Downloaded from genome.cshlp.org on January 31, 2021 - Published by Cold Spring Harbor Laboratory Press

A new emu genome illuminates the evolution of genome configuration and nuclear architecture of avian chromosomes

Jing Liu[1,2*], Zongji Wang[1,2,3*], Jing Li[1], Luohao Xu[1,2], Jiaqi Liu[4], Shaohong Feng[5], Chunxue Guo[5], Shengchan Chen[6], Zhanjun Ren[7], Jinpeng Rao[8], Kai Wei[8], Yuezhou Chen[9], Erich D. Jarvis[10,11], Guojie Zhang[12,13,14,15], Qi Zhou[1,2,8†]

1. MOE Laboratory of Biosystems Homeostasis & Protection and Zhejiang Provincial Key Laboratory for Cancer Molecular Cell Biology, Life Sciences Institute, Zhejiang University, Hangzhou 310058, China
2. Department of Neuroscience and Developmental Biology, University of Vienna, Vienna 1090, Austria
3. Institute of Animal Sex and Development, Zhejiang Wanli University, Ningbo, 315100, China
4. Wuhan Gooalgene Technology Co, China
5. BGI-Shenzhen, Beishan Industrial Zone, Shenzhen 518083, China
6. Longteng Ecological Culture Co., LTD, Zhashui 711400, China
7. Key Laboratory of Animal Genetics, Breeding and Reproduction of Shaanxi Province, College of Animal Science and Technology, Northwest A&F University, Yangling 712100, China
8. Center for Reproductive Medicine, The 2nd Affiliated Hospital, School of Medicine, Zhejiang University, Hangzhou 310052, China
9. Jianzhou Poultry Industry Co., LTD, Yong'an 366000, China

Downloaded from genome.cshlp.org on January 31, 2021 - Published by Cold Spring Harbor Laboratory Press

10. Laboratory of Neurogenetics of Language, The Rockefeller University, New York 10065, USA

11. Howard Hughes Medical Institute, Chevy Chase, Maryland 20815, USA

12. China National GeneBank, BGI-Shenzhen, Jinsha Road, Shenzhen, 518120, China

13. State Key Laboratory of Genetic Resources and Evolution, Kunming Institute of Zoology, Chinese Academy of Sciences, Kunming 650223, China

14. Section for Ecology and Evolution, Department of Biology, University of Copenhagen, DK-2100 Copenhagen, Denmark

15. Center for Excellence in Animal Evolution and Genetics, Chinese Academy of Sciences, Kunming 650223, China

[*]These authors contributed equally to the work.

[†]Corresponding author: zhouqi1982@zju.edu.cn

Downloaded from genome.cshlp.org on January 31, 2021 - Published by Cold Spring Harbor Laboratory Press

Abstract

Emu and other ratites are more informative than any other birds in reconstructing the evolution of the ancestral avian or vertebrate karyotype because of their much slower rate of genome evolution. Here we generated a new chromosome-level genome assembly of a female emu, and estimated the tempo of chromosome evolution across major avian phylogenetic branches, by comparing it to chromosome-level genome assemblies of 11 other bird and one turtle species. We found ratites exhibited the lowest numbers of intra- and inter-chromosomal changes among birds since their divergence with turtles. The small-sized and gene-rich emu microchromosomes have frequent interchromosomal contacts that are associated with housekeeping genes, which appears to be driven by clustering their centromeres in the nuclear interior, away from the macrochromosomes in the nuclear periphery. Unlike non-ratite birds, only less than one third of the emu W Chromosome regions has lost homologous recombination and diverged between the sexes. The emu W is demarcated into a highly heterochromatic region (WS0), and another recently evolved region (WS1) with only moderate sequence divergence with the Z Chromosome. WS1 has expanded its inactive chromatin compartment, increased chromatin contacts within the region, and decreased contacts with the nearby regions, possibly influenced by the spreading of heterochromatin from WS0. These patterns suggest that alteration of chromatin conformation comprises an important early step of sex chromosome evolution. Overall, our results provide novel insights into the evolution of avian genome structure and sex chromosomes in three-dimensional space.

Keywords: emu, chromosome-level assembly, interchromosomal contacts, sex chromosomes

Downloaded from genome.cshlp.org on January 31, 2021 - Published by Cold Spring Harbor Laboratory Press

Introduction

Most birds have about 10 pairs of relatively large size macrochromosomes, including one pair of sex chromosomes (male ZZ, female ZW), and about 30 pairs of smaller microchromosomes, some of which can be hardly discerned by light microscopy (Takagi and Sasaki 1974). It was hypothesized that microchromosomes represent archaic linkage groups of ancestral vertebrates (Ohno et al. 1969; Tegelström and Ryttman 1981; Burt 2002; Uno et al. 2012). This was implicated by reconstruction of the ancestral vertebrate karyotype first using the genomes of a few (Nakatani et al. 2007), and recently many more available species (Sacerdot et al. 2018). Direct supporting evidence for the hypothesis came from genomic comparisons of chicken vs. the amphibians Mexican axolotl and Western clawed frog (Voss et al. 2011), the spotted gar fish (Braasch et al. 2016) and the invertebrate amphioxus (Simakov et al. 2020). These studies found one-to-one correspondence between many but not all chicken microchromosomes vs. (micro-)chromosomes of amphibians and gar fish or the reconstructed chromosomes of the vertebrate/gnathostome common ancestor using amphioxus (Simakov et al. 2020), sea lamprey (Smith et al. 2018) or elephant shark (Venkatesh et al. 2014) genome, dating the likely existence of microchromosomes at least to the ancestor of jawed vertebrates.

Chicken microchromosomes tend to be gene-rich, have higher recombination rate and GC content than macrochromosomes (ICGSC 2004). Such a distinct genomic composition probably dictates the segregated nuclear architecture in chicken cells that might also have existed in the vertebrate ancestor. Similar to the small-sized human chromosomes (Cremer et al. 2001), the chicken microchromosomes predominantly occupy the nuclear interior (Habermann et al. 2001), which corresponds to the transcriptionally active or A compartments revealed by Hi-C analysis (Lieberman-Aiden et al. 2009). While macrochromosome regions are mainly located at the

Downloaded from genome.cshlp.org on January 31, 2021 - Published by Cold Spring Harbor Laboratory Press

nuclear periphery (Habermann et al. 2001) and correspond to the inactive heterochromatin or B compartments. The scarcity of high-quality chromosome-level genomes of birds, particularly underrepresentation of identified microchromosome sequences except for chicken, has hampered the reconstruction of the evolutionary trajectories of avian and thus vertebrate chromosome architectures.

We previously generated a near chromosome-level genome of ostrich (Zhang et al. 2015) with Illumina reads and optical-mapping, and found that its Z Chromosome (ChrZ) harbors much fewer inversions than chicken and zebra finch, when all three species were compared to the autosomal counterparts of lizard and snake (Zhou et al. 2014). As ChrZ usually exhibits disproportionately more inversions than any other chromosomes in birds (Hooper and Price 2017), this highly conserved intrachromosomal synteny between ratites and reptiles is very likely a genome-wide pattern. This was confirmed by recent refinements of our ostrich genome (O'Connor et al. 2018). These results are consistent with a much lower genome-wide substitution rate in ratites, associated with their larger body size and longer generation time (Bromham 2011; Jarvis et al. 2014; Wang et al. 2019). Particularly, ratites have a pair of homomorphic sex chromosomes with a much lower pairwise substitution rate, in contrast to the heteromorphic sex chromosomes of most other birds and mammals (Cortez et al. 2014; Zhou et al. 2014). We recently determined over two thirds of the ratite (except for kiwis) sex-linked regions are still recombining with each other as the pseudoautosomal region (PAR) (Wang et al. 2019; Xu et al. 2019). In the remaining non-recombining sexually differentiated regions (SDR), ratites have undergone at least one ancestral recombination suppression (RS) shared by all birds, and another lineage specific RS (Wang et al. 2019), which demarcated the SDRs into two regions of ‘evolutionary strata’ (the older stratum is named as stratum S0, and the younger one as S1) by

Downloaded from genome.cshlp.org on January 31, 2021 - Published by Cold Spring Harbor Laboratory Press

104 their different levels of Z/W pairwise sequence divergence (Zhou et al. 2014; Wang et al. 2019).
105 Among the studied ratites, emu and cassowary are even less differentiated between sex
106 chromosomes than ostrich and any other Neognathae species, thus best preserving the ancestral
107 status of avian sex chromosomes (Zhou et al. 2014).

108 Here we chose emu, a unique model for studying the evolution of vertebrate chromosome
109 architectures and avian sex chromosomes, for high-quality genome assembly using the cutting-
110 edge third-generation long-read sequencing and Hi-C technologies with the pipeline of the
111 Vertebrate Genomes Project (Rhie et al. 2020). By comparing the new emu genome to other
112 chromosome-level genome assemblies of 11 bird and one turtle species, we estimated the tempo
113 of inter- and intra-chromosomal rearrangements in major lineages of birds, in order to test the
114 hypothesis that emu and other ratites have the lowest lineage-specific chromosome evolution rate
115 among birds, therefore best representing the ancestral avian genome configuration. We further
116 compared the chromatin architectures between the emu macro- and microchromosomes, and
117 between the Z and W Chromosomes using the liver Hi-C data, so that to gain insights into the
118 nuclear architecture of the avian ancestor, and the evolutionary process of avian sex
119 chromosomes in the three-dimensional (3D) nuclear space.
120
121 **Results**
122 **Chromosomal assembly of a female emu**
123 To generate a high-quality reference genome of emu, including the repetitive ChrW, we
124 produced 70-fold genomic coverage of PacBio reads (subread N50 length 15.5 kb), and long-
125 range linkage data of 124-fold 10x Genomics linked reads, 154-fold Dovetail Chicago data, and
126 46-fold Hi-C data from three female emu individuals (**Fig. 1A**). Our preliminary assembly

Downloaded from genome.cshlp.org on January 31, 2021 - Published by Cold Spring Harbor Laboratory Press

127 derived from PacBio reads alone produced 1,389 gapless contig sequences with an N50 size of
128 13.3 Mb. Using the three different types of linkage data followed by manual curation, these
129 contigs were oriented and connected into 802 scaffolds with an N50 size of 82.7 Mb. We
130 performed a final gap filling step and assembly polishing using the raw reads of both PacBio and
131 Illumina data that are compensatory to each other, and assembled the entire genome into 1.26 Gb
132 (97.9% of the 1.29 Gb estimated genome size by GenomeScope (Vurture et al. 2017)). About
133 1.20 Gb sequences or 94.9% of the assembled genome have been anchored into 29 autosomes
134 (macrochromosomes Chr1-Chr10 and microchromosomes Chr11-Chr28 and Chr33), one ChrZ
135 (82.7 Mb), and one ChrW SDR region (7.92 Mb or 29.1% of the homologous Z-linked region,
136 which is 27.2Mb long) (**Supplemental Table S1; Supplemental Fig. S1**).
137 This version of the emu genome (droZJU1.0) has a 77-fold increase of continuity
138 measured by N50 size relative to a recent Illumina-based assembly (droNov1) (Sackton et al.
139 2019), including a 7.6-fold increase of the size of assembled ChrW SDR sequences (Wang et al.
140 2019). The distribution of the scaffold lengths was shifted toward much longer sequences in the
141 droZJU1.0 assembly relative to droNov1 or the most recent ostrich chromosomal assembly
142 (Zhang et al. 2015) (**Fig. 1B**). A major improvement came from the resolution of repetitive
143 sequences that were enriched within or between the previous Illumina scaffolds, by the long
144 reads and long-range linkage information (**Fig. 1C**). This was reflected by the increased
145 annotated repeat content (9.9% vs. 7.2%) of droZJU1.0 vs. droNov1 genomes, which is mainly
146 concentrated at several long terminal repeat (LTR) retrotransposon families (**Supplemental Fig.**
147 **S2**). We annotated a total of 20,823 consensus gene models combining homology-based
148 predictions, 14 tissue-specific transcriptomes, and *de novo* predictions. The structural accuracy
149 of our chromosome assembly was demonstrated by its highly consistent synteny with another

Downloaded from genome.cshlp.org on January 31, 2021 - Published by Cold Spring Harbor Laboratory Press

emu chromosomal genome independently produced by DNAzoo from Illumina reads and Hi-C data (Dudchenko et al. 2017) (**Supplemental Fig. S3**). Its completeness was reflected by a high BUSCO gene value (95%), and our annotation of the putative centromeric regions of 26 chromosomes, and telomeres of eight chromosomes as well as the interstitial telomeric repeats (ITR) (**Supplemental Table S2; Supplemental Fig. S4-5**), which were corroborated by the published cytogenetic results of the emu (Nanda et al. 2002).

The new emu genome exhibited distinctive genomic features between sex chromosomes and autosomes, and between macro- and microchromosomes. Because of RS, ChrW is expected to accumulate transposable elements (TEs) and diverge from ChrZ in genomic sequences (Charlesworth et al. 2005). This was confirmed by a 3.9-fold increase of TE content, particularly various subfamilies of LTR retrotransposons and DNA transposons on ChrW (10.5%) compared to the rest of the genome (0.88%) (**Fig. 1D**). We managed to assemble the Z- and W-linked regions of S1 into two separate sequences (ZS1 and WS1) of similar lengths (about 4.8Mb) that showed the same level of female read depth, but a much higher level of female single-nucleotide polymorphism (SNP) density because of Z/W divergence, compared with autosomes and the juxtaposed long PAR (0Mb-55.5Mb, 67.1% of the ChrZ). The ancient stratum S0 (60.3Mb-82.7Mb on the ChrZ) that suppressed recombination in the ancestor of birds has become much more repetitive or diverged on the ChrW than S1. Therefore, its W-linked sequence assembly is quite fragmented, and its Z-linked region exhibited half the female read coverage relative to autosomes (**Fig. 1E**). The two evolutionary strata constituted the SDR of emu sex chromosomes and showed a generally male-biased gene expression pattern of Z-linked genes because of lack of global dosage compensation (**Fig. 1E; Supplemental Table S3**; **Supplemental Fig. S6)**. (Wang et al. 2014).

Downloaded from genome.cshlp.org on January 31, 2021 - Published by Cold Spring Harbor Laboratory Press

The emu macrochromosomes had a significantly higher content of LINE and DNA transposon families (P=1.16 × 10^{-5}, Wilcoxon test), but lower content of simple repeats (P=0.036, Wilcoxon test) than the microchromosomes, with an apparent gradient of change in some TE subfamilies (e.g., L2) (**Fig. 1D; Supplemental Fig. S7**). Meanwhile, the emu macrochromosomes exhibited lower GC content (P=9.89 × 10^{-6}, Wilcoxon test) and gene density (P=3.85 × 10^{-4}, Wilcoxon test) than the microchromosomes (**Fig. 1E; Supplemental Fig. S7-9**). These findings indicated that the autosome organization of and relative genomic features in macro- and microchromosomes of the emu are similar to other birds (Burt 2002; ICGSC 2004; Zhou et al. 2014), but revealed here at higher and more complete resolution due to the more complete assembly.

Emu and other ratites have best preserved the ancestral avian chromosome configuration

Avian chromosome evolution has been proposed to be dominated by intrachromosomal changes, but this was based on cytogenetic methods of low resolution, with a limited number of species (Griffin et al. 2007), and with major underrepresentation on the microchromosomes. A recent genomic investigation covered nearly 30 bird species but for the Palaeognathae lineage, with 22 of the 31 studied species having only a scaffold-level Illumina-based genome assembly (Damas et al. 2018). With the newly produced more complete emu genome, we were motivated to compare its chromosome evolution rate vs. those of other Neognathous birds, to quantitatively delineate a finer picture of the tempo of avian chromosome evolution. Using the sea turtle (C*helonia mydas*) (Wang et al. 2013; Dudchenko et al. 2017) as an outgroup, we identified the inter- and intra-chromosomal rearrangements of 12 birds with all chromosome-level genomes (**Supplemental Table S4-5**).

Downloaded from genome.cshlp.org on January 31, 2021 - Published by Cold Spring Harbor Laboratory Press

All but eight of the assembled emu chromosomes together accounted for 87.3% of the genome that mapped to one single homologous chromosome in the turtle (**Fig. 2A**). The eight outlier emu chromosomes are still aligned to four turtle chromosomes, which can be the result of either four chromosome fusions in the turtle lineage or four fissions in the avian lineage. To discriminate between the two scenarios, we inspected other reptile species and found five emu outlier chromosomes (Chr4, Chr10, Chr14, Chr23 and Chr25) were mapped to one single chromosome in rattlesnake (Schield et al. 2019) (**Fig. 2B**). While among others, the same combination of two emu chromosomes that mapped to one turtle chromosome could not be found in rattlesnake or alligator (Dudchenko et al. 2017; Rice et al. 2017) (**Fig. 2C**). For example, emu Chr6 and Chr12 were homologous to turtle Chr7, but were homologous to parts of snake Chr5 and Chr2, respectively. Therefore, it is likely that the eight outlier emu chromosomes were also ancestrally single chromosomes that have undergone independent fusions or translocations in other amniote species. Similar to emu, all the other investigated birds showed few fusions or fissions of chromosomes compared to the turtle (**Supplemental Fig. S10**), except for the golden eagle (*Aquila chrysaetos*) (Dudchenko et al. 2017; Van Den Bussche et al. 2017), where we identified 4 fissions and 10 fusions/translocations (**Fig. 2D**) after it diverged from the California condor (*Gymnogyps californianus*) (Dudchenko et al. 2017). A lack of any identified fusions or fissions in the condor indicated that the extensive interchromosomal changes is not a universal feature of birds-of-prey. Overall, these results provided evidence that the common ancestor (*Archelosauria*) of birds and turtles had almost the same karyotype as that of emu, where certain other bird lineages went on to evolve more chromosomal rearrangements.

We found much greater variations of intrachromosomal rearrangements (inversions and translocations) among the studied birds using the turtle as an outgroup (**Fig. 2E; Supplemental**

Downloaded from genome.cshlp.org on January 31, 2021 - Published by Cold Spring Harbor Laboratory Press

Table S6). By determining the presence or absence of orthologous rearrangement regions of the focal species in its related species, we inferred the rate of intrachromosomal rearrangements on each phylogenetic branch based on parsimony (**Methods**). Overall ratites have undergone significantly less genomic rearrangements ($P<2.20 \times 10^{-16}$, Wilcoxon test) including both intra- (**Fig. 2F**) and inter-chromosomal changes than Neognathae species after they diverged from the avian ancestor, supporting that the ratites have better preserved the ancestral avian genomic configuration than any other birds. The intrachromosomal evolution rates at the external lineages are significantly lower ($P=0.022$, *t*-test) than those at the internal lineages of ratites. This reflects the impacts of independently evolved gigantism and elongated generation time among ratites (Sackton et al. 2019). In contrast, the intrachromosomal evolution rate was greatly accelerated at the ancestor of Neognathae, accompanied with their lineage species radiation (Jarvis et al. 2014), was maintained at a high level along the internal branches, and then became decelerated in most of the external branches (**Fig. 2E; Supplemental Fig. S11-12**).

Some chromosomes (e.g., Chr3) seem to be a hotspot for rearrangements across all the investigated birds, while some (e.g., Chr10) seems to be highly conserved for their gene synteny between all species (**Fig. 2E**). The cause for such variations of the numbers of rearrangements between chromosomes remains unclear: the variations do not seem to correlate with those of chromosome-wide expression levels, gene density, GC content or repeat content (**Supplemental Table S7**). The chromosome evolution rate of Chr10 seems to have slowed down in the ancestor of alligator and emu, i.e., archosaurs; while that of Chr3 has accelerated in the ancestor of reptiles and independently slowed down in the alligator lineage (**Supplemental Fig. S13**). ChrZ of Neognathae but not Paleoganathae have fixed a significantly higher number of rearrangements ($P=1.21 \times 10^{-5}$, Wilcoxon test) than other macrochromosomes compared to the homologous

Downloaded from genome.cshlp.org on January 31, 2021 - Published by Cold Spring Harbor Laboratory Press

turtle autosome. This is consistent with the recent cytogenetic examination of over 400 passerine species which characterized the ChrZ with more fixed inversions than any other macrochromosomes (Hooper and Price 2017; Damas et al. 2018). The faster evolution of ChrZ genomic structure (the 'fast-Z' effect) (Meisel and Connallon 2013) can be explained by a hemizygous ChrZ that is more likely to fix genomic rearrangements by genetic drift, due to a reduced effective population size; or it is driven by selection for incompatible inversion alleles between species. As expected, ratites do not exhibit such a fast-Z pattern because the most parts of their ChrZs are evolving predominantly like autosomes (**Fig. 1E**). Microchromosomes generally had a higher rate of intrachromosomal rearrangements than the macrochromosomes after scaling for chromosome size or removing the outlier species (e.g., pigeon) (**Fig. 2G**, P=0.004, Wilcoxon test; **Supplemental Fig. S14**). This may be influenced by the higher recombination rate and GC content of the microchromosomes leading to more frequent DNA double-strand breaks (DSBs). To test this hypothesis, we examined the rearrangement breakpoint regions, i.e., evolutionary breakpoints in three species representing each major avian group, and found that the breakpoints indeed have significantly higher GC content than the genome average (**Supplemental Fig. S15**, $P=1.17 \times 10^{-10}$, Wilcoxon test), probably driven by the GC-biased gene conversion caused by a high local recombination rate (Weber et al. 2014).

We hypothesized that the different GC content of macro- vs. microchromosomes (**Supplemental Fig. S7**) is also associated with their different contributions to the genome size reduction of birds, relative to their reptile relatives. To quantitatively measure this chromosomal difference, we calculated the length difference between turtle and emu in their syntenic blocks, whose aligned sizes together accounted for 95.9% of all the investigated emu chromosomes (Chr1-Chr28, ChrZ) (**Supplemental Table S6**). Repetitive regions within the syntenic blocks

Downloaded from genome.cshlp.org on January 31, 2021 - Published by Cold Spring Harbor Laboratory Press

exhibited a much larger length difference between the two species ($P<2.20 \times 10^{-16}$, Wilcoxon test) than any other genomic regions, while the exonic regions have maintained around the same lengths with slightly larger exons in emu (**Fig. 2H; Supplemental Fig. S16**). These findings confirm that the reduction of avian genome size is mainly attributed to the genome-wide contraction of TE content (Zhang et al. 2014), possibly related to the evolution of flight (Kapusta et al. 2017). We found the smaller an emu chromosome is, the larger the GC content ($P=7.45 \times 10^{-5}$, Pearson's correlation r=-0.67) and the more sequence loss, particularly repeat sequence loss ($P<2.20 \times 10^{-16}$, Pearson's correlation r=0.71) relative to the turtle (**Fig. 2I; Supplemental Fig. S17**). This pattern is consistent among all the studied birds where microchromosomes exhibited significantly more extensive sequence loss ($P<2.20 \times 10^{-16}$, Wilcoxon test) than macrochromosomes, when compared to turtle or crocodile genomes (**Fig. 2J; Supplemental Fig. S18-19**).

These results suggest that due to the higher GC content on microchromosomes resulting from the higher recombination rate than macrochromosomes, their non-coding sequences are more prone to deletions, triggered by for example replication slippage (Kiktev et al. 2018). Using the available population genomic data of duck (Zhou et al. 2018), we found a significant positive correlation ($P=4.73 \times 10^{-8}$, Pearson’s correlation r=0.31) between the recombination rate and the extent of sequence loss within its syntenic region with turtle (**Supplemental Fig. S20**). This is consistent with the reported negative correlation between the GC content and the genome size among mammals (Romiguier et al. 2010). Birds have maintained a higher number of microchromosomes since the divergence with other related reptile species (Uno et al. 2012). This may also have contributed to their genome size reduction.

Downloaded from genome.cshlp.org on January 31, 2021 - Published by Cold Spring Harbor Laboratory Press

Microchromosomes have an excess of interchromosomal contacts associated with housekeeping genes

The nuclear arrangement of chicken microchromosomes in the interior and macrochromosomes at the periphery (Habermann et al. 2001; Maslova et al. 2015) should give rise to more frequent interchromosomal (*trans*-) contacts between microchromosomes. To test this hypothesis, we measured the *trans*- and *cis*-chromosomal contacts by the emu normalized Hi-C read pairs that are derived from the different and the same chromosomal regions, respectively. These contacts quantify the frequency of spatial proximity between two distant genomic regions captured by the Hi-C technique that may be related to but not necessarily demonstrate functional association between these regions (Lieberman-Aiden et al. 2009). Similar to the reported patterns of mammals and *Drosophila* (Szabo et al. 2019), *cis*-contacts were the dominant type of chromatin interactions (**Fig. 3A**). There were much more abundant and stronger *trans*-contacts between microchromosomes than between macrochromosomes, and the overall *trans*-contact frequencies were negatively correlated with the chromosome size (**Fig. 3B**, P=9.13 × 10^{-11}, Pearson's correlation r=-0.88). A previous study showed that human Chr18 and Chr19, albeit having a similar size, occupy distinct nuclear territories with the gene-poor Chr18 located at the nuclear periphery while the gene-rich Chr19 was at the interior (Croft et al. 1999; Cremer and Cremer 2001). This suggested that the gene content rather than the chromosome size is underlying the segregated nuclear territories. Indeed, we found that the gene density and GC content was positively correlated with the *trans*-contact numbers per chromosome (**Supplemental Fig. S21-23**). As with humans, we noted that some emu macrochromosomal regions with high gene densities also showed robust *trans*-contacts with other chromosomes (**Supplemental Fig. S24,** P<2.20 × 10^{-16}, Wilcoxon test). Such high *trans*-contacts between microchromosomes were also

Downloaded from genome.cshlp.org on January 31, 2021 - Published by Cold Spring Harbor Laboratory Press

found in our companion study on the duck genome, and were reported previously for the rattlesnake (Schield et al. 2019), as well as for small-sized chromosomes in human cells (Lieberman-Aiden et al. 2009) and therefore it is probably a conserved chromosome territorial feature of vertebrates (Perry et al. 2020).

To explore the functional significance of these *trans*-contacts, we divided the entire emu genome according to their Hi-C interaction profiles into the active (A) and inactive (B) compartments (Lieberman-Aiden et al. 2009), and then compared the *trans*-contact frequencies within or between the two types of compartments. We found that the *trans*-contacts more frequently involved two regions that were both from active compartments (AA contacts) than the *cis*-contacts ($P<2.20 \times 10^{-16}$, Fisher's Exact test), and microchromosomes had more frequent AA *trans*-contacts than the macrochromosomes (**Fig. 3C**, $P<2.20 \times 10^{-16}$, Fisher's Exact test). Further, genes exhibiting high frequencies (ranked top 10%) of *trans*-contacts detected by our liver Hi-C data had significantly higher expression levels ($P<2.20 \times 10^{-16}$, Wilcoxon test), particularly in the liver (**Supplemental Fig. S25**), and broader tissue expression patterns ($P<2.20 \times 10^{-16}$, Wilcoxon test) than the other genes (**Fig. 3D**). These *trans*-contacting genes defined were enriched in cell regulatory and metabolic functional categories (**Supplemental Fig. S26**). Therefore, *trans*-contacts between active compartments of different chromosomes probably play an important role in regulating housekeeping gene expression. One caveat about this conclusion is that our *cis*- or *trans*-contacts were calculated from Hi-C data of only the liver tissue, and the conclusion needs to be verified in other tissues in future.

The frequency of *trans*-contacts but not *cis*-contacts were on average the highest at the centromeric regions of micro- but not macrochromosomes, and decayed by the distance away from the centromeres (**Fig. 3E-F**). The chromosomal distribution of *trans*-contacts is consistent

Downloaded from genome.cshlp.org on January 31, 2021 - Published by Cold Spring Harbor Laboratory Press

with the radial 3D conformation of the predominantly acrocentric microchromosomes of birds, whose pericentromeric heterochromatin associates with the interior nucleolus (Habermann et al. 2001; Maslova et al. 2015). To search for the candidate genomic determinants of such nuclear conformation, we compared the centromeric sequences between the emu macro- and microchromosomes. We identified two GC-rich (GC content>55%) repeat monomers of 65-bp and 81-bp long enriched at the putative centromeres, whose copy numbers are among the most abundant throughout the entire emu genome. The 65-bp repeats were more enriched in microchromosomes than macrochromosomes (**Fig. 3G-H; Supplemental Fig. S4**). Similar microchromosome centromere enriched repeats, but with different sequences have also been reported in chicken (chicken nuclear membrane associated repeats, CNM) (Matzke et al. 1990; Shang et al. 2010) and other bird and turtle species (Yamada et al. 2002; Yamada et al. 2005; Nishida et al. 2013). Thus, we named this repeat as emu nuclear interior associated (ENI) repeat. Whether the differential chromosomal distribution of ENI repeats drives the segregated nuclear architecture of emu macro- vs. microchromosomes remains a question for future functional investigation.

Emu sex chromosome evolution involves alteration of chromatin conformation

The newly assembled ChrW of emu comprises a model for studying the stepwise evolution of genome and chromatin conformation under a non-recombining environment. The long PAR is shared between sex chromosomes and represents the ancestral autosome state before recombination was suppressed, while S1 and S0 respectively represent the early and late phases of sex chromosome differentiation. This was demonstrated by the gradient of accumulated TEs and functional gene loss formed by the W-linked S0 (WS0), S1 (WS1) and PAR. The greatly

Downloaded from genome.cshlp.org on January 31, 2021 - Published by Cold Spring Harbor Laboratory Press

different TE content between WS0 (43.5%), WS1 (9.3%) and PAR (7.3%) has demarcated the three regions (**Supplemental Fig. S27**), with a small part of WS0 having been reshuffled between WS1 and PAR (**Fig. 4A**). The highly heterochromatic WS0 had RS over 150 million years ago (MYA) in the avian ancestor, and had only about 3.6 Mb sequences assembled, compared to the 22.4 Mb long Z-linked S0 (ZS0) region. The ZS0 also seemed to have undergone much more intrachromosomal rearrangements than other Z-linked regions when being compared to the homologous autosomal region of the sea turtle (**Supplemental Fig. S28**), presumably due to the reduction of recombination rate. The few alignable sequences between ZS0 and WS0 hampered our inference of how RS occurred within S0. In contrast, the younger S1 emerged 23 MYA (Wang et al. 2019) and exhibited a low level of average sequence divergence (5%) between the Z/W, confirming the low rate of emu sex chromosome evolution. The entire WS1 formed a large inversion compared to its Z-linked homolog, as well as the ostrich ChrZ and the homologous turtle autosome (**Fig. 4B**). This suggests that the emu S1 likely evolved RS through a W-linked chromosome inversion.

247 out of 273 (90.5%) WS0 genes have become deleted or accumulated nonsense mutations, compared to 27 out of 42 (64.3%) WS1 genes (**Fig. 4C**). Among the retained W-linked genes with intact open reading frames, they were expressed at a significantly lower level (P=0.020, Wilcoxon test) than their Z-linked homologs, with the WS0 region showing more severe downregulation of gene expression (P=8.16 × 10^{-9}, Wilcoxon test) than the WS1 (**Fig. 4D**). Despite the low level of sequence divergence, WS1 nevertheless showed clear signatures of downregulation of gene expression, suggesting regulatory changes simultaneously occurred with or even preceded the amino acid changes during W Chromosome evolution.

Downloaded from genome.cshlp.org on January 31, 2021 - Published by Cold Spring Harbor Laboratory Press

We hypothesized that the global change of chromatin conformation induced by accumulation of TEs may have important contributions to such broad downregulation of W-linked genes. This was previously shown on the young Y Chromosome of *Drosophila miranda*, where TE accumulation has increased the level of silencing histone modifications (H3K9me3), and thus induced gene downregulation (Zhou et al. 2013). Although the TE content of emu WS1 is comparable to that of the PAR and autosomes (9.3% vs. 7.3% and 9.8%), it has become 1.6-fold higher than that of Z-linked S1 (ZS1, 5.8%), suggesting an ongoing process of heterochromatinization. To test this hypothesis, we compared between sex chromosomes for their frequencies of *cis*-contacts within S1 and between S1 and the neighboring PAR. We expected that if the WS1 was becoming globally heterochromatic, its chromatin configuration would become more compact and incur more *cis*-contacts, suggested by the significantly more *cis*-contacts within the B compartments than the A compartments at the genome-wide level (**Supplemental Fig. S29-30**, $P<2.20 \times 10^{-16}$, Wilcoxon test). Indeed, ChrW had significantly higher numbers of *cis*-contacts (**Fig. 4E**, $P=0.035$, Wilcoxon test) and a moderately increased inactive compartment strength ($P=0.051$, Fisher's exact test) than ChrZ within the S1 region, but on average decreased *cis*-contacts between WS1 and PAR (**Fig. 4E**, $P=7.99 \times 10^{-13}$, Wilcoxon test).

The chromatin compartments of WS1 also have become more segregated: compared to the randomly distributed A and B compartments of the ZS1, the WS1 is divided into one large B compartment (WS1B, 3.24Mb-5.28Mb) bordering the S0, and one large A compartment (WS1A, 5.28Mb-7.68Mb), with expectedly higher *cis*-contacts of WS1B ($P<2.20 \times 10^{-16}$, Wilcoxon test) than that of WS1A (**Supplemental Fig. S31**). There are, however, few significant differences in the structure and strength of the finer chromatin units enclosed in the A or B compartments,

Downloaded from genome.cshlp.org on January 31, 2021 - Published by Cold Spring Harbor Laboratory Press

topologically associated domains (TADs) (**Fig. 4F; Supplemental Fig. S32**). We found WS1B exhibited a significantly higher level of Z/W pairwise sequence divergence ($P<2.20 \times 10^{-16}$, Wilcoxon test) and repeat content ($P=0.027$, Wilcoxon test) than WS1A, a pattern similar to the evolutionary strata. Such a 'strata within strata' pattern clearly formed without secondary inversions within S1 (**Fig. 4B**). We suggest that such pronounced changes of chromatin conformation may be caused by the spreading of heterochromatin from the highly heterochromatic WS0 region to the nearby WS1B, similar to the mechanism of positional effect variegation (PEV) (Elgin and Reuter 2013). As the homologous regions of emu ChrZ and the turtle autosomes of WS1B have a lower gene density (32% vs. 54% genic region) and a higher proportion of B compartment (60% vs. 51%) than those of WS1A, the ancestrally more inactive WS1B probably have undergone a weaker natural selection against the spreading of S0 heterochromatin after its recombination was suppressed. In addition, the boundary between WS1A/B was overlapped with a TAD boundary with the lowest insulation score, i.e., the highest boundary strength within the entire S1 region (**Fig. 4F**). Selection on such a strong TAD boundary may prevent the further invasion of heterochromatin into WS1A.

Finally, some WS1A genes like *ALDH1A1* and *ANXA1* have unexpectedly higher expression levels than their Z-linked homologs (**Fig. 4D**). We recently showed that *ANXA1* has a conserved ovary-biased gene expression pattern in various birds and the green anole lizard, and was even restored on the W Chromosomes of some songbirds through transposition after the loss of its original copy (Xu and Zhou 2020). And *ALDH1A1* has been shown to function at the onset of female meiosis in mice (Bowles et al. 2016). We proposed once the WS1 became specifically transmitted in females after recombination suppression, the female-specific selection targeting

Downloaded from genome.cshlp.org on January 31, 2021 - Published by Cold Spring Harbor Laboratory Press

these genes with pre-existing female-related functions may account for their upregulation and curb the further degeneration or the spreading of heterochromatin from WS1B into WS1A.

Discussion

Among vertebrates, birds have one of the most conserved karyotypes with many of the microchromosomes that have been recently suggested to exist before the divergence of the vertebrate/gnathostome ancestor about 500 MY ago (Simakov et al. 2020). Our study here showed that among birds, emu and other ratites are the most informative for the evolution of the ancestral avian or vertebrate karyotype, because they have experienced much less lineage-specific chromosomal changes. It is possible that the vertebrate ancestor had a radial nuclear conformation of chromosomes similar to that of birds (**Fig. 5A**), with the gene-rich microchromosomes and gene-poor macrochromosomes segregated respectively to the transcriptionally active nuclear interiors and more silent peripheries (Habermann et al. 2001). This is supported by cytogenetic studies of primates, birds and frogs, which showed that their chromosomes or partial chromosomal regions are all compartmentalized by their gene or GC content in the interphase nuclei (Cremer and Cremer 2001; Federico et al. 2005; Federico et al. 2006). There are some variations between different cell types (Stadler et al. 2004) and an exception of inverted nuclei in rod photoreceptors of nocturnal mammals (Feodorova et al. 2020). The functional significance of such higher-order nuclear architecture (Cremer et al. 2006) is suggested by the interspecific conservation of chromosome territories, regardless of the chromosome rearrangements. For example, in both birds and primates, fusions of microchromosomes or gene-dense chromosomes with other chromosomes do not seem to alter their nuclear positions (Tanabe et al. 2002; Tanabe et al. 2005; O'Connor et al. 2019). Our

Downloaded from genome.cshlp.org on January 31, 2021 - Published by Cold Spring Harbor Laboratory Press

finding that clustered emu microchromosomes have more *trans*-contacts between the active compartments of housekeeping genes, and presumably cluster in the nuclear center like other birds, is consistent with the colocalization of 'splicing speckle' nuclear structure enriched for splicing factors with microchromosomes in chicken neuronal cells (Berchtold et al. 2011). It remains to be examined in the future with more complete genome assemblies and Hi-C contact maps of more vertebrates, whether such an association between *trans*-contacts and housekeeping genes in the nuclear center is conserved and represent an ancestral gene regulation mechanism in vertebrates.

On the other hand, the nuclear peripheries are preferentially occupied by the macrochromosomes and possibly the W chromosome, through tethering their heterochromatin to the nuclear lamina (Falk et al. 2019). The conventional model of sex chromosome evolution of birds and mammals involves suppression of recombination through chromosomal inversions, forming the pattern of evolutionary strata (Lahn and Page 1999; Cortez et al. 2014; Zhou et al. 2014). Our recent study in the bird-of-paradise (Xu et al. 2019), however, suggested an alternative PEV-like model not dependent on inversions, which is demonstrated here by the 'strata within strata' pattern of S1. The stratified Z/W sequence divergence levels and repeat content between WS1A and WS1B were likely not caused by inversions indicated by their sequence alignments with the Z Chromosome. It possibly involved the spreading of heterochromatin from the highly heterochromatic S0 to the nearby WS1B region, attracting it closer to the heterochromatin domains of the nuclear periphery (**Fig. 5B**). Under this scenario, the chromatin conformation change would have initiated the heterochromatinization process first in the WS1B region. The complex suit of selective forces still acting on the W Chromosome, including selection to maintain or even upregulate the gene expression levels of female-related

Downloaded from genome.cshlp.org on January 31, 2021 - Published by Cold Spring Harbor Laboratory Press

genes (e.g., *ANXA1*), and that on the TAD boundaries, may have blocked the further spreading of heterochromatin.

Overall, we demonstrated that emu and other ratite species have better preserved the ancestral avian chromosome composition and nuclear architecture than other birds, probably due to their slower evolutionary rate. We speculated that the ancestor of birds or even vertebrates probably had a segregated nuclear architecture like that of most extant birds, with microchromosomes concentrated at the nuclear center, and the macrochromosomes mainly at the nuclear periphery. We showed emu sex chromosome evolution involved alteration of chromatin architecture in the absence of large genomic rearrangements, which may also comprise a critical step at the early stage of sex chromosome evolution of many other species.

Methods

Sample collection and sequencing

All female samples were derived from Copenhagen zoo or emu farms at Fujian and Shaanxi provinces of China. We extracted the high molecular weight (HMW) DNA from the blood of a female emu (*Dromaius novaehollandiae*) and constructed the libraries for SMRT sequencing. In total, 89 SMRT cells were generated on a PacBio RS II platform (Pacific Biosciences), and 88 Gb subreads with an N50 read length of 15.5 kb were produced. HMW DNA of another female emu individual was used to generate a linked-reads library following the protocol on the 10x Genomics Chromium platform. This 10x library was subjected to MGISEQ-2000 sequencing and 156 Gb PE150 reads were collected. HMW DNA of a third female individual was used for constructing Chicago and Hi-C libraries at Dovetail as described previously (Putnam et al. 2016). Briefly, ~500ng of HMW emu gDNA (mean fragment length = 48 kb) was reconstituted into

Downloaded from genome.cshlp.org on January 31, 2021 - Published by Cold Spring Harbor Laboratory Press

chromatin and digested with DpnII. The DNA was then sheared to ~350 bp mean fragment size and sequencing libraries were generated using NEBNext Ultra enzymes and Illumina-compatible adapters. The libraries were sequenced on an Illumina HiSeq X to produce 492 million 2 ×151 bp paired-end reads, which provided 297.50 × physical coverage of the genome (1-100 kb pairs). The libraries were sequenced on an Illumina HiSeq X to produce 211 million 2×151 bp paired-end reads, which provided 13,633.10× physical coverage of the genome (10-10,000 kb pairs).

Genome assembly and annotation

The chromosome-length genome assemblies for the ostrich, greater rhea, southern cassowary, greater prairie chicken, double-crested cormorant, spotted owl, golden eagle, California condor and the green sea turtle as well as the associated Hi-C datasets were downloaded from the DNAzoo Consortium website (dnazoo.org), where they were shared ahead of publication. The assemblies incorporated data from (Wang et al. 2013; Zhang et al. 2014; Burga et al. 2017; Van Den Bussche et al. 2017; Sackton et al. 2019) as well as unpublished datasets. More information is available on the corresponding assembly pages on dnazoo.org. The genomes were assembled using methods described in (Dudchenko et al. 2017).

For the genome assembly of emu, we produced the contig sequences derived from the PacBio subreads with FALCON (20171207) (Chin et al. 2016), using the option '-k24 -e.96 -l2500' for ovlp_daligner, '-e0.75 -l3200' for pa_daligner. We used Purge Haplotigs (20180325) (Roach et al. 2018) to remove the haplotigs after mapping the raw reads with minimap2 (2.10) (Li 2018) against the contigs to estimate the coverage distribution. The contigs were then polished by Racon (1.3.0) (Vaser et al. 2017) with default parameters. Because the S1 of the sex chromosome evolved recently with sequence similarity between the ZW as high as 95%, we

Downloaded from genome.cshlp.org on January 31, 2021 - Published by Cold Spring Harbor Laboratory Press

resolved the ZW haploid assembly by partitioning the Z- and W-derived long-reads for separate assemblies. To do so, first we identified the sex-linked contigs by aligning the contigs to the previous reference Z Chromosome of emu (Xu et al. 2019) and extracted the associated reads. Then we used FALCON to assemble the ZW-linked reads with more stringent overlapping (-k18 -e0.81 -l3000 for pa_daligner and -k24 -e.96 -l4000 for ovlp_daligner) to avoid haploid collapse, where therefore both Z- and W-linked haploid sequences could be assembled. We distinguished the Z- and W-linked contigs according to their sequence similarity with the reference Z and whether they could be mapped with male reads. Then we extracted the reads derived from the Z- and W-linked contigs and assembled them with Canu (1.8) (Koren et al. 2017) respectively. The parameters corOutCoverage=200 correctedErrorRate=0.15 was used for Canu haploid assembly.

The contigs were then scaffolded first with 10x linked reads using Scaff10X (https://github.com/wtsi-hpag/Scaff10X https://github.com/wtsi-hpag/Scaff10X), then with ARCS+LINKS (Warren et al. 2015; Coombe et al. 2018). Finally, the input *de novo* assembly, Chicago library reads, and Dovetail Hi-C library reads were used as input data for HiRise, a software pipeline designed specifically for using proximity ligation data to scaffold genome assemblies (Putnam et al. 2016). The Dovetail Chicago scaffolding was performed with HiRise (version 2.1.2-4e9d295dd196) and the Hi-C scaffolding was performed with HiRise (version v2.1.6-072ca03871cc). A previous version of the Dovetail Genomics HiRise assembler (Putnam et al. 2016) is available as an open-source distribution at https://github.com/DovetailGenomics/HiRise_July2015_GR; however, Dovetail Genomics has not made the HiRise versions used on this assembly available as open-source software at this time. An iterative analysis was conducted: first, PacBio and Chicago library sequences were aligned to the draft input assembly using a modified SNAP read mapper

Downloaded from genome.cshlp.org on January 31, 2021 - Published by Cold Spring Harbor Laboratory Press

(http://snap.cs.berkeley.edu). The Chicago read pairs spanning different scaffolds were analyzed by HiRise to produce a likelihood model for estimating the genomic distances between read pairs, and the model was also used to identify and break putative misjoins, to score prospective joins, and make joins above a threshold. After aligning and scaffolding steps with the Chicago data, Dovetail Hi-C library sequences were aligned for scaffolding the assembly following the similar method. To curate and correct putative assembly errors, we remapped the Hi-C reads and used Juicer tools (Durand et al. 2016) to trim the draft assembly manually. After scaffolding, PacBio long reads were used to close gaps between contigs using the Arrow-corrected PacBio subreads and PBJelly software. The assembly was polished with Illumina reads by Pilon (Walker et al. 2014). Genome completeness was evaluated by BUSCO v3.0.2 (Simão et al. 2015).

For repeat annotation, we first used RepeatModeler (open-1.0.10) to construct the consensus repeat sequence library of the emu. Then the *de novo* library and the repeat consensus library in Repbase (Bao et al. 2015) were merged to annotate all repetitive elements in the emu genome using RepeatMasker (open-4.0.9). We integrated evidence of protein homology, transcriptome, and *de novo* prediction to annotate the protein-coding genes with the MAKER v2.31.10 (Cantarel et al. 2008) pipeline to obtain complete gene models. For the protein homology-based evidence, protein sequences of *Gallus gallus*, *Struthio camelus*, and *Alligator mississippiensis* were downloaded from NCBI. For the transcriptome-based evidence, a genome-guided method was applied to transcriptome assembly. To do so, RNA-seq reads were mapped to the genome with HISAT2 v2.1.0 (Kim et al. 2015) and assembled with StringTie v1.3.4 (Pertea et al. 2015). For the *ab initio* gene predictions, we used AUGUSTUS v3.3 (Stanke et al. 2006) and SNAP (Korf 2004) to predict gene models using the parameters that were trained based on the results of protein homology and transcriptome predictions. Gene functions were assigned

Downloaded from genome.cshlp.org on January 31, 2021 - Published by Cold Spring Harbor Laboratory Press

using DIAMOND v0.9.24 (Buchfink et al. 2015) against UniProtKB (SWISS-PROT + TrEMBL) database with a sensitive mode and an e-value threshold of 1×10^{-5} (--more-sensitive -e 1×10^{-5}).

To annotate the putative centromeres, we searched for tandem repeats across the genome using TRFinder v4.09 (Benson 1999) with the parameters: 2 5 7 80 10 50 2000. We overlapped these findings with the prediction that centromeric regions tend to have lower Hi-C contacts (Muller et al. 2019; Tao et al. 2020), where two putative centromeric units (65bp and 81bp) were identified. The centromere positions for all chromosomes were further manually checked with the reported karyotype of emu chromosomes (Takagi and Sasaki 1974; Kabir 2012). For telomeres, we used the known vertebrate consensus sequence 'TTAGGG/CCCTAA' to search for the clusters of consensus sequences on both strands. Consensus sequence enriched genomic blocks in a 50kb window were then defined as the putative telomere regions.

Comparative genomic analyses

12 chromosome-level genomes of bird species including common ostrich (*Struthio camelus*) (Zhang et al. 2014), greater rhea (*Rhea americana*) (Sackton et al. 2019), Southern cassowary (*Casuarius casuarius*) (Sackton et al. 2019), emu (*Dromaius novaehollandiae*), Peking duck (*Anas platyrhynchos*), greater prairie chicken (*Tympanuchus cupido*) (Johnson et al. 2015), chicken (*Gallus gallus*) (Warren et al. 2017), band-tailed pigeon (*Patagioenas fasciata*) (Murray et al. 2017), double-crested cormorant (*Phalacrocorax auritus*) (Burga et al. 2017), spotted owl (*Strix occidentalis*) (Dudchenko et al. 2017), golden eagle (*Aquila chrysaetos*) (Van Den Bussche et al. 2017), California condor (*Gymnogyps californianus*) (Dudchenko et al. 2017) were aligned to the chromosome-level genome of green sea turtle (C*helonia mydas*) using LAST (http://last.cbrc.jp/). For each investigated species, small syntenic blocks were first merged into

Downloaded from genome.cshlp.org on January 31, 2021 - Published by Cold Spring Harbor Laboratory Press

larger blocks using the custom Perl script. After two rounds of the merging process, blocks whose lengths were shorter than 50kb were discarded. Genomic rearrangements including inversions, translocations and inverted translocations were then detected based on the orientation and position of retained blocks, using the custom Perl scripts. To infer the rearrangement rate for each node of the phylogenetic tree (Claramunt and Cracraft 2015), we followed the principle that if any rearrangement was shared by two closely related species, i.e. the overlap length ratio was larger than 80%, it was present in their common ancestral node. This tracing process was performed iteratively until we reached the ancestral node of all birds. Regarding tolerance for assembly errors, a maximum ratio of one rearrangement missing out of five species was allowed for tracing back to the ancestral node. The rearrangement rate was calculated with the published divergence times for each node (Claramunt and Cracraft 2015).

Hi-C contact analyses

Hi-C read mapping, filtering, correction, binning and normalization were performed by HiC-Pro v2.10.0 (Servant et al. 2015) with the default parameters. In brief, Hi-C reads from the liver of the female were mapped to the respective reference genome and only uniquely mapped reads were kept. Then each uniquely mapped reads were assigned to a restriction enzyme fragment and invalid ligation products were discarded. The interaction contacts were then binned to generate the genome-wide interaction matrix at 5kb, 10kb, 20kb, 40kb, 100kb, 500kb, 1Mb and 10Mb resolution. The ICE (iterative correction and eigenvector decomposition) normalization was then applied to the interaction matrix (Imakaev et al. 2012). Then the *cis*-contacts and *trans*-contacts for each 40kb window were calculated using 40kb normalized interaction matrix. We identified the A/B compartments using the pca.hic function from HiTC package (Servant et al. 2012) with

Downloaded from genome.cshlp.org on January 31, 2021 - Published by Cold Spring Harbor Laboratory Press

default parameters. TADs were identified by HiCExplorer v3.0 (Wolff et al. 2018) with the application hicFindTADs.

Sex chromosomes analyses

The pseudoautosomal region (PAR) was identified based on an equal ratio of male vs female genomic read coverage. S0 was identified based on the half ratio of female vs male genomic read coverage. S1 was identified using the different levels of nucleotide diversity between male and female after both reads of both sexes were mapped to Z Chromosome. To identify the gene repertoire of Chromosome W, we performed a TBLASTN search (Altschul et al. 1990) with the cutoff E-value of 1×10^{-5} against the Chromosome W sequences using the Z-linked genes as query sequences. Aligned sequence fragments were combined into one predicted gene if they belonged to the same query protein. Then each candidate gene region was extended for 5 kb from both ends to predict its open reading frame by GeneWise v2.4.1 (Birney et al. 2004). We annotated the open reading frames as disrupted when GeneWise reported at least one premature stop codon or frame-shift mutation.

Software availability

The software and pipeline for genome *de novo* assembly, annotation and Hi-C analysis, and the custom code and scripts for comparative genomics and sex chromosome evolutionary analysis in this work have been deposited at https://github.com/JhinAir/Emu and in the **Supplemental code**. All figures were plotted in R (R Core Team 2020).

Data access

Downloaded from genome.cshlp.org on January 31, 2021 - Published by Cold Spring Harbor Laboratory Press

The emu assembly has been deposited at DDBJ/ENA/Genbank (https://www.ncbi.nlm.nih.gov/nuccore/JABVCD000000000) under the accession number JABVCD000000000.1. The raw PacBio long reads, 10x linked reads, Chicago and Hi-C linked reads generated in this study have been deposited at the NCBI BioProject database (https://www.ncbi.nlm.nih.gov/bioproject/) under the accession number PRJNA638233. All the DNAzoo genome assemblies are available at https://www.dnazoo.org.

Acknowledgments

We thank Dr. Irina Solovei from Ludwig Maximilian University of Munich for her inspiring discussion and constructive comments on the manuscript, Valentina Peona for sharing the curated repeat library of the emu genome. We thank Erez Lieberman Aiden and Olga Dudchenko and other staff members from DNAzoo for sharing the chromosomal genomes. We thank BGI-Shenzhen for providing the 10x linked reads data of emu. Q.Z. is supported by the National Natural Science Foundation of China (31722050, 31671319, 32061130208), the Natural Science Foundation of Zhejiang Province (LD19C190001) and the European Research Council Starting Grant (grant agreement 677696).

Competing interests

The authors declare that they have no competing interests.

Downloaded from genome.cshlp.org on January 31, 2021 - Published by Cold Spring Harbor Laboratory Press

Figure Legends

Figure 1. Genome assembly of a female emu. a, The PacBio long reads were first used to generate contigs, then used various linkage data of 10x linked reads, Chicago and Hi-C reads to connect the contigs into chromosomal sequences, and finally polished the assembly with corrected PacBio long reads and Illumina reads. **b,** Comparison of scaffold length distribution between droZJU1.0 and droNov1 emu assemblies and ostrich assembly, the latter two of which were generated by Illumina reads. **c,** An example showing that the new droZJU1.0 assembly is more continuous than droNov1. In almost all the cases, one droZJU1.0 contig corresponds to multiple droNov1 contigs, and regions of high repeat content (e.g., LTR/EVRK) coincide with the breakpoints between contigs. **d,** The abundance of each repeat family was normalized to range from 0 and 1 for each chromosome, respectively. **e,** The genomic landscape of emu chromosomes. We showed two macrochromosomes (Chr5 & Chr9), two microchromosomes (Chr18 & Chr20) and the Z/W sex chromosomes for their genomic compositions. I: The PAR (blue), autosomes (green), S1 (orange) show a 2-fold higher female read coverage than the S0 and ChrW (red). II: S1 shows a higher female SNP density than any other genomic regions. III: S0 shows a male-biased expression pattern. IV: Microchromosomes have a higher GC content and a lower TE content than the microchromosomes.

Figure 2. Tempo of avian chromosome evolution. a-d, Genomic synteny between green sea turtle, rattlesnake, American alligator vs. emu and green sea turtle vs. golden eagle, respectively. Chromosome names with blue/red color denote the fused chromosomes in reptiles and the homologous chromosomes of emu, respectively. Each chromosome is

Downloaded from genome.cshlp.org on January 31, 2021 - Published by Cold Spring Harbor Laboratory Press

674 indicated by the first letter of species name and chromosome number. **e,** Chromosomal
675 rearrangements across all major phylogenetic branches of birds. The phylogenetic branches
676 (Jarvis et al. 2014; Claramunt and Cracraft 2015) are colored coded according to the
677 respective average rate of intrachromosomal changes, and the numbers with different colors
678 indicate those of detected chromosomal fissions (red) and fusions (blue). The
679 intrachromosomal rearrangement number per chromosome of birds compared to sea turtle is
680 shown in the heatmap with a different color scale for macro- and microchromosomes given
681 their drastically different size and rearrangement numbers. **f,** Paleognaths show more
682 rearrangements per chromosome than neognaths. ***: $P<0.0005$. **g,** Microchromosomes have
683 a higher rearrangement number per 10Mb length than macrochromosomes. ***: $P<0.0005$. **h,**
684 The distributions of the length ratios of syntenic blocks comparing turtle vs. emu across
685 different types of genomic regions, which indicate the major source of sequence loss in birds
686 is from repeat regions. **i,** The outer dot plot shows the correlation between the overall turtle vs.
687 emu syntenic length ratio per chromosome vs. the size of the chromosome (blue for
688 microchromosomes, red for macrochromosomes). The size of the dots is scaled to the average
689 GC content of each chromosome. The inner plot shows the positive correlation between GC
690 content and turtle vs. emu syntenic length ratio. Each dot represents one syntenic block, with
691 the red ones for the macrochromosome blocks, and the blue ones for the microchromosome
692 blocks. **j**, Microchromosomes have higher turtle vs. birds syntenic length ratios than
693 macrochromosomes, suggesting that microchromosomes experienced more severe sequence
694 loss in birds.
695

Downloaded from genome.cshlp.org on January 31, 2021 - Published by Cold Spring Harbor Laboratory Press

Figure 3. 3D chromatin contacts of macro- and microchromosomes. a, The upper right panel: each blue triangle shows the *cis*-contacts of each chromosome measured by the numbers of Hi-C reads connecting any of the two 40kb regions of the same chromosome. Microchromosomes exhibit more frequent *trans*-contacts measured by the number of Hi-C reads connecting any of the two 40kb regions of two different chromosomes. The color is scaled to the contact strength, i.e., the Hi-C read numbers. The lower left heatmap shows the chromosome-wide average strength of trans-contacts between any of the two chromosomes. The dashed lines demarcate macro- and microchromosomes.**b,** Microchromosomes show more *trans*-contacts than macrochromosomes, after being scaled by chromosome size, and the SDR of ChrW (black dot) shows very few *trans*-contacts. Each dot represents one chromosome (blue for microchromosomes, red for macrochromosomes). **c,** Comparing different types of contacts connecting two active compartments (AA), two inactive compartments (BB), or active and inactive compartments (AB) between macro- and microchromosomes. **d,** Genes that are overlapped with any of the 40kb windows exhibiting the top 10% high levels of trans-contacts ('High_trans') have significantly higher expression levels and lower *tau* values (the lower the *tau* value is, the broader tissue expression the gene has) than the other genes, suggesting these are likely housekeeping genes. ***: $P<0.0005$. **g-h,** From the outer to inner circles: I, Hi-C contacts where the black lines indicate the punctuation of such contacts; II, genomic distribution of 65bp- putative centromeric repeats (blue); III, genomic distribution of 81bp- (red) centromeric repeats. The putative centromeres were annotated by one or two of these three sources of information, corroborated with karyotype information, and then color-coded accordingly on the chromosomes. **e-f,** The average distributions of *trans*- or *cis*-contacts along the chromosomes with the distance away from the

Downloaded from genome.cshlp.org on January 31, 2021 - Published by Cold Spring Harbor Laboratory Press

centromeres, suggesting centromeres have more impacts on the *trans*-contacts of microchromosomes than those of macrochromosomes.

Figure 4. Emu sex chromosome evolution. a, Dot plot of SDR of ChrW, which is segregated into a highly repetitive S0 (red), and only moderately repetitive S1 (orange). Part of the S0 region was reshuffled between the PAR (blue) and the S1, which was inferred based on its homology with the Z-linked S0 region (**Fig. 1E**). **b,** Syntenic plot between the ZW Chromosomes of emu and ostrich and the homologous autosome of green sea turtle, which suggests a W-linked inversion created the emu S1. **c,** There are more genes that have become deleted (grey) or pseudogenes (blue, those containing premature stop codons or frameshift mutations) in S0 than in S1 on the W Chromosome. **d,** The log expression levels (TPM) of single-copy homologous genes in the S0 and S1 regions of Z and W Chromosomes. B: brain; EB: embryonic brain; K: kidney; EK: embryonic kidney; S: spleen; O: ovary. The genes with blue color indicate the pseudogenes. **e,** The left panel plot shows the *cis*-contacts between the PAR and the Z- (blue) and W- (red) linked S1 regions, which exhibit a reduced W-linked contacts than the Z-linked contacts. The x-axis shows the distance to the S1/PAR boundary. The right panel plot indicates higher *cis*-contacts within the W-linked S1 region compared to its homologous Z-linked region. ***: $P<0.0005$. **f,** The W-linked S1 is segregated into two compartments WS1A (left, blue) and WS1B (right, red). From upper to lower: A(blue)/B(red) compartment division based on the Hi-C contact profiles, TAD insulation scores (the lower values correspond to the TAD boundaries), ZW pairwise sequence divergence level, repeat and GC content. The black line indicates the boundary (58.22Mb on ChrZ and 5.28Mb on

Downloaded from genome.cshlp.org on January 31, 2021 - Published by Cold Spring Harbor Laboratory Press

741 ChrW) between WS1A and WS1B. Note, the WS1 has been reversed for the convenience of
742 ZW comparison.
743
744 **Figure 5. Nuclear architectures of avian chromosomes. a,** Microchromosomes of birds,
745 possibly those of vertebrate ancestors are clustered around the nucleolus in the nuclear center,
746 which might be associated with specific centromeric repeats. Such a radial chromosome
747 conformation can promote *trans*-contacts between microchromosomes. By contrast,
748 macrochromosomes are distributed in nuclear peripheries with few trans-contacts. **b,** 3D
749 evolution of emu sex chromosomes. About 150 MY ago, S0 formed in the ancestor of birds,
750 and WS0 started to degenerate. During the heterochromatinization process, WS0 became
751 anchored to the nuclear lamina like any other heterochromatic regions. About 23 MY ago, one
752 W-linked inversion has produced S1. Possibly due to the spreading of heterochromatin of S0
753 (red), the S1 region (WS1B, orange) adjacent to the WS0 underwent heterochromatinization
754 earlier than the other region (WS1A), and evolved larger inactive/B compartments. This
755 increased the *cis*-contacts within S1, but decreased *cis*-contacts between the S1 and PAR. The
756 further spreading of heterochromatin into WS1A may be also halted by the selection on the
757 female-related genes (red dots) located in the WS1A, or the natural selection acting to
758 preserve the TAD boundary between the WS1A and WS1B.
759

Downloaded from genome.cshlp.org on January 31, 2021 - Published by Cold Spring Harbor Laboratory Press

References

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. *J Mol Biol* **215**: 403-410.

Bao W, Kojima KK, Kohany O. 2015. Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mob DNA* **6**: 11.

Benson G. 1999. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res* **27**: 573-580.

Berchtold D, Fesser S, Bachmann G, Kaiser A, Eilert JC, Frohns F, Sadoni N, Muck J, Kremmer E, Eick D et al. 2011. Nuclei of chicken neurons in tissues and three-dimensional cell cultures are organized into distinct radial zones. *Chromosome Res* **19**: 165-182.

Birney E, Clamp M, Durbin RJGr. 2004. GeneWise and genomewise. *Genome Res* **14**: 988-995.

Bowles J, Feng C-W, Miles K, Ineson J, Spiller C, Koopman P. 2016. ALDH1A1 provides a source of meiosis-inducing retinoic acid in mouse fetal ovaries. *Nat Commun* **7**: 10845.

Braasch I, Gehrke AR, Smith JJ, Kawasaki K, Manousaki T, Pasquier J, Amores A, Desvignes T, Batzel P, Catchen J et al. 2016. The spotted gar genome illuminates vertebrate evolution and facilitates human-teleost comparisons. *Nat Genet* **48**: 427-437.

Bromham L. 2011. The genome as a life-history character: why rate of molecular evolution varies between mammal species. *Philos Trans R Soc Lond, B, Biol Sci* **366**: 2503-2513.

Buchfink B, Xie C, Huson DH. 2015. Fast and sensitive protein alignment using DIAMOND. *Nat Methods* **12**: 59-60.

Burga A, Wang W, Ben-David E, Wolf PC, Ramey AM, Verdugo C, Lyons K, Parker PG, Kruglyak L. 2017. A genetic signature of the evolution of loss of flight in the Galapagos cormorant. *Science* **356**: eaal3345.

Burt D. 2002. Origin and evolution of avian microchromosomes. *Cytogenet Genome Res* **96**: 97-112.

Cantarel BL, Korf I, Robb SM, Parra G, Ross E, Moore B, Holt C, Alvarado AS, Yandell M. 2008. MAKER: an easy-to-use annotation pipeline designed for emerging model organism genomes. *Genome Res* **18**: 188-196.

Charlesworth D, Charlesworth B, Marais G. 2005. Steps in the evolution of heteromorphic sex chromosomes. *Heredity* **95**: 118-128.

Chin C-S, Peluso P, Sedlazeck FJ, Nattestad M, Concepcion GT, Clum A, Dunn C, O'Malley R, Figueroa-Balderas R, Morales-Cruz A. 2016. Phased diploid genome assembly with single-molecule real-time sequencing. *Nat Methods* **13**: 1050-1054.

Claramunt S, Cracraft J. 2015. A new time tree reveals Earth history's imprint on the evolution of modern birds. *Sci Adv* **1**: e1501005.

Coombe L, Zhang J, Vandervalk BP, Chu J, Jackman SD, Birol I, Warren RL. 2018. ARKS: chromosome-scale scaffolding of human genome drafts with linked read kmers. *BMC Bioinform* **19**: 1-10.

Cortez D, Marin R, Toledo-Flores D, Froidevaux L, Liechti A, Waters PD, Grützner F, Kaessmann H. 2014. Origins and functional evolution of Y chromosomes across mammals. *Nature* **508**: 488-493.

Cremer M, Von Hase J, Volm T, Brero A, Kreth G, Walter J, Fischer C, Solovei I, Cremer C, Cremer T. 2001. Non-random radial higher-order chromatin arrangements in nuclei of diploid human cells. *Chromosome Res* **9**: 541-567.

Cremer T, Cremer C. 2001. Chromosome territories, nuclear architecture and gene regulation in mammalian cells. *Nat Rev Genet* **2**: 292-301.

Cremer T, Cremer M, Dietzel S, Müller S, Solovei I, Fakan S. 2006. Chromosome territories–a functional nuclear landscape. *Curr Opin Cell Biol* **18**: 307-316.

Croft JA, Bridger JM, Boyle S, Perry P, Teague P, Bickmore WA. 1999. Differences in the localization and morphology of chromosomes in the human nucleus. *J Cell Biol* **145**: 1119-1131.

Downloaded from genome.cshlp.org on January 31, 2021 - Published by Cold Spring Harbor Laboratory Press

Damas J, Kim J, Farre M, Griffin DK, Larkin DM. 2018. Reconstruction of avian ancestral karyotypes reveals differences in the evolutionary history of macro- and microchromosomes. *Genome Biol* **19**: 155.

Dudchenko O, Batra SS, Omer AD, Nyquist SK, Hoeger M, Durand NC, Shamim MS, Machol I, Lander ES, Aiden AP. 2017. De novo assembly of the Aedes aegypti genome using Hi-C yields chromosome-length scaffolds. *Science* **356**: 92-95.

Durand NC, Shamim MS, Machol I, Rao SS, Huntley MH, Lander ES, Aiden EL. 2016. Juicer provides a one-click system for analyzing loop-resolution Hi-C experiments. *Cell Syst* **3**: 95-98.

Elgin SC, Reuter G. 2013. Position-effect variegation, heterochromatin formation, and gene silencing in Drosophila. *Cold Spring Harb Perspect Biol* **5**: a017780.

Falk M, Feodorova Y, Naumova N, Imakaev M, Lajoie BR, Leonhardt H, Joffe B, Dekker J, Fudenberg G, Solovei I. 2019. Heterochromatin drives compartmentalization of inverted and conventional nuclei. *Nature* **570**: 395-399.

Federico C, Cantarella CD, Scavo C, Saccone S, Bed'Hom B, Bernardi G. 2005. Avian genomes: different karyotypes but a similar distribution of the GC-richest chromosome regions at interphase. *Chromosome Res* **13**: 785-793.

Federico C, Scavo C, Cantarella CD, Motta S, Saccone S, Bernardi G. 2006. Gene-rich and gene-poor chromosomal regions have different locations in the interphase nuclei of cold-blooded vertebrates. *Chromosoma* **115**: 123-128.

Feodorova Y, Falk M, Mirny LA, Solovei I. 2020. Viewing Nuclear Architecture through the Eyes of Nocturnal Mammals. *Trends Cell Biol* **30**: 276-289.

Griffin DK, Robertson LB, Tempest HG, Skinner BM. 2007. The evolution of the avian genome as revealed by comparative molecular cytogenetics. *Cytogenet Genome Res* **117**: 64-77.

Habermann FA, Cremer M, Walter J, Kreth G, Von Hase J, Bauer K, Wienberg J, Cremer C, Cremer T, Solovei I. 2001. Arrangements of macro-and microchromosomes in chicken cells. *Chromosome Res* **9**: 569-584.

Hooper DM, Price TD. 2017. Chromosomal inversion differences correlate with range overlap in passerine birds. *Nat Ecol Evol* **1**: 1526-1534.

ICGSC. 2004. Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. *Nature* **432**: 695-716.

Imakaev M, Fudenberg G, McCord RP, Naumova N, Goloborodko A, Lajoie BR, Dekker J, Mirny LA. 2012. Iterative correction of Hi-C data reveals hallmarks of chromosome organization. *Nat Methods* **9**: 999-1003.

Jarvis ED, Mirarab S, Aberer AJ, Li B, Houde P, Li C, Ho SY, Faircloth BC, Nabholz B, Howard JT. 2014. Whole-genome analyses resolve early branches in the tree of life of modern birds. *Science* **346**: 1320-1331.

Johnson J, Novak B, Athrey G, Shapiro B, Phelan R, Brand S. 2015. Whole genome sequence analysis reveals evolutionary history of extinct Heath Hen. In *Proceedings of the Joint Meeting of the American Ornithologist's Union & Cooper Ornithological Society, Norman, OK, USA*, Vol 28.

Kabir A. 2012. Karyotypic analysis of chickens and other birds. *Global Advanced Research Journal of Biotechnology* **1**: 023-032.

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

Kiktev DA, Sheng Z, Lobachev KS, Petes TD. 2018. GC content elevates mutation and recombination rates in the yeast Saccharomyces cerevisiae. *Proc Natl Acad Sci* **115**: E7109-E7118.

Kim D, Langmead B, Salzberg SL. 2015. HISAT: a fast spliced aligner with low memory requirements. *Nat Methods* **12**: 357-360.

Koren S, Walenz BP, Berlin K, Miller JR, Bergman NH, Phillippy AM. 2017. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Res* **27**: 722-736.

Korf I. 2004. Gene finding in novel genomes. *BMC Bioinform* **5**: 59.

Downloaded from genome.cshlp.org on January 31, 2021 - Published by Cold Spring Harbor Laboratory Press

Lahn BT, Page DC. 1999. Four evolutionary strata on the human X chromosome. *Science* **286**: 964-967.
Li H. 2018. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**: 3094-3100.
Lieberman-Aiden E, Van Berkum NL, Williams L, Imakaev M, Ragoczy T, Telling A, Amit I, Lajoie BR, Sabo PJ, Dorschner MO. 2009. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* **326**: 289-293.
Maslova A, Zlotina A, Kosyakova N, Sidorova M, Krasikova A. 2015. Three-dimensional architecture of tandem repeats in chicken interphase nucleus. *Chromosome Res* **23**: 625-639.
Matzke M, Varga F, Berger H, Schernthaner J, Schweizer D, Mayr B, Matzke A. 1990. A 41–42 bp tandemly repeated sequence isolated from nuclear envelopes of chicken erythrocytes is located predominantly on microchromosomes. *Chromosoma* **99**: 131-137.
Meisel RP, Connallon T. 2013. The faster-X effect: integrating theory and data. *Trends Genet* **29**: 537-544.
Muller H, Gil Jr J, Drinnenberg IA. 2019. The impact of centromeres on spatial genome architecture. *Trends Genet* **35**: 565-578.
Murray GG, Soares AE, Novak BJ, Schaefer NK, Cahill JA, Baker AJ, Demboski JR, Doll A, Da Fonseca RR, Fulton TL. 2017. Natural selection shaped the rise and fall of passenger pigeon genomic diversity. *Science* **358**: 951-954.
Nakatani Y, Takeda H, Kohara Y, Morishita S. 2007. Reconstruction of the vertebrate ancestral genome reveals dynamic genome reorganization in early vertebrates. *Genome Res* **17**: 1254-1265.
Nanda I, Schrama D, Feichtinger W, Haaf T, Schartl M, Schmid M. 2002. Distribution of telomeric (TTAGGG)(n) sequences in avian chromosomes. *Chromosoma* **111**: 215-227.
Nishida C, Ishijima J, Ishishita S, Yamada K, Griffin DK, Yamazaki T, Matsuda Y. 2013. Karyotype reorganization with conserved genomic compartmentalization in dot-shaped microchromosomes in the Japanese mountain hawk-eagle (Nisaetus nipalensis orientalis, Accipitridae). *Cytogenet Genome Res* **141**: 284-294.
O'Connor RE, Farre M, Joseph S, Damas J, Kiazim L, Jennings R, Bennett S, Slack EA, Allanson E, Larkin DM et al. 2018. Chromosome-level assembly reveals extensive rearrangement in saker falcon and budgerigar, but not ostrich, genomes. *Genome Biol* **19**: 171.
O'Connor RE, Kiazim L, Skinner B, Fonseka G, Joseph S, Jennings R, Larkin DM, Griffin DK. 2019. Patterns of microchromosome organization remain highly conserved throughout avian evolution. *Chromosoma* **128**: 21-29.
Ohno S, Muramoto J, Stenius C, Christian L, Kittrell W, Atkin N. 1969. Microchromosomes in holocephalian, chondrostean and holostean fishes. *Chromosoma* **26**: 35-40.
Perry BW, Schield DR, Adams RH, Castoe TA. 2020. Microchromosomes exhibit distinct features of vertebrate chromosome structure and function with underappreciated ramifications for genome evolution. *Mol Biol Evol* doi:10.1093/molbev/msaa253.
Pertea M, Pertea GM, Antonescu CM, Chang T-C, Mendell JT, Salzberg SL. 2015. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat Biotechnol* **33**: 290-295.
Putnam NH, O'Connell BL, Stites JC, Rice BJ, Blanchette M, Calef R, Troll CJ, Fields A, Hartley PD, Sugnet CW. 2016. Chromosome-scale shotgun assembly using an in vitro method for long-range linkage. *Genome Res* **26**: 342-350.
R Core Team. 2020. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.
Rhie A, McCarthy SA, Fedrigo O, Damas J, Formenti G, Koren S, Uliano-Silva M, Chow W, Fungtammasan A, Gedman GL. 2020. Towards complete and error-free genome assemblies of all vertebrate species. *bioRxiv* doi:10.1101/2020.05.22.110833.
Rice ES, Kohno S, John JS, Pham S, Howard J, Lareau LF, O'Connell BL, Hickey G, Armstrong J, Deran A. 2017. Improved genome assembly of American alligator genome reveals conserved architecture of estrogen signaling. *Genome Res* **27**: 686-696.
Roach MJ, Schmidt SA, Borneman AR. 2018. Purge Haplotigs: allelic contig reassignment for third-gen diploid genome assemblies. *BMC Bioinform* **19**: 460.

Downloaded from genome.cshlp.org on January 31, 2021 - Published by Cold Spring Harbor Laboratory Press

Romiguier J, Ranwez V, Douzery EJ, Galtier N. 2010. Contrasting GC-content dynamics across 33 mammalian genomes: relationship with life-history traits and chromosome sizes. *Genome Res* **20**: 1001-1009.

Sacerdot C, Louis A, Bon C, Berthelot C, Roest Crollius H. 2018. Chromosome evolution at the origin of the ancestral vertebrate genome. *Genome Biol* **19**: 166.

Sackton TB, Grayson P, Cloutier A, Hu Z, Liu JS, Wheeler NE, Gardner PP, Clarke JA, Baker AJ, Clamp M. 2019. Convergent regulatory evolution and loss of flight in paleognathous birds. *Science* **364**: 74-78.

Schield DR, Card DC, Hales NR, Perry BW, Pasquesi GM, Blackmon H, Adams RH, Corbin AB, Smith CF, Ramesh B et al. 2019. The origins and evolution of chromosomes, dosage compensation, and mechanisms underlying venom regulation in snakes. *Genome Res* **29**: 590-601.

Servant N, Lajoie BR, Nora EP, Giorgetti L, Chen C-J, Heard E, Dekker J, Barillot E. 2012. HiTC: exploration of high-throughput 'C'experiments. *Bioinformatics* **28**: 2843-2844.

Servant N, Varoquaux N, Lajoie BR, Viara E, Chen C-J, Vert J-P, Heard E, Dekker J, Barillot E. 2015. HiC-Pro: an optimized and flexible pipeline for Hi-C data processing. *Genome Biol* **16**: 259.

Shang WH, Hori T, Toyoda A, Kato J, Popendorf K, Sakakibara Y, Fujiyama A, Fukagawa T. 2010. Chickens possess centromeres with both extended tandem repeats and short non-tandem-repetitive sequences. *Genome Res* **20**: 1219-1228.

Simakov O, Marletaz F, Yue JX, O'Connell B, Jenkins J, Brandt A, Calef R, Tung CH, Huang TK, Schmutz J et al. 2020. Deeply conserved synteny resolves early events in vertebrate evolution. *Nat Ecol Evol* **4**: 820-830.

Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. 2015. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31**: 3210-3212.

Smith JJ, Timoshevskaya N, Ye C, Holt C, Keinath MC, Parker HJ, Cook ME, Hess JE, Narum SR, Lamanna F et al. 2018. The sea lamprey germline genome provides insights into programmed genome rearrangement and vertebrate evolution. *Nat Genet* **50**: 270-277.

Stadler S, Schnapp V, Mayer R, Stein S, Cremer C, Bonifer C, Cremer T, Dietzel S. 2004. The architecture of chicken chromosome territories changes during differentiation. *BMC Cell Biol* **5**: 1-17.

Stanke M, Keller O, Gunduz I, Hayes A, Waack S, Morgenstern B. 2006. AUGUSTUS: ab initio prediction of alternative transcripts. *Nucleic Acids Res* **34**: W435-W439.

Szabo Q, Bantignies F, Cavalli G. 2019. Principles of genome folding into topologically associating domains. *Sci Adv* **5**: eaaw1668.

Takagi N, Sasaki M. 1974. A phylogenetic study of bird karyotypes. *Chromosoma* **46**: 91-120.

Tanabe H, Küpper K, Ishida T, Neusser M, Mizusawa H. 2005. Inter-and intra-specific gene-density-correlated radial chromosome territory arrangements are conserved in Old World monkeys. *Cytogenet Genome Res* **108**: 255-261.

Tanabe H, Müller S, Neusser M, von Hase J, Calcagno E, Cremer M, Solovei I, Cremer C, Cremer T. 2002. Evolutionary conservation of chromosome territory arrangements in cell nuclei from higher primates. *Proc Natl Acad Sci* **99**: 4424-4429.

Tao W, Xu L, Zhao L, Zhu Z, Wu X, Min Q, Wang D, Zhou Q. 2020. High-quality chromosome-level genomes of two tilapia species reveal their evolution of repeat sequences and sex chromosomes. *Mol Ecol Resour* doi:10.1111/1755-0998.13273.

Tegelström H, Ryttman H. 1981. Chromosomes in birds (Aves): evolutionary implications of macro-and microchromosome numbers and lengths. *Hereditas* **94**: 225-233.

Uno Y, Nishida C, Tarui H, Ishishita S, Takagi C, Nishimura O, Ishijima J, Ota H, Kosaka A, Matsubara K et al. 2012. Inference of the protokaryotypes of amniotes and tetrapods and the evolutionary processes of microchromosomes from comparative gene mapping. *PLoS One* **7**: e53027.

Downloaded from genome.cshlp.org on January 31, 2021 - Published by Cold Spring Harbor Laboratory Press

Van Den Bussche RA, Judkins ME, Montague MJ, Warren WC. 2017. A resource of genome-wide single nucleotide polymorphisms (Snps) for the conservation and management of golden eagles. *J Raptor Res* **51**: 368-377.

Vaser R, Sović I, Nagarajan N, Šikić M. 2017. Fast and accurate de novo genome assembly from long uncorrected reads. *Genome Res* **27**: 737-746.

Venkatesh B, Lee AP, Ravi V, Maurya AK, Lian MM, Swann JB, Ohta Y, Flajnik MF, Sutoh Y, Kasahara M. 2014. Elephant shark genome provides unique insights into gnathostome evolution. *Nature* **505**: 174-179.

Voss SR, Kump DK, Putta S, Pauly N, Reynolds A, Henry RJ, Basa S, Walker JA, Smith JJ. 2011. Origin of amphibian and avian chromosomes by fission, fusion, and retention of ancestral chromosomes. *Genome Res* **21**: 1306-1312.

Vurture GW, Sedlazeck FJ, Nattestad M, Underwood CJ, Fang H, Gurtowski J, Schatz MC. 2017. GenomeScope: fast reference-free genome profiling from short reads. *Bioinformatics* **33**: 2202-2204.

Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, Sakthikumar S, Cuomo CA, Zeng Q, Wortman J, Young SK. 2014. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PloS One* **9**: e112963.

Wang Z, Pascual-Anaya J, Zadissa A, Li W, Niimura Y, Huang Z, Li C, White S, Xiong Z, Fang D. 2013. The draft genomes of soft-shell turtle and green sea turtle yield insights into the development and evolution of the turtle-specific body plan. *Nat Genet* **45**: 701-706.

Wang Z, Zhang J, Xu X, Witt C, Deng Y, Chen G, Meng G, Feng S, Szekely T, Zhang G. 2019. Phylogeny, transposable element and sex chromosome evolution of the basal lineage of birds. *bioRxiv* doi:10.1101/750109.

Wang Z, Zhang J, Yang W, An N, Zhang P, Zhang G, Zhou Q. 2014. Temporal genomic evolution of bird sex chromosomes. *BMC Evol Biol* **14**: 250.

Warren RL, Yang C, Vandervalk BP, Behsaz B, Lagman A, Jones SJ, Birol I. 2015. LINKS: Scalable, alignment-free scaffolding of draft genomes with long reads. *Gigascience* **4**: s13742-13015-10076-13743.

Warren WC, Hillier LW, Tomlinson C, Minx P, Kremitzki M, Graves T, Markovic C, Bouk N, Pruitt KD, Thibaud-Nissen F et al. 2017. A New Chicken Genome Assembly Provides Insight into Avian Genome Structure. *G3: Genes, Genomes, Genetics* **7**: 109-117.

Weber CC, Boussau B, Romiguier J, Jarvis ED, Ellegren H. 2014. Evidence for GC-biased gene conversion as a driver of between-lineage differences in avian base composition. *Genome Biol* **15**: 1-16.

Wolff J, Bhardwaj V, Nothjunge S, Richard G, Renschler G, Gilsbach R, Manke T, Backofen R, Ramírez F, Grüning BA. 2018. Galaxy HiCExplorer: a web server for reproducible Hi-C data analysis, quality control and visualization. *Nucleic Acids Res* **46**: W11-W16.

Xu L, Auer G, Peona V, Suh A, Deng Y, Feng S, Zhang G, Blom MP, Christidis L, Prost S. 2019. Dynamic evolutionary history and gene content of sex chromosomes across diverse songbirds. *Nat Ecol Evol* **3**: 834-844.

Xu L, Zhou Q. 2020. Female-specific and dosage selections restore genes through transpositions onto the degenerated songbird W chromosomes. *Genes* **11**: 1126.

Yamada K, Nishida-Umehara C, Matsuda Y. 2002. Characterization and chromosomal distribution of novel satellite DNA sequences of the lesser rhea (Pterocnemia pennata) and the greater rhea (Rhea americana). *Chromosome Res* **10**: 513-523.

Yamada K, Nishida-Umehara C, Matsuda Y. 2005. Molecular and cytogenetic characterization of site-specific repetitive DNA sequences in the Chinese soft-shelled turtle (Pelodiscus sinensis, Trionychidae). *Chromosome Res* **13**: 33-46.

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

Downloaded from genome.cshlp.org on January 31, 2021 - Published by Cold Spring Harbor Laboratory Press

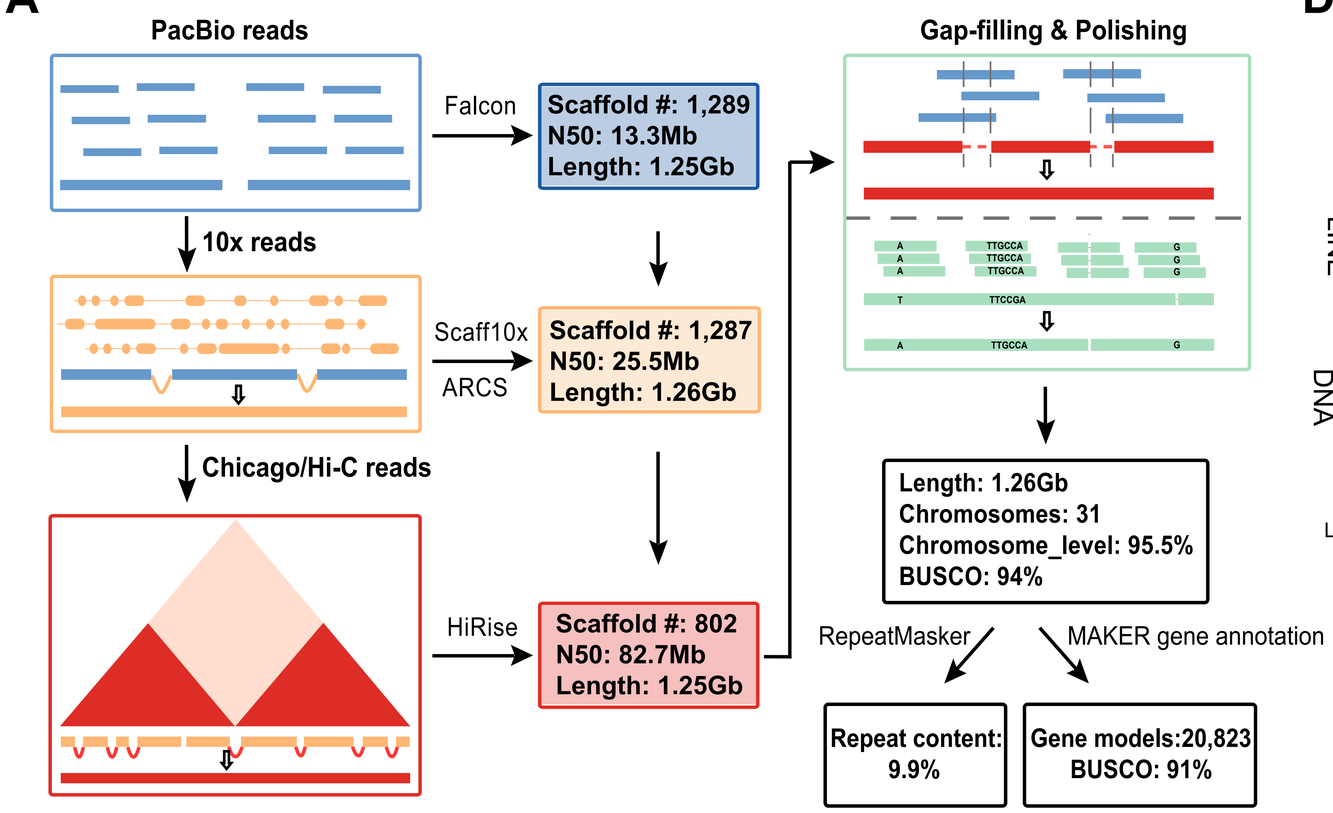
Zhang J, Li C, Zhou Q, Zhang G. 2015. Improving the ostrich genome assembly using optical mapping data. *Gigascience* **4**: s13742-13015-10062-13749.

Zhou Q, Ellison CE, Kaiser VB, Alekseyenko AA, Gorchakov AA, Bachtrog D. 2013. The epigenome of evolving Drosophila neo-sex chromosomes: dosage compensation and heterochromatin formation. *PLoS Biol* **11**: e1001711.

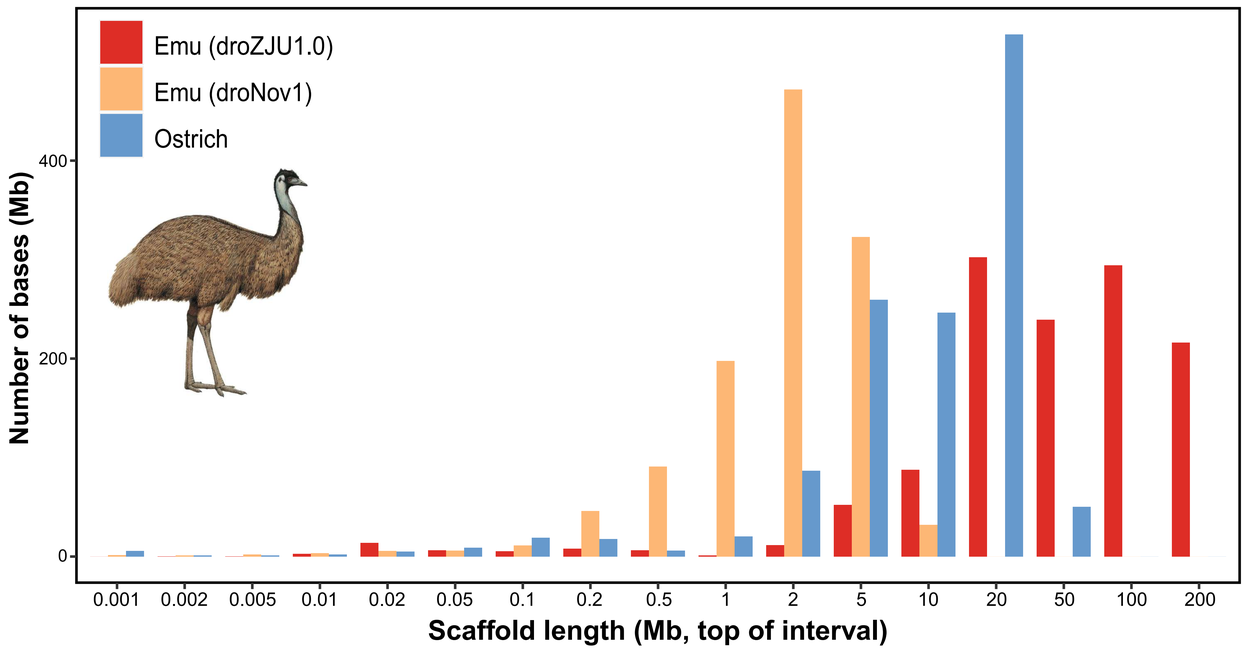
Zhou Q, Zhang J, Bachtrog D, An N, Huang Q, Jarvis ED, Gilbert MTP, Zhang G. 2014. Complex evolutionary trajectories of sex chromosomes across bird taxa. *Science* **346**.

Zhou Z, Li M, Cheng H, Fan W, Yuan Z, Gao Q, Xu Y, Guo Z, Zhang Y, Hu J et al. 2018. An intercross population study reveals genes associated with body size and plumage color in ducks. *Nat Commun* **9**: 2648.

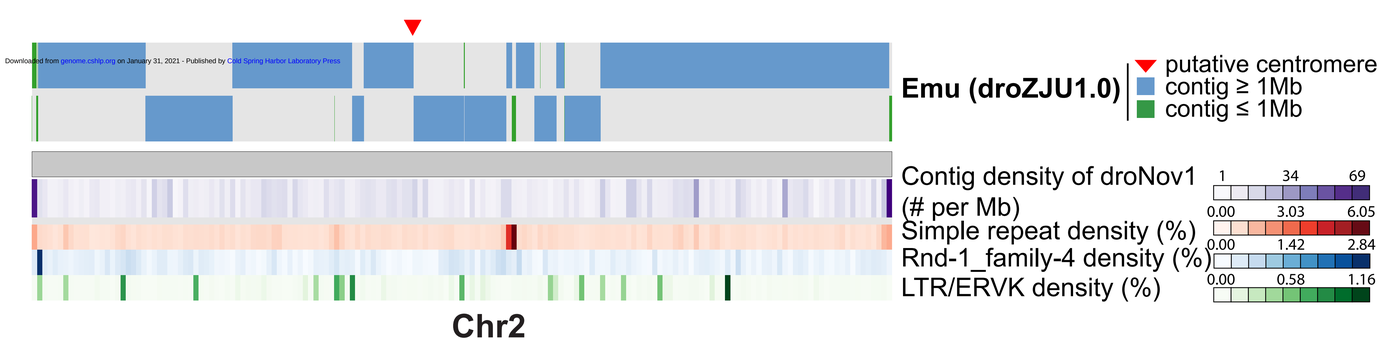
A


PacBio reads
Falcon
Scaffold #: 1,289
N50: 13.3Mb
Length: 1.25Gb
10x reads
Scaff10x
ARCS
Scaffold #: 1,287
N50: 25.5Mb
Length: 1.26Gb
Chicago/Hi-C reads
HiRise
Scaffold #: 802
N50: 82.7Mb
Length: 1.25Gb
Gap-filling & Polishing
Length: 1.26Gb
Chromosomes: 31
Chromosome_level: 95.5%
BUSCO: 94%
RepeatMasker
MAKER gene annotation
Repeat content:
9.9%
Gene models:20,823
BUSCO: 91%

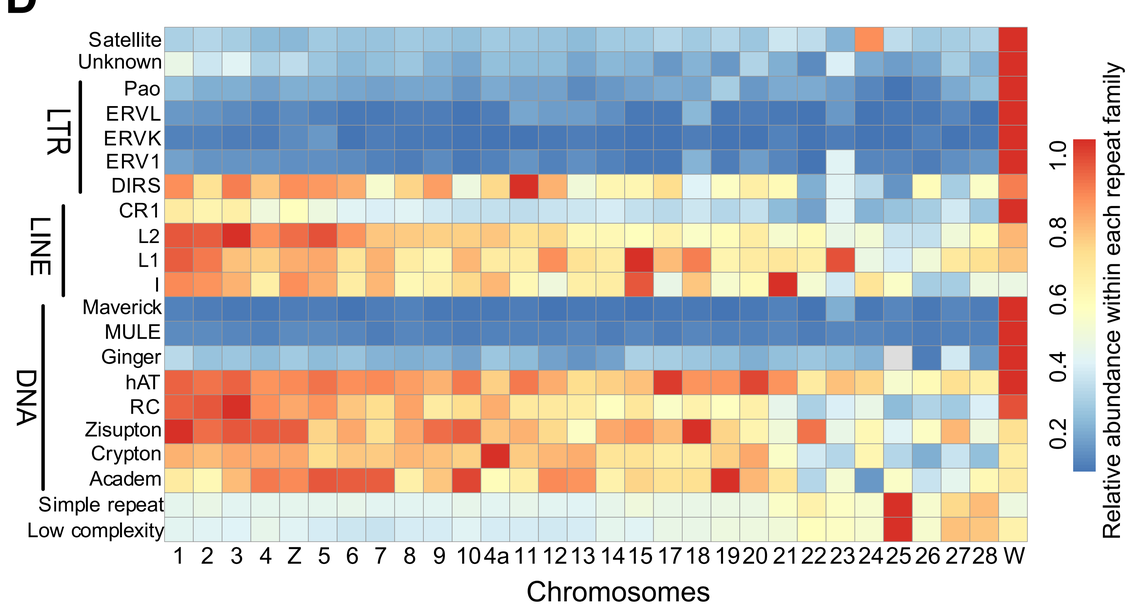
B


Emu (droZJU1.0)
Emu (droNov1)
Ostrich
Number of bases (Mb)
0
200
400
0.001
0.002
0.005
0.01
0.02
0.05
0.1
0.2
0.5
1
2
5
10
20
50
100
200
Scaffold length (Mb, top of interval)

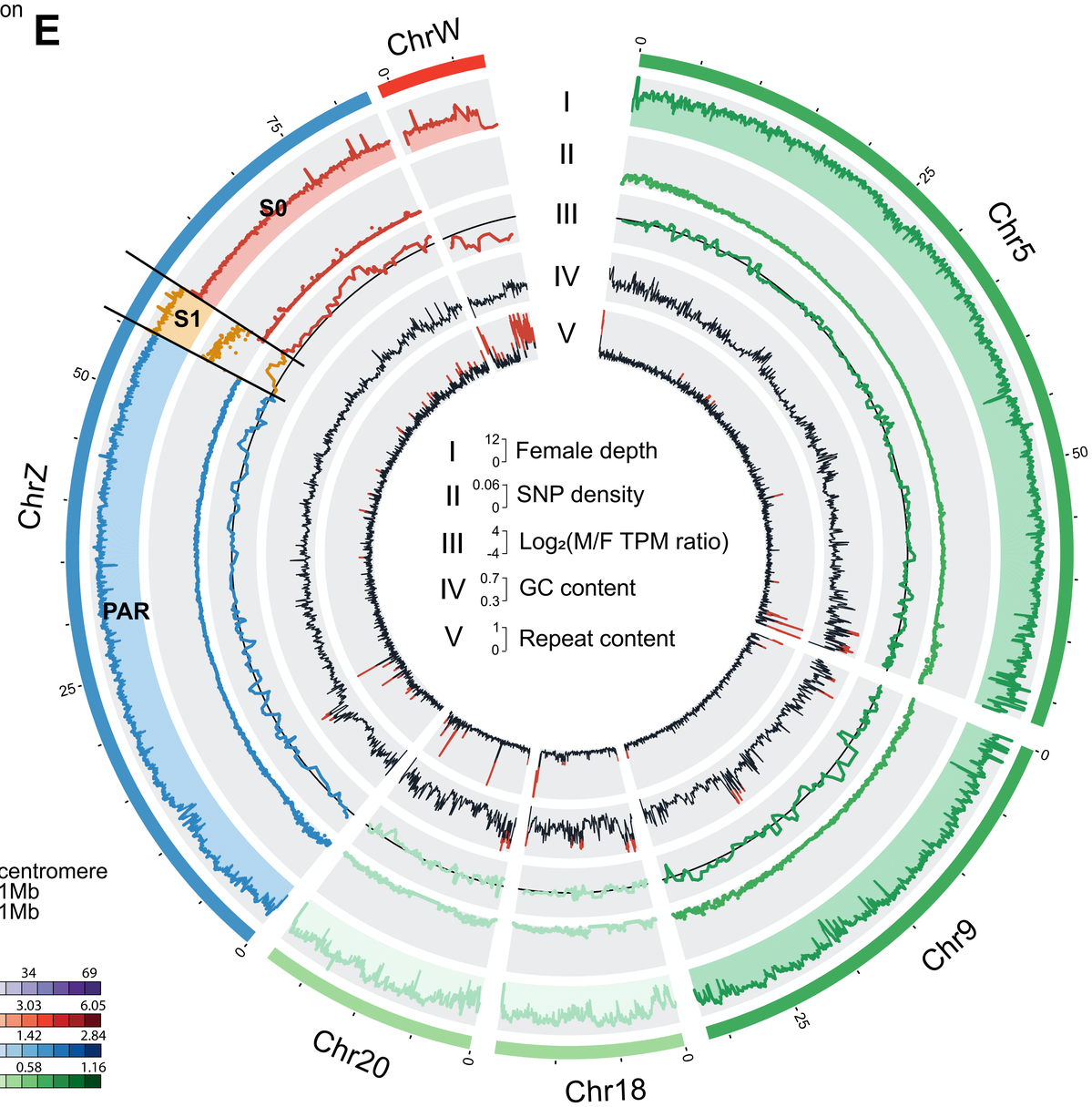
C

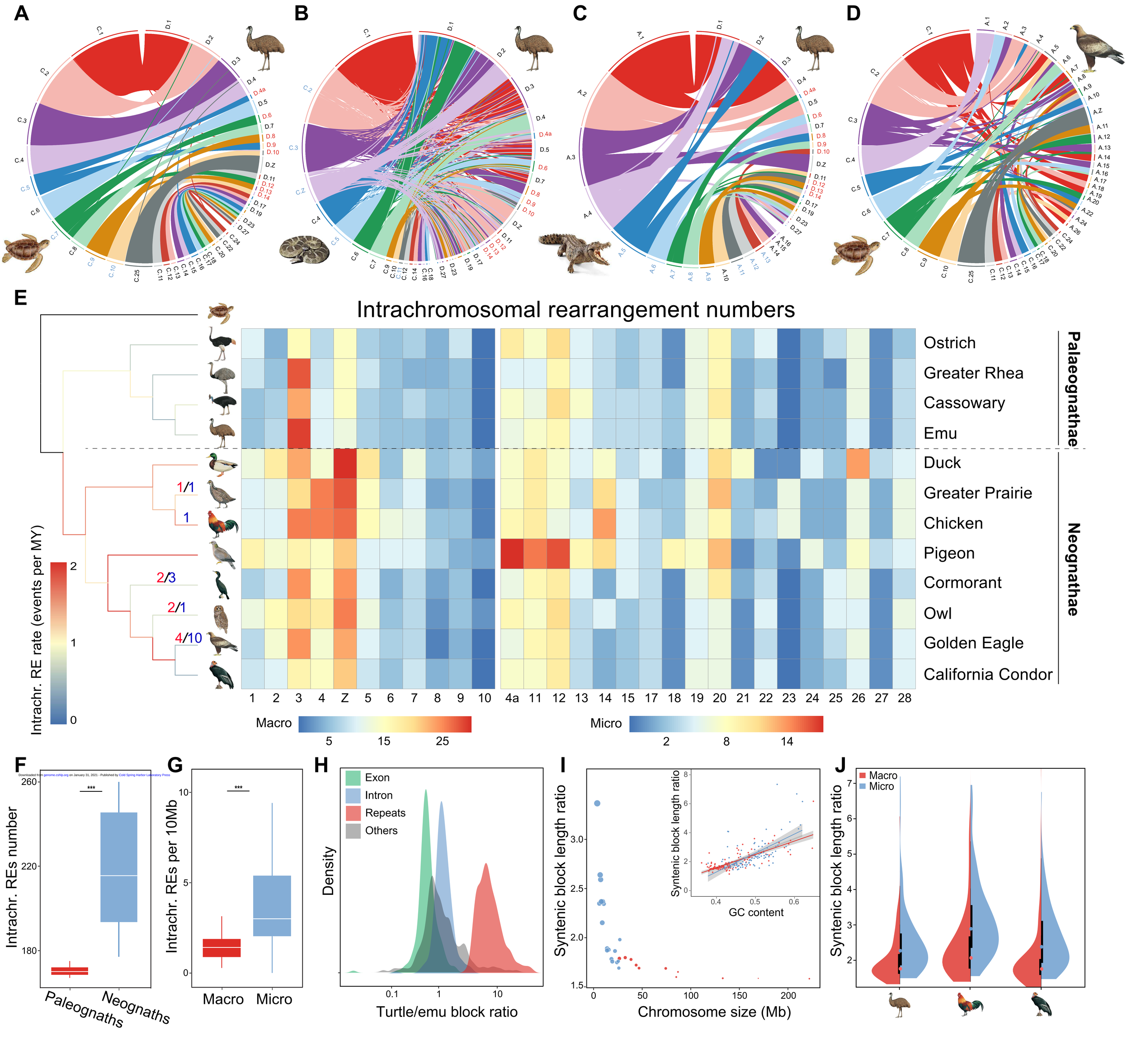

Emu (droZJU1.0)
putative centromere
contig ≥ 1Mb
contig ≤ 1Mb
Contig density of droNov1
(# per Mb)
1
34
69
Simple repeat density (%)
0.00
3.03
6.05
Rnd-1_family-4 density (%)
0.00
1.42
2.84
LTR/ERVK density (%)
0.00
0.58
1.16
Chr2

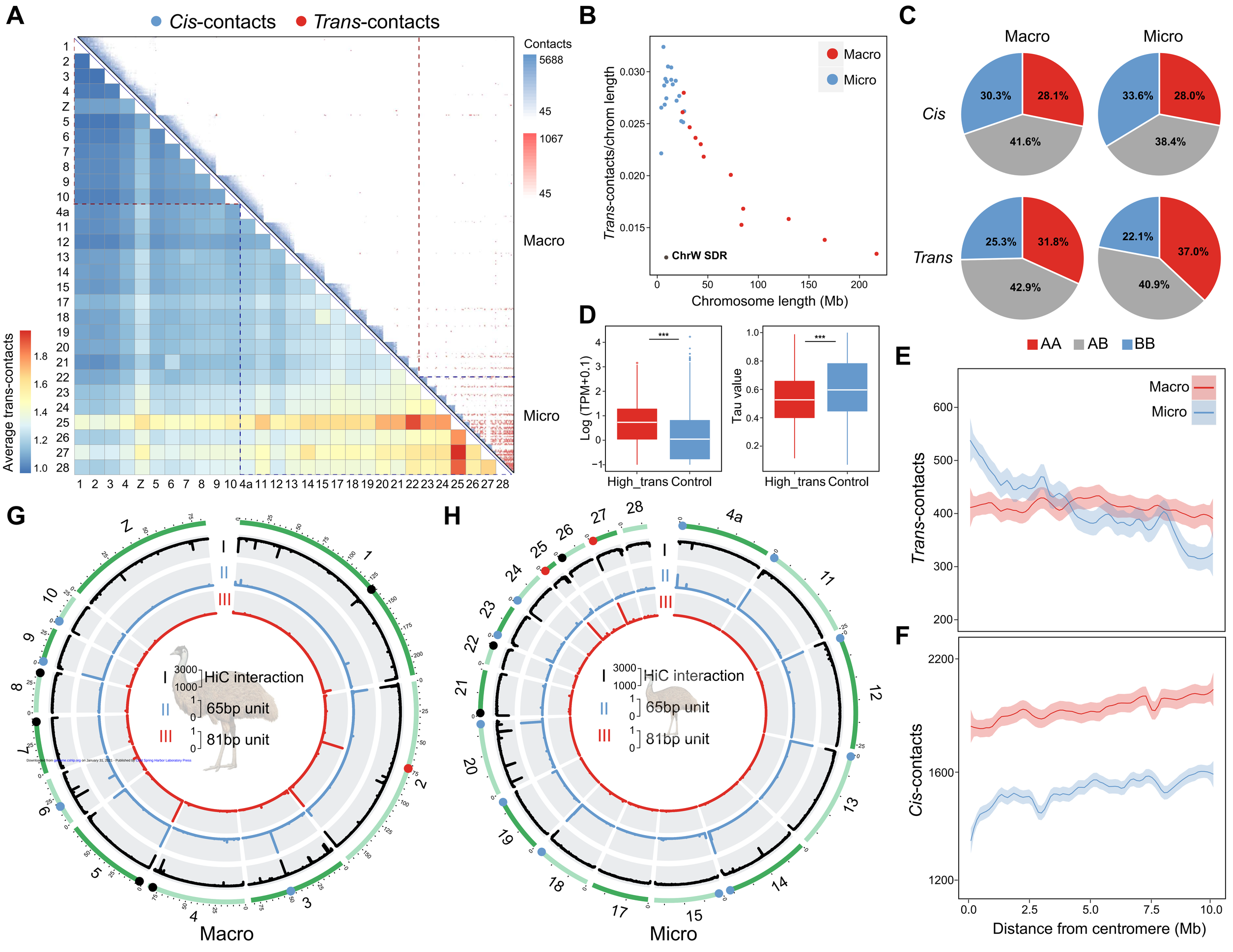
D


Satellite
Unknown
LTR
Pao
ERVL
ERVK
ERV1
DIRS
LINE
CR1
L2
L1
I
DNA
Maverick
MULE
Ginger
hAT
RC
Zisupton
Crypton
Academ
Simple repeat
Low complexity
1 2 3 4 Z 5 6 7 8 9 10 4a 11 12 13 14 15 17 18 19 20 21 22 23 24 25 26 27 28 W
Chromosomes
Relative abundance within each repeat family
0.2
0.4
0.6
0.8
1.0

E

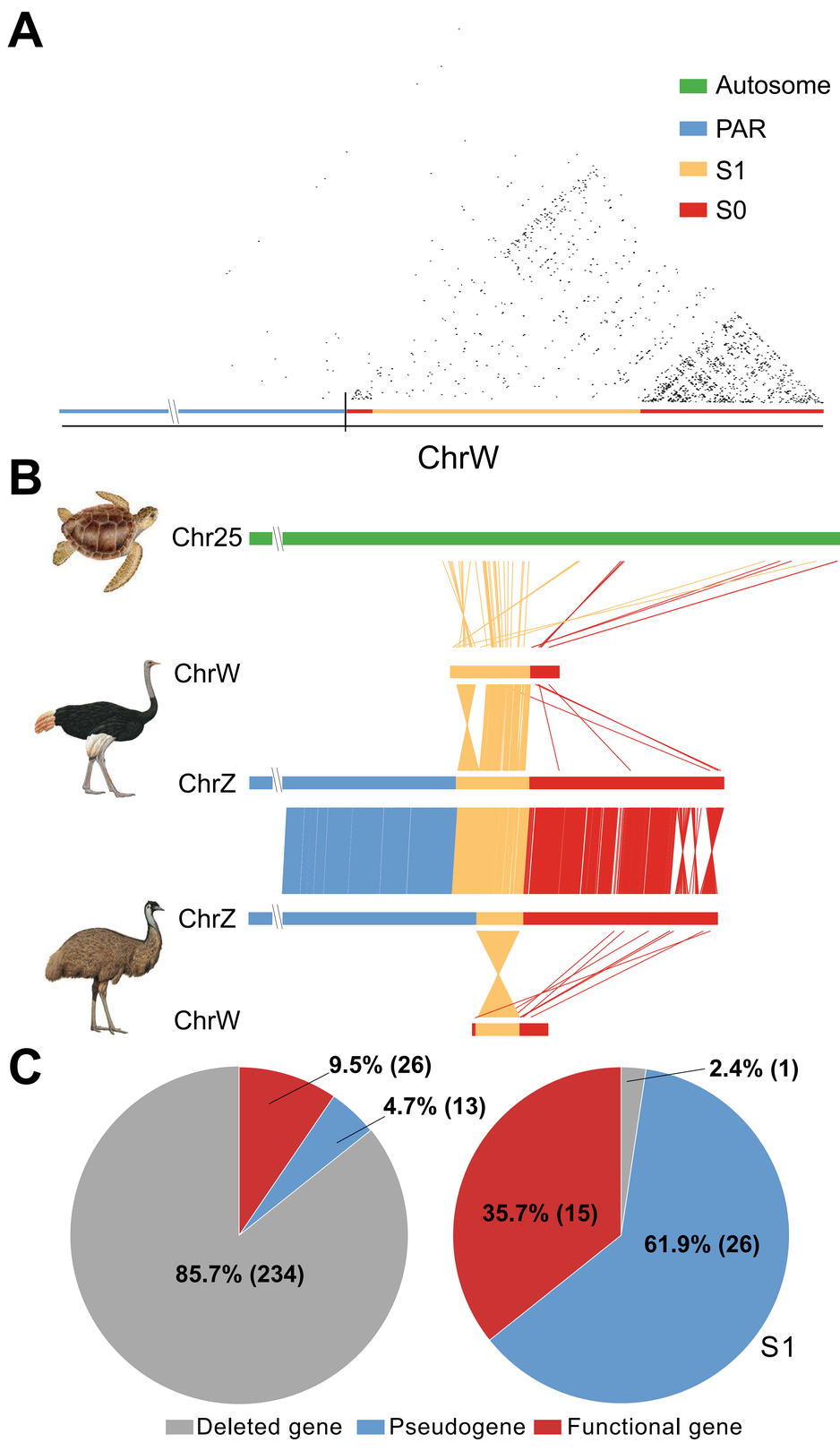

ChrW
S0
S1
ChrZ
PAR
Chr5
Chr9
Chr18
Chr20
I
II
III
IV
V
Female depth
SNP density
Log₂(M/F TPM ratio)
GC content
Repeat content

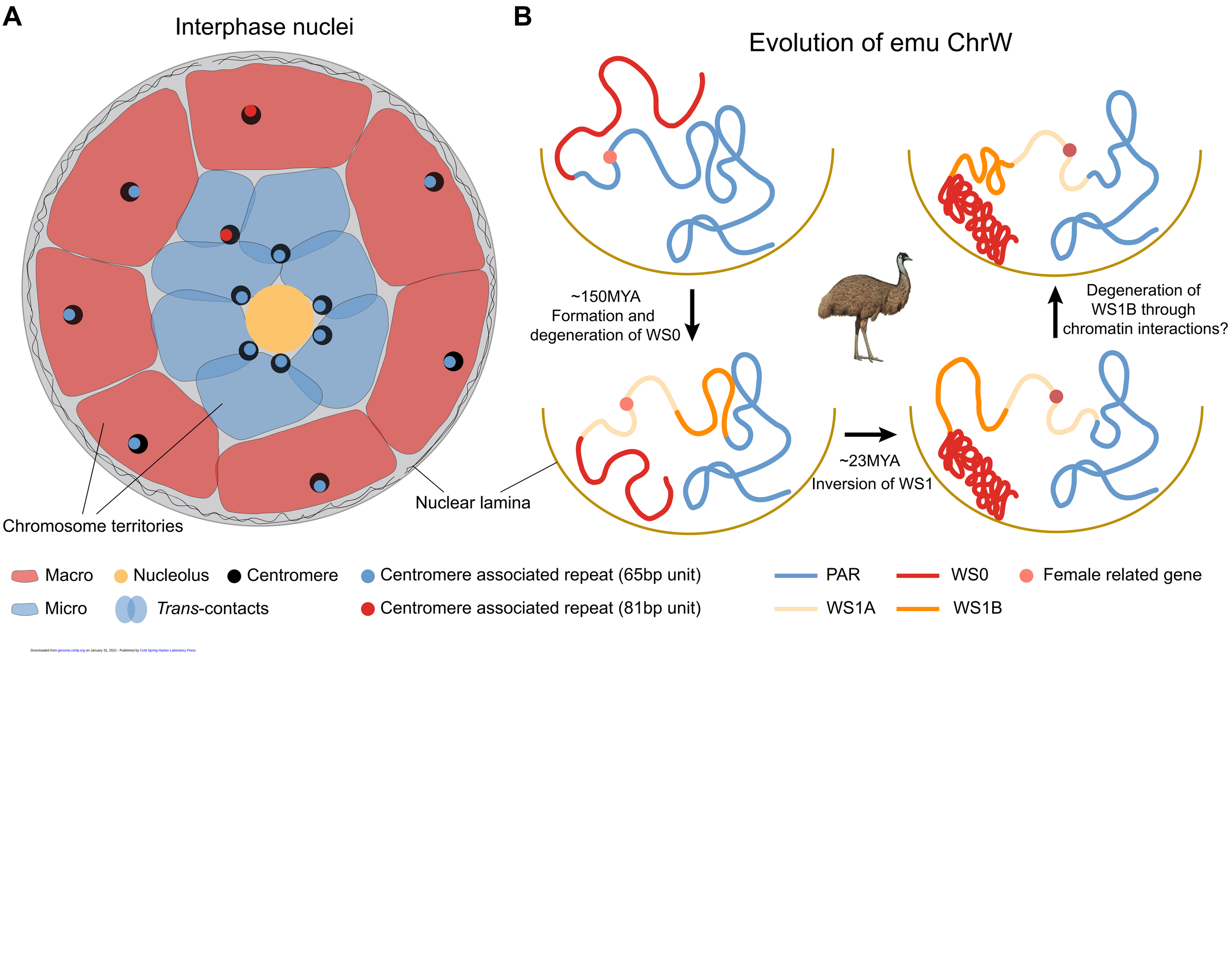

A
B
C
D
E
Intrachromosomal rearrangement numbers
Ostrich
Greater Rhea
Cassowary
Emu
Duck
Greater Prairie
Chicken
Pigeon
Cormorant
Owl
Golden Eagle
California Condor
Palaeognathae
Neognathae
1/1
1
2/3
2/1
4/10
Intrachr. RE rate (events per MY)
Macro
Micro
F
Intrachr. REs number
Paleognaths
Neognaths
G
Intrachr. REs per 10Mb
H
Exon
Intron
Repeats
Others
Density
Turtle/emu block ratio
I
Syntenic block length ratio
GC content
Chromosome size (Mb)
J
Macro
Micro
Syntenic block length ratio



Downloaded from genome.cshlp.org on January 31, 2021 - Published by Cold Spring Harbor Laboratory Press

Downloaded from genome.cshlp.org on January 31, 2021 - Published by Cold Spring Harbor Laboratory Press





Downloaded from genome.cshlp.org on January 31, 2021 - Published by Cold Spring Harbor Laboratory Press

Downloaded from genome.cshlp.org on January 31, 2021 - Published by Cold Spring Harbor Laboratory Press

A new emu genome illuminates the evolution of genome configuration and nuclear architecture of avian chromosomes

Jing Liu, Zongji Wang, Jing Li, et al.

Genome Res. published online January 6, 2021
Access the most recent version at doi:10.1101/gr.271569.120

P<P Published online January 6, 2021 in advance of the print journal.

Accepted Manuscript Peer-reviewed and accepted for publication but not copyedited or typeset; accepted manuscript is likely to differ from the final, published version.

Creative Commons License This article is distributed exclusively by Cold Spring Harbor Laboratory Press for the first six months after the full-issue publication date (see http://genome.cshlp.org/site/misc/terms.xhtml). After six months, it is available under a Creative Commons License (Attribution-NonCommercial 4.0 International), as described at http://creativecommons.org/licenses/by-nc/4.0/.

Email Alerting Service Receive free email alerts when new articles cite this article - sign up in the box at the top right corner of the article or **click here.**

Advance online articles have been peer reviewed and accepted for publication but have not yet appeared in the paper journal (edited, typeset versions may be posted when available prior to final publication). Advance online articles are citable and establish publication priority; they are indexed by PubMed from initial publication. Citations to Advance online articles must include the digital object identifier (DOIs) and date of initial publication.

To subscribe to *Genome Research* go to:
http://genome.cshlp.org/subscriptions

Published by Cold Spring Harbor Laboratory Press