance – molecular calibrations versus the fossil record. *Systematics and Biodiversity* **11**: 7–13.
Melo M, Warren BH, Jones PJ. **2011**. Rapid parallel evolution of aberrant traits in the diversification of the Gulf of Guinea white-eyes (Aves, Zosteropidae). *Molecular Ecology* **20**: 4953–4967.
Milá B, Warren BH, Heeb P, Thébaud C. **2010**. The geographic scale of diversification on islands: genetic and morphological divergence at a very small spatial scale in the Mascarene grey white-eye (Aves: *Zosterops borbonicus*). *BMC Evolutionary Biology* **10**: 158.
Moyle RG, Filardi CE, Smith CE, Diamond J. **2009**. Explosive Pleistocene diversification and hemispheric expansion of a “great speciator”. *Proceedings of the National Academy of Sciences* **106**: 1863.
Murphy WJ, Sun S, Chen Z-Q, Pecon-Slattery J, O’Brien SJ. **1999**. Extensive Conservation of Sex Chromosome Organization Between Cat and Human Revealed by Parallel Radiation Hybrid Mapping. *Genome Research* **9**: 1223–1230.
Nabholz B, Künstner A, Wang R, Jarvis ED, Ellegren H. **2011**. Dynamic Evolution of Base Composition: Causes and Consequences in Avian Phylogenomics. *Molecular Biology and Evolution* **28**: 2197–2210.
Nabholz B, Lanfear R, Fuchs J. **2016**. Body mass-corrected molecular rate for bird mitochondrial DNA. *Molecular Ecology* **25**: 4438–4449.
Nam K, Munch K, Hobolth A, Dutheil JY,

Veeramah KR, Woerner AE, Hammer MF, Mailund T, Schierup MH. **2015**. Extreme selective sweeps independently targeted the X chromosomes of the great apes. *Proceedings of the National Academy of Sciences* **112**: 6413.

Nanda I, Karl E, Volobouev V, Griffin DK, Schartl M, Schmid M. **2006**. Extensive gross genomic rearrangements between chicken and Old World vultures (Falconiformes: Accipitridae). *Cytogenetic and Genome Research* **112**: 286–295.

Nanda I, Schlegelmilch K, Haaf T, Schartl M, Schmid M. **2008**. Synteny conservation of the Z chromosome in 14 avian species (11 families) supports a role for Z dosage in avian sex determination. *Cytogenetic and Genome Research* **122**: 150–156.

Nguyen L-P, Galtier N, Nabholz B. **2015a**. Gene expression, chromosome heterogeneity and the fast-X effect in mammals. *Biology letters* **11**: 20150010–20150010.

Nguyen L-T, Schmidt HA, von Haeseler A, Minh BQ. **2015b**. IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies. *Molecular Biology and Evolution* **32**: 268–274.

Oatley G, De Swardt DH, Nuttall RJ, Crowe TM, Bowie RCK. **2017**. Phenotypic and genotypic variation across a stable white-eye (*Zosterops* sp.) hybrid zone in central South Africa. *Biological Journal of the Linnean Society* **121**: 670–684.

Oatley G, Voelker G, Crowe TM, Bowie RCK. **2012**. A multi-locus phylogeny reveals a complex pattern of diversification related to climate and habitat heterogeneity in southern African white-eyes. *Molecular Phylogenetics and Evolution* **64**: 633–644.

O'Connor RE, Romanov MN, Kiazim LG, Barrett PM, Farré M, Damas J, Ferguson-Smith M, Valenzuela N, Larkin DM, Griffin DK. **2018**. Reconstruction of the diapsid ancestral genome permits chromosome evolution tracing in avian and non-avian dinosaurs. *Nature Communications* **9**: 1883.

Ohta T. **1992**. The Nearly Neutral Theory of Molecular Evolution. *Annual Review of Ecology and Systematics* **23**: 263–286.

de Oliveira EHC, Habermann FA, Lacerda O, Sbalqueiro IJ, Wienberg J, Müller S. **2005**. Chromosome reshuffling in birds of prey: the karyotype of the world's largest eagle (Harpy eagle, Harpia harpyja) compared to that of the chicken (Gallus gallus). *Chromosoma* **114**: 338–343.

Pala I, Naurin S, Stervander M, Hasselquist D, Bensch S, Hansson B. **2012a**. Evidence of a neo-sex chromosome in birds. *Heredity* **108**: 264–272.

Pala I, Hasselquist D, Bensch S, Hansson B. **2012b**. Patterns of Molecular Evolution of an Avian Neo-sex Chromosome. *Molecular Biology and Evolution* **29**: 3741–3754.

Papadopulos AST, Chester M, Ridout K, Filatov DA. **2015**. Rapid Y degeneration and dosage compensation in plant sex chromosomes. *Proceedings of the National Academy of Sciences* **112**: 13021.

Paradis E, Claude J, Strimmer K. **2004**. APE: Analyses of Phylogenetics and Evolution in R language. *Bioinformatics* **20**: 289–290.

Pedersen BS, Quinlan AR. **2018**. Mosdepth: quick coverage calculation for genomes and exomes. *Bioinformatics* **34**: 867–868.

Peona V, Weissensteiner MH, Suh A. **2018**. How complete are "complete" genome assemblies?—An avian perspective. *Molecular Ecology Resources* **18**: 1188–1195.

Pertea M, Pertea GM, Antonescu CM, Chang T-C, Mendell JT, Salzberg SL. **2015**. StringTie enables improved reconstruction of a transcriptome from

RNA-seq reads. *Nature Biotechnology* **33**: 290.
Pessia E, Popa A, Mousset S, Rezvoy C, Duret L, Marais GAB. **2012**. Evidence for widespread GC-biased gene conversion in eukaryotes. *Genome biology and evolution* **4**: 675–682.
Philippe H, Vienne D, Ranwez V, Roure B, Baurain D, Delsuc F. Pitfalls in supermatrix phylogenomics. *European Journal of Taxonomy* **283**: 1–25.
Poelstra JW, Vijay N, Bossu CM, Lantz H, Ryll B, Müller I, Baglione V, Unneberg P, Wikelski M, Grabherr MG, *et al.* 2014. The genomic landscape underlying phenotypic integrity in the face of gene flow in crows. *Science* **344**: 1410.
Pool JE, Nielsen R. **2007**. Population size changes reshape genomic patterns of diversity. *Evolution; international journal of organic evolution* **61**: 3001–3006.
Prum RO, Berv JS, Dornburg A, Field DJ, Townsend JP, Lemmon EM, Lemmon AR. **2015**. A comprehensive phylogeny of birds (Aves) using targeted next-generation DNA sequencing. *Nature* **526**: 569.
R core team. **2018**. *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.
Rannala B, Yang Z. **2007**. Inferring Speciation Times under an Episodic Molecular Clock. *Systematic Biology* **56**: 453–466.
Ranwez V, Douzery EJ, Cambon C, Chantret N, Delsuc F. 2018. MACSE v2: toolkit for the alignment of coding sequences accounting for frameshifts and stop codons. Molecular Biology and Evolution 35:2582–2584.
Raudsepp T, Santani A, Wallner B, Kata SR, Ren C, Zhang H-B, Womack JE, Skow LC, Chowdhary BP. **2004**. A detailed physical map of the horse Y chromosome. *Proceedings of the National Academy of Sciences of the United States of America* **101**: 9321.
Revell LJ. **2012**. phytools: an R package for phylogenetic comparative biology (and other things). *Methods in Ecology and Evolution* **3**: 217–223.
Rousselle M, Faivre N, Ballenghien M, Galtier N, Nabholz B. **2016**. Hemizygosity Enhances Purifying Selection: Lack of Fast-Z Evolution in Two Satyrine Butterflies. *Genome Biology and Evolution* **8**: 3108–3119.
Roux C, Fraïsse C, Romiguier J, Anciaux Y, Galtier N, Bierne N. **2016**. Shedding Light on the Grey Zone of Speciation along a Continuum of Genomic Divergence. *PLOS Biology* **14**: e2000234.
Salmela L, Rivals E. **2014**. LoRDEC: accurate and efficient long read error correction. *Bioinformatics* **30**: 3506–3514.
Schwartz S, Kent WJ, Smit A, Zhang Z, Baertsch R, Hardison RC, Haussler D, Miller W. **2003**. Human–Mouse Alignments with BLASTZ. *Genome Research* **13**: 103–107.
She R, Chu JS-C, Uyar B, Wang J, Wang K, Chen N. **2011**. genBlastG: using BLAST searches to build homologous gene models. *Bioinformatics* **27**: 2141–2143.
Sigeman H, Ponnikas S, Videvall E, Zhang H, Chauhan P, Naurin S, Hansson B. **2018**. Insights into Avian Incomplete Dosage Compensation: Sex-Biased Gene Expression Coevolves with Sex Chromosome Degeneration in the Common Whitethroat. *Genes* **9**.
Sigeman H, Ponnikas S, Chauhan P, Dierickx E, Brooke ML, Hansson B. 2019. Genomics of expanded avian sex chromosomes shows that certain chromosomes are predisposed towards sex-linkage in vertebrates. *bioRxiv* 10.1101/585059
Smeds L, Warmuth V, Bolivar P, Uebbing S, Burri R, Suh A, Nater A, Bureš S, Garamszegi LZ,

Hogner S, *et al.* 2015. Evolutionary analysis of the female-specific avian W chromosome. *Nature Communications* **6**: 7330.

Suh A, Smeds L, Ellegren H. 2018. Abundant recent activity of retrovirus-like retrotransposons within and among flycatcher species implies a rich source of structural variation in songbird genomes. *Molecular Ecology* **27**: 99–111.

Smit A, Hubley R. **2008**. *RepeatModeler Open-1.0*.

Smit A, Hubley R, Green P. **2013**. *RepeatMasker Open-4.0*.

Stanke M, Waack S. **2003**. Gene prediction with a hidden markov model and a new intron submodel. *Bioinformatics* **19**: 215–225.

Sun S, Heitman J. **2012**. Should Y stay or should Y go: the evolution of non-recombining sex chromosomes. *BioEssays : news and reviews in molecular, cellular and developmental biology* **34**: 938–942.

Tajima F. 1983. Evolutionary relationship of DNA sequences in finite populations. *Genetics* **105**:437–460.

Tilak M-K, Botero-Castro F, Galtier N, Nabholz B. **2018**. Illumina Library Preparation for Sequencing the GC-Rich Fraction of Heterogeneous Genomic DNA. *Genome Biology and Evolution* **10**: 616–622.

Tomaszkiewicz M, Medvedev P, Makova KD. **2017**. Y and W Chromosome Assemblies: Approaches and Discoveries. *Trends in Genetics* **33**: 266–282.

Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, Salzberg SL, Wold BJ, Pachter L. **2010**. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature biotechnology* **28**: 511–515.

Vicoso B, Charlesworth B. **2006**. Evolution on the X chromosome: unusual patterns and processes. *Nature Reviews Genetics* **7**: 645.

Völker M, Backström N, Skinner BM, Langley EJ, Bunzey SK, Ellegren H, Griffin DK. **2010**. Copy number variation, chromosome rearrangement, and their association with recombination during avian evolution. *Genome research* **20**: 503–511.

Warren WC, Clayton DF, Ellegren H, Arnold AP, Hillier LW, Künstner A, Searle S, White S, Vilella AJ, Fairley S, *et al.* 2010. The genome of a songbird. *Nature* **464**: 757.

Warren WC, Hillier LW, Tomlinson C, Minx P, Kremitzki M, Graves T, Markovic C, Bouk N, Pruitt KD, Thibaud-Nissen F, *et al.* 2016. A New Chicken Genome Assembly Provides Insight into Avian Genome Structure. *G3 (Bethesda, Md.)* **7**: 109–117.

Waterhouse RM, Seppey M, Simão FA, Manni M, Ioannidis P, Klioutchnikov G, Kriventseva EV, Zdobnov EM. **2017**. BUSCO applications from quality assessments to gene prediction and phylogenomics. *Molecular biology and evolution* **35**: 543–548.

Weber CC, Boussau B, Romiguier J, Jarvis ED, Ellegren H. **2014**. Evidence for GC-biased gene conversion as a driver of between-lineage differences in avian base composition. *Genome biology* **15**: 549–549.

Weingartner LA, Delph LF. 2014. Neo-sex chromosome inheritance across species in Silene hybrids? Journal of Evolutionary Biology. 27:1491-1499**Weissensteiner MH, Pang AWC, Bunikis I, Höijer I, Vinnere-Petterson O, Suh A, Wolf JBW**. **2017**. Combination of short-read, long-read, and optical mapping assemblies reveals large-scale tandem repeat arrays with population genetic implications. *Genome Research* **27**: 697–708.

Wickham H. **2016**. *ggplot2: Elegant Graphics for*

Data Analysis. Springer-Verlag New York.
Wilke C. **2016**. *cowplot: Streamlined Plot Theme and Plot Annotations for 'ggplot2'*.
Wilson Sayres MA. **2018**. Genetic Diversity on the Sex Chromosomes. *Genome Biology and Evolution* **10**: 1064–1078.
Wolf JBW, Künstner A, Nam K, Jakobsson M, Ellegren H. **2009**. Nonlinear dynamics of nonsynonymous (dN) and synonymous (dS) substitution rates affects inference of selection. *Genome biology and evolution* **1**: 308–319.
Wright AE, Harrison PW, Zimmer F, Montgomery SH, Pointer MA, Mank JE. **2015**. Variation in promiscuity and sexual selection drives avian rate of Faster-Z evolution. *Molecular ecology* **24**: 1218–1235.
Yang Z. 2007. PAML 4: Phylogenetic Analysis by Maximum Likelihood. Molecular Biology and Evolution 24:1586–1591.
Yang Z, Rannala B. **2006**. Bayesian Estimation of Species Divergence Times Under a Molecular Clock Using Multiple Fossil Calibrations with Soft Bounds. *Molecular Biology and Evolution* **23**: 212–226.
Yoshida K, Makino T, Yamaguchi K, Shigenobu S, Hasebe M, Kawata M, Kume M, Mori S, Peichel CL, Toyoda A, *et al.* 2014. Sex Chromosome Turnover Contributes to Genomic Divergence between Incipient Stickleback Species. *PLOS Genetics* **10**: e1004223.
Zhang G, Li C, Li Q, Li B, Larkin DM, Lee C, Storz JF, Antunes A, Greenwold MJ, Meredith RW, *et al.* 2014. Comparative genomics reveals insights into avian genome evolution and adaptation. *Science* **346**: 1311.
Zhang G, Parker P, Li B, Li H, Wang J. **2012**. The genome of Darwin's Finch (Geospiza fortis). *GigaScience*.
Zhou Q, Bachtrog D. **2012**. Sex-Specific Adaptation Drives Early Sex Chromosome Evolution in *Drosophila*. *Science* **337**: 341.
Zimin AV, Puiu D, Luo M-C, Zhu T, Koren S, Marçais G, Yorke JA, Dvořák J, Salzberg SL. **2017**. Hybrid assembly of the large and highly repetitive genome of Aegilops tauschii, a progenitor of bread wheat, with the MaSuRCA mega-reads algorithm. *Genome research* **27**: 787–792.

SUPPLEMENTARY INFORMATION

Table S1: Summary statistics of the *Zosterops* genome assemblies as computed by Assemblathon2. Values in bold
2505 correspond to the summary statistics of the publicly available sequences.

Species	Assembler	Reference	Assembly size	# scaffolds	Longest scaff.	N50 scaff.	L50 scaff.	N%	# contigs
Z. borbonicus	SOAPdenovo	this study	1188818773	130278	2861829	477419	714	6.43	213846
Z. borbonicus	**SOAPdenovo + SSPACE**	**this study**	**1222452283**	**97503**	**11984413**	**1762991**	**174**	**8.52**	**207674**
Z. borbonicus	SOAPdenovo + SSPACE + Synteny	this study	1222650938	96503	155428647	71485074	6	8.54	207709
Z. borbonicus	MaSuRCA	this study	1077182487	4487	11330673	1864885	153	0.41	9504
Z. pallidus	**SOAPdenovo**	**this study**	**1163666926**	**170557**	**3877680**	**374698**	**785**	**4.57**	**227542**
Z. lateralis	**AllPath_LG**	**Cornetti et al. 2015**	**1036003386**	**2933**	**15146312**	**3581248**	**83**	**3.32**	**55972**

Table S2: List of the 27 avian species used in DeCoSTAR. Number of CDS translated to protein and used for orthology detection.

Species	Accession (NCBI) or URL	# Proteins
Acanthisitta chloris	http://dx.doi.org/10.5524/101015	15052
Anas platyrhynchos	http://dx.doi.org/10.5524/101001	15053
Aptenodytes forsteri	http://dx.doi.org/10.5524/100005	14767
Calypte anna	http://dx.doi.org/10.5524/101004	14543
Chaetura pelagica	http://dx.doi.org/10.5524/101005	14111
Charadrius vociferous	http://dx.doi.org/10.5524/101007	14465
Columba livia	http://dx.doi.org/10.5524/100007	14982
Corvus brachyrhynchos	http://dx.doi.org/10.5524/101008	14927
Corvus cornix	GCA_000738735.2_ASM73873v2	13622
Cuculus canorus	http://dx.doi.org/10.5524/101009	14727
Egretta garzetta	http://dx.doi.org/10.5524/101002	14127
Falco peregrinus	http://dx.doi.org/10.5524/101006	14839
Ficedula albicollis	GCA_000247815.2_FicAlb1.5	15382
Gallus gallus	ftp://climb.genomics.cn/pub/10.5524/100001_101000/101000/chicken/	14728
Geospiza fortis	http://dx.doi.org/10.5524/100040	14180
Haliaeetus leucocephalus	http://dx.doi.org/10.5524/101027	15212
Manacus vitellinus	http://dx.doi.org/10.5524/101010	16402
Melopsittacus undulates	http://dx.doi.org/10.5524/100059	14231
Nipponia Nippon	http://dx.doi.org/10.5524/101003	15018
Opisthocomus hoatzin	http://dx.doi.org/10.5524/101011	13333
Picoides pubescens	http://dx.doi.org/10.5524/101012	14136
Pygoscelis adeliae	http://dx.doi.org/10.5524/100006	13734
Struthio camelus	http://dx.doi.org/10.5524/101013	14577
Taeniopygia guttata	ftp://climb.genomics.cn/pub/10.5524/100001_101000/101000/zebrafinch/	15693
Tinamus guttatus	http://dx.doi.org/10.5524/101014	15723
Zonotrichia albicollis	GCF_000385455.1_Zonotrichia_albicollis-1.0.1	14376
Zosterops borbonicus	this study	22558

2510

Table S3: Fossil calibration combinations used in the molecular dating analyses

Calibration sets	Node	Maximum bound (Myr)	Minimum bound (Myr)
1,2,3	Neognathae /Palaeognathae	86.5	66
1,2,3,4	Galloanserae / Neoaves	Free	66
1,2,3,4	Apodidae / Trochilidae	Free	51
1,2,3,4	Sphenisciformes / Threskiornithidae	Free	60.5
1,3	Passerines / Psittaciformes	65.5	53.3
2,4	Passerines / Psittaciformes	Free	53.3
3	Oscines / Suboscines	34	28
4	Neognathae / Palaeognathae	140	66

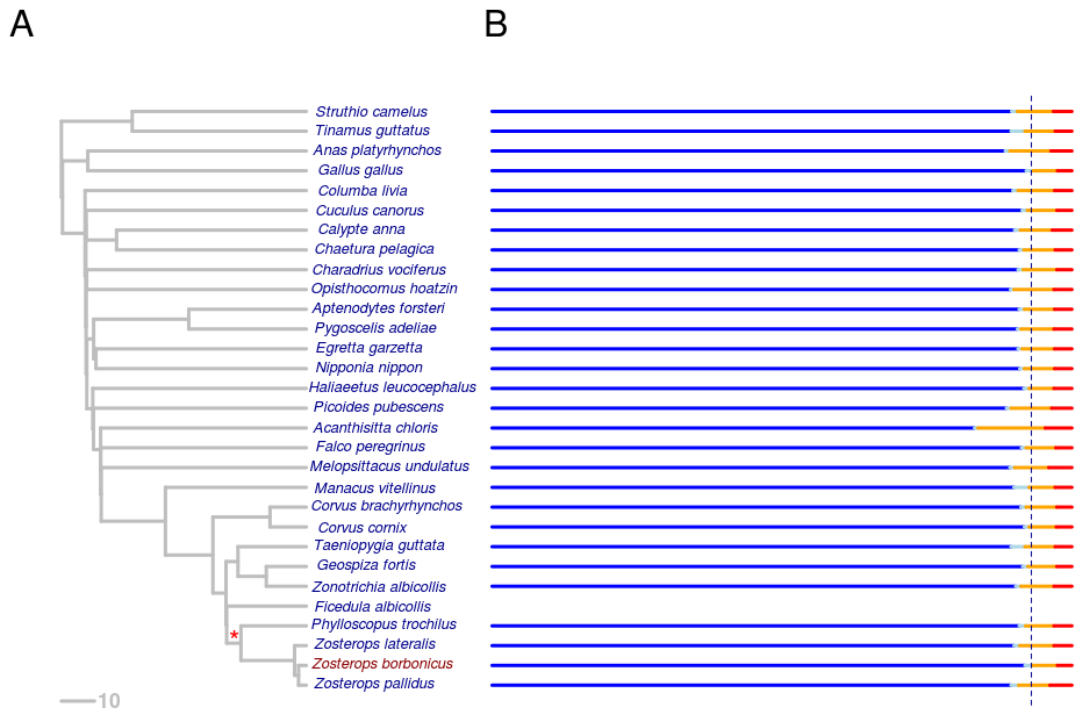


Figure S1: A) Phylogenetic tree based on all investigated avian species. The red star indicates the origin of the two neo-sex chromosomes (ancestor of Sylvioidea). B) Summary statistics of the BUSCO analysis based on the 27 genome assemblies used as input for DeCoSTAR. Blue, light blue, orange and red colors indicate single copy, duplicated, fragmented, missing genes, respectively. The blue dotted line corresponds to the cumulative proportion of complete genes (single copy and duplicated ones) found in *Z. borbonicus*. For unknown reasons, BUSCO analysis failed for the *F. albicollis* assembly.

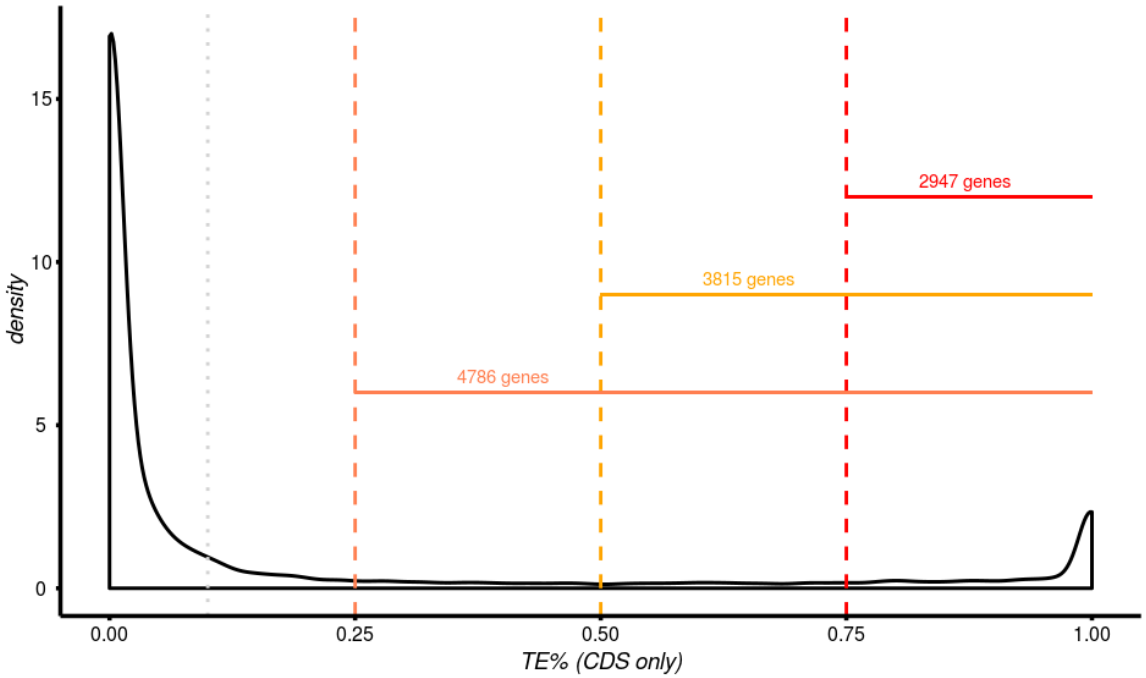


Figure S2: Distribution of the TE content in coding regions observed in the 22,558 *Z. borbonicus* gene models. The grey line shows the threshold used for identifying the most accurate genes (see Fig. 1).

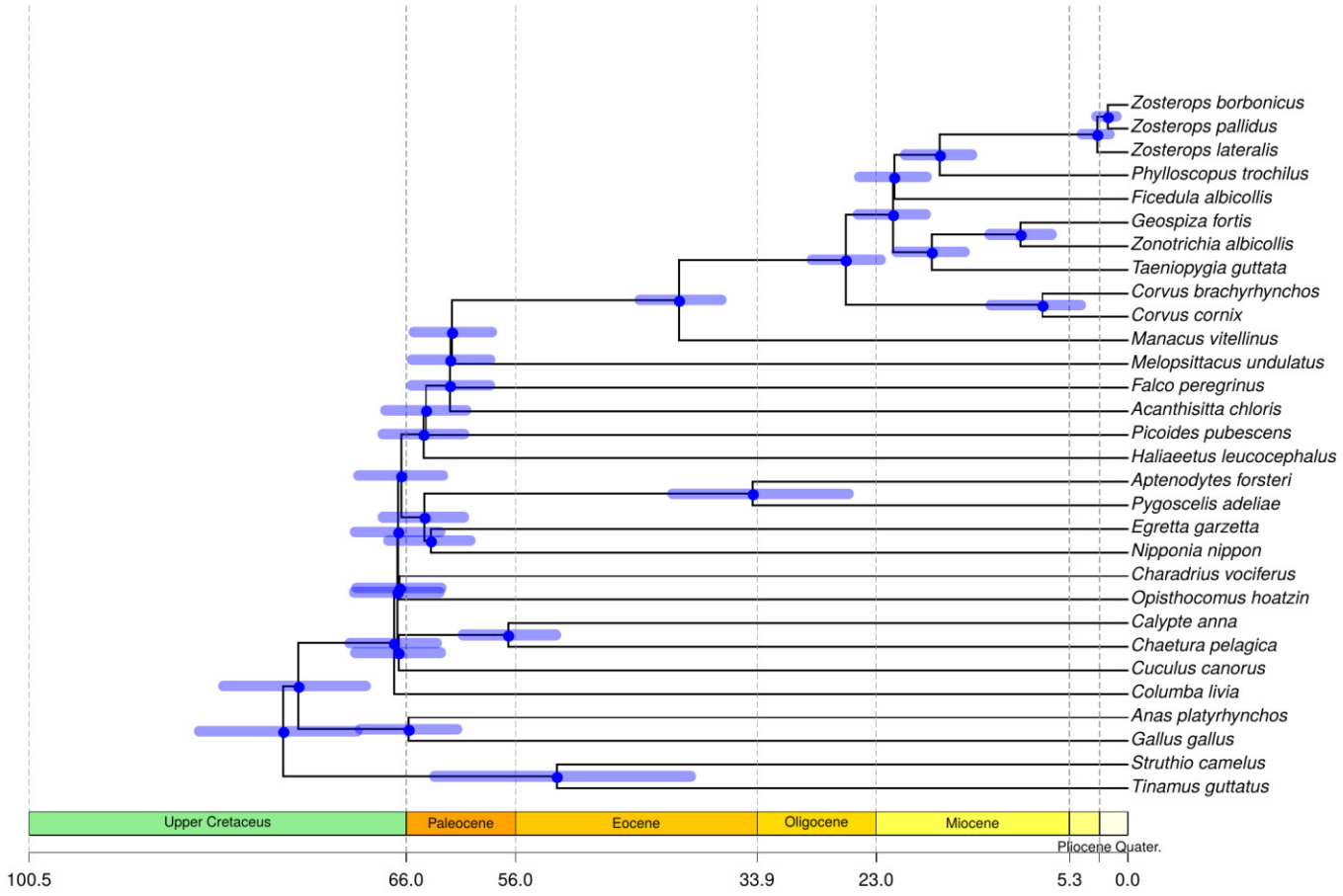
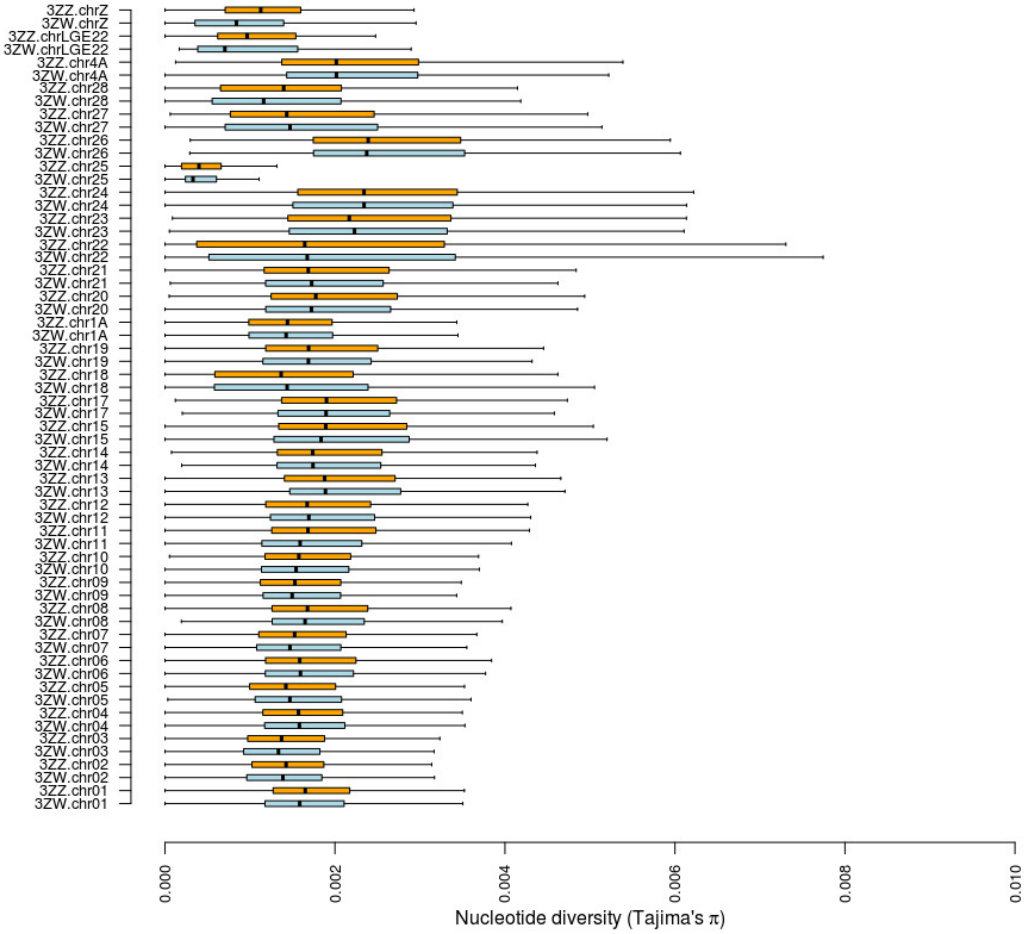


Figure S3: ***Molecular dating of the 30 birds species.*** *Molecular dating estimates are based on the dataset 1 with CAT GTR substitution model, log-normal molecular rate model and calibration set 1 (Table S3).*



2530 **Figure S4: Interchromosomal and interdataset variation in Tajima's $\pi_{females}$ (light blue) and π_{males} (orange) over non-overlapping 10-kb sliding windows.**

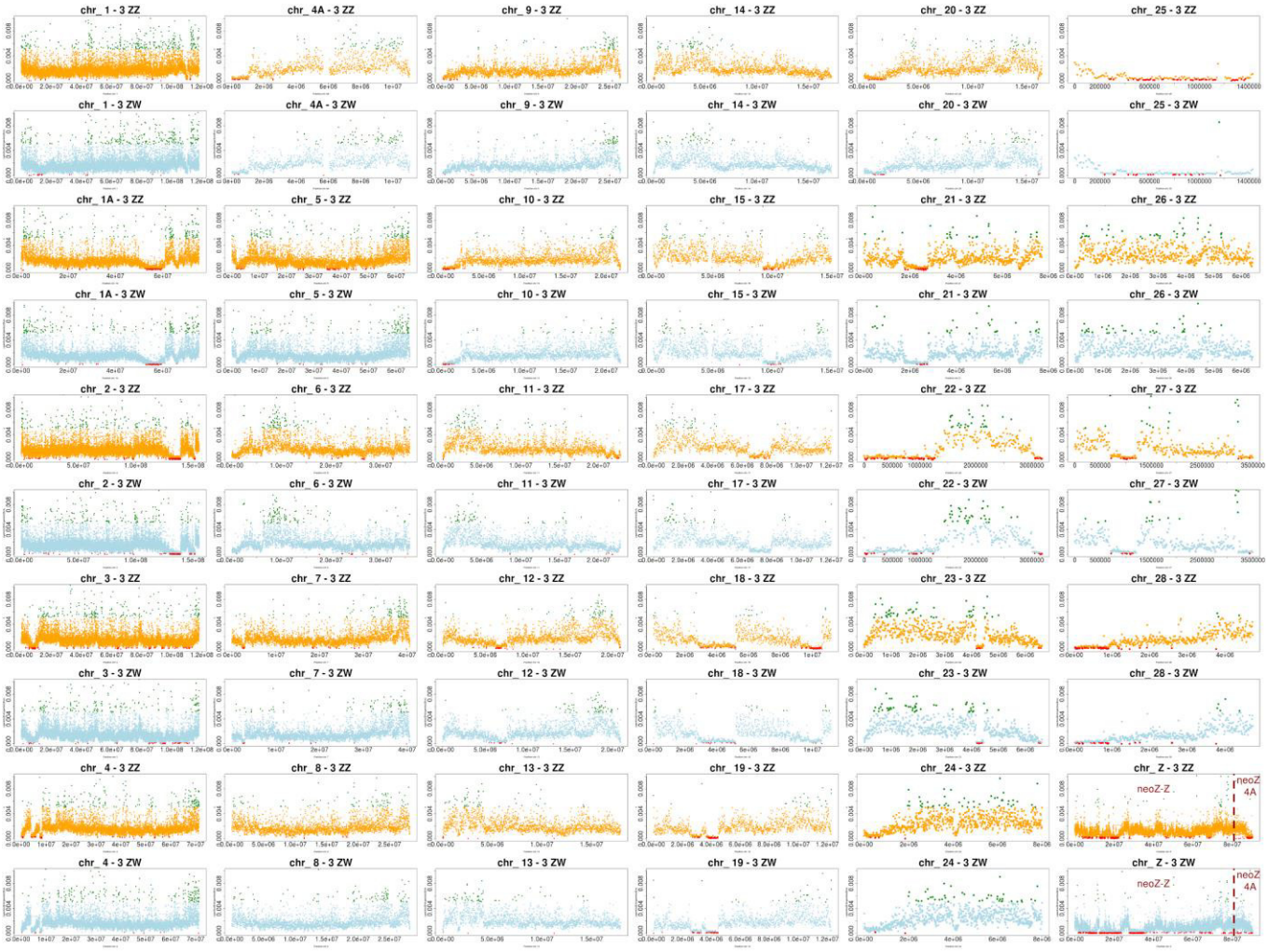
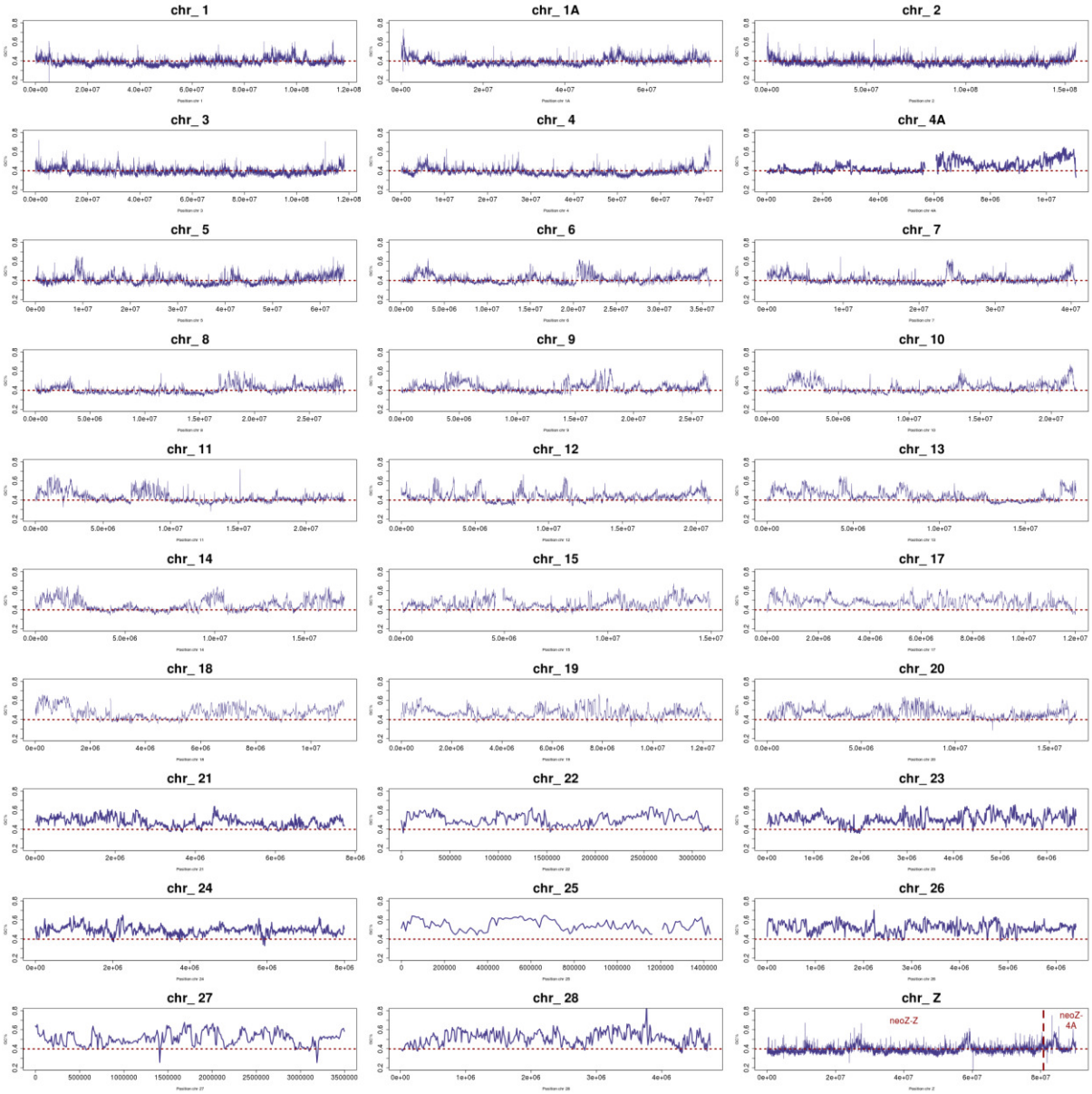


Figure S5: Nucleotide diversity variations along 30 *Z. borbonicus* chromosomes as estimated using genetic information from three males (π_{males}, orange) and three females ($\pi_{females}$, light blue). For each dataset, top 2.5% and bottom 2.5% of π values among windows scanning all chromosomes are shown in green and red, respectively. The chromosomal breakpoint on the neoZ chromosome is shown with a red line.

Figure S6: Tajima's D variations along 30 *Z. borbonicus* chromosomes as estimated using genetic information from three males (orange) and three females (light blue). Top 2.5% and bottom 2.5% of D values among windows scanning all chromosomes are shown in green and red, respectively (baseline for bars is for D=0). The chromosomal breakpoint on the neoZ chromosome is shown with a red line.



2545 **Figure S7: Variation in G+C content along 30 *Z. borbonicus* chromosomes.** The red dotted line indicated the median GC value over non-overlapping 10kb sliding windows for scaffolds assigned to chromosomes only. As expected, strong departures from this median values are observed on minichromosomes.

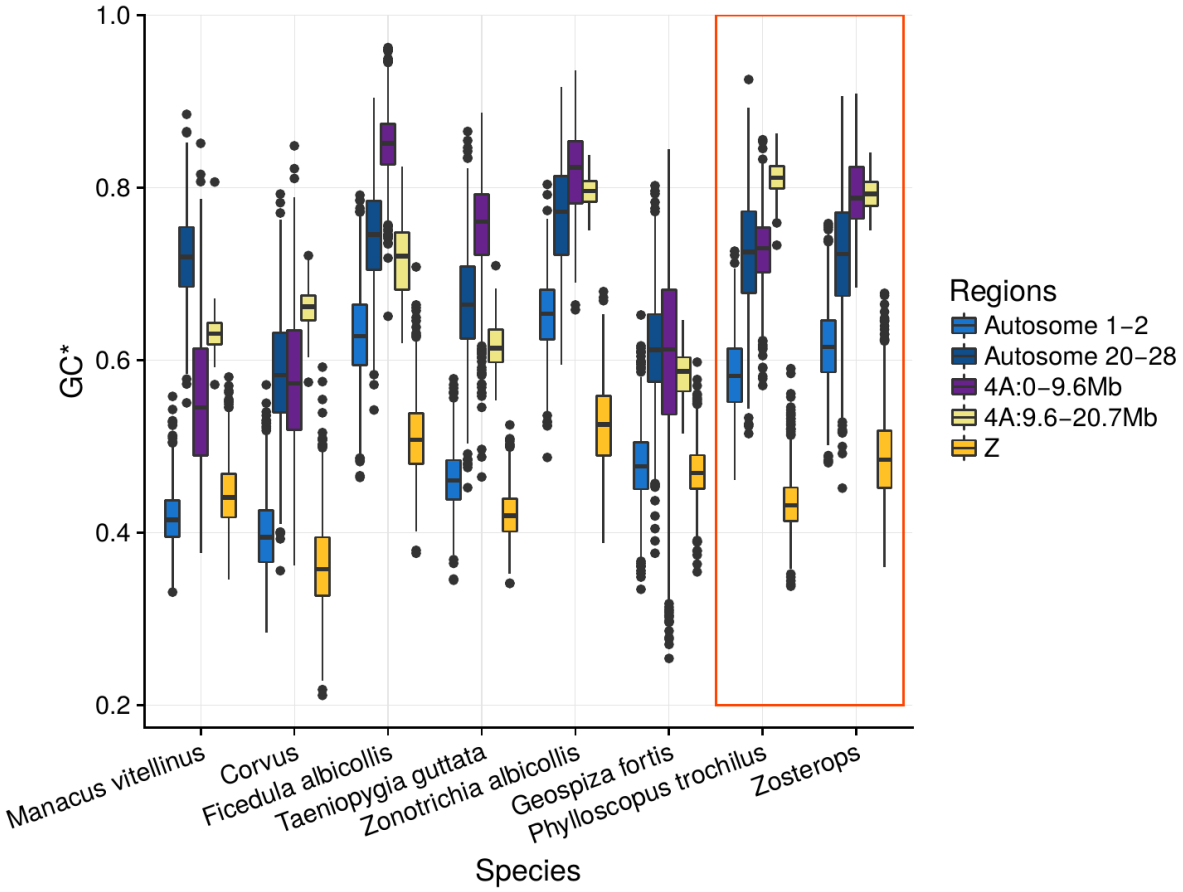


Figure S8: Variations in GC content at equilibrium (GC*) at third codon positions. Variability was obtained by bootstrapping genes within each genomic region.

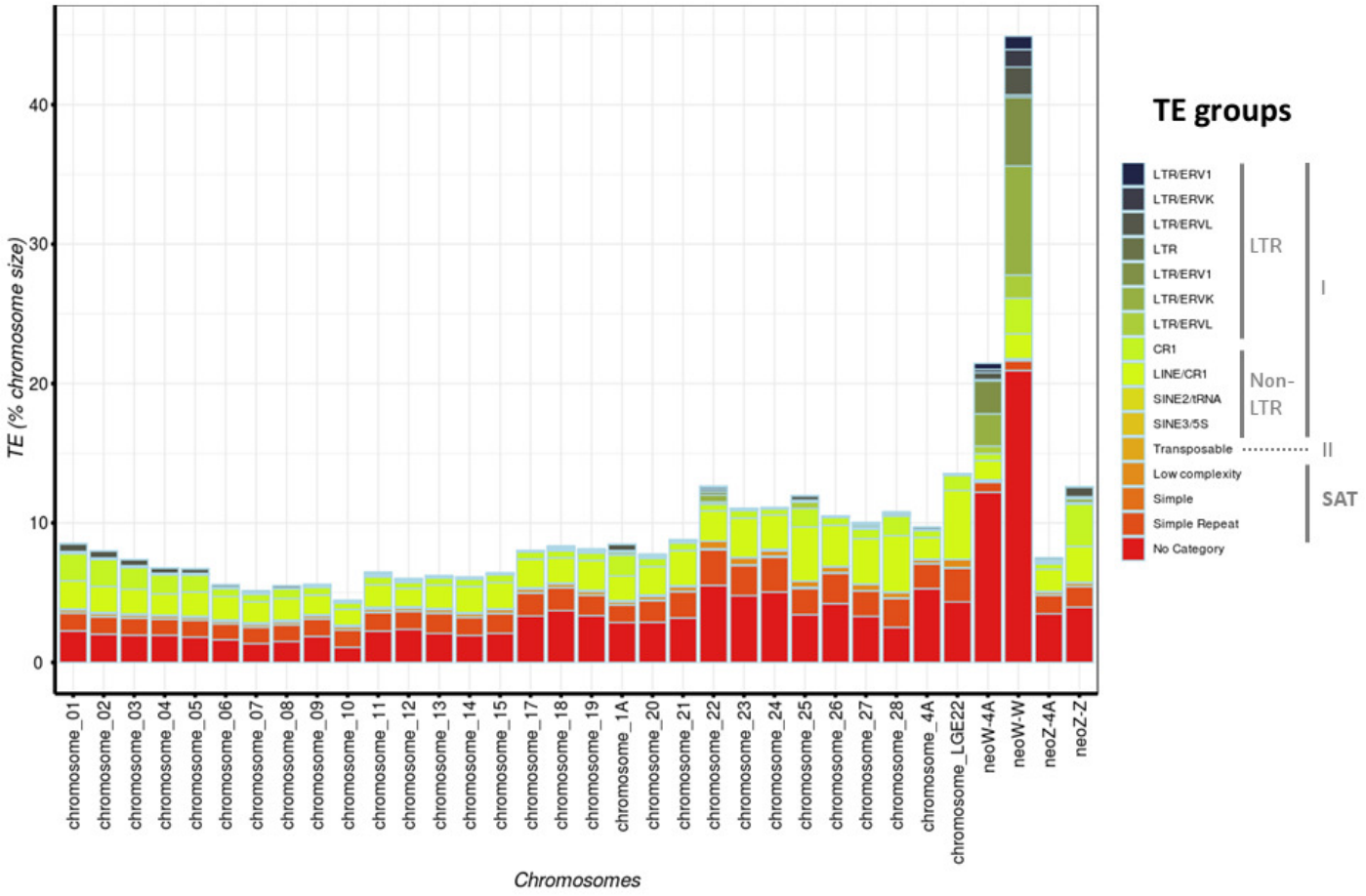


Figure S9: TE density for each chromosome and for each RepBase TE family. Class I TE elements were categorized following the two subclasses: LTR and non-LTR retrotransposons.