Ineffective silencing of transposable elements on an avian W Chromosome

Vera M Warmuth, Matthias H Weissensteiner and Jochen Wolf

Genome Res. published online February 11, 2022
Access the most recent version at doi:10.1101/gr.275465.121

Title: Accumulation and ineffective silencing of transposable elements on an avian W Chromosome

Vera M. Warmuth[1*], Matthias H. Weissensteiner[1,2], Jochen B. W. Wolf[1]

1 Division of Evolutionary Biology, Faculty of Biology, LMU Munich, Planegg-Martinsried, Germany

2 Department of Biology, Penn State University, University Park, PA, USA

*Author to whom correspondence should be addressed:

Vera M. Warmuth
LMU Biozentrum
Department Biologie II
Großhaderner Str. 2
82152 Planegg-Martinsried
Germany

Running title: Transposable element expression in birds

Keywords: Transcription, RNA-seq, sex bias, W chromosome, Transposable element, silencing, LTR

One of the defining features of transposable elements (TEs) is their ability to move to new locations in the host genome. To minimise the potentially deleterious effects of de novo TE insertions, hosts have evolved several mechanisms to control TE activity, including recombination-mediated removal and epigenetic silencing; however, increasing evidence suggests that silencing of TEs is often incomplete. The crow family experienced a recent radiation of LTR retrotransposons (LTRs), offering an opportunity to gain insight into the regulatory control of young, potentially still active TEs. We quantified the abundance of TE-derived transcripts across several tissues in 15 Eurasian crows (*Corvus* (*corone*) *spp.*) raised under common garden conditions and find evidence for ineffective TE suppression on the female-specific W Chromosome. Using RNA-seq data, we show that ~ 9.5% of all transcribed TEs had considerably greater (average: 16-fold) transcript abundance in female crows, and that more than 85% of these female-biased TEs originated on the W Chromosome. After accounting for differences in TE density among chromosomal classes, W-linked TEs were significantly more highly expressed than TEs residing on other chromosomes, consistent with ineffective silencing on the former. Together, our results suggest that the crow W Chromosome acts as a source of transcriptionally active TEs, with possible negative fitness consequences for female birds analogous to *Drosophila* (an X/Y system), where overexpression of Y-linked TEs is associated with male-specific aging and fitness loss ('toxic Y').

INTRODUCTION

Transposable elements (TEs) are mobile genetic elements that are characterised by their ability to move to new genomic locations (McClintock 1948; McClintock 1950). As sources of molecular variation, TEs can have beneficial effects on organismal fitness (Trizzino et al. 2017; Schrader and Schmitz 2019; Rutter et al. 2020; Sundaram and Wysocka 2020). However, TE insertions can also disrupt gene function or initiate chromosomal re-arrangements thereby reducing host fitness. In humans, for instance, more than 120 diseases are caused by TE insertions (Hancks and Kazazian 2016). Hosts can reduce the impact of deleterious TEs through removal or silencing. Removal of TEs by purifying selection is less effective when recombination rates are low, as tends to be the case in sex-limited chromosomes (Y or W; Dolgin and Charlesworth 2008). Removal efficacy is additionally affected by the effective population size of the target which is highest in autosomes, intermediate in sex-chromosomes with at least one copy in both sexes (X or Z), and lowest in the sex-limited chromosomes (Charlesworth and Charlesworth 1983; Charlesworth and Langley 1989). An alternative way of limiting the potentially negative effects of TEs is to suppress their activity (Deniz et al. 2019). In vertebrates, this primarily involves chemical modifications at TE loci ('epigenetic silencing') such as cytosine methylation (Deniz et al. 2019).

TE activity was long thought to be restricted to the early developing germline, where epigenetic repression is relaxed (Zamudio and Bourc'his 2010). However, in several taxa, ubiquitous TE activity has been shown to occur in healthy adult tissue, suggesting that TE silencing is often incomplete (Esteve-Codina et al. 2011; Dong et al. 2017; Brunet et al. 2018). It has been suggested that incomplete TE silencing may reflect a pervasive trade-off between epigenetic TE silencing and host genome function, as the former can interfere with the expression of nearby genes (reviewed in Choi and Lee 2020). If occurring on sex-limited chromosomes, such trade-offs have important consequences for the heterogametic sex. For example, in *Drosophila* (an X/Y system), incomplete suppression of TEs on the Y chromosome is thought to increase the mutational burden in males ('toxic' Y) (Wei et al. 2020), potentially contributing to male-specific ageing (Brown et al. 2020). In X/Y systems, TEs on the sex-limited chromosome thus determine general sex-specific differences. In female-heterogametic Z/W systems, such as birds, the sex-limited W Chromosome often shows exceptionally high TE densities (Smeds et al. 2015; Warren et al. 2016; Bertocchi et al. 2018; Peona et al. 2021). It is conceivable that W-linked TEs have broad sex-specific (fitness) effects in Z/W systems analogous to X/Y systems. However, whether the avian W Chromosome acts as a toxic reservoir of transcriptionally active TE copies remains poorly understood, with only one recent study reporting evidence for expression of female-specific,

likely W-linked, endogenous retroviruses in gonadal tissue of emu, chicken, and zebra finch (Peona et al. 2021).

Avian genomes harbour comparatively few TEs (~4 to 10%, compared to up to 60% in other tetrapod vertebrates) (Zhang et al. 2014; Gao et al. 2017; Sotero-Caio et al. 2017). However, the large clade of passerine birds appears to have experienced a recent diversification of LTR retrotransposons (Kapusta and Suh 2017; Suh et al. 2018). Two observations point to a particularly high activity of LTRs in the crow family. First, in a genome-wide comparison of 48 bird species, the American crow (*Corvus brachyrhynchos*) had more than twice as many copies of endogenous retroviruses (ERVs, an LTR superfamily) than all the other species (1,032 compared to an average 335; Cui et al. 2014). Second, an in-depth annotation of the first version of the hooded crow (*Corvus* (*corone*) *cornix*) assembly revealed a high diversity of lineage-specific, evolutionarily young LTRs in this species (Vijay et al. 2016). Young TEs are particularly likely to still be intact and therefore capable of transposition, offering the possibility to gain insight into their regulatory control.

In this study, we exploit the high contiguity of the current hooded crow genome, and added newly assembled W chromosomal sequence to quantify the genomic abundance and transcription of TEs in 15 Eurasian crows raised under common garden conditions.

RESULTS

W chromosomal sequence

To identify W Chromosome-linked sequences in the hooded crow genome, we generated a *de novo* genome assembly for a female individual using Oxford Nanopore Technologies (ONT) long-read sequencing (1.12 Gb total size and 9.3 Mb contig N50 after filtering, see Methods). We then combined a sex-specific coverage-based approach with synteny information from the W Chromosome of the closely related New Caledonian crow (*C. moneduloides*) to assign to the W Chromosome. The resulting 101 contigs had a total sequence length of 12.3 Mb and were ordered by synteny to the New Caledonian crow W Chromosome (21.5 Mb), the only other corvid W Chromosome to date. This W assembly was added to the hooded crow reference genome (GCA_000738735.5) and was used for all subsequent analyses.

Recent expansion of LTR retrotransposons on the W

RepeatMasker identified a total of 189,169 distinct repeat sequences (**Supplemental Table S1**). Of these, 187,409 were bona fide TEs (Class I or Class II), with the remainder

consisting of simple repeats, low complexity repeats, satellite DNA, rRNA and tRNA (**Table 1**). Repeat content, expressed as the proportion of genomic bases belonging to repeat sequence, varied among chromosomes type and was lowest on autosomal sequence (~6.5%), intermediate on the Z Chromosome (~11.9%), and highest by far on the W Chromosome (~84.8%). Genome-wide, the vast majority of TE copies were LINE elements (72.4%), followed by LTR retrotransposons (14.6 %), DNA transposons (6.5%) and SINE elements (3.3%) (**Table 1, Supplemental Table S1**), which is consistent with previous findings in other avian taxa (Warren et al. 2016; Kapusta and Suh 2017; Suh et al. 2018). LTR retrotransposons occupied a similar amount of genomic sequence as LINE elements on the autosomes and the Z Chromosome, but were by far the most dominant type of TE on the W Chromosome (**Fig. 1A, Supplemental Fig. S1**).

Transcription of LINES and young LTR retrotransposons

Of the 187,409 bona fide TE copies annotated in the crow reference genome, 13,940 (7.4%) were transcribed in our population sample as per our definition (≥ 0.5 transcripts per million in at least four of the 15 samples, **Supplemental Table S2**). This constitutes more than half of all transcribed genomic features (26,714) in our dataset. The proportions of autosomal, Z-, and W-chromosomal TEs that were transcribed were ~ 6.3%, 6.9%, and 23.4%, respectively (**Fig. 1B**). Despite its relatively small size, the W Chromosome was therefore not only the most TE-rich (**Fig. 1A**), but also had the highest proportion of transcriptionally active TEs (**Fig. 1B**). On the autosomes and the Z Chromosome, LINEs were the most frequently expressed TE type; by contrast, on the W Chromosome, LTRs were both the most abundant and the most frequently expressed (**Supplemental Fig. S2**). A number of autosomes also contained regions with a high density of transcribed TEs (e.g., Chromosome 13, **Supplemental Fig. S2**); however, these were often located near sub-telomeric regions, and the maximum density of transcribed TEs on any autosome was lower than the average density of transcribed TEs on the W Chromosome.

Assigning all annotated TE copies to bins of increasing divergence from their respective, family-level, consensus sequences (Kapusta et al. 2017) shows that, across all chromosome types, a large proportion of LTRs are young (0-1% divergence, **Fig. 2A**), whereas LINEs are predominantly older (> 10% divergence; **Fig. 2A**). On all three chromosome types, older LTRs (> 10% divergence) were rarely transcribed, whereas transcription of LINEs appeared to have no such age restriction (**Fig. 2B**).

Influence of tissue and sex on TE expression

We used Tau as an indicator of expression specificity (Yanai et al. 2005). The Tau index ranges between 0 and 1, where 0 indicates broad expression, and 1 indicates expression

specific to one tissue (Kryuchkova-Mostacci and Robinson-Rechavi 2017). Of all 13,940 transcribed TEs, 6,641 were expressed only in one tissue (Tau=1). A further 2,192 TEs were highly tissue specific (Tau ≧ 0.8), but were expressed in more than one tissue. Instances of TEs with Tau values ≧ 0.8 were recorded 487 times in liver, 3,625 in spleen, 2,459 in testis, and 2,276 in ovary (**Supplemental Table S3**).

To decompose the variance of transcript abundance across tissues, we performed principal components analyses (PCA) on the normalized and variance-stabilized counts of reads mapped to annotated genes and TEs, respectively. As expected, PCA of genes showed a clear partitioning of variance by tissue, with the first two PC axes accounting for 92 % of the variation (**Fig. 3A**). Except for gonadal tissue (ovary, testis), sex contributed only marginally to the overall variation in genes (**Fig. 3A, Supplemental Fig. S3)**. By contrast, TE transcript counts showed a clear partitioning of variance among the sexes for all tissues, with the first two PC axes explaining 72% of the variation (**Fig. 3B**). This is consistent with a signal not only of tissue-specificity, but also of sex-specificity. Neither genes nor TEs showed any clear separation of *corone* and *cornix* sub-species (see Supplemental Text for taxonomic considerations), consistent with previous findings of extremely low expression divergence between these two taxa (Poelstra et al. 2015).

To further explore the extent of taxon and sex-bias in TE expression, we carried out a differential expression (DE) analysis using DESeq2 with a linear model including terms for sub-species, tissue, and sex. The repetitive nature of TE elements means that an unknown number of short reads may map to multiple locations in the genome, thereby biasing estimates of TE transcript abundance. Using only uniquely mapping reads increases the rate of true positives but underestimates the signal of transcription associated with younger TE families (Teissandier et al. 2019). One solution is to include multi-mapping reads in analyses of TE expression but to randomly report only one position (Teissandier et al. 2019). To guard against mis-interpretation of the data owing to the choice of one or the other mapping method, we employed both mapping strategies (see Methods).

In both random and unique mapping mode, only six TEs (four CR1 LINEs and two LTRs) were differentially expressed between sub-species (**Supplemental Table S4**). This observation is consistent with the observed lack of separation by sub-species in our principal components analyses as well as previous gene expression studies in this system (Poelstra et al. 2014; Poelstra et al. 2015).

Contrary to taxon, sex had a strong effect on TE expression. Almost 10% of all 13,940 transcribed TEs showed differential expression between the sexes (random mode: 1,328; unique mode: 1,247; **Supplemental Table S5**). Nearly all of these differentially expressed TEs (DETEs) showed higher transcript abundance in females compared to males (random

mode: 1,285 or 96.8%; unique mode: 1,193 or 95.6%). The vast majority of female-biased DETEs originated on the W Chromosome (unique mode: 1,135 or 91.0%) or had a copy on the W Chromosome indistinguishable from copies elsewhere in the genome (random mode: 1,136 or 88.4%) (Fig. 4A, B). Autosomes and the Z Chromosome also contributed to female-biased TE expression, but to a much lower extent: in random mode, 146 DETEs mapped to the autosomes, and 46 to the Z Chromosome (**Fig. 4A;** Supplemental Table S5). In unique mode, the numbers were similar, with 110 DETEs mapping to the autosomes, and 51 to the Z Chromosome (**Fig. 4B**; **Supplemental Table S5**). Even in unique mode, a handful of male DNA-seq reads mapped to the W Chromosome (**Fig. 4B**; **Supplemental Table S5**). This likely reflects the presence of TE copies in the male genome that are highly similar to one or more W-linked TEs.

Ineffective silencing of W-linked TEs

We were interested to assess, whether overexpression of W-linked TEs was independent of both ploidy (only one copy of the Z in females; no copy of the W in males) and TE abundance (highest relative TE abundance on the W Chromosome). We tested the effect of ploidy on TE expression using a linear model, in which we asked whether TE transcript abundance (RNA-seq) was higher in females than in males after accounting for TE abundance (DNA-seq) (**Supplemental Methods**). We found that TE expression (RNA-seq counts) was significantly higher in females than in males across all chromosomal classes and tissues, including from TEs residing on the Z Chromosome (**Supplemental Fig. S4**). This suggests that TE transcript abundance is generally higher in females, and that TE copies on the W Chromosome are driving this pattern.

Next, we asked whether TE transcription from the W Chromosome was elevated above and beyond what is expected based on the high abundance, particularly of (young) LTRs, on the W Chromosome compared to autosomes and the Z Chromosome (**Fig. 1B, 2B**). To examine this question, we limited our considerations to females, where direct comparisons of abundance *vs*. expression are possible for all three chromosomal classes (A, Z, W) (**Supplemental Methods**). Specifically, we asked whether transcript levels from W-linked TEs exceeded expectations based on their abundance on this chromosome. We conducted this analysis separately for LTR elements, of which only young copies were expressed (**Fig. 2B**), and for the substantially older LINE elements. We found evidence for upregulation of W-chromosomal, but not autosomal or Z-chromosomal, TEs of both types in liver and spleen. In gonads, where sexual selection, particularly on the Z Chromosome, is expected to confound general patterns, this effect was not observed (**Supplemental Methods**). Overall, these

results suggest that female genomes, and in particular the W Chromosome, constitutes a permissive environment for TE transcription.

Mixed effect of TEs on the expression of neighbouring genes

TE insertions can disrupt coding sequence or modify the transcriptional activity of neighbouring genes. To assess whether TE insertions in or near coding sequence are selected against, we first determined the number of TEs that reside within 10 kb of annotated protein coding genes. Of the 187,409 bona fide TEs in the hooded crow genome, 370 (0.2%) overlapped with sequence of annotated genes (**Supplemental Table S6**), and a further 31,789 (17.0%) were found within 10kb upstream or downstream of a gene. Of the 370 genes containing TEs, 38 contained more than one (up to six) TE(s), though some of those instances might represent nested TEs rather than individual copies.

To assess if the vicinity of TEs influenced gene expression, we compared the normalised read counts of genes containing TE sequence 0-2 kb, 2-10 kb, and >20 kb away from transcription units. There was no statistically significant difference in expression levels of genes with TEs in their immediate vicinity (0-2 kb; median = 0.05, N= 10788) and genes with TEs > 20 kb away (median = 0.09, N=888) (Mann–Whitney U, $p > 0.05$, two-tailed). By contrast, genes with TEs located 2-10 kb away (N=13568) had significantly lower expression levels (median=0.02) than genes in the other two distance bins (Mann–Whitney U, $p < 0.0001$, two-tailed).

Next, we assessed whether sex bias in TE expression might be driven by sex-biased expression of nearby genes. For example, the autosomal and Z-chromosomal components of female bias in TE expression may either be due to ineffective TE silencing in females or result from co-expression between TEs and neighbouring, female-biased genes. Of the sex-specific DETEs identified in unique or random mode, respectively, only three were located within the transcription unit of genes (**Supplemental Table S7**), and none of them were differentially expressed between males and females. A further 44 genes were located within 10 kb of a sex-biased DETE (**Supplemental Table S7**); of these, only one was itself differentially expressed between the sexes (*C17orf58* homolog). Assuming equal power for inferring sex-bias in protein coding genes and TEs, we infer that sex-bias in TE expression is independent of sex-bias in protein coding genes in our study system.

DISCUSSION

In this study, we characterized the genomic and transcriptomic repertoire of repetitive elements in a population sample of an avian lineage experiencing recent TE expansion.

Inferences are based on short-read sequencing data from DNA and RNA mapped to a reference genome. This approach comes with the limitation that many TEs segregating at low frequencies in single or few individuals will go undetected (Weissensteiner et al. 2020). Despite this limitation, our work revealed strong tissue, sex and chromosome effects on TE expression.

Repeat content

We report an average repeat content of ~ 6.5% for the autosomes of the hooded crow genome. This figure falls toward the higher end of the range observed in birds (~4 to 10%; Zhang et al. 2014), but corresponds closely with the repeat content of the American crow (*C. brachyrhynchos*) genome (7.4%; Zhang et al. 2014). Consistent with theoretical predictions (Charlesworth 1991), and as shown for sex-limited chromosomes in other taxa (Śliwińska et al. 2016), TE content was highest (84.8%) on the W Chromosome. Higher tolerance of TE insertions on the W likely results from a combination of reduced effective population size, absence of recombination, and low gene density (3/6 genes per Mb on the W vs. 12/14 on the Z, and an average of 26/39 on autosomes in *Corvus* (*corone*) *spp*. / *Corvus moneduloides* (Weissensteiner et al. 2020). Smeds et al (2015) reported a TE content of ~ 48.5% across 6.9 Mb of the non-recombining part of the collared flycatcher W Chromosome, a short-read based assembly. Repetitive DNA, including transposable elements, are often under-represented in short-read assemblies (Peona et al. 2018). Higher levels of missing data in the short-read based flycatcher genome compared to the long-read-based crow genome examined here could explain the higher TE content found in crows; however, species-specific differences in the density of TEs on avian W chromosomes could also explain the difference: in six recently analysed, reference-quality, avian W assemblies, TE densities ranged between 22 and 80% (Kapusta and Suh 2017; Peona et al. 2021).

A comparison of the dominant TE classes on the different chromosomes revealed an overabundance of CR1 LINE elements on the autosomes and the Z Chromosome, and of LTR retrotransposons on the W Chromosome, a pattern previously also seen in flycatchers (Suh et al. 2018). The observation that ~ 85 % of the crow W Chromosome consisted of TEs, and that most of the W-based TEs were young LTR retrotransposons, lends indirect support to the hypothesis that the avian W Chromosome may act as a refugium for transcriptionally active TEs (Smeds et al. 2015; Warren et al. 2016; Suh et al. 2018; Peona et al. 2021).

TE distribution

In the hooded crow genome, only ~ 0.2% of TEs reside within transcriptional units of coding genes. This figure increases to ~20% if we include sequence 10 kb up-and downstream of

coding genes. Overall, these figures correspond closely with those reported for other avian genomes. For example, in zebra finch, 16% of all TEs are reported to reside within 10 kb of either end of transcription units; similarly, in the chicken genome, about 25% of endogenous retroviruses (ERVs) were either within transcriptional units or within 10 kb of each end (Bolisetty et al. 2012).

TE activity

Uncontrolled TE activity can interrupt functional sequence and result in deleterious, large-scale genome re-arrangements. TE activity in healthy organisms has therefore long been assumed to be restricted to the early developing germline, embryonic tissue, and the placenta (Slotkin and Martienssen 2007). However, it is becoming increasingly clear that TE regulation is often incomplete, with important implications for host genome stability and fitness (Hollister and Gaut 2009; Brown et al. 2020; Wei et al. 2020). Using a third-generation reference genome and two different pipelines to quantify TE transcript abundance, we show that ~ 10% of the annotated TEs in the crow genome escape silencing.

The successful transposition of TEs depends on the full transcription of a number of TE-associated genes. The transposition of LTR retrotransposons, in particular, requires a full length, sense-strand transcript including the proteins encoded by the gag and pol genes. Over time, mutations in the host genome will lead to a deterioration of TE sequence and a corresponding decrease in activity. Transcribed LTR retrotransposons in our system fall predominantly into young age bins (<8% divergence from the consensus sequence), supporting the expectation whereby transcribed TEs are mostly young. CR1 LINEs did not follow this pattern, showing evidence for transcription of far older elements instead.

TE expression was highly tissue-specific, corroborating findings from studies in other vertebrates. In humans, for instance, multiple studies have reported tissue-specific expression of TEs, with a recent study showing that TE expression was as predictive of tissue groupings as gene expression (Chung et al. 2019). Similarly, in rats, separation by tissue/organ explained ~95% of the variance in TE expression (Dong et al. 2017). In birds, tissue-specific TE expression has also been reported for chicken (Bolisetty et al. 2012). In a study of the human TE transcriptome, tissue-specific TE expression co-varied with the expression of nearby genes and was due to the presence on TEs of binding sites for transcription factors that regulate expression in a given tissue (Trizzino et al. 2017). As we found no evidence for strong patterns of gene TE co-expression in our study system, tissue-specific TE expression may best be explained by trans-acting regulatory mechanisms.

A substantial number of TEs were expressed in the gonadal tissue of both sexes, consistent with TEs being generally expressed in gonads (Dechaud et al. 2019). Gonadal tissue

contains germline cells at various stages of their development, including the earliest stages, when epigenetic repression is relaxed (Zamudio and Bourc'his 2010). In this study, ovarian tissue had among the highest level of transcription, mirroring results in an XY system (rats), where testis tissue had among the highest levels of TE expression (Dong et al. 2017).

Sex-biased TE expression

Approximately 10% of transcribed TEs were differentially expressed between the sexes. Recent studies of TE expression in mammals (Dong et al. 2017; Trizzino et al. 2017) and birds (Peona et al. 2021) have also reported sex-differences in TE expression. In contrast to our finding, however, neither study observed a pattern of consistent overexpression in only one of the sexes across several tissues, and the number of expressed TEs with a detectable sex bias was small. In rats, for example, only 26 TEs were found to be differentially expressed between the sexes, compared to more than 1,000 in this study. As in our study, most (84.6%) of the TEs showing sex-biased expression in rats were LTR retrotransposons. Due to the way in which they propagate, LTR retrotransposons are thought to have a greater potential for deleterious effects on host fitness than other TE classes (Faulkner et al. 2009). Elevated activity of W-based LTRs in female crows might therefore increase the mutational load and negatively affect female fitness analogous to the negative fitness effects of Y-linked TEs observed in *Drosophila* males (Brown et al. 2020).

High transcript abundance of TEs in females may simply reflect excess abundance of young, transcriptionally active TE copies on the female W Chromosome. However, we found TE transcription to be elevated even when controlling for TE abundance and TE age effects: The probability of a TE being expressed is elevated across the entire genome in females and particularly so on the W Chromosome. This finding suggests trans-acting effects on TE expression that are likely driven by the W Chromosome. An analogous pattern of male-biased expression of Y-linked TEs has recently been described in two *Drosophila* species, *D. pseudoobscura* and *D. miranda.* As in our study, the observed sex-bias in TE expression was largely due to elevated expression of Y-linked TEs in the heterogametic sex (males in *Drosophila*; females in crows) (Wei et al. 2020), and while it was confined to early developmental stages in *D. pseudoobscura*, it persisted into adulthood in *D. miranda*. In *Drosophila*, the availability of a well-annotated Y Chromosome proved essential in identifying the conflict between transcription of functional Y-linked genes and suppression of Y-linked TEs as a potential reason for the observed incomplete silencing of Y-linked TEs: euchromatin, required for transcription, antagonises the formation of heterochromatin, a major mechanism of TE silencing. The avian W Chromosome contains dozens of functional genes (Xu and Zhou 2020). It is therefore conceivable, that a similar mechanism might

372 explain the incomplete silencing of W-based TEs in female birds. Our study contributes to
373 accumulating evidence whereby sex-limited chromosomes in general, and the W
374 Chromosome in particular, may have roles beyond sex determination and gonadal
375 development, and that Y-/ W-linked repeats, and polymorphisms thereof, can have genome-
376 wide epistatic effects (Lemos et al. 2010; Kutch and Fedorka 2017).

377 The trans-regulating capabilities of transposable elements are known from several systems,
378 where TE-derived RNAs regulate either distant genes or the activity of TEs themselves
379 (Piergentili 2010; Kawaoka et al. 2011; McCue and Slotkin 2012; Wang et al. 2017; Cho
380 2018). Trans-regulatory activity of W-based TEs in the crow would explain our observation of
381 a female bias in TE expression generally, including hundreds of autosomal and even Z-
382 chromosomal TEs. This result may thus be a first indication of a tentative regulatory role of
383 W-linked repeats in birds.

384 Overall, our work provides evidence for a female heterogametic analogue of the toxic Y in
385 X/Y systems with important implications for our understanding of transcriptional control of
386 TEs, TE induced fitness effects, and long-term TE propagation.

387

388 ## Methodological considerations

389 For most protein coding genes, each individual contains at least a single gene copy. In the
390 absence of mRNA molecule counts, we can conclude with some confidence that this gene is
391 not expressed. More generally, we expect a direct relationship between mRNA abundance
392 and transcription activity (except for a minority of recently duplicated genes). By contrast,
393 TEs segregate in a population with strongly skewed site frequency spectra. As a
394 consequence, the majority of TE copies is unique to a single or few individuals. Only a small
395 minority of copies will be shared between all individuals (Weissensteiner et al. 2020). As a
396 consequence, lack of an mRNA signal may either result from lack of expression or from the
397 absence of a syntenic copy in the reference genome. Using a single genome as mapping
398 reference does not allow us to distinguish between these possibilities. Moreover, RNA-seq
399 reads from a copy that is absent in the reference genome runs an increased risk of
400 multimapping to similar copies elsewhere in the genome.

401 This population genetic reality has consequences for the interpretation of our results. First,
402 the reference genome will only represent a fraction of the TE inserts present in the
403 population sample. Our estimate that 7.4% of all TEs are transcribed therefore likely
404 constitutes an underestimate. Second, multimapping will be increased beyond levels already
405 prevalent with short-read sequencing data. This expectation motivated us to use different
406 mapping approaches (unique vs. random mode, see Methods) to assess the robustness of
407 the results. Furthermore, statistical inference of factors modulating transcription activity

required the inclusion of a DNA-seq reference experiencing the same reference bias. Third, results exploring the effect of TEs on expression of nearby genes are to be interpreted with caution.

To circumvent the abovementioned reference bias for insertion-deletion polymorphisms in future studies of TE activity will require a data set combining transcriptome data with haplotype-resolved *de novo* assemblies for all individuals under investigation. Both data types would strongly benefit from the use of long-read data to assure reliable inference of structural variation (Tusso et al. 2019) and minimize the effect of multimapping. This 'pangenome' approach has become a reality for the study of TEs in small eukaryotic genomes (Tusso et al. 2021) and is starting to be within reach for vertebrate-size genomes (Weissensteiner et al. 2020; De Coster et al. 2021).

METHODS

Sampling and data generation

In May 2014, crow hatchlings of an approximate age of 21 days were obtained directly from the nest using an unmanned aerial vehicle to assess nest status (Weissensteiner et al. 2015). Hooded crows (*C.* (*corone*) *cornix*) were sampled in the area around Uppsala, Sweden (59°52'N, 17°38'E), and carrion crows (*C.* (*corone*) *corone*)) in the area around Konstanz, Germany (47°45N', 9°10'E) (**Supplemental Table S7**). To avoid any confounding effects of relatedness, only a single individual was selected from each nest. After transfer of carrion crows to Sweden by airplane, all crows were hand-raised indoors at Tovetorp field station, Sweden (58°56'55"N, 17°8'49"E). When starting to feed by themselves they were released to large roofed outdoor enclosures (6.5 x 4.8 x 3.5 m), specifically constructed for the purpose. All crows were maintained under common garden conditions in groups of a maximum of six individuals separated by sub-species and sex. In October 2016, at an age of approximately 2.5 years, individuals were euthanized by cervical dislocation. Tissues were immediately harvested and stored at -80°C until extraction.

Regierungspräsidium Freiburg granted permission for the sampling of wild carrion crows in Germany (Aktenzeichen 55- 8852.15/05). Import into Sweden was registered with the *Veterinäramt Konstanz* (Bescheinigungsnummer INTRA.DE.2014.0047502) and its Swedish counterpart *Jordbruksverket* (Diarienummer 6.6.18-3037/14). Sampling permission in Sweden was granted by *Naturvårdsverket* (Dnr: NV-03432-14) and *Jordbruksverket* (Diarienummer 27-14). Animal husbandry and experimentation was authorised by *Jordbruksverket* (Diarienummer 5.2.18-3065/13, Diarienummer 27-14) and ethically

approved under the Directive 2010/63/EU on the Protection of Animals used for Scientific Purposes by the *European Research Council* (ERCStG-336536).

DNA and RNA Extraction and Sequencing

RNA was extracted from liver, spleen, and gonadal tissue of seven *C. (c.) corone* (*N*=4 males, and *N* = 3 females) and eight *C. (c.) cornix* (*N*=3 males, and *N* = 5 females), respectively, using the RNeasy Plus Universal extraction kit (Qiagen) and following the manufacturer's guidelines. Libraries were prepared from 500 ng total RNA using the TruSeq stranded mRNA library preparation kit (Illumina Inc), which includes removal of ribosomal RNA using poly(A) selection. Sequencing of 50 bp, single-end reads was performed by the SNP&SEQ Technology Platform at Uppsala University, Sweden, using Illumina HiSeq 2500 v. 2 chemistry and a target sequencing depth of 20x.

High-molecular weight DNA was extracted from a female *C. (c.) cornix* individual (NCBI BioSample SAMN13509843), sampled from the same sub-population as the (male) individual that was used for the reference genome. Four flow cells of the Oxford Nanopore Technologies (ONT) minION sequencer yielded 29.55 Gb of raw data (read N50: 13.47 kb), and one flow cell of the ONT promethION sequencer yielded 38.67 Gb of raw data (read N50: 13.11 kb).

Identification and assembly of W-linked contigs

We then assembled individual long sequence reads with *flye* assembly algorithm (Kolmogorov et al. 2019). The initial whole genome assembly consisted of 8,875 contigs with an average size of 128 kb (median: 3.2 kb) and a contig N50 of 9.21 Mb. For further downstream analysis, we restricted the minimum contig length to 10 kb, leading to an exclusion of 6,778 contigs and a total assembly size of 1.117 Gb. We then used two complementary approaches to identify W-linked contigs: a synteny-based approach and a coverage-based approach. For the synteny-based approach, we chose the Vertebrate Genomes Project's (VGP) assembly of the New Caledonian crow (*Corvus moneduloides*) which includes among the best W Chromosome assemblies available for birds from the corvid family. We downloaded the *C. moneduloides* primary assembly from the VGP website (https://vgp.github.io/genomeark/Corvus_moneduloides/, accessed 15 November 2020), and aligned all contigs > 10 kb (2,197 out of 8,875) against the *C. moneduloides* W Chromosome using LASTZ (Harris 2007). This approach yielded 227 potentially W-linked contigs.

The coverage-based approach relies on the fact that the W Chromosome is female-limited. Sequencing reads from male crows should therefore not map against W-chromosomal sequence. To identify potential W-linked contigs, we mapped both Illumina short reads and Pacific Bioscience long reads of a male *C. (c.) cornix* from the Swedish population (BioSample SAMN02439830, accession number SRS602284) to the 8,875 contigs of our assembly using NGMLR (Sedlazeck et al. 2018). We limited the maximum number of segments per kb of reference a read can align to 3 (--max_segments = 3). We used mosdepth (Pedersen and Quinlan 2018) to calculate coverage for each contig. Excluding contigs with a mean coverage < 3, and retaining only contigs > 50 kb yielded 302 potentially W-linked contigs. After excluding all contigs that were identified by only one of the two approaches, our W assembly consisted of 101 contigs with a combined sequence of 12.27 Mb. Comparing these 101 with the previously identified pseudo-autosomal region (PAR) in the crow genome yielded no conclusive linear alignments (Catalán et al. 2021). We therefore conclude that the PAR is absent in our set of W-linked sequences. To further illustrate that the 101 contigs are of W Chromosome origin, we mapped both the female long-read sequencing data described above and Illumina short read sequencing data of four more hooded crows (two males and two females; SRA accession numbers ERR2900302, ERR2900303, ERR2900305, ERR2900306) to the hooded crow reference assembly and the W-linked contigs identified in this study. We then calculated read depth in 10,000 randomly subsampled 10-kb windows on chromosome 1 and compared it to the same number of subsampled windows on W-linked contigs, the Z Chromosome and the autosomes. **Supplemental Fig. S5** shows the relative read depths of all chromosomes and demonstrates that the W-linked contigs meet the expectation of half of autosome coverage in females and approaching zero in male individuals.

TE content

Pipelines for the discovery and annotation of TEs were run on the combined sequence of the (male) crow reference genome (NCBI accession number: GCA_000738735.5) and the newly generated assembly of W-linked contigs. We used three different pipelines for the discovery and annotation of TEs: RepeatMasker (http://www.repeatmasker.org), CARP (Zeng et al. 2018), and RetroTector (Sperber et al. 2007). CARP uses pairwise alignment and single linkage clustering to identify families of repeats, with a bias towards low-divergence, potentially still active, TEs (Zeng et al. 2018). RetroTector specializes in the identification and characterisation of retroviral sequences, and is potentially also biased towards the detection of evolutionarily younger elements (Sperber et al. 2007). Both RetroTector and CARP were run using default settings. Following the authors' recommendations, we used a score cut-off of 300 for the retroviral sequences identified by RetroTector. To annotate the (evolutionarily

young) TEs identified repeats identified by CARP, we used zebra finch consensus sequences available on Repbase (http://www.girinst.org/repbase/index.html; accessed 09 April, 2018), consensus sequences of TEs identified in the flycatcher genome (Suh et al. 2018), and consensus sequences of TEs identified in an earlier version of the hooded crow genome (Vijay et al. 2016). We then ran RepeatMasker on the combined genomic sequence using our TE library together with the inbuilt chicken database and under default settings.

TE age estimates

We used the degree of sequence divergence from the consensus sequence as a rough estimate of the age of TEs. Sequence divergence of TEs from their consensus was computed from the RepeatMasker output files (.align) following (Kapusta et al. 2017) and using the Perl script 'parseRM.pl' (https://github.com/4ureliek/Parsing-RepeatMasker-utputs/blob/master/). Estimates of sequence divergence obtained using this approach are sufficiently accurate for our purpose, which is to compare the broad age ranges of LINEs and LTRs across the different chromosomal classes.

Quantification of TE-derived transcripts

Following removal of Illumina adaptors and reads with a Phred quality score below 20 in Trim Galore v. 0.4.4 (https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/), RNA-seq reads were mapped against the merged sequence of the hooded crow reference genome and the W-linked contigs using STAR v2.5.2b (Dobin et al. 2013). We followed recent recommendations by (Teissandier et al. 2019) for the study of transposon expression and regulation (see also Supplemental Text, Section 2). Specifically, we quantified TE transcript abundance using the 'random' counting method, in which each multi-mapping read contributes one randomly chosen valid alignment (Teissandier et al. 2019). To assess how including *vs*. excluding multi-mapping reads affected the conclusions drawn from our analyses, we compared this approach with one that only considers uniquely mapping reads ('unique' mode; but see Supplemental Text, Section 2). For read mapping, we adjusted the flags in STAR as follows: --outFilterMultimapNmax 1000 --outSAMmultNmax 1 --outFilterMismatchNmax 3 --outMultimapperOrder Random --winAnchorMultimapNmax 1000 --alignEndsType EndToEnd --alignIntronMax 1 --alignMatesGapMax 350. We used the *featureCounts* function of the R (R Core Team, 2017) package Subread (Liao et al. 2019) for read counting, setting countMultiMappingReads to TRUE in random mode, and to FALSE in unique mode. For mapping summary statistics see **Supplemental Table S8**.

Counts were normalised using the Median of Ratios Normalisation (MRN) implemented in the R package DESeq2 (Love et al. 2014). We considered TEs to be transcribed if transcript abundance exceeded 0.5 transcripts per million in at least four samples. Tests for differential expression were carried out following the standard differential expression analysis steps of DESeq2 and using the design formula ~ tissue + sub-species + sex. Differentially expressed genomic features were identified at a $\log_2$ fold change threshold (LFC) > 1, corresponding to a greater than twofold expression difference between test groups, and a false discovery rate (FDR) ≤ 10%.

Quantification of TE abundance in the crow genome

We were interested to assess the genomic abundance of individual TEs to relate genomic TE abundance to abundance of transcribed TE copies. The number of normalised DNA-seq reads mapping to a particular TE is expected to be roughly proportional to the number of times this TE appears in the DNA sample, thus reflecting its genomic abundance (Magi et al. 2012). To compare RNA- and DNA-seq read abundance for the same TE, we normalised counts from both types of data using Median of Ratios Normalisation. While principally used for RNA-seq data, the Median of Ratios method has recently been shown to be suitable for the normalisation of DNA-seq data for the purpose of gene abundance estimation (Pereira et al. 2018).

In the absence of re-sequencing data for the 15 individuals used for the RNA-seq study, we used Illumina re-sequencing data of ten individuals (*N*=5 females and *N*=5 males) from an Italian crow population (NCBI Sequencing Read Archive accession number PRJEB9057; **Supplemental Table S7**). Assuming an excess of low frequency TE insertions segregating in the population (Bourgeois and Boissinot 2019), not all TEs identified in the reference genome will be present in non-reference individuals. As such, we will likely underestimate *absolute* abundances. However, because site-frequency spectra of crows sampled across Europe are near-identical (Vijay et al. 2016), the same bias will apply to all samples, therefore allowing *relative* comparisons between DNA-seq (Italian samples) and RNA-seq data (German, Swedish samples).

DNA-seq reads, initially 100bp long, were trimmed to a final length of 50bp, the same length as the RNA-seq reads. Reads were then mapped against our crow assembly (including the W-linked contigs) using NextGenMap v.0.5.4 (Sedlazeck et al. 2013) with a minimum identity i=0.88, corresponding to the minimum identity used for the RNA-seq data, and default values otherwise. Reads were counted using the function *featureCounts* in the R package subReads with countMultiMappingReads set to FALSE. Raw DNA-seq read counts were normalised using Median of Ratios Normalisation. While principally used for RNA-seq data,

581 the Median of Ratios method has recently been shown to be suitable for the normalisation of
582 DNA-seq data for the purpose of gene abundance estimation (Pereira et al. 2018).

583

584 ## Comparison of TE abundance and transcription

585 For each individual crow, read counts (DNA-seq for Italian crows, RNA-seq for
586 Swedish/German crows) were averaged across the DETEs identified in unique mode. A
587 linear model was fit to the log-transformed counts using sex (male or female), tissue (liver,
588 spleen or gonad) and nucleic acid type (RNA or DNA) as categorical explanatory variables
589 (formula: log(counts) ~ sex + tissue + type + sex * type). To test whether transcription of W-
590 linked TEs remains higher than that of autosomal and Z-linked TEs after accounting for
591 differences in TE abundance, we limited our considerations to the female dataset, where a
592 direct comparison of abundance *vs*. expression is possible between all three chromosomal
593 classes (A, Z, W). We asked whether TE transcript levels (RNA-seq) correspond to TE
594 abundance (DNA-seq) across all chromosome classes, or whether transcripts from W-linked
595 TEs are elevated (statistical interaction RNA-seq/DNA-seq x A/Z/W). To account for the
596 higher relative abundance specifically of (young) LTR elements on the W chromosome, we
597 conducted this analysis separately for LTR elements, of which only young copies were
598 expressed, and for the substantially older LINE elements. For each tissue separately, we
599 then fit a linear model to the female data using the formula: log(counts) ~ type +
600 chromosomal class + type * chromosomal class. All analyses were carried out in R version
601 4.0.0 (R Core Team, 2017).

602

603 ## Distance to transcription units

604 The distance of TEs from the nearest annotated transcription unit (coding sequence, 3' UTRs
605 and 5' UTRs) was determined using the function *closest* in BEDTools v2.29.2 (Quinlan and
606 Hall 2010), and the data were catalogued in 2 kb intervals. A more detailed binning of TE
607 locations with respect to different classes of coding sequence was not possible due to the
608 preliminary state of the annotation for the crow genome version used here.
609
610
611
612
613

DATA ACCESS

All raw sequencing data generated in this study and the newly assembled W-linked sequence have been submitted to the NCBI BioProject database (https://www.ncbi.nlm.nih.gov/bioproject/) under accession number PRJNA772570. All custom scripts used for the analysis of the data presented here are available at https://github.com/EvoBioWolf/2022_Warmuth_GenomeRes.

COMPETING INTEREST STATEMENT

The authors declare no conflict of interest.

ACKNOWLEDGMENTS

We express our gratitude to Sven Jakobsson for providing the infrastructure for animal husbandry at Tovetorp research station. We would also like to thank Christen Bossu for her contribution in obtaining samples. Thomas Giegold, Nils Andbjer, Tamara Volkmer, Barbara Martin Schitsch, Luisa Sontheimer and Joanna Schinner were of invaluable support in raising and maintaining the captive crow population. Martin Wikelski, Inge Müller and additional staff from the Max-Planck-Institute for Ornithology in Radolfzell facilitated sampling in Germany and transport to Sweden. We would further like to thank Dirk Metzler for advice on statistical analyses, Saurabh Pophaly for assistance with the MYSQL database required for RetroTector and Justin Meröndun for producing Suppl. Fig. 5. We acknowledge funding from European Research Council (ERCStG-336536 to JBWW) and LMU Munich (to JBWW) and financial contribution from Stockholm University to Tovetorp field station.

AUTHOR CONTRIBUTIONS

VW and JBWW conceived of the study and wrote the paper with input from MHW. VW conducted all analyses except for the W assembly, which was carried out by MHW. JBWW and MHW obtained crow hatchlings in the field; JBWW was responsible for the common garden experiment and sampled the tissue material.

REFERENCES

Bertocchi NA, de Oliveira TD, del Valle Garnero A, Coan RLB, Gunski RJ, Martins C, Torres FP. 2018. Distribution of CR1-like transposable element in woodpeckers (Aves Piciformes): Z sex chromosomes can act as a refuge for transposable elements. *Chromosome Res.* 26:333–343.

Bolisetty M, Blomberg J, Benachenhou F, Sperber G, Beemon K. 2012. Unexpected Diversity and Expression of Avian Endogenous Retroviruses. *mBio* 3. Available from: https://mbio.asm.org/content/3/5/e00344-12

Bourgeois Y, Boissinot S. 2019. On the Population Dynamics of Junk: A Review on the Population Genomics of Transposable Elements. *Genes* 10:419.

Brown EJ, Nguyen AH, Bachtrog D. 2020. The Y chromosome may contribute to sex-specific ageing in Drosophila. *Nat. Ecol. Evol.* 4:853–862.

Brunet F, Roche A, Chalopin D, Naville M, Klopp C, Vizziano-Cantonnet D, Volff J-N. 2018. Analysis of Transposable Elements Expressed in the Gonads of the Siberian Sturgeon. In: Williot P, Nonnotte G, Vizziano-Cantonnet D, Chebanov M, editors. The Siberian Sturgeon (Acipenser baerii, Brandt, 1869) Volume 1 - Biology. Cham: Springer International Publishing. p. 115–130. Available from: https://doi.org/10.1007/978-3-319-61664-3_7

Catalán A, Merondun J, Knief U, Wolf JBW. 2021. Epigenetic mechanisms of partial dosage compensation in an avian, female heterogametic system. Available from: https://www.biorxiv.org/content/10.1101/2021.08.17.456618v1

Charlesworth B. 1991. The evolution of sex chromosomes. *Science* 251:1030–1033.

Charlesworth B, Charlesworth D. 1983. The population dynamics of transposable elements. *Genet. Res.* 42:1–27.

Charlesworth B, Langley CH. 1989. The population genetics of Drosophila transposable elements. *Annu. Rev. Genet.* 23:251–287.

Cho J. 2018. Transposon-Derived Non-coding RNAs and Their Function in Plants. *Front. Plant Sci.* 9. Available from: https://www.frontiersin.org/articles/10.3389/fpls.2018.00600/full

Choi JY, Lee YCG. 2020. Double-edged sword: The evolutionary consequences of the epigenetic silencing of transposable elements. *PLOS Genet.* 16:e1008872.

Chung N, Jonaid GM, Quinton S, Ross A, Sexton CE, Alberto A, Clymer C, Churchill D, Navarro Leija O, Han MV. 2019. Transcriptome analyses of tumor-adjacent somatic tissues reveal genes co-expressed with transposable elements. *Mob. DNA* 10:39.

Cui J, Zhao W, Huang Z, Jarvis ED, Gilbert MTP, Walker PJ, Holmes EC, Zhang G. 2014. Low frequency of paleoviral infiltration across the avian phylogeny. *Genome Biol.* 15:539.

De Coster W, Weissensteiner MH, Sedlazeck FJ. 2021. Towards population-scale long-read sequencing. *Nat. Rev. Genet.* 22:572–587.

682 Dechaud C, Volff J-N, Schartl M, Naville M. 2019. Sex and the TEs: transposable elements
683 in sexual development and function in animals. *Mob. DNA* 10:42.

684 Deniz Ö, Frost JM, Branco MR. 2019. Regulation of transposable elements by DNA
685 modifications. *Nat. Rev. Genet.* 20:417–431.

686 Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M,
687 Gingeras TR. 2013. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* 29:15–
688 21.

689 Dolgin ES, Charlesworth B. 2008. The Effects of Recombination Rate on the Distribution and
690 Abundance of Transposable Elements. *Genetics* 178:2169–2177.

691 Dong Y, Huang Z, Kuang Q, Wen Z, Liu Z, Li Y, Yang Y, Li M. 2017. Expression dynamics
692 and relations with nearby genes of rat transposable elements across 11 organs, 4
693 developmental stages and both sexes. *BMC Genomics* 18:666.

694 Esteve-Codina A, Kofler R, Palmieri N, Bussotti G, Notredame C, Pérez-Enciso M. 2011.
695 Exploring the gonad transcriptome of two extreme male pigs with RNA-seq. *BMC*
696 *Genomics* 12:552.

697 Faulkner GJ, Kimura Y, Daub CO, Wani S, Plessy C, Irvine KM, Schroder K, Cloonan N,
698 Steptoe AL, Lassmann T, et al. 2009. The regulated retrotransposon transcriptome of
699 mammalian cells. *Nat. Genet.* 41:563–571.

700 Gao B, Wang S, Wang Y, Shen D, Xue S, Chen C, Cui H, Song C. 2017. Low diversity,
701 activity, and density of transposable elements in five avian genomes. *Funct. Integr.*
702 *Genomics* 17:427–439.

703 Hancks DC, Kazazian HH. 2016. Roles for retrotransposon insertions in human disease.
704 *Mob. DNA* 7:1–28.

705 Harris RS. 2007. Improved pairwise alignment of genomic DNA. PhD Dissertation. Available
706 from: https://etda.libraries.psu.edu/catalog/7971

707 Hollister JD, Gaut BS. 2009. Epigenetic silencing of transposable elements: A trade-off
708 between reduced transposition and deleterious effects on neighboring gene
709 expression. *Genome Res.* 19:1419–1428.

710 Kapusta A, Suh A. 2017. Evolution of bird genomes-a transposon's-eye view. *Ann. N. Y.*
711 *Acad. Sci.* 1389:164–185.

712 Kapusta A, Suh A, Feschotte C. 2017. Dynamics of genome size evolution in birds and
713 mammals. *Proc. Natl. Acad. Sci.* 114:E1460–E1469.

714 Kawaoka S, Kadota K, Arai Y, Suzuki Y, Fujii T, Abe H, Yasukochi Y, Mita K, Sugano S,
715 Shimizu K, et al. 2011. The silkworm W chromosome is a source of female-enriched
716 piRNAs. *RNA N. Y. N* 17:2144–2151.

717 Kojima KK. 2018. Human transposable elements in Repbase: genomic footprints from fish to
718 humans. *Mob. DNA* [Internet] 9.

719 Kolmogorov M, Yuan J, Lin Y, Pevzner PA. 2019. Assembly of long, error-prone reads using
720 repeat graphs. *Nat. Biotechnol.* 37:540–546.

721 Kryuchkova-Mostacci N, Robinson-Rechavi M. 2017. A benchmark of gene expression
722 tissue-specificity metrics. *Brief. Bioinform.* 18:205–214.

Kutch IC, Fedorka KM. 2017. A test for Y-linked additive and epistatic effects on surviving bacterial infections in Drosophila melanogaster. *J. Evol. Biol.* 30:1400–1408.

Lemos B, Branco AT, Hartl DL. 2010. Epigenetic effects of polymorphic Y chromosomes modulate chromatin components, immune response, and sexual conflict. *Proc. Natl. Acad. Sci.* 107:15826–15831.

Liao Y, Smyth GK, Shi W. 2019. The R package Rsubread is easier, faster, cheaper and better for alignment and quantification of RNA sequencing reads. *Nucleic Acids Res.* 47:e47–e47.

Love MI, Huber W, Anders S. 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* 15:550.

Magi A, Tattini L, Pippucci T, Torricelli F, Benelli M. 2012. Read count approach for DNA copy number variants detection. *Bioinformatics* 28:470–478.

McClintock B. 1948. Mutable loci in maize. In: Annual Report of the Director of the Department of Genetics, Carnegie Institution of Washington Year Book No. 47, 1947-1948. Cold Spring Harbor, New York: Carnegie Institution of Washington. p. 155–169.

McClintock B. 1950. The origin and behavior of mutable loci in maize. *Proc. Natl. Acad. Sci.* 36:344–355.

McCue AD, Slotkin RK. 2012. Transposable element small RNAs as regulators of gene expression. *Trends Genet. TIG* 28:616–623.

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

Peona V, Palacios-Gimenez OM, Blommaert J, Liu J, Haryoko T, Jønsson KA, Irestedt M, Zhou Q, Jern P, Suh A. 2021. The avian W chromosome is a refugium for endogenous retroviruses with likely effects on female-biased mutational load and genetic incompatibilities. *Philos. Trans. R. Soc. B* [Internet]. Available from: http://royalsocietypublishing.org/doi/abs/10.1098/rstb.2020.0186

Peona V, Weissensteiner MH, Suh A. 2018. How complete are “complete” genome assemblies?—An avian perspective. *Mol. Ecol. Resour.* 18:1188–1195.

Pereira MB, Wallroth M, Jonsson V, Kristiansson E. 2018. Comparison of normalization methods for the analysis of metagenomic gene abundance data. *BMC Genomics* 19:274.

Piergentili R. 2010. Multiple roles of the Y chromosome in the biology of Drosophila melanogaster. *ScientificWorldJournal* 10:1749–1767.

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* 344:1410–1414.

Poelstra JW, Vijay N, Hoeppner MP, Wolf JBW. 2015. Transcriptomics of colour patterning and coloration shifts in crows. *Mol. Ecol.* 24:4617–4628.

Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* 26:841–842.

764 R Core Team (2017). R: A language and environment for statistical computing. R Foundation
765 for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/

766 Rutter MT, Bisner AM, Kohler C, Morgan K, Musselman O, Pickel J, Tan J, Yamasaki Y,
767 Willson J, Callahan HS, et al. 2020. Disrupting the disruptors: the consequences of
768 mutations in mobile elements for ecologically important life history traits. *Evol. Ecol.*
769 34:363–377.

770 Schrader L, Schmitz J. 2019. The impact of transposable elements in adaptive evolution.
771 *Mol. Ecol.* 28:1537–1549.

772 Sedlazeck FJ, Rescheneder P, von Haeseler A. 2013. NextGenMap: fast and accurate read
773 mapping in highly polymorphic genomes. *Bioinforma. Oxf. Engl.* 29:2790–2791.

774 Sedlazeck FJ, Rescheneder P, Smolka M, Fang H, Nattestad M, von Haeseler A, Schatz
775 MC. 2018. Accurate detection of complex structural variations using single molecule
776 sequencing. *Nat. Methods* 15:461–468.

777 Śliwińska EB, Martyka R, Tryjanowski P. 2016. Evolutionary interaction between W/Y
778 chromosome and transposable elements. *Genetica* 144:267–278.

779 Slotkin RK, Martienssen R. 2007. Transposable elements and the epigenetic regulation of
780 the genome. *Nat. Rev. Genet.* 8:272–285.

781 Smeds L, Warmuth V, Bolivar P, Uebbing S, Burri R, Suh A, Nater A, Bureš S, Garamszegi
782 LZ, Hogner S, et al. 2015. Evolutionary analysis of the female-specific avian W
783 chromosome. *Nat. Commun.* 6:1–10.

784 Sotero-Caio CG, Platt RN, Suh A, Ray DA. 2017. Evolution and Diversity of Transposable
785 Elements in Vertebrate Genomes. *Genome Biol. Evol.* 9:161–177.

786 Sperber GO, Airola T, Jern P, Blomberg J. 2007. Automated recognition of retroviral
787 sequences in genomic data—RetroTector©. *Nucleic Acids Res.* 35:4964–4976.

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

791 Sundaram V, Wysocka J. 2020. Transposable elements as a potent source of diverse cis-
792 regulatory sequences in mammalian genomes. *Philos. Trans. R. Soc. B Biol. Sci.*
793 375:20190347.

794 Teissandier A, Servant N, Barillot E, Bourc'his D. 2019. Tools and best practices for
795 retrotransposon analysis using high-throughput sequencing data. *Mob. DNA* 10:52.

796 Trizzino M, Park Y, Holsbach-Beltrame M, Aracena K, Mika K, Caliskan M, Perry GH, Lynch
797 VJ, Brown CD. 2017. Transposable elements are the primary source of novelty in
798 primate gene regulation. *Genome Res.* 27:1623–1633.

799 Tusso S, Nieuwenhuis BPS, Sedlazeck FJ, Davey JW, Jeffares DC, Wolf JBW. 2019.
800 Ancestral Admixture Is the Main Determinant of Global Biodiversity in Fission Yeast.
801 *Mol. Biol. Evol.* 36:1975–1989.

802 Tusso S, Suo F, Liang Y, Du L-L, Wolf JBW. 2021. Reactivation of transposable elements
803 following hybridization in fission yeast. *Genome Res*. doi: 10.1101/gr.276056.121.

804 Vijay N, Bossu CM, Poelstra JW, Weissensteiner MH, Suh A, Kryukov AP, Wolf JBW. 2016.
805 Evolution of heterogeneous genome differentiation across multiple contact zones in a
806 crow species complex. *Nat. Commun.* 7:1–10.

807 Wang L, Rishishwar L, Mariño-Ramírez L, Jordan IK. 2017. Human population-specific gene
808 expression and transcriptional network modification with polymorphic transposable
809 elements. *Nucleic Acids Res.* 45:2318–2328.

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

813 Wei KH-C, Gibilisco L, Bachtrog D. 2020. Epigenetic conflict on a degenerating Y
814 chromosome increases mutational burden in Drosophila males. *Nat. Commun.*
815 11:5537.

816 Weissensteiner MH, Bunikis I, Catalán A, Francoijs K-J, Knief U, Heim W, Peona V, Pophaly
817 SD, Sedlazeck FJ, Suh A, et al. 2020. Discovery and population genomics of
818 structural variation in a songbird genus. *Nat. Commun.* 11:3403.

819 Weissensteiner MH, Poelstra JW, Wolf JBW. 2015. Low-budget ready-to-fly unmanned aerial
820 vehicles: an effective tool for evaluating the nesting status of canopy-breeding bird
821 species. *J. Avian Biol.* 46:425–430.

822 Xu L, Zhou Q. 2020. The Female-Specific W Chromosomes of Birds Have Conserved Gene
823 Contents but Are Not Feminized. *Genes* 11.

824 Yanai I, Benjamin H, Shmoish M, Chalifa-Caspi V, Shklar M, Ophir R, Bar-Even A, Horn-
825 Saban S, Safran M, Domany E, et al. 2005. Genome-wide midrange transcription
826 profiles reveal expression level relationships in human tissue specification.
827 *Bioinformatics* 21:650–659.

828 Yang WR, Ardeljan D, Pacyna CN, Payer LM, Burns KH. 2019. SQuIRE reveals locus-
829 specific regulation of interspersed repeat expression. *Nucleic Acids Res.* 47:e27–e27.

830 Zamudio N, Bourc'his D. 2010. Transposable elements in the mammalian germline: a
831 comfortable niche or a deadly trap? *Heredity* 105:92–104.

832 Zeng L, Kortschak RD, Raison JM, Bertozzi T, Adelson DL. 2018. Superior ab initio
833 identification, annotation and characterisation of TEs and segmental duplications from
834 genome assemblies. *PLOS ONE* 13:e0193588.

835 Zhang G, Li C, Li Q, Li B, Larkin DM, Lee C, Storz JF, Antunes A, Greenwold MJ, Meredith
836 RW, et al. 2014. Comparative genomics reveals insights into avian genome evolution
837 and adaptation. *Science* 346:1311–1320.

838

839

840

841

TABLES

Table 1. Summary of the genomic abundance and transcription of annotated repeats by repeat type. The three most abundant TE types in terms of copy numbers are highlighted in bold. TE classification according to (Kojima 2018). Percentages refer to individual insertions of a particular repeat type in relation to all repeats.

Repeat Class	Copies in genome (%)	Copies transcribed (%)
LINE	**137,022 (72.4)**	**8,367 (59.7)**
LTR	**27,624 (14.6)**	**4,831 (34.5)**
DNA	**12,357 (6.5)**	**359 (2.6)**
SINE	6,170 (3.3)	183 (1.3)
UCON	2,299 (1.2)	70 (0.5)
Eulor	1,087 (0.6)	30 (0.2)
MER	558 (0.3)	17(0.1)
Unclassified	292 (0.2)	83 (0.6)
Total TEs	**187,409**	**13,940**
non-TEs	1760	69
Total repeats	**185,649**	**13,871**

FIGURE LEGENDS

Figure 1. Abundance and transcription of TEs annotated in the hooded crow reference genome shown by chromosomal class and by TE type. **A**) TE content expressed as the proportion of TE sequence relative to total sequence length. **B**) Proportion of transcribed TEs relative to the total number of TE copies. A: autosomes.

Figure 2. Frequency distribution of TE sequence in bins of increasing divergence from the consensus sequence shown separately for autosomes (left panels), the Z (middle panels) and W Chromosome (right panels). Panel **A**) shows a histogram of all annotated TEs, whereas panel **B**) depicts all TE copies that show evidence for transcription in our population sample. Note the shift in abundance *vs*. transcription for LTRs.

862 **Figure 3.** Principal components analysis of RNA-seq data. PC1 and PC2 of normalized and
863 vst-transformed read counts for **A**) coding genes and **B**) TEs in the hooded crow genome
864 (including W-linked contigs).

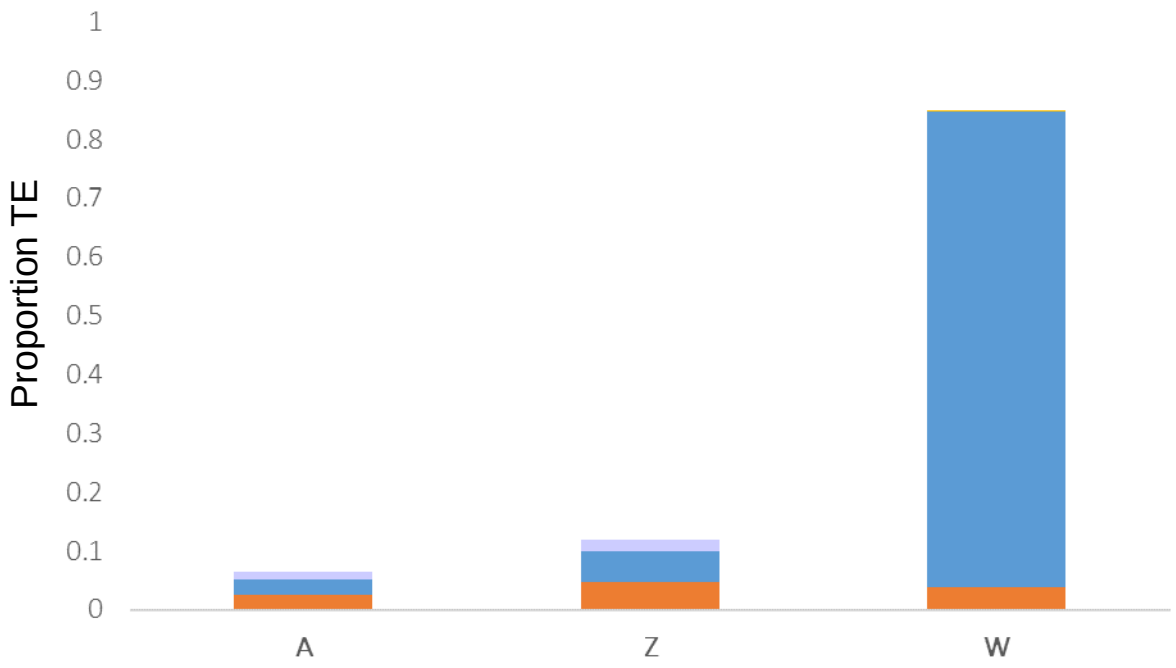
865

866 **Figure 4**. Normalised and log-transformed RNA-seq read counts of autosomal, Z-
867 chromosomal, and W- chromosomal DETEs across three different tissues in females *vs*.
868 males using the **A**) 'random' and **B**) 'unique' approaches, respectively. Note that even in
869 unique mode, some of the W-linked DETEs attracted male RNA-seq reads. This observation
870 is consistent with some of the W-based TEs having highly similar copies elsewhere in the
871 genome.

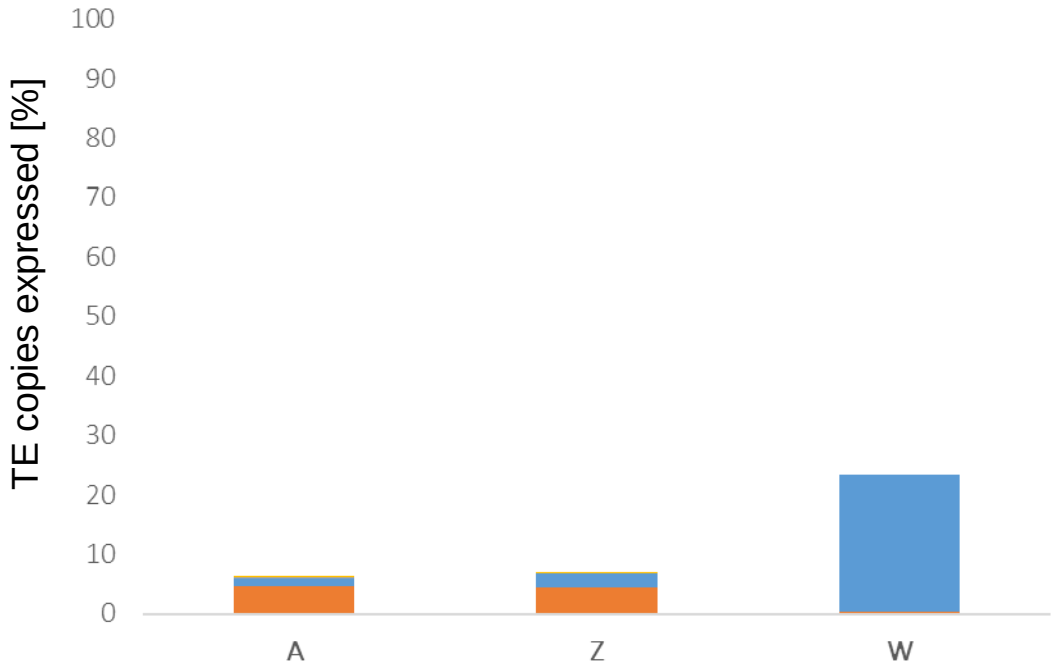
872

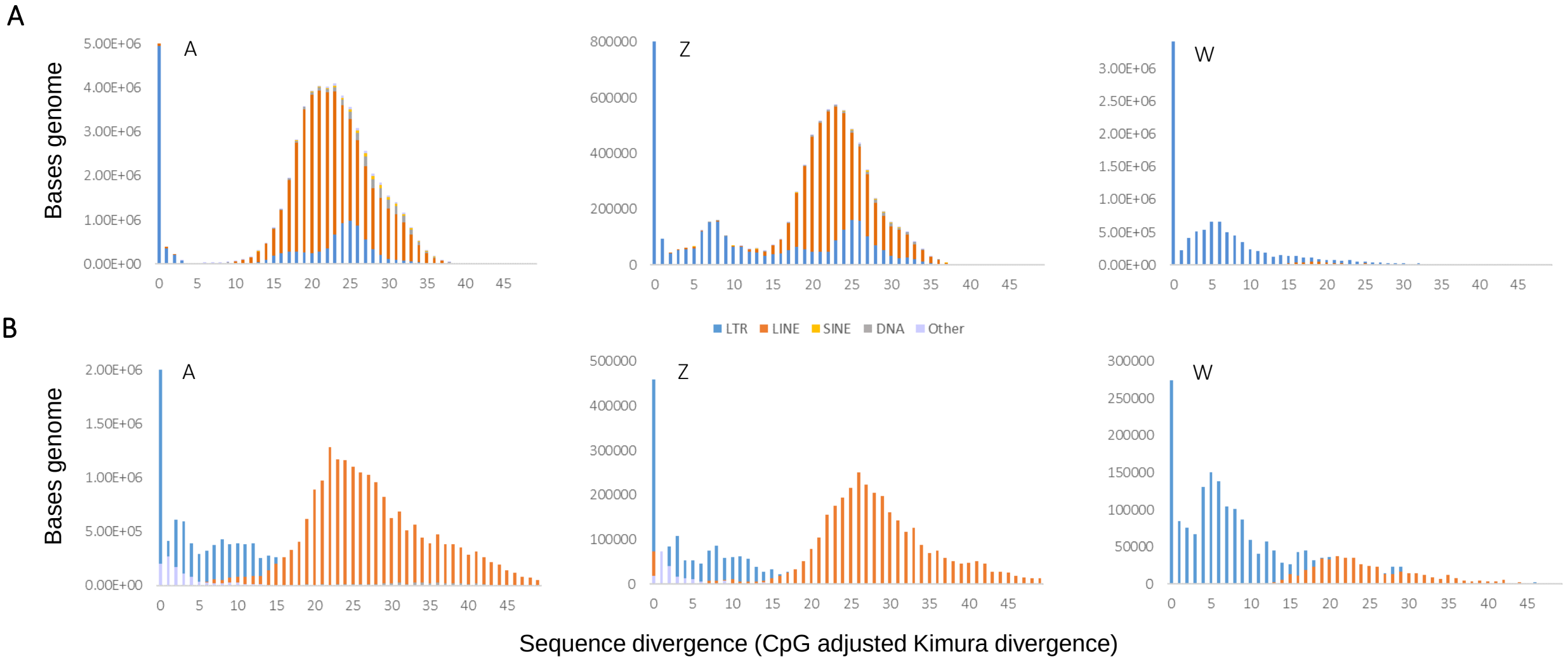
873

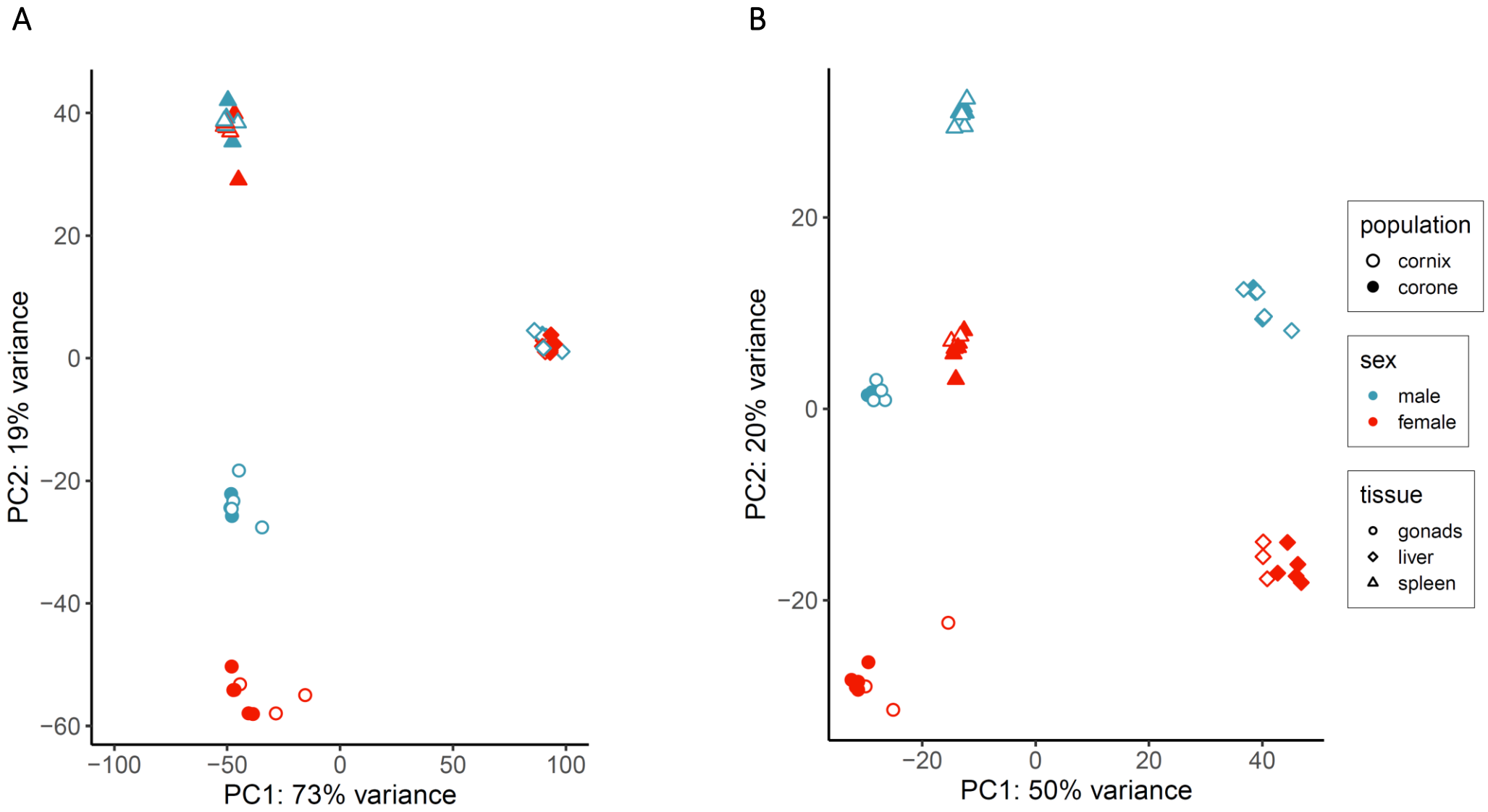
A

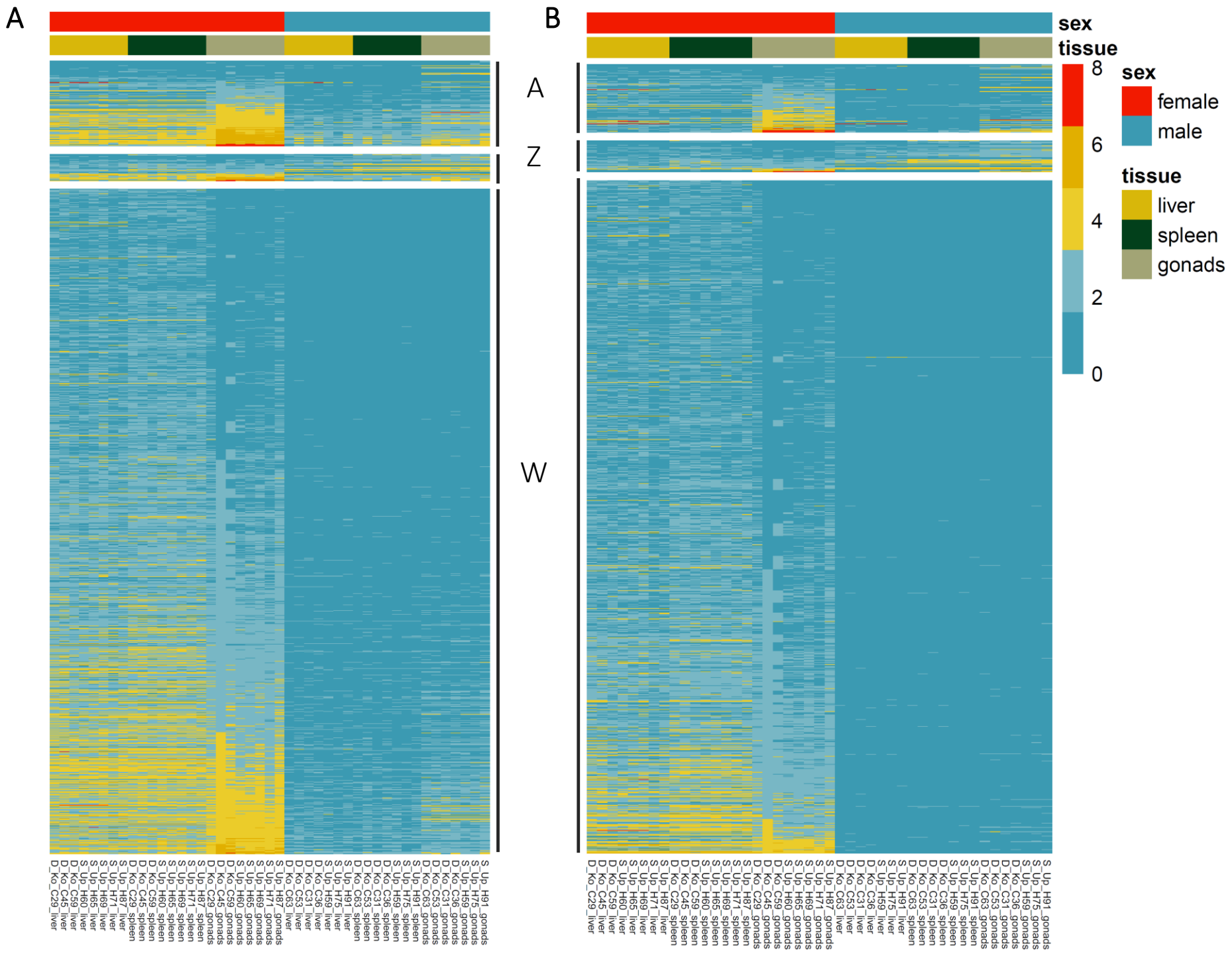

Proportion TE
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
A
Z
W

B


TE copies expressed [%]
100
90
80
70
60
50
40
30
20
10
0
A
Z
W


A
Bases genome
A
Z
W
LTR
LINE
SINE
DNA
Other
B
Bases genome
A
Z
W
Sequence divergence (CpG adjusted Kimura divergence)


A
B
PC2: 19% variance
PC1: 73% variance
PC2: 20% variance
PC1: 50% variance
population
cornix
corone
sex
male
female
tissue
gonads
liver
spleen


A
B
A
Z
W
sex
tissue
8
6
4
2
0
sex
female
male
tissue
liver
spleen
gonads