The avian W chromosome is a refugium for endogenous retroviruses with likely effects on female-biased mutational load and genetic incompatibilities

Peona, Valentina; Palacios-Gimenez, Octavio M.; Blommaert, Julie; Liu, Jing; Haryoko, Tri; Jønsson, Knud A.; Irestedt, Martin; Zhou, Qi; Jern, Patric; Suh, Alexander

Published in:
Philosophical Transactions of the Royal Society B: Biological Sciences

DOI:
10.1098/rstb.2020.0186

Publication date:
2021

Document version
Publisher's PDF, also known as Version of record

Document license:
CC BY

Citation for published version (APA):
The avian W chromosome is a refugium for endogenous retroviruses with likely effects on female-biased mutational load and genetic incompatibilities

Valentina Peona1, Octavio M. Palacios-Gimenez1,†, Julie Blommaert1, Jing Liu3,4, Tri Haryoko5, Knud A. Jønsson6, Martin Irestedt7, Qi Zhou3,4,8, Patric Jern2 and Alexander Suh1,9

1Department of Organismal Biology—Systematic Biology, and 2Science for Life Laboratory, Department of Medical Biochemistry and Microbiology, Uppsala University, Uppsala, Sweden
3MOE Laboratory of Biosystems Homeostasis and Protection, Life Sciences Institute, Zhejiang University, Hangzhou, People’s Republic of China
4Department of Neuroscience and Development, University of Vienna, Vienna, Austria
5Museum Zoologicum Bogoriense, Research Centre for Biology, Indonesian Institute of Sciences (LIPI), Cibinong, Indonesia
6Natural History Museum of Denmark, University of Copenhagen, Copenhagen, Denmark
7Department of Bioinformatics and Genetics, Swedish Museum of Natural History, Stockholm, Sweden
8Center for Reproductive Medicine, The 2nd Affiliated Hospital, School of Medicine, Zhejiang University, Hangzhou 310052, People’s Republic of China
9School of Biological Sciences—Organisms and the Environment, University of East Anglia, Norwich, UK

VP, 0000-0001-5119-1837; OMP-G, 0000-0002-1472-9949; JB, 0000-0003-1411-2313; TH, 0000-0002-8540-3662; KAJ, 0000-0002-1875-9504; MI, 0000-0003-1680-6861; QZ, 0000-0002-7419-2047; PJ, 0000-0003-3393-5825; AS, 0000-0002-8979-9992

1. Introduction

Many organisms exhibit a genetic sex determination system where a pair of sex chromosomes guides sex development [1]. There are two major genetic sex-determining systems: the XY system with male heterogygamy (XX females and
XY males) and the ZW system with female heterogamety (ZW females and ZZ males), whereby the Y and W are the sex-limited chromosomes (SLCs).

Sex chromosomes generally evolve from a pair of autosomes [2] that acquire a sex-determining locus and locally suppressed recombination around that locus [3,4]. The non-recombining region may remain very small, keeping the two sex chromosomes largely homomorphic. Conversely, in heteromorphic sex chromosomes, the non-recombining region may expand over time until only a small pseudo-autosomal region remains recombining, while the rest of the SLC diverges, degenerates or loses genes, and accumulates repeats [5]. The evolution of the non-recombining region of the SLC is mostly shaped by its low recombination rate. Its associated low effective population size drastically decreases the efficacy of selection [6] (i.e. accentuating the effects of drift and linked selection) and makes these chromosomes vulnerable to the accumulation of slightly deleterious mutations (e.g. through Muller’s ratchet and Hill–Robertson interference mechanisms), such as repeats [3,7].

Because of their low gene content and high repeat density, SLCs were thought to not have any effect beyond sex determination and gonadal development, remaining largely understudied or even absent in the majority of the genome assemblies and studies [8]. However, recent studies on SLCs, especially in humans and other model organisms, have shown that they play roles in human diseases [9,10], male infertility [11], determining sex-specific traits [12], shaping the genome-wide heterochromatic landscape [13], exerting epistatic effects [14–16], reproductive isolation [17] and suppressing meiotic drivers on other chromosomes (e.g. through RNAi pathways) [18].

While Y chromosomes of mammals and flies have recently received considerable attention, the evolutionary implications of W chromosomes in any organism are still poorly understood. Here, we provide, to our knowledge, the first evidence that the avian W chromosome is not merely a graveyard of repetitive elements but a refugium of potentially active transposable elements (TEs) that probably have sex-specific implications. Bird genomes are known to be repeat-poor with a mean TE content of less than 10% [19], but the first female assemblies based on short [20] or long reads [21–23] showed that the non-recombinating W chromosome is over 50% repetitive and especially rich in endogenous retroviruses (ERVs). By analysing reference-quality genome assemblies [23,26], we annotated full-length ERVs (fl-ERVs; electronic supplementary material, table S12) and detected a large proportion of fl-ERVs on the W chromosome compared to the rest of the genome (figure 1c and table 1; electronic supplementary material, tables S3–S12). Despite the fact that the W chromosome accounted only for the 1–3% of the total length of assembled chromosomes (figure 1c; electronic supplementary material, table S3), this chromosome carried the same or higher numbers of fl-ERVs than the autosomes altogether, with the exception of emu with half the number on W than autosomes together (figure 1d and table 1). The distribution of fl-ERVs deviated significantly ($\chi^2$-test, $p$-values<0.01) from a random distribution across all chromosomes (electronic supplementary material, table S4), with an impoverishment of total ERV-derived bp on the autosomes (0–0.4 times fewer bp than expected) and an extreme accumulation on the W (12–54 times more bp than expected; electronic supplementary material, table S12). By contrast, we identified a negligible amount of other full-length TEs per genome (0–11 DNA transposons, 0–8 CR1 LINEs; electronic supplementary material, table S4).

We propose a ‘refugium index’ (equation (4.1)) to quantify the excess accumulation of TE-derived bp on an SLC relative to the rest of the genome by comparing the observed and expected abundance of TEs. Positive values of the refugium index indicate an excess of TEs, while negative values a depletion of TEs. Because only a subset of TE copies are usually capable of (retro)transposition, we propose a ‘toxicity index’ as a quantitative measure for the excess of intact TE copies in the heterogametic versus homogametic sex through the presence of an SLC (equation (4.2)). The excess is calculated by comparing the number of full-length TE in the diploid state in the two sexes. The toxicity index indicates a non-toxic SLC when equal to 0, toxicity of SLCs when positive and toxicity of Z or X when negative. The term ‘toxicity’ pays tribute to the recently proposed ‘toxic-Y’ hypothesis in Drosophila [13], which suggested that an excess of Y-specific active TEs can lead to male-biased

2. Results and discussion
(a) Enrichment of endogenous retroviruses on the W chromosome
We analysed six avian genomes spanning the avian Tree of Life (figure 1a) and representing the current standard for reference-quality genome assemblies [23,26]. Autosomes had between 6 and 12% TEs on average (figure 1b; electronic supplementary material, table S2 and S1) and the Z chromosome had similar or slightly higher TE densities (5–17%), while the W chromosome stood out as having approximately 22–80% TEs (electronic supplementary material, table S2). Notably, we also found the homomorphic W chromosome of emu to be richer in TEs than the autosomes and Z (22 versus 6.4 and 5.6%). Generally, the Z chromosome exhibited a TE landscape more similar to the autosomes than to the W chromosome, both regarding abundances and types of TEs (figure 1b; electronic supplementary material, table S2). While long interspersed elements (LINEs) from the Chicken Repeat 1 (CR1) superfamily were the dominant repeats on autosomes and Z (cf. [19,27]), ERVs were the major component of the W chromosome and accounted for more than 50% of the assembled chromosome itself (electronic supplementary material, table S2).

ERVs are long terminal repeat (LTR) retrotransposons deriving from germ-line-inherited retrovirus integrations and exist mainly in two genomic forms [28,29]: (i) full-length elements with terminal repeats (likewise called LTRs) flanking its protein-coding genes necessary for retrotransposition; and (ii) solo-LTRs resulting from homologous recombination between the two flanking LTRs. Only full-length elements are capable of autonomous retrotransposition. Using RETROTECTOR and LTRHarvest/LTRdigest [30–32], we annotated full-length ERVs (fl-ERVs; electronic supplementary material, S2 and S3) and detected a large proportion of fl-ERVs on the W chromosome compared to the rest of the genome (figure 1c and table 1; electronic supplementary material, tables S3–S12).
transposition and genome instability, together probably detrimental to the genome and the organism. For birds, we calculated the toxicity index as the excess of fl-ERVs carried by diploid females compared to diploid males (table 1), suggesting that females with heteromorphic sex chromosomes carried between 20 and 90% more fl-ERVs than males, and that even the emu has 7–16% more fl-ERVs in females than males despite largely homomorphic sex chromosomes [25,33]. We assume this phenomenon to reflect that the non-recombining region of the W, no matter how big or small, constantly accumulates large quantities of new TEs. It is important to note that, given the difficulties in assembling SLCs even with long-read sequencing technologies [8,23,26,34], the W chromosome models are likely to be less complete than the other chromosomes. We thus consider our W repeat annotations as well as indexes to be conservative estimates for the true repeat content.

Our results suggest that the avian W chromosome is acting as a refugium for intact and thus potentially active TEs, particularly ERVs, which may have numerous implications. We thus propose the ‘refugium hypothesis’ for SLCs in general: the accumulation of TEs on the SLC leads to an excess of intact TEs in the heterogametic sex, with a toxic effect absent from the SLC-lacking homogametic sex. This sex-specific toxic effect may manifest itself as sex-biased mutational load, genomic instability, ageing and genetic incompatibilities as a result of SLC-linked TE activity and heterochromatin dynamics (explained below). To quantify and test the refugium hypothesis in any sex chromosome system of interest, we introduced two indexes above: the refugium index to measure the density of TE-derived bp on the SLC relative to the remaining chromosomes; and the toxicity index to measure the number of intact TEs (i.e. full-length copies of LTRs, LINEs and DNA transposons) in the heterogametic sex relative to the other sex.

(b) Transcription and translation of W-linked endogenous retroviruses

Considering the exceptionally high number of W-linked fl-ERVs, we tested whether the avian W chromosome harbours a potentially active load of ERVs specific to females. In the absence of available retrotransposition assays for birds, we regarded the transcription and translation of W-linked ERVs as proxies of their activity. We identified W-linked single-nucleotide variants (SNVs) within ERVs by mapping genome re-sequencing data from male and female individuals, as well as female transcriptome data, to consensus sequences of our repeat library (electronic supplementary material, S5). We consider this to be a conservative subset of W-linked SNVs because we required each SNV to be present in all females and absent in all males per species. However, the paradise crow dataset that contained only one male probably gave rise to false positive W-linked SNVs. We then traced the presence of ERV proteins in the male and female proteome data available for white leghorn chicken.

We analysed zebra finch, paradise crow, chicken and emu for W-linked SNVs in genome re-sequencing and RNA-seq data mapped against ERV consensus sequences. In each species, we found between 52 and 332 ERV subfamilies with

Figure 1. Massive accumulation of ERVs on W chromosomes of six female reference-quality genome assemblies spanning the avian Tree of Life. (a) Avian time tree after [24] with schematic homomorphic or heteromorphic sex chromosomes [25]. (b) TE landscapes of autosomes and sex chromosomes as stacked bar plots. Abundance of interspersed repeats (bp occupied) normalized by chromosome size plotted against percentage of divergence calculated as Kimura 2-parameter distance to consensus. The refugium index (RI) for interspersed repeats on autosomes, Z and W is indicated for each species. (c) Comparison of full-length TE numbers (mainly ERVs) on autosomes and sex chromosomes as pie charts. (d) Comparison of full-length TE numbers (mainly ERVs) on autosomes and sex chromosomes as pie charts. The toxicity index (TI) of the W chromosome is reported for each sex as the range between estimates from RetriTe or LTRAUST + LTRGHOST (table 1).
W-linked SNVs (table 2), with ERVL subfamilies being the most represented (electronic supplementary material, table S13) and found evidence for the transcription of between 12 and 182 ERV subfamilies in female gonads or female pectoral muscle (table 2; electronic supplementary material, table S13). Our estimates of transcribed W-linked ERVs are probably just the tip of the iceberg, because we expect to identify W-linked SNVs only if those ERVs have not yet spread in the genome (e.g. very recent variants) or if they accumulated exclusively on the W chromosome (e.g. fl-ERV only existing as solo-LTRs on other chromosomes). Alongside ERVs, we also identified W-linked SNVs in CR1 LINEs and DNA transposons (electronic supplementary material, table S13).

Although there is evidence for their transcription, their scarcity of full-length elements makes CR1 LINEs and DNA transposons an unlikely source of mutational load for females.

Next, we analysed the overall RNA expression level of ERVs in male and female gonads of emu, chicken and zebra finch via RNA-seq read mapping to genomic regions annotated as LTR or ERV fragments by REPEATMAKER (figure 2a). Overall, females expressed such ERVs more highly than males, with the single Z chromosome of females showing expression levels that matched the two male Z chromosomes. This pattern contrasts with incomplete dosage compensation of Z-linked genes in birds indicated by the usual twofold higher expression level of Z-linked genes in males [35,36]. Assuming that some of these ERVs are full-length and capable of retrotransposition, female gonads would thus be exposed to a greater mutational load. Furthermore, many of the autosomal and Z-linked ERVs showed differential expression towards females (electronic supplementary material, figure S1).

Finally, we analysed protein mass spectrometry data of white leghorn chicken gonads [37] with MAXQUANT [38] for the presence of TE-related proteins and found a higher quantity of more of these proteins expressed in females than in males as indicated by a high H/L SILAC ratio (figure 2b; electronic supplementary material, S6). Together, these results demonstrate that some W-linked ERVs are transcribed and that females have more ERV translation than males, and that W chromosomes thus feature fl-ERV's potentially able to retrotranspose.

Given our present data, we cannot distinguish whether this higher ERV translation stems solely from the W chromosome but it is plausible that the presence of an SLC causes a higher TE activity (similarly to what happens in Drosophila [13]).

(c) Sex-biased implications for mutational load

SLCs have been largely considered inert chromosomes with few effects beyond sex determination and gonadal development because of their low gene content (e.g. only 13 genes on Drosophila Y [39] and 28 genes on chicken W [21]). However, accumulating evidence shows that SLCs can have additional effects [12,40,41]. For example, it is important to highlight that the Y-linked regulatory variation within populations of Drosophila can have genome-wide epistatic effects [14–16,42]. This Y-linked regulatory variation cannot be explained simply by regulatory variation of the protein-coding genes and it has been proposed that the variability in Y repetitive content and structural variation are responsible for re-shaping the genome-wide heterochromatin landscape [43]. This hypothesis is known as the heterochromatin sink model, suggesting that large heterochromatin blocks on SLCs act as a sink for the heterochromatin machinery and thereby reduce the efficiency of...
heterochromatin maintenance elsewhere relative to the SLC-lacking sex [13,43].

Recently, the Y chromosome repeat content has been linked to the destabilization and loss of heterochromatin, which in turn is correlated to the shorter lifespan of the heterogametic sex [13,44]. By using Drosophila melanogaster experimental lines with different Y dosages (XO males, XXY females, XYY males), Brown et al. [13] showed that the presence and number of Y chromosomes carried are correlated with shorter lifespans. It was thus suggested that the Y itself is ‘toxic’ for the entire genome and organism, and this toxicity is caused by the Y-linked load of active TEs [13,45,46] whose expression is unleashed by heterochromatin loss. Possibly, the dysregulation of TEs owing to heterochromatin loss is also associated with laminopathic diseases in Drosophila and humans [47]. According to the refugium hypothesis proposed here, we predict that in species with a high toxicity index (i.e. excess of intact TEs on the SLCs and/or paucity thereof in the rest of the genome), this toxic effect will be more accentuated (figure 3a,b). The toxic-Y hypothesis has been recently investigated from a theoretical point of view in vertebrates with both XY and ZW systems [48] and put in contrast with the classic ‘unguarded-X’ hypothesis [49–51], which proposes that the expression of recessive mutations on X/Z chromosomes is the cause of the shorter lifespan in the heterogametic sex. It is important to note that reduced female lifespan in birds has been documented in many species [52–55]. Sultnova et al. [48] used the sizes of Y and W relative to X and Z as a proxy for toxicity, i.e. assuming that smaller SLCs are more repetitive. Although the correlation between the Y size and relative lifespan in mammals was strong, the authors did not find such a correlation for the W in birds. We note that while SLC size relative to X/Z size might indeed correlate negatively with the overall repeat content (i.e. satellites and fragmented TEs), this might not necessarily be informative for the number of intact TEs. Therefore, we propose that our toxicity index could be a more suitable proxy for toxicity because it considers the sex differences in the load of intact and (potentially) active TEs. Among the six birds compared here, emu and Anna’s hummingbird would be those with the lowest and highest toxicity indexes, and it remains to be tested if this indeed is a better predictor of female lifespans.

(d) Sex-biased implications for genetic incompatibilities

In addition to TE mutational load and heterochromatin maintenance influencing organismal physiology, SLC-linked TEs can also play an important role during hybridization. This

<table>
<thead>
<tr>
<th>species</th>
<th>no. of W-linked SNVs in ERVs</th>
<th>no. of ERV subfamilies with W-linked SNVs</th>
<th>no. of transcribed SNVs in ERVs</th>
<th>no. of transcribed ERV subfamilies</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dromaius novaehollandiae</td>
<td>764</td>
<td>58</td>
<td>82</td>
<td>12</td>
</tr>
<tr>
<td>Gallus gallus</td>
<td>2088</td>
<td>52</td>
<td>671</td>
<td>28</td>
</tr>
<tr>
<td>(red junglefowl)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gallus gallus (white leghorn)</td>
<td>6385</td>
<td>166</td>
<td>3534</td>
<td>102</td>
</tr>
<tr>
<td>Lycoorax pyrrhopterus</td>
<td>1591</td>
<td>198</td>
<td>42</td>
<td>21</td>
</tr>
<tr>
<td>Taeniopygia guttata</td>
<td>15 012</td>
<td>332</td>
<td>3306</td>
<td>182</td>
</tr>
</tbody>
</table>

Table 2. Number of female-specific and thus W-linked SNVs relative to ERV consensus sequences detected at the genomic and transcriptomic levels for each species. (More details about SNVs in ERVs, LINEs and DNA transposons are in the electronic supplementary material, table S13.)
point may be not overly surprising in the context of Haldane’s rule, which states that upon hybridization, if there is a sterile or inviable sex, it will be the heterogametic one. Accumulating evidence suggests that hybrid genome stability can be compromised during mitosis and meiosis by species-specific differences in heterochromatin landscapes leading to uncontrolled TE activity (reviewed by Serrato-Capuchina & Matute [56]). Furthermore, species-specific families of repeats can induce lagging chromatin at cell division during early embryogenesis (when heterochromatin is first established), leading to chromosome mis-segregation and F2 hybrid embryo death [57]. In the context of the refugium hypothesis, it is important to consider that new and active TEs are one of the main targets of heterochromatinization [58,59], and SLCs could be a source for both sex-specific and species-specific heterochromatin differences.

Figure 3. Synthesis of the consequences of the refugium hypothesis on micro- (a,b) and macroevolutionary (c,d) time scales. For simplicity, a schematic example of avian sex chromosomes is shown, but we expect these consequences for any ZW or XY system with SLC-linked intact TEs. (a) Simplified karyotypes of male and female birds indicating that the W chromosome is a TE refugium and heterochromatin sink. Grey circles: centromeres and heterochromatin blocks; red lines: intact TEs; red arrows: TEs spreading away from the W chromosome; grey arrows: heterochromatin deposition on the W chromosome. (b) The ‘toxic’ effect of the gradual de-repression of TEs during an organism’s lifetime is more accentuated in females carrying more intact TEs than males. The W-linked activity of intact TEs could explain the shorter lifespan of the heterogametic sex. The toxicity of active TEs is represented by the increasing number of transcripts in red as a proxy for genome-wide TE insertions. (c) Simplified karyotypes of two bird species with species-specific TEs (blue and orange lines) and sequence-specific TE repressors (blue star and orange hexagon). Assuming a rapid accumulation and sequence turnover of TEs especially on the W, diverging species or populations may quickly acquire different TE/repressor repertoires. (d) Genetic incompatibility owing to the W chromosome in a female F1 hybrid between species 1 and species 2 of (c). Schematic example of four possible meiotic products (oocytes) of the F1 hybrid, two of which lack the blue repressor of blue TEs because of meiotic recombination between the autosomes. The TE/repressor mismatch may lead to de-repression of W-linked TEs in gametes or embryos and thereby a female-biased reduction in hybrid fitness.
TEs generally evolve very rapidly in their sequence and usually only few elements remain intact and capable of transposition [60]. In addition, many TE repressor systems are in a sequence-specific arms race (e.g. piRNAs or KRAB-zinc finger proteins [58,61,62]); therefore, TE sequences and their repressors can both diverge rapidly between populations and species. Because SLCs rapidly evolve and accumulate repeats [5,18], SLCs are probably sex-specific refugia of species-specific active TEs. Hybrid incompatibility owing to TE/repressor mismatches can arise when new TE families are introduced into a naive genomic background (lacking specific repressors), which can lead to the uncontrolled proliferation of such TEs, followed by gene disruption, genome instability [63] and hybrid dysgenesis [17]. TE/repressor mismatches can already occur during meiosis in the F1 hybrids, when recombination can separate the repressor from the controlled TEs (figure 3c,d) [64]. Although this scenario can occur in both sexes, we expect that in species with a high number of intact TEs on the SLCs relative to the rest of the genome (i.e. high toxicity index, highly heteromorphic SLCs), there are more chances for a mismatch between a repressor and intact TEs on the SLCs than for other chromosomes (figure 3d).

For the birds analysed here, the W chromosome is probably the main source for genome-wide new TE insertions because it contains 16%–50% of all intact TEs in a diploid female. Furthermore, potential TE/repressor mismatches stemming from the W chromosome would also reinforce the observation of reduced mitochondrial (maternally inherited as the W) introgression during hybridization in birds [65]. Mitonuclear incompatibilities, i.e. mismatches between mitochondrial and nuclear alleles, play a disproportion role in both intra-specific and hybrid incompatibilities [66–69]. In ZW systems, these mitonuclear incompatibilities may be even more exacerbated because of the co-inheritance with the SLCs, and especially when SLCs feature active repeats. Thus, in addition to the preservation of dosage-sensitive genes [21,70], the W represents a reservoir of many different and intact TEs that, through their potential for de-repression in hybrids, may constitute an additional explanatory variable for Haldane’s rule.

3. Conclusion

We suggest that the avian W chromosome, no matter how heteromorphic or homomorphic, is a refugium for TEs and specifically fl-ERVs, some of which are expressed and thus potentially capable of retrotransposition. This pattern should be generalizable for all birds given our broad sampling of Palaearctinae, Galloanserae and Neoaves. We propose that ERVs are continuously shaping W evolution and are one of the major contributors of structural changes of this chromosome. If so, it is reasonable to speculate that ERVs have played a relevant role in the expansion of the non-recombining region of the W (cf. [71]), for example, by contributing to the heterochromatinization of euchromatic regions through new ERV insertions.

We hope that the refugium and toxicity indexes proposed here will help testing these hypotheses in avian W chromosom es, and SLCs in general. The toxicity index measures the excess of intact TEs on an SLC, which represents the potential for genome-wide sex-specific mutational load as well as sex-specific genome instability. On the short time scale of individuals, a high toxicity index could lead to larger physiological differences between the two sexes [13]. In the long term, e.g. between populations and species, the accumulation of TEs as measured by the refugium index can have effects on reproductive isolation through TE/repressor mismatches, similar to the situation in Drosophila [17,57]. It is important to underline that the toxicity of SLCs should be linked to the number of intact TEs rather than to the general repetitiveness of the chromosome. Furthermore, the refugium and toxicity indexes can be useful to predict and test hybrid incompatibilities, in addition to measuring the genetic distance between nuclear and mitochondrial genes [72]. We predict that with the increasing availability of genome assemblies based on long reads, these indexes will find applicability across SLCs in general. For birds and their W chromosomes, the possible toxic effect of the W on lifespan requires additional tests in vivo that exclude the effects of the phenotypic sex (e.g. developing systems similar to the four core genotypes in mice [73] or the attached-X/attached-Y karyotypes in Drosophila [74,75]) and account for confounding ecological factors (e.g. intense sexual competition and predations especially of males).

To conclude, SLCs are not merely refugia for repeats with usually neutral or slightly deleterious effects on SLCs themselves, but SLC-linked intact TEs may have genome-wide effects that could effectively turn SLCs into ‘toxic wastelands’.

4. Material and methods

(a) Samples, DNA, RNA and proteome data

We used the female reference-quality genome assemblies of chicken (Gallus gallus; GCA_000002315.5; galGalGa), paradise crow (Lycocornix pyrrhothorpus; GCA_014706295.1) [23], emu (Dromaius novaehollandiae; GCA_01612833.31) [76], Anna’s hummingbird (Calypte anna; GCA_003957555.2) [26], kakāpō (Strigops habroptila; GCA_004027225.2; bStrHab1.2.pri) [26] and zebra finch (Taeniopygia guttata; GCA_009859065.2; bTaeGut2. pri.v2) [26]. All these six assemblies have chromosome models and we carried out all analyses considering only using assembled chromosomes, i.e. discarding unplaced contigs and scaffolds.

For chicken, Illumina genome re-sequencing libraries were collected for two females and three males of Gallus gallus gallus (red junglefowl) from [77] (originally uploaded on NCBI as of undetermined sex) and a female library of Gallus gallus bankiva (red junglefowl from Java) from [78]. The sexes of the individuals from [77] were determined using the SEXCMD with default sex markers [79]. Red junglefowl RNA-seq libraries of a female (ovary) and of a male (testes) were retrieved from [80]. We also collected publicly available data for the chicken breed white leghorn, i.e. Illumina genome re-sequencing libraries of one female and three males from [70,81], RNA-seq libraries and protein mass spectrometry libraries for five ovaries and five testes [37].

For paradise crow, we used one 10X Genomics Chromium linked-read library of DNA from a pectoral muscle sample of a female from [23]. We also newly generated such data for three females and one male using the same methods [23] and generated RNA-seq data from female pectoral muscle (preserved in RNAlater). RNA was extracted with phenol-based phase separation using the TRZol reagent (ThermoFisher Scientific) following the standard protocol recommended by the supplier, followed by DNase treatment for 30 min using the DNA-free DNA removal kit (ThermoFisher Scientific). Sequencing libraries were prepared according to the TruSeq stranded total library preparation kit with RibonZero Gold treatment (Illumina, Inc., cat. no. 20020598/9). Paired-reads (150 bp) were sequenced on the NovaSeq SP flowcell (Illumina, Inc.).

For zebra finch, we used Illumina genome re-sequencing libraries of four females and four males from [82], and RNA-seq libraries of two ovaries and one testis from [83,84].
Finally, for emu, we collected Illumina genome re-sequencing libraries of two females and two males from [85–87], and RNA-seq libraries for seven ovaries and five testes from [86,88].

More details and accession numbers for all the libraries and genomic sequences used here can be found in the electronic supplementary material, table S1.

(b) Repeat annotation
To best annotate repeats in all six avian species, we made sure to have species-specific repeat prediction tools for each. The repeat libraries of chicken, paradise crow and zebra finch were already manually curated elsewhere [23,89,90] while species-specific repeat libraries did not exist for emu, Anna’s hummingbird and kākāpō. Therefore, we de novo characterized repetitive elements in these last three species using RepeatMiner v2 [91] and manually curated those sequences labelled as ‘LTR’ and ‘unknown’ following the same method as in [23]. We also inspected consensus sequences with unusual classification for being avian repeats like many DNA transposon superfamilies [19]. We then concatenated the newly curated libraries with the avian consensus sequences from Repbase [92], hooded crow [93], blue-capped cordon bleu [94], collared fly-catcher [95] and paradise crow [23], and used this final library to mask all six genomes with RepeatMasker [96]. The new repeat libraries and notes on their classification are given in the electronic supplementary material, S7.

(c) Quantity of endogenous retrovirus transcription and their differential expression
We used Illumina RNA-seq reads from adult gonads from emu, chicken and zebra finch (electronic supplementary material, table S1) mapped against genomic copies/fragments annotated as LTR or ERV by RepeatMasker to quantify ERV transcription levels and investigate whether the ERVs were differentially expressed across available tissues. For these species, three to five biological replicates for every tissue were used.

Raw RNA-seq data were quality controlled using FastQC [97] and trimmed with TrimGalore [98] using default settings, then mapped to the respective reference genomes using STAR [99]. The alignment was filtered by running featureCounts function from the package Subread v2.0.0 in paired-end mode [100], and only uniquely mapping reads were retained. We provided featureCounts with a filtered RepeatMasker .out file containing only repeat copies annotated as LTR or ERV. Per-genome counts were obtained using read counts and lengths of corresponding ERVs.

DESeq2 1.20.0 [101] implemented in the R Bioconductor package was used for relative quantification of the ERV transcripts and for calculating the TPM (transcripts per million), giving a normalized ERV expression level. Male reads that mapped to the W chromosome represented low counts and were, therefore, removed during the normalization step. The values from replicates of each sample were averaged for the final plots of ERV expression. To identify biased ERVs per chromosome type (i.e. autosomes, Z and W), we compared adult gonads from male and female individuals. The statistical analysis of differentially expressed ERVs was performed using DESeq2. All p-values were adjusted (padj) using the Wald test. The degree of bias was determined by the log2 fold-change (log2FC) difference between conditions. Therefore, the ERVs with log2FC > 0 and log2FC < 0 together with a padj < 0.05 were considered as biased ERVs in the conditions.

(d) Full-length transposable element detection and abundance
Here, we define full-length TEs as possible (retro)transposition-competent elements with relatively complete structures and the potential to produce transcripts. We identified fl-TEs in all the six avian genomes by adopting different methods for DNA transposons, LINES (e.g. CR1) and LTR retrotransposons (ERV). For DNA transposons and LINES, we first identified open reading frames (ORFs) in the insertions annotated by RepeatMasker, then translated such ORFs and aligned with RPS-BLAST [102] against a custom Pfam [103] database containing transposon-related proteins (similar approach to [104]). ORFs from LINES of at least 600 bp that spanned 90% of both endonuclease and reverse transcriptase domains were considered as full-length elements. Likewise, ORFs belonging to DNA transposons of at least 1 kb that spanned 90% of the transposase protein domain were considered full length.

In order to detect and quantify fl-ERV, we used RetroTector [30] as well as LTRharvest [31] together with LTRdigest [32]. RetroTector results were filtered for scores over 300 and presence of 5’-LTR and 3’-LTR, as well as ORFs with complete or partly complete gag, pol and env genes as previously described in [30,105]. LTRharvest results were filtered for false positive using LTRdigest in combination with hidden Markov models profiles of LTR retrotransposon-related proteins downloaded from Pfam [103] and GyDB [106].

(e) Identification of single-nucleotide variants of W-linked endogenous retroviruses and their transcription and translation
To verify the hypothesis that the W chromosome is a refugium of intact and potentially active ERVs, we identified W-linked SNVs within ERVs and traced their transcription in RNA-seq data and translation in protein mass spectrometry data wherever possible. W-linked ERV transcription was analysed in G. gallus, L. pyrrhopos and T. guttata (electronic supplementary material, table S1). ERV translation was analysed in G. gallus white leghorn breed [37]. RNA-seq and proteome libraries selected for this analysis were from gonad tissue with the exception of L. pyrrhopos for which the RNA-seq data were generated from female pectoral muscle.

To identify W-linked SNVs from male/female read mapping, we used the WhatGene pipeline developed by Ruiz-Ruano et al. [107] for SNV analyses of B chromosomes and germline-restricted chromosomes [108] where we mapped male and female genome re-sequencing reads to the consensus sequences of our repeat library. We considered variants to be W-linked if they were present in all females but absent in males. We then checked for the presence of these W-linked variants in the RNA-seq data always following the WhatGene pipeline. Variants that were called W-linked from genomic data but were present in male transcriptomic data were discarded as false positives owing to sample size.

To check for the presence of ERV-related proteins in white leghorn chicken proteome data, we extracted the ORFs from ERV consensus sequences and translated them into peptides using ORFFinder [109]. The peptide sequences were used as query database for MaxQuant 1.6.17.0 [38]. We used the experimental parameters described in [37] (electronic supplementary material, S3); search results were filtered with a false discovery rate of 0.01. Second peptides, dependent peptides and match between runs parameters were enabled.

(f) Refugium index and toxicity index
To test whether intact TEs are uniformly distributed throughout the genome, we compared the observed total number of fl-ERVs (assuming that the numbers of other intact TEs are negligible in avian genomes [19]) on autosomes and sex chromosome to their expected values with a χ²-test with 2 degrees of freedom. We calculated the expected values of TE densities on the chromosomes by assuming a uniform density of these elements across chromosomes (electronic supplementary material, tables S3 and S4). Next, we calculated the refugium and toxicity indexes, which are described below for SLCs in general.
The refugium index (equation (4.1)) calculates the percentage of excess or depletion of observed TE-derived bp (%TEobs) with respect to the genome-wide average of the total TE-derived bp of a haploid genome assembly (%TEexp). We recommend estimating TE densities in RepeatMasker or similar homology-based annotations using a species-specific repeat library combined with libraries of related species in Repbase or similar databases:

$$\text{refugium index} = \frac{\%\text{TE}_\text{obs} - \%\text{TE}_\text{exp}}{\%\text{TE}_\text{exp}} \quad (4.1)$$

The refugium index indicates whether an SLC shows an excess (RI > 0) or a depletion of TEs (%TEobs < %TEexp) in the genome. Here, 2nhet and 2nhom are the total numbers of intact TEs in the diploid state in the homogametic sex (2 × autosomes + 2 × Z or X) and the heterogametic sex (2 × autosomes + 1 × Z or X + 1 × W or Y), respectively. We recommend quantifying intact TEs as the sum of the number of full-length LTR retrotransposons (incl. ERVs) in RepeatModel/LTRHarvest or similar structure-based approaches and the number of copies spanning greater than 90% of the ORFs of DNA transposons (i.e. transposase) and LINEs (i.e. ORF1 or ORF2) in RPS-BLAST or similar homology-based searches:

$$\text{toxicity index} = \frac{2n_{\text{het}} - 2n_{\text{hom}}}{2n_{\text{hom}}} \quad (4.2)$$

The toxicity index indicates whether there is no sex difference in toxicity (TI = 0), toxicity of the W or Y chromosome (TI > 0) or even toxicity of the Z or X chromosome (TI < 0). Consequently, we expect the toxicity index to be applicable not only to XY and ZW systems, but also XO systems.

Ethics. All research in Indonesia was carried out in compliance with the ethics guidelines supplied by the Research Center for Biology, Indonesian Institute of Sciences (RCB-LIPI), the Bogor Zoological Museum and the State Ministry of Research and Technology (RISTEK) under research permit numbers: 491/SIP/FRP/SMT/XI/2013 and 421/SIP/FRP/SMT/XI/2013. Data accessibility. All the data are publicly available, and the code is accessible at http://github.com/ValentinaBOP/Wefugium. All newly generated data were deposited in BioProject PRJNA64967.

Authors’ contributions. V.P. and A.S. designed the study, wrote and revised the subsequent drafts with input from all authors. V.P. analysed the data and wrote the first manuscript draft. K.A.J. and T.H. provided paradise crow samples. M.I. extracted paradise crow DNA. O.M.P.-G. extracted paradise crow RNA and helped with the analysis of repeat transcription and their differential expression. J.B. helped with repeat annotation. P.J. ran RepeatModel analyses. Q.Z. and J.L. provided the emu genome assembly. A.S. supervised the study. All authors read and approved the manuscript.

Competing interests. We declare we have no competing interests.

Funding. This research was supported by grants from the Swedish Research Council Formas (2017-01597 to A.S.; 2018-01008 to P.J.), the Swedish Research Council Vetenskaprådet (2016-05139 to A.S.; 2018-03017 to P.J.; 621-2014-5113 and 2019-03900 to M.I.), the SciLife Lab Swedish Biodiversity Program (2015-R14 to A.S.), Viltum Foundation (Young Investigator Programme, project no. 15560 to K.A.J.) and from the Carlsberg Foundation (Distinguished Associate Professor Fellowship, project no. CF17-0248 to K.A.J.). K.A.J. acknowledges a National Geographic Research and Exploration grant (8853-10), the Dybrøn Hoff’s Foundation and the Corrıt Foundation for financial support for fieldwork in Indonesia.

Acknowledgements. We thank Francisco Ruiz-Ruano for help with running WhatGen, Philipp Pottmier for help with RNA extractions, Alexander J. Charles for help with running MaQc-qR, and the Sib laboratory and Johannesson laboratory for helpful discussions. We thank Max Käller, Phil Ewels, Remi-André Olsen, Joel Gruselius and Fanny Taboraks-Lines for generating 10X data and assemblies at SciLifeLab Stockholm. We thank Erich Jarvis and the Vertebrate Genomes Project (VGP) for making their zebra finch, kakāpō and hummingbird assemblies available prior to publication, and Guojie Zhang and the 10 000 Bird Genomes (10K) project for doing the same with emu short read data. We thank Marco Ricci, Ivar Westerberg, Jesper Boman, Diem Nguyen and three anonymous reviewers for their comments on the manuscript. The Swedish Biodiversity Program has been made available by support from the Knut and Alice Wallenberg Foundation. Sequencing was performed by the SNPA&SEQ Technology Platform in Uppsala, which is part of the National Genomics Infrastructure (NGI) Sweden and Science for Life Laboratory, and by the National Genomics Infrastructure in Stockholm. Both facilities are funded by Science for Life Laboratory, the Knut and Alice Wallenberg Foundation and the Swedish Research Council. Computations were performed on resources provided by the Swedish National Infrastructure for Computing (SNIC) through Uppsala Multidisciplinary Center for Advanced Computational Science (UPPMAX). We thank the State Ministry of Research and Technology (RISTEK); the Ministry of Forestry, Republic of Indonesia; the Research Center for Biology, Indonesian Institute of Sciences (RCB-LIPI); the Bogor Zoological Museum for providing permits to carry out fieldwork in Indonesia and to export select samples; the Natural Resources and Conservation Agency (BKSDA) Maluku, Ministry of Environment and Forestry-Republic of Indonesia.

References


57. Ferree PM, Barbash DA. 2009 Species-specific heterochromatin prevents mitotic chromosome segregation to cause hybrid lethality in Drosophila.


