Non-canonical DNA and sequencing challenges in bird genomes

Linnéa Smeds[1], Jacob Sieg[1], Simona Secomandi[2,3], Chul Lee[3], Marco Sollitto[2,4], Jack A. Medico[2], Francesca Chiaromonte[5,6], Erich D. Jarvis[2,3,7], Giulio Formenti[2], Kateryna D. Makova[1*]

Affiliations:
[1]Department of Biology, Penn State University, University Park, PA 16802, USA
[2]The Vertebrate Genome Laboratory, The Rockefeller University, NY 10065, USA
[3]The Laboratory of Neurogenetics of Language, The Rockefeller University, NY 10065, USA
[4]Department of Biology, University of Florence, 50019, Sesto Fiorentino, FI, Italy
[5]Department of Statistics, Penn State University, University Park, PA 16802, USA
[6]L'EMbeDS, Sant'Anna School of Advanced Studies, 56127 Pisa, Italy
[7]Howard Hughes Medical Institute, Chevy Chase, MD 20815, USA
*Corresponding author

Email addresses:
Linnéa Smeds: lbs5874@psu.edu
Jacob Sieg: jus841@psu.edu
Simona Secomandi: ssecomandi@rockefeller.edu
Chul Lee: clee03@rockefeller.edu
Marco Sollitto: msollitto@rockefeller.edu
Jack A. Medico: amedico@rockefeller.edu
Francesca Chiaromonte: fxc11@psu.edu
Erich D. Jarvis: ejarvis@rockefeller.edu
Giulio Formenti: gformenti@rockefeller.edu
Kateryna D. Makova: kdm16@psu.edu

Abstract

Non-canonical (non-B) DNA motifs are sequences that can fold into structures (e.g., G-quadruplexes and Z-DNA) distinct from the canonical right-handed helix. In mammals, these structures regulate gene expression, act as mutation hotspots, and are associated with cancer, yet they remain undercharacterized in other species. Because non-B DNA motifs are difficult to sequence, many are absent from incomplete genome assemblies, limiting functional analyses. Here, we present the first comprehensive analysis of non-B DNA motifs in birds, using the telomere-to-telomere genome of zebra finch, the near-complete chicken genome, and high-quality genomes of six additional bird species. We show that, first, unlike in mammals, the non-B DNA landscape in birds differs markedly among chromosome groups: gene-rich and extremely small dot chromosomes show the highest coverage (15.1–30.1% in zebra finch), microchromosomes—intermediate coverage (6.4–18.1%), and macrochromosomes—the lowest (5.9–6.9%). Non-B DNA motif coverage on dot chromosomes negatively correlates with PacBio sequencing depth, potentially explaining their assembly challenges. Second, similar to mammals, in zebra finch, G-quadruplexes are enriched at promoters and 5′UTRs, implying regulatory roles. We experimentally validated four common G-quadruplexes and predicted others using long-read methylation data. Overall, non-B DNA distribution reflects distinct features of avian genome architecture, suggests a role in gene regulation, and informs strategies for complete bird genome sequencing.

Introduction

Recent improvements in genome assembly, driven by advances in long-read sequencing and computational algorithms, have advanced the goal of generating *complete and gapless* reference genomes, i.e., telomere-to-telomere (T2T) assemblies[1]. The first published T2T genome was that of a human female[2], followed by the complete human Y chromosome[3], as well as the sex chromosomes[4] and autosomes[5] of several ape species. T2T genomes have proven to be essential for studies of segmental duplications[6,7], transposable elements[8], satellites[9], and non-B DNA motifs[10,11]. It was demonstrated that non-B DNA motifs—sequences that have the potential to form non-canonical, or non-B, DNA—were greatly enriched in the newly added sequences of ape T2T genomes compared to previous assemblies[10]. Non-B DNA structures have recently emerged as novel functional elements[12] and drivers of genome evolution[13]. Determining the complete repertoire of non-B DNA predicted motifs is crucial for understanding the roles of these structures in genomes, which range from being mutation hotspots and promoting genome instability[13,14] to regulating gene expression[15,16]. Despite substantial progress made in analyzing non-B DNA in human and great ape T2T genomes[10,11], the occurrence of non-B DNA motifs in complete genomes of other species has remained largely unexplored, in part due to the lack of T2T assemblies. Moreover, most of our knowledge of non-B DNA function comes from the analysis of genomes of mammals and model organisms, such as yeast[15]. Non-B DNA functions in the genomes of other species are still critically understudied.

Bird genomes have been included in several large-scale evolutionary studies of non-B DNA motifs, investigating either certain motif types or particular genomic regions. For example, a recent study investigated non-B DNA in UnTranslated Regions (UTRs) across hundreds of eukaryote species including 23 bird species and reported that G-quadruplexes (G4s) were the predominant non-B DNA motif type in bird UTRs, just as in mammals and reptiles[17]. Another study that analyzed over 150 bird species, found that Z-DNA motifs were most common in bird promoter regions, and established a negative relationship between Z-DNA motif density and developmental time[18]. Yet another study annotated non-B DNA motifs in promoters from over a thousand species across the tree of life, including six bird species, and made comparisons over large taxonomic groups[19]. However, to our knowledge, a detailed study of the non-B DNA motif landscape across bird genomes and among their chromosomes has been lacking so far.

More than 1,700 bird genomes are now publicly available[20]. Yet almost none of them have been assembled to completeness. Bird genomes are smaller than mammalian genomes (~1 Gbp vs. ~3 Gbp for the haploid genome), with a much lower repeat content (~10% vs. ~25-50%)[21–23]. Nevertheless, they have proven challenging to sequence in full, likely due to the unique organization and content of their genomes, which consist of large *macrochromosomes* as well as multiple small GC-rich *microchromosomes* that contain many genes and have high recombination rates[24]. Most of the bird genome assemblies released in the 2010s were produced using short-read sequencing technologies, which have difficulties sequencing templates with high GC content[25]. As a result, a substantial fraction of the bird microchromosomes have been almost entirely missing from the assemblies[26]. Long-read sequencing technologies have led to a vast increase in high-quality chromosome-level

assemblies for birds[27]. Such assemblies are available for 269 bird species on NCBI at the time of writing (NCBI, March 2026). While long-read technologies do not have the same GC bias as short-read ones, sequencing and assembling small bird chromosomes have proven to be continuously difficult even with long reads[28,29]. Indeed, the first bird genome assembly with all microchromosomes represented—that of a chicken (*Gallus gallus*)—was published only recently[30]. This publication suggested a further division of the small chromosomes into microchromosomes and *dot chromosomes*, the latter of which are the most extreme in terms of small size, high GC content, and enrichment for housekeeping genes[30].

Recently, analyzing the newly published T2T genome of zebra finch (*Taeniopygia guttata*), we found that the coverage of non-B DNA motifs increases as chromosome size decreases and that introns in the euchromatic compartment of dot chromosomes have a particularly high non-B DNA enrichment[31]. Here, we expand these results and present a comprehensive analysis of the non-B DNA landscape in zebra finch and other bird species. For zebra finch, we study the distribution of non-B DNA in functionally important regions (e.g., different genic compartments) in detail, as well as in the sequences added to the T2T vs. the previous assembly (e.g., centromeres and repeats). Using the lack of methylation as a proxy for G4 formation[32–34], we predict where in the genome such structures are likely to fold. Using Circular Dichroism (CD) analysis, an experimental spectroscopy technique, we further examine the *in vitro* folding of the most common G-quadruplex (G4) sequences in the zebra finch genome. We compare our results to the non-B DNA motif coverage in the nearly complete chicken genome[30] as well as in six other high-quality assemblies spanning ~108 million years (MY) across the bird phylogeny[35]. These findings allow us to formulate hypotheses about functions of non-B DNA in bird genomes. Finally we provide a potential explanation as to why the smallest bird chromosomes have been notoriously challenging to sequence.

Results

Non-B DNA motifs are distributed unevenly across the zebra finch genome, with contrasting coverage among chromosome categories

We annotated seven types of motifs with the potential of forming non-canonical (non-B) DNA structures in the recently available zebra finch diploid T2T assembly[31]. These included A-phased repeats, direct repeats, short tandem repeats, inverted repeats, triplex motifs, G4s, and Z-DNA (Fig. 1A, see Methods). Compared to previous studies[10,31], here we applied novel algorithms for annotating G4s and Z-DNA[11,36], using scoring thresholds based on experimental data (such as G4-seq[37]) rather than solely pattern match. We also used stricter thresholds for repeat and spacer lengths for the other motifs, specifically shortening the repeat and spacer length for inverted repeats (maximum lengths of 30 bp and 10 bp respectively)[38], and requiring 100% homopurine content and shorter spacer length (10 bp) for mirror repeats to be considered triplex-forming[39]. As a result, we conservatively focused on motifs most likely to form non-B DNA structures (see Table S1 for a detailed comparison of our results with those obtained with earlier methods). Taken together, the non-B DNA motif coverage across the genome was 7.6% (Table S1).

The zebra finch genome consists of 11 macrochromosomes (including Z and W), 19 microchromosomes, and 11 dot chromosomes[31]. On the dot chromosomes, there are two well-separated compartments: A compartment (consisting of euchromatin) and B compartment (consisting of heterochromatin). The chromosomal distribution of non-B DNA motifs in the zebra finch assembly was, as we previously noted[31], strikingly skewed towards high densities on small chromosomes, especially for the 11 dot chromosomes, with an overall motif coverage of 22% (Figs. 1B, 2A, and S1, Tables S1 and S2). Some dot chromosomes had non-B DNA motif coverage as high as 30% (Table S1). While we expected to observe a particularly high coverage of G4s due to the high GC content at micro- and dot chromosomes (Fig. S2), we found that not only G4s but also A-phased repeats, direct repeats, and short tandem repeats had significant enrichment at dot chromosomes compared to macrochromosomes, whereas Z-DNA showed a significant depletion (Fig. 2A). Microchromosomes sometimes had intermediate coverage when compared to macro- and dot chromosomes, e.g., for short tandem repeats, but often had coverage more similar to those of macrochromosomes (Fig. 2A). The non-B DNA coverage on the two sex chromosomes—the Z and the W—was 6.5% and 6.9%, respectively, which was typical of the macrochromosomes (Table S1). Thus, we observed substantial differences in non-B DNA motif coverage among chromosome categories, with starkest contrasts between macro- and dot chromosomes.

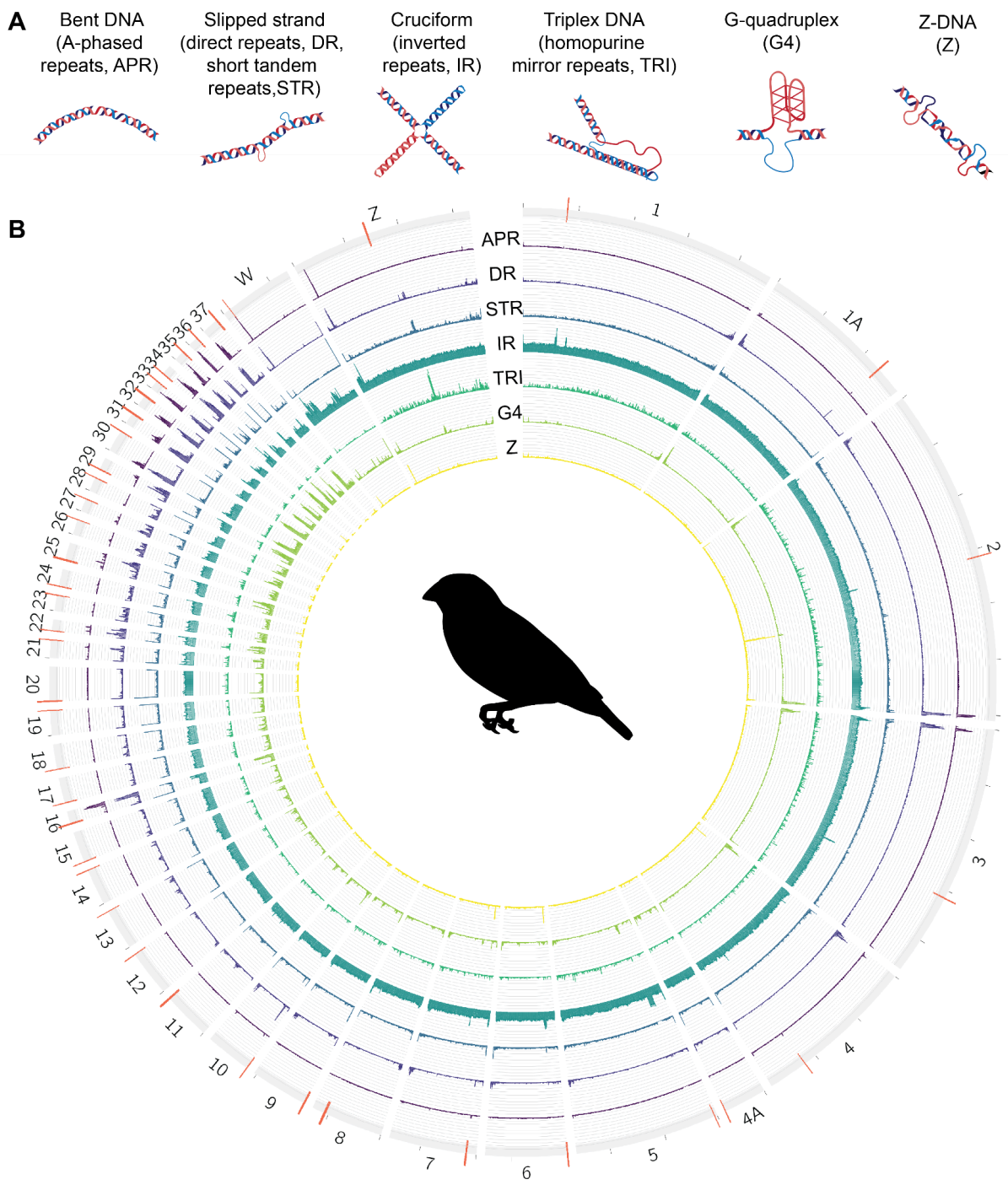

Figure 1. Examples of non-B DNA structures and their motif landscape in the zebra finch T2T assembly. A Non-B DNA structures: bent DNA formed by A-phased repeats (APR), slipped-strand structures formed by direct repeats (DR) or short tandem repeats (STR), cruciforms formed by inverted repeats (IR), triplex DNA formed by homopurine mirror repeats (TRI), G-quadruplexes (G4) formed by guanine stems separated by loops, and Z-DNA (Z) formed by purine-pyrimidine stretches. **B** A circos plot with non-B motif coverage for the zebra finch primary haplotype. Abbreviations as in (A). Bird silhouette is from https://www.phylopic.org.

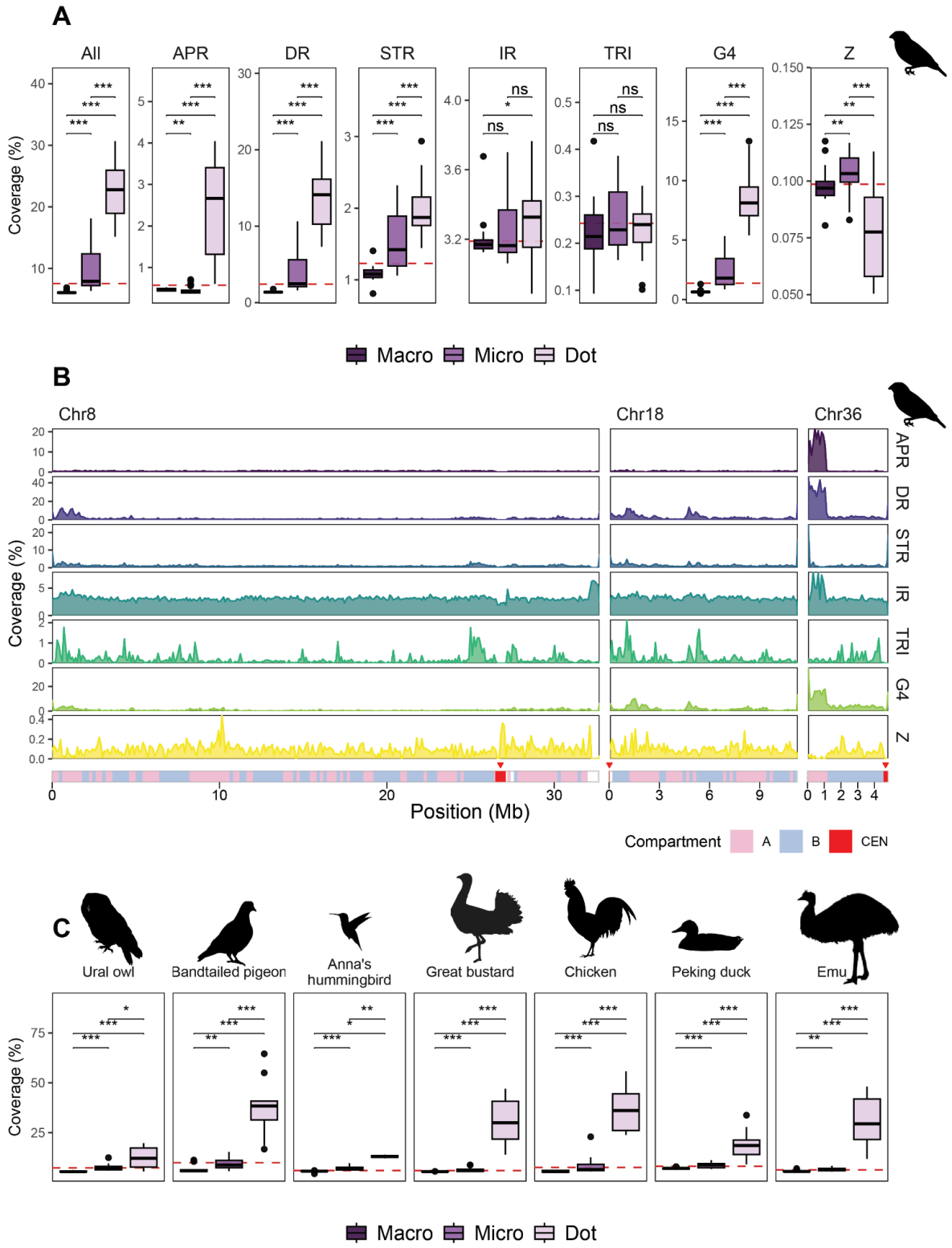

Figure 2. Non-B DNA motif abundance in different chromosome categories in birds. A Overall differences in various non-B DNA motif coverage among macro-, micro-, and dot chromosomes in the diploid zebra finch T2T genome. "All" means all non-B DNA motif types taken together. ***P* <0.01, ****P* <0.001, Wilcoxon test adjusted for multiple testing using FDR. Comparisons marked with 'ns' are not significant. Note that the scale on the y-axis is different for each panel. **B** Examples of non-B DNA motif coverage in 100-kb windows along one zebra finch

macrochromosome (chr8), one microchromosome (chr18), and one dot chromosome (chr36), provided as examples. Maternal haplotypes were used in each case. The scale of the y-axis is different across motif types but is the same for the three chromosomes. The centromeres (red, marked with arrowheads) and compartments (pink: A compartment; and blue: B compartment) are shown below the panels. All chromosomes are shown in Fig. S3. **C** Overall non-B motif coverage in seven other bird species spanning more than 100 million years of evolution, comparable to "All" coverage plot for zebra finch in A. Statistical tests as in A. All bird silhouettes except the great bustard (drawn by the authors) are from https://www.phylopic.org.

Visual inspection indicated that the non-B DNA motif landscape was highly variable within chromosomes, with peak regions for different motif types sometimes colocated (Fig. 2B). Peaks were generally higher on dot chromosomes compared to macrochromosomes, with a tendency for Z-DNA motif peaks to be near some of the centromeres (Figs. 2B and S3). Interestingly, one high peak in the pseudo-autosomal region (PAR), and hence present on both Z and W, was enriched in all non-B DNA motifs except triplex motifs (Fig S3, last panel). In dot chromosomes (e.g., chr36 shown in Fig. 2B; Figure S3), A compartments were enriched in A-phased, direct and inverted repeats, as well as G4s, whereas B compartments were enriched in Z-DNA motifs.

We also annotated non-B DNA motifs in the nearly complete chicken genome[30], which showed a remarkably similar non-B DNA motif coverage to zebra finch, both genome-wide (7.6%, taking all motifs together) and per chromosome category, with the lowest non-B DNA motif coverage on macrochromosomes and the highest on dot chromosomes (Fig. S4, Table S3). Similar to the zebra finch genome, non-B DNA coverage in the chicken genome displayed a negative correlation with chromosome size and a positive one with GC content (Figs. S5B and S6B). Interestingly, dot chromosomes displayed an even more extreme total non-B DNA coverage in chicken than in zebra finch, with some chromosomes exceeding 50% (Table S3). This is consistent with the fact that the smallest chicken dot chromosomes are shorter and display a higher GC content than those of the zebra finch. While this may be explained by their evolutionary distance (~100 MY[35]), caution should be exercised in interpreting this difference since chicken micro- and dot chromosomes are still missing some telomeres and are therefore likely incomplete[31].

To examine whether the skewed non-B DNA motif coverage between chromosome categories is a general feature of bird genomes, we repeated the analysis in six other bird genomes: Ural owl (*Strix uralensis*)[40], band-tailed pigeon (*Patagioenas fasciata,* VGP genome), Anna's hummingbird (*Calypte Anna,* VGP genome)[1], great bustard (*Otis tarda*)[41], Pekin duck (*Anas platyrhynchos*), and emu (*Dromaius novaehollandiae,* VGP genome). These were chosen based on assembly quality as well as phylogeny; they are all from different orders and evenly spread across the bird phylogenetic tree (Fig. S7)[35]. All six showed the same pattern as zebra finch and chicken, with dot chromosomes displaying the highest non-B DNA motif content (Fig. 2C, Figs. S5, S6 and S8, Tables S4-S9), strongly suggesting that this is a general trend across the bird genomes. However, the absolute non-B DNA motif coverage varied between the species, which could reflect true differences and/or differences in assembly quality/completeness (see Discussion).

We also assessed overlaps between non-B DNA motif types (i.e., when the same position was annotated as part of two or more motif types; Fig. S9). In zebra finch, 15% (25 Mb) of non-B DNA annotated bases were annotated as at least two different motif types, and 5% (8 Mb) were annotated as more than two types. Genomewide, the most common overlap was between G4s and direct repeats (3.5% of all non-B DNA annotated bases, or 5.8 Mb). Interestingly, almost 60% of all overlaps between G4s and direct repeats occur on the dot chromosomes, even though the combined length of these chromosomes represents only ~6% of the genome. On macrochromosomes, overlaps between non-B DNA motif types were sparse (Fig. S9A, top panels), with the most common overlaps occurring between A-phased repeats, direct repeats, and short tandem repeats. Microchromosomes displayed an intermediate pattern of overlaps between dot and macrochromosomes (Fig. S9A, middle panels), and the sex chromosomes showed levels similar to those of the other macrochromosomes (Fig. S9A bottom panels, Table S10). Overlaps between different motif types in the other bird genomes were largely similar to those observed in the zebra finch (Fig. S9B-H). The chicken W (Fig. S9B) was a notable exception, as it is an incompletely assembled microchromosome in chicken and had many overlapping motif annotations.

Non-B DNA sequences are enriched and are likely to fold in non-canonical structures at several functional genomic regions

Non-B DNA motifs were enriched at functional regions in the zebra finch T2T genome. Macro- and microchromosomes showed moderate enrichment in almost all non-B DNA motif types (between 1.1-6.0× compared to genome-wide coverage; except for A-phased repeats and inverted repeats) at putative promoter regions (1 kb from the transcription start site, TSS) and 5'UTRs (Fig. 3A). The fold-enrichment was usually higher for microchromosomes compared to macrochromosomes, especially for G4s (6.0× vs. 3.9×, respectively). Dot chromosomes displayed a remarkably different pattern, with higher fold enrichment observed overall, and significant enrichment at intronic regions, which contained more than eight times the content of A-phased repeats, direct repeats, and G4s compared to genome-wide densities (Fig. 3A, bottom panel). A-phased repeats, direct repeats, and G4s also showed a higher enrichment at promoters and 5'UTRs on dot chromosomes than on macro- and microchromosomes (Fig. 3A). G4s were enriched at all functional categories on dot chromosomes, reflecting the higher overall G4 coverage on these chromosomes compared to the genome-wide average. Thus, the coverage of non-B DNA motifs at functional regions varied among the different chromosome categories.

We also assessed which strands the G4s preferentially occurred on, relative to the direction of gene transcription. For 5'UTRs, protein-coding regions (CDS), and 3'UTRs, there were significantly more G4s annotated on the template than on the coding strand (Fig. 3B). However, the opposite was observed for introns. This contrasting pattern was most prominent on the dot chromosomes. These observations suggest that G4s are less common at the level of the mRNA.

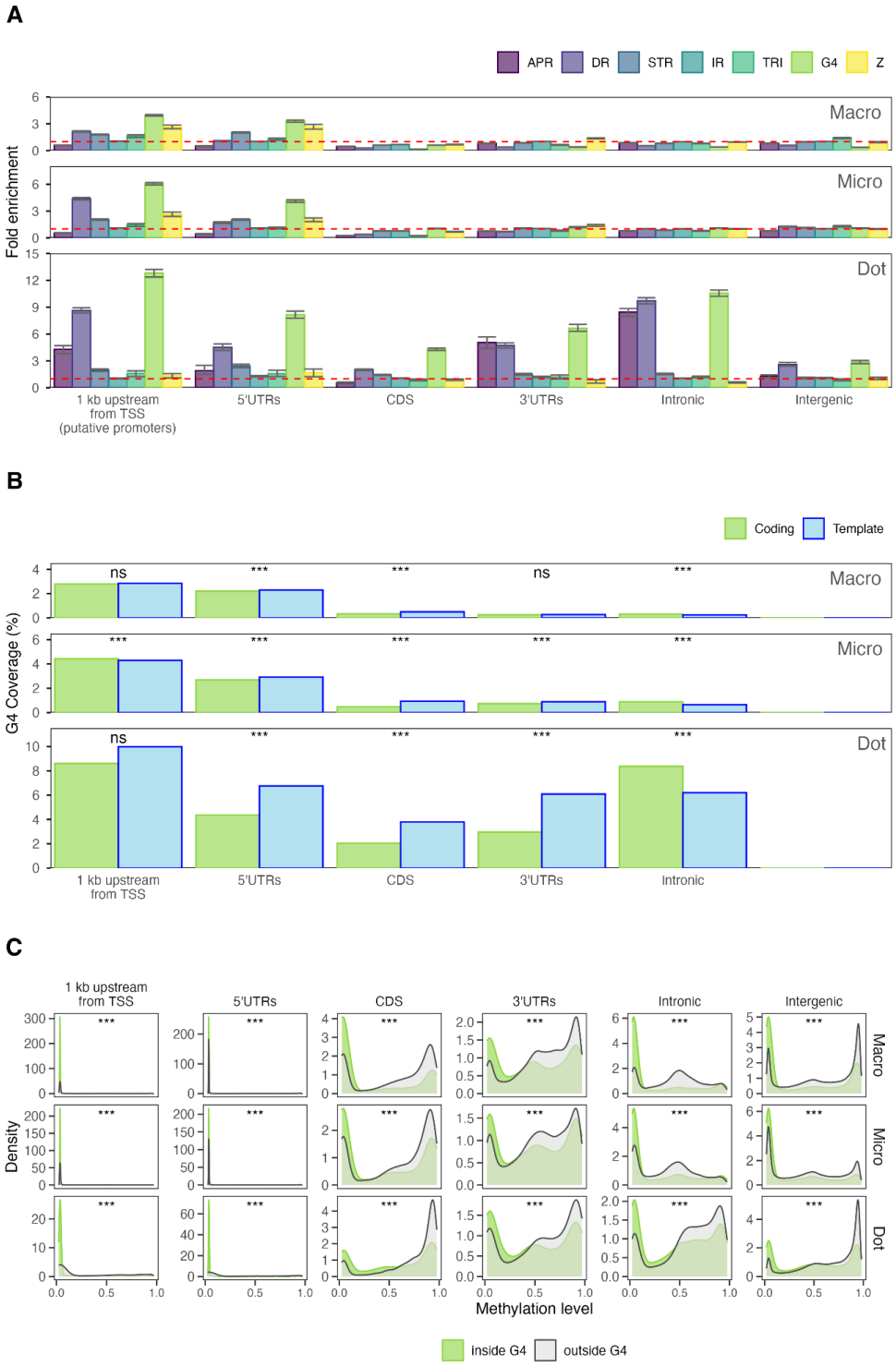

Figure 3. Non-B DNA motif enrichment and methylation distribution at zebra finch functional regions. A Non-B DNA enrichment at functional regions on zebra finch macro-, micro-, and dot chromosomes, as compared to genome-wide non-B DNA motif coverage (red dashed line). For each bar, an interval was constructed by downsampling the data to 50% 100 times, and excluding the two highest and the two lowest enrichment values obtained; enrichment is considered non-significant if such interval overlaps the red dashed line (see Methods for

details). **B** Coverage of G4 motifs on the coding and template strand in zebra finch. Bars are compared using a paired Wilcoxon test on the coverage per gene for each group, corrected with FDR; *** $P<0.001$, ** $P<0.01$, * $P<0.05$, 'ns' non-significant. Note that intergenic regions are strand-ignorant and therefore were excluded. **C** Distribution of gene median methylation levels at CpG sites outside (gray) and inside (green) of G4s in zebra finch. Gene regions are separated into chromosome categories, and G4s are strand-ignorant (for strand-specific G4s, see Fig. S10). *** $P<0.001$, Mann-Whitney U test**.** TSS: transcription start site; UTR: untranslated region, CDS: protein-coding region.

To assess whether G4s fold at functional regions, we used zebra finch blood 5mC methylation levels as a proxy, because G4 formation is inversely correlated with methylation[32–34]. Genic regions overlapping with G4s had significantly lower median methylation levels than genic regions not overlapping with G4s (Fig. 3C; methylation patterns for G4s located on the template vs. the coding strand were similar; Fig. S10). This pattern was most prominent at putative promoter regions and 5'UTRs, where the vast majority of CpG sites overlapping with a G4 motif were unmethylated, and suggests that G4s are likely to fold in these regions (Fig. 3C). Interestingly, for dot chromosomes, there was less methylation at 5'UTR G4s as compared to the putative promoter G4s, potentially indicating that their gene regulatory elements are closer to the TSSs, reflecting the compact nature of these chromosomes. We also observed that a significantly smaller proportion of 5'UTRs and putative promoters on the dot chromosome, as compared to macrochromosomes, have CpG sites at G4s, which can be methylated and hence interfere with G4 formation (Fig. S11). This is despite the high overall GC content on the dot chromosomes[31]. Therefore, our methylation analysis suggests that G4s are likely to fold and be involved in regulating functional genic regions in the zebra finch genome.

The non-B DNA enrichment patterns at functional regions in the other seven bird genomes were remarkably similar to that of zebra finch, with putative promoters and 5'UTRs enriched especially in G4s for all chromosome categories, and the most extreme enrichment levels in dot chromosomes; sometimes reaching 15-20× (Fig. S12). Dot chromosome introns were especially enriched in direct repeats, mirror repeats and G4s—just as in zebra finch but often to an even higher degree (up to 20× for G4s in great bustard, compared to genome-average, Fig. S12I). Contrary to observations for zebra finch, direct repeats, mirror repeats, and G4s were enriched in intergenic regions of dot chromosomes in the other bird species. We cannot exclude that this is due to differences in annotation completeness (Table S11). Most of the other assemblies are lacking telomeres, which will affect the genome-wide density of G4 motifs.

Non-B DNA motifs are overrepresented at many tandem repeats and at some transposable elements and satellites

Different repeat classes in the zebra finch T2T genome assembly were analyzed for enrichment or depletion of non-B DNA motifs per chromosome category, in comparison to the whole-genome motif coverage. The enrichment patterns were mainly shared across the chromosome categories, with some exceptions (Fig. 4A). Among transposable elements (TEs), multiple classes of DNA transposons displayed enrichment in several non-B DNA motif types, particularly in direct repeats and G4s on dot chromosomes, though some TE classes (DNA/Kolobok, LTR/Copia) were entirely missing from these chromosomes (Fig. 4A, bottom panel). The most striking enrichment among TEs was observed for Ngaro elements, a distinct

group of retrotransposons[42], which exhibited a ~300-fold enrichment of Z-DNA motifs compared to the genome-wide average coverage. Ngaro elements occupy ~70 kb in the diploid zebra finch genome and are present there in small clusters on 56 of the 80 chromosomes. We found that long interspersed nuclear elements (LINEs), short interspersed nuclear elements (SINEs), and long terminal repeats (LTRs) were usually not enriched in non-B DNA motifs (Fig. 3A).

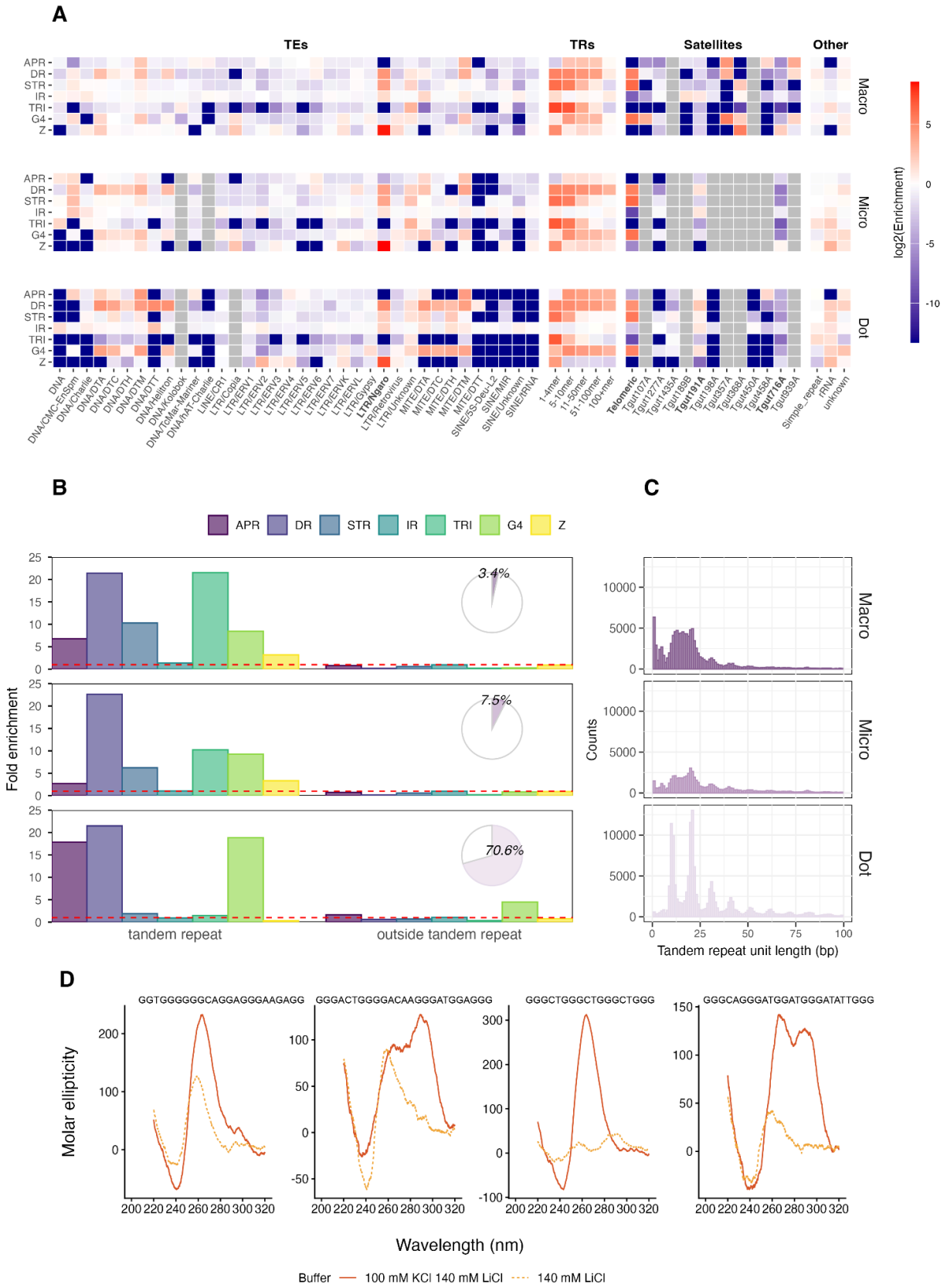
A
TEs
TRs
Satellites
Other
APR
DR
STR
IR
TRI
G4
Z
Macro
Micro
Dot
log2(Enrichment)
5
0
-5
-10
DNA
DNA/CMC_EnSpm
DNA/Charlie
DNA/DTA
DNA/DTC
DNA/DTH
DNA/DTM
DNA/DTT
DNA/Helitron
DNA/Kolobok
DNA/TcMar-Mariner
DNA/hAT-Charlie
LINE/CR1
LTR/Copia
LTR/ERV1
LTR/ERV2
LTR/ERV3
LTR/ERV4
LTR/ERV5
LTR/ERV6
LTR/ERV7
LTR/ERVK
LTR/ERVL
LTR/Gypsy
LTR/Ngaro
LTR/Retrovirus
LTR/Unknown
MITE/DTA
MITE/DTC
MITE/DTH
MITE/DTM
MITE/DTT
SINE/5S-Deu-L2
SINE/MIR
SINE/Unknown
SINE/RNA
1-4mer
5-10mer
11-50mer
51-100mer
100-mer
Telomeric
Tgut107A
Tgut1277A
Tgut1435A
Tgut188B
Tgut191A
Tgut198A
Tgut357A
Tgut368A
Tgut450A
Tgut458A
Tgut715A
Tgut939A
Simple_repeat
rRNA
unknown
B
APR
DR
STR
IR
TRI
G4
Z
Fold enrichment
3.4%
7.5%
70.6%
tandem repeat
outside tandem repeat
C
Counts
Macro
Micro
Dot
Tandem repeat unit length (bp)
D
GGTGGGGGGCAGGAGGGAAGAGG
GGGACTGGGGACAAGGGATGGAGGG
GGGCTGGGCTGGGCTGGG
GGGCAGGGATGGATGGGATATTGGG
Molar ellipticity
Wavelength (nm)
Buffer
100 mM KCl 140 mM LiCl
140 mM LiCl

Figure 4. Non-B DNA motif enrichment in zebra finch repetitive regions. A Enrichment of non-B DNA motifs at different repeat classes for each chromosome category. Only repeat classes occupying at least 10 kb in the diploid genome are shown. TE: Transposable elements; TR: tandem repeats. White denotes no enrichment compared to the genome-wide average motif coverage; red and blue denote enrichment and depletion, respectively (color coding on the log scale). Repeat classes that are missing from a chromosome category are marked in gray. Ngaro elements and telomeric and centromeric/pericentromeric satellites are marked in bold. **B** Enrichment of non-B DNA motifs within introns annotated vs. not annotated as tandem repeats. Inset piecharts show the proportion of introns that overlaps with tandem repeats. The dashed red line denotes genome-wide average. **C** Number of annotated tandem repeats in introns for each repeat unit length. **D** Circular Dichroism (CD) spectra for four commonly occurring G4 motifs in the zebra finch genome, listed in Table 1.

Tandem repeats of all lengths showed enrichment for essentially all non-B DNA motif types (Fig. 4A), and this enrichment was negatively correlated with the repeat unit length (Fig. S13). We further investigated overlaps between introns and tandem repeats. We found that parts of introns annotated as tandem repeats exhibited notably higher enrichment for non-B DNA motifs compared to the rest of the introns (Fig. 4B). This enrichment inside tandem repeats was high for all chromosome categories. We noticed that tandem repeats were highly abundant on dot chromosomes (where they constituted over 70% of introns) but rare on macro- and microchromosomes (where they constituted 3.4% and 7.5% of introns, respectively; Fig. 4B, inset). An exception was the high non-B DNA motif peak in PAR, which contained two genes with many tandem repeats in their introns (Fig. S14). The most common tandem repeats on dot chromosome introns were minisatellites, especially with repeated units of multiples of 10-11 bp (Fig. 4C).

The telomeres were enriched for direct repeats and G4s (as expected, because the telomeric motif corresponds to a G4 stem-loop combination). Satellites showed a highly mosaic pattern of enrichment and depletion in different non-B DNA motif types, with many low-frequency satellites enriched in A-phased repeats, direct repeats, and G4s, but often depleted in short tandem repeats, triplex repeats, and Z-DNA motifs (Fig. S15). One of the most common satellites, Tgut716A, was enriched for Z-DNA motifs on macrochromosomes (Fig. 4A). The other common satellite, Tgut191A, only showed a slight enrichment in inverted repeats on macrochromosomes and in direct repeats on micro- and dot chromosomes. Both of these satellites were previously suggested to be associated with centromeres in the zebra finch[43]. Tgut716A has been identified as the dominant centromeric satellite repeat[31]. At the same time, Tgut191A appears to play a different role, potentially associated with Z-W conjugation in the PAR region[31]. The 18S/28S rRNA, almost exclusively present on dot chromosome pair 37 in the zebra finch[31], was enriched in all motif types except for APRs. The 5S rRNA (named Tgut368A and shown in Fig. 4A), located in two clusters on macrochromosome pair 2[31], was highly enriched in Z-DNA, STRs, and G4s (45.3×, 4.9×, and 4.4×, respectively, compared to the genome-wide average).

Circular Dichroism analysis demonstrates that common G4s fold *in vitro*

We summarized the most common G4 motifs in the zebra finch genome (Table S12). The top two motifs, GGGTTAGGGTTAGGGTTAGGG (telomeric) and GGGAAGGGAAGGGAAGGG (in LTR retrostransposons on 73/80 chromosomes), occurring >69,000 and >33,000 times, respectively, in the diploid genome, have already been verified to form G4 structures *in vitro* in previous studies[44–46]. We further tested four additional G4 motifs experimentally: one from the

5S rRNA (Tgut368A), two from LINE/CR1 elements (transposable elements unique to birds), and one from a tandem repeat that also had 10% overlap with putative promoters. All four motifs produced circular dichroism (CD) and UV-absorbance spectra that were consistent with known signatures of G4 structure[47,48] (see Table 1 and Fig. 4D; Fig. S16 includes thermal UV absorbance difference spectra). Two of the tested G4 motifs also showed signs of multiple conformations on native gels (Fig. S17). These results indicate that these common G4 motifs can indeed form G4 structures *in vitro*.

Table 1. Experimentally verified common G4s in zebra finch

Sequence	Location	N occurrences[a]	Form G4 *in vitro*[b]	Structure[c]
GGTGGGGGGCAGGAGGGAAGAGG	5S rRNA (Tgut368A)	>1,900	Yes	Parallel, intermolecular
GGGACTGGGGACAAGGGATGGAGGG	LINE/CR1	>800	Yes	Hybrid, intramolecular
GGGCTGGGCTGGGCTGGG	Tandem repeat (5-mer), promoters	>800	Yes	Parallel, intramolecular
GGGCAGGGATGGATGGGATATTGGG	LINE/CR1	>600	Yes	Hybrid, intramolecular and intermolecular

[a]In the diploid zebra finch T2T genome

[b]On the basis of potassium-sensitive CD spectra, thermal UV-absorbance hypochromism near 295 nm, and electrophoretic mobilility[48].

[c]On the basis of CD peaks at near 265 nm, 295 nm, or both, for parallel, anti-parallel, or hybrid topologies, respectively[48].

Non-B DNA enrichment at bird centromeres

In the newly released T2T zebra finch genome[31], the centromeres were fully resolved and consist mainly of the highly conserved satellite Tgut716A, either flanked by the satellite Tgut191A or directly adjacent to the telomere. Z-DNA motifs were significantly overrepresented in six centromeres (2.9-19.9× compared to the genome-wide average; Figs. 5A and S18). The strongest overrepresentation was observed for macrochromosome 7, which likely drives the enrichment in satellite Tgut716, as documented in the repeat analysis above (Fig. 4A). Because the average non-B DNA motif coverage differed strikingly among chromosome categories (Fig. 1B), we also assessed non-B DNA enrichment at each centromere in relation to the rest of the chromosome it was located on, rather than to the genome-wide average coverage, but this did not change the overall patterns (Fig. S19). However, we note that the choice of algorithm for annotating Z-DNA mattered; using gfa's pattern match of $(RY)_{5+}$ utilized in[10,17–19] resulted in Z-DNA motif enrichment on 77 of 80 centromeres (Fig. S20). This is caused by a large amount of short (10-bp) purine-pyrimidine stretches that are not found by Z-DNA Hunter (which uses the minimum threshold of 12 bp)[36]. It is unclear whether these small motifs can form Z-DNA. Notably, almost all chicken centromeres showed enrichment in non-B DNA motifs, mainly A-phased repeats, direct repeats, short tandem repeats, and G4s (Fig. 5B).

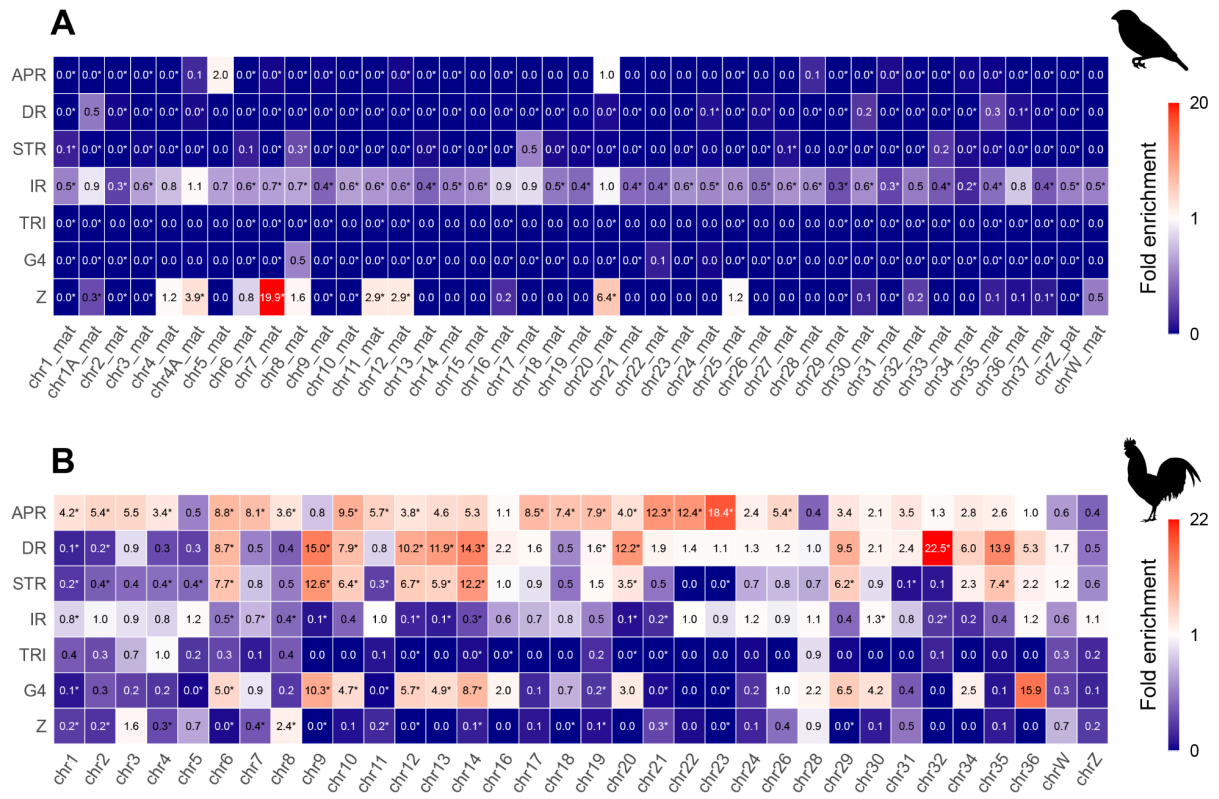

Figure 5. Non-B DNA motif enrichment at bird centromeres. A Enrichment of non-B DNA motifs at centromeres in the haploid zebra finch genome (maternal+Z), compared to the genome-wide average. Red and blue denote, respectively, enrichment and depletion compared to the genome-wide non-B DNA content. Values marked with "*" are significantly enriched or depleted as compared to the genome-wide average (two-sided randomization test compared to background of 100 windows, $P < 0.05$, see Methods). Abbreviations are as in Fig. 1. See Fig. S18 for a full version including all haplotypes, and Fig. S19 for fold comparisons to the chromosome average instead of the genome-wide average. **B** Centromere enrichment in the chicken genome. Note that centromeric positions were not available for all chicken chromosomes.

Dot chromosomes display a unique non-B DNA motif landscape

Because dot chromosomes differed significantly in non-B DNA motif content compared to micro- and macrochromosomes, we examined their motif distribution in zebra finch in greater detail. The euchromatic A compartment of dot chromosomes was significantly more enriched in non-B DNA motifs taken together than the heterochromatic B compartment (Fig. 6A). When analyzing different motif types separately, we found that almost all non-B DNA motif types were more prevalent in A than B compartments, whereas Z-DNA showed the opposite pattern. A-phased repeats, direct repeats, and G4 motifs had particularly contrasting fold enrichment between A and B compartments (Fig. 6A). The high amount of G4s in the A compartment may be attributed to its high content of protein-coding genes, which in turn leads to an abundance of promoters and 5'UTRs, where G4s are enriched (Fig. 3A). Indeed, the A and B compartments have a gene coverage of 73% and 19%, respectively (exon coverage of 9.2% and 3.1%, respectively)[31]. Moreover, the A compartment has many minisatellites in the introns, in which A-phased repeats, direct repeats, and G4s are highly enriched (Fig. S21). A similar enrichment in non-B DNA

motifs was also observed at intronic minisatellites in the B compartment, but such intronic minisatellites were much less prevalent in this compartment than in the A compartment (Fig. S21). In contrast, the B compartment contains many macrosatellites, including the Z-DNA-rich centromere satellite Tgut716A[31], which contributes to its overall enrichment of Z-DNA.

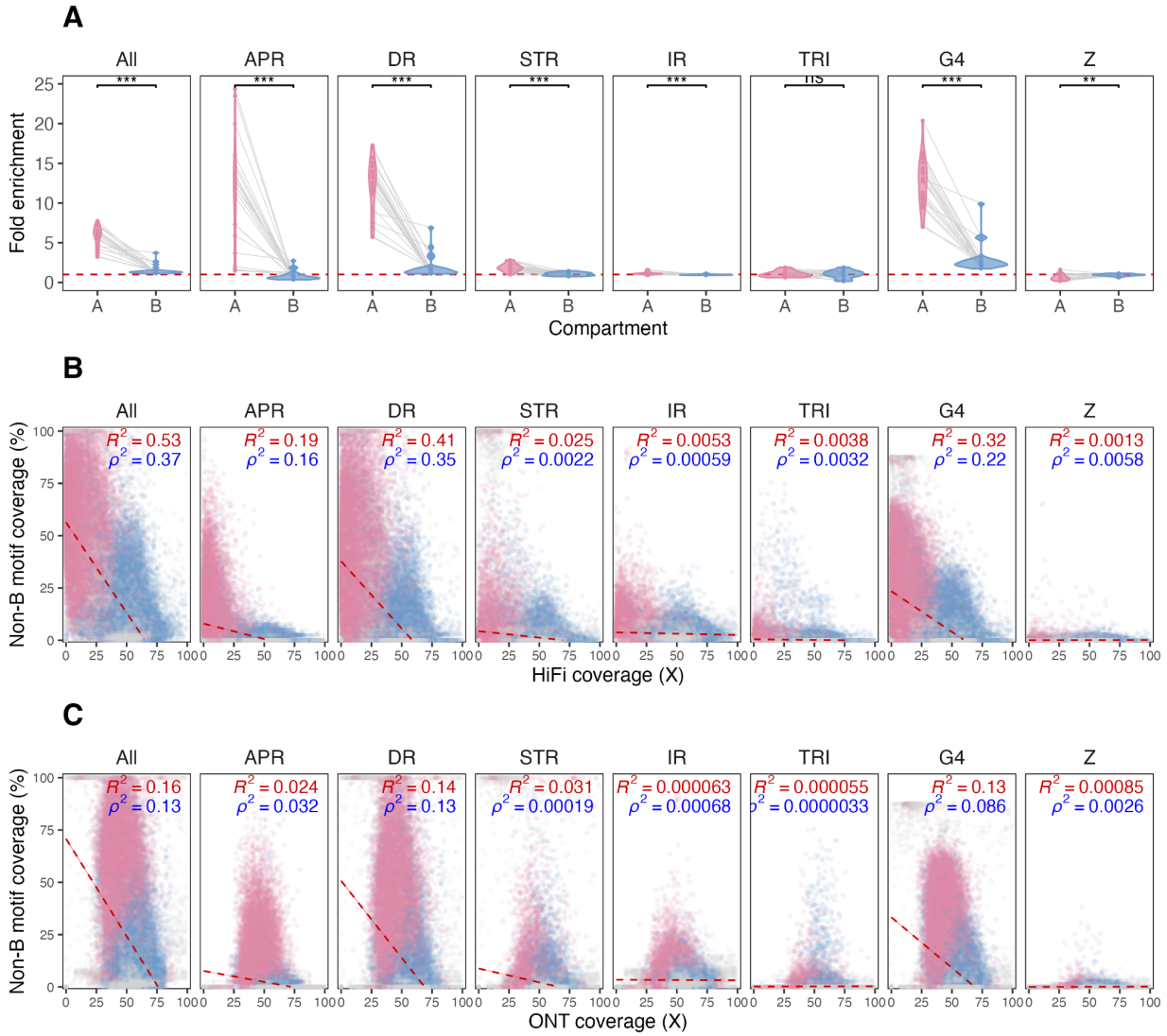

Figure 6. Non-B DNA motif enrichment and sequencing coverage at dot chromosomes. A Enrichment of non-B DNA at the A compartments (euchromatic) and the B compartments (heterochromatic) for the 11 dot chromosomes in zebra finch (averaged across the compartments for each chromosome and haplotype separately) compared to the genome-wide average (red dashed line). Gray lines between violins connect the compartments from the same dot chromosome. *P <0.05; ***P <0.001; Wilcoxon test adjusted for multiple testing using FDR. Note that some parts of the chromosomes were not assigned to A or B compartments; the telomeres and some satellites were excluded in the original annotations[31]. **B** Non-B DNA motif coverage and PacBio HiFi coverage on dot chromosomes in non-overlapping windows of length 1,024 bp, taken from[31]. Windows in A compartment are shown in pink, windows in B compartment are shown in blue, and windows not assigned to compartments are shown in gray. Red dashed lines show linear regression fits, and Pearson's R^2 (red) and Spearman's ρ^2 (blue) are shown for each correlation. Windows with sequence coverage greater than 100× are not displayed for visualization purposes. **C** Same as B but for ONT coverage data.

Low sequencing depth on dot chromosomes can, in part, be explained by high non-B DNA content

To test whether there is a link between non-B DNA and reduced sequencing depth on dot chromosomes, we evaluated the relationship between non-B DNA coverage and PacBio HiFi or Oxford Nanopore Technology (ONT) sequencing depth in 1,024-bp windows. This was performed separately for dot, micro-, and macrochromosomes. For PacBio HiFi, we found a negative relationship, which was particularly strong for dot chromosomes, with non-B DNA content explaining 53% of the variability in sequencing depth (Fig. 6B). When considered separately, the content of direct repeats and G4s explained 41% and 32% of the variability in sequencing depth, respectively. Similar, but much weaker, patterns were observed for micro- and macrochromosomes (Fig. S22). We also found a negative relationship between non-B DNA and ONT sequencing coverage (Figs. 6C and S23), but this was much less pronounced. Indeed, non-B DNA content explained only 16% of the variability in ONT sequencing depth on dot chromosomes.

Discussion

We conducted a comprehensive analysis of non-B DNA coverage in the zebra finch T2T genome[31] and compared it to that of the second most complete bird genome, the chicken[49], as well as to those of six other high-quality bird genomes[1,40,41]. We found that in both zebra finch and chicken, 7.6% of the genomes contain sequences likely to form non-B DNA structures. This similarity, despite approximately 100 MY of divergence[50], is remarkable. The other six species contained 6.1-10% of non-B DNA motifs, but we note that these fractions may change (most likely increase) when these genomes are assembled to the T2T level.

Here, we chose to use more stringent, conservative non-B annotations than previous studies[10,31], only selecting motifs with the highest potential to fold. Notably, using more lenient definitions produces a non-B motif coverage of ~11% in zebra finch and chicken, a level very similar to the non-B DNA motif coverage observed in ape T2T genomes (9.2-14.9%)[10], which were also annotated with more lenient definitions. This is despite approximately 300 MY of divergence between mammals and birds[51].

Micro- and dot chromosomes

The non-B DNA motif coverage is highly skewed towards small (micro- and dot) chromosomes in all eight bird genomes analyzed here. Small chromosomes in bird genomes tend to have high GC-content, elevated recombination rates, high gene density in euchromatin (A compartments), and high repeat content in heterochromatin (B compartments)[30]. At the same time, small bird chromosomes appear to represent dynamic, functionally important genomic units, requiring active regulation of transcription and recombination. Non-B DNA likely contributes to this regulation in birds, consistent with previous evidence of its involvement in gene expression regulation in humans[15], its facilitation of recombination in yeast (e.g.,[52]), as well as our observations of high non-B DNA motif frequency on the small gene-rich chromosomes in birds.

A particularly strong enrichment in non-B DNA was found on the dot chromosomes. While their high GC content (related to their gene-richness) could explain enrichment of certain non-B DNA motif types such as G4s, our data indicate that most non-B DNA motif types, including A-rich motifs (i.e., A-phased repeats), were overrepresented. This pattern is largely explained by the abundance of minisatellite-like repeats that make up most of the dot chromosomes' non-coding (i.e., intronic) euchromatic sequence. The presence of minisatellites on dot chromosomes was recently observed in chicken, where the euchromatic sequences on dot chromosomes were proposed to evolve via a local instability mechanism named sequence stuttering[53]. Because these minisatellites contain multiple non-B motifs, they substantially increase non-B DNA motif enrichment for the euchromatic A compartment whereas the heterochromatic B compartment typically has non-B DNA coverage close to genome-wide levels.

The association between high recombination rates, GC-biased gene conversion, and open chromatin might provide a mechanistic explanation for the elevated non-B DNA content in A-compartments observed on dot chromosomes. Dot chromosomes display unusually high interchromosomal interactions, as observed from Hi-C maps (e.g.,[31,54,55]), presumably because they are clustered in the interior of interphase nuclei[56], and it is possible that the different 3D conformations of non-B motifs could play an important role in this phenomenon[57]. Dot chromosomes experience a high recombination rate per unit length because each chromosome typically requires at least one crossover during meiosis. Elevated recombination rates increase the frequency of GC-biased gene conversion events. Heterochromatin (B compartments) generally recombines substantially less than euchromatin due to its highly condensed structure, which is often enriched in repetitive DNA and limits the accessibility of the recombination machinery to DNA and suppresses crossover formation.

The role of non-B DNA in regulating bird genes

Promoters and 5'UTRs are enriched in non-B DNA motifs, especially in G4s, for genes in all bird chromosome categories. Great ape non-B DNA motifs, particularly G4s and Z-DNA, have strong enrichment at promoters and 5'UTRs as well[10]. This aligns with previous findings that G4s are important promoter elements[16] and evolve under purifying selection at UTRs[58] in the human genome. Another study, using a different methodology and data sets, demonstrated that G4s located at human promoter and UTR regions are overrepresented, subject to purifying selection, and are predicted to form particularly stable structures[12]. This line of evidence strongly points towards the functional role of non-B DNA at gene regulatory regions in both birds and primates.

Our observation that G4s located at bird promoters and 5'UTRs are usually unmethylated, and are less methylated than the non-G4 parts of the same functional regions, suggests that these G4s indeed fold, given the negative correlation between G4 folding and methylation[32,33]. These patterns of low methylation at promoters and 5'UTRs are also consistent with observations for human and other great apes[11]. Just as in great apes[11], bird protein-coding sequences, introns, and 3'UTRs display a bimodal distribution of methylated vs. unmethylated G4s. However, the zero methylation peak is taller with respect to the methylation peak in introns for birds than for great apes, suggesting that a higher proportion of G4s fold in the former than in the latter.

We observed that G4 motifs in birds are more prevalent on the template than the coding strand in UTRs and CDS, suggesting that G4 structures are avoided at the level of mRNA. This is consistent with our observations in humans and other apes[11,12]. Interestingly, we observed no differences in methylation patterns whether the G4 is annotated on the coding (non-transcribed) or template (transcribed) strand, suggesting that, while differing in abundance, G4s on both strands can fold and affect gene regulation. This finding, too, aligns with results for human and ape G4s[11], suggesting similar regulation patterns between these two taxonomic groups.

Additionally, our results suggest that, compared to macro- and microchromosomes, dot chromosomes have evolved to contain fewer CpGs at G4s in their regulatory regions, potentially to ensure that these G4s always fold and the genes they regulate are consistently expressed. This agrees with the fact that dot chromosomes carry many housekeeping genes[30].

Non-B DNA at repeats and centromeres

We found that various repeat elements exhibit distinct patterns of enrichment and depletion of non-B DNA motifs in birds, a phenomenon also observed in great apes[10]. Similar to our observations in apes[10], we did not find a general enrichment of non-B DNA motifs at TEs as a group; however, we observed a high enrichment for certain motifs at specific TEs. The most striking enrichment was of Z-DNA motifs in Ngaro elements, a distinct type of retrotransposon that resembles LTRs and is found in many animals and fungi[42]. LTRs that contain Z-DNA-forming sequences have previously been suggested to act as alternative promoters for genes[59], and the Ngaro elements containing Z-DNA motifs in zebra finch—present on 70% of the zebra finch chromosomes—could have a similar function. More generally, non-B DNA has been suggested to play a role in the life cycle of some mammalian transposable elements, such as LINE-1s (e.g.,[60]); Z-DNA at Ngaro might have a similar function. Experimental validation *in vitro* of four highly abundant G4 sequences, together with previous validation[44–46] of the two most commonly found G4 motifs, further strengthens our hypothesis that non-B DNA can indeed form and play functional roles in bird genomes, just as in mammals.

Non-B DNA was found to be enriched at bird tandem repeats (i.e., micro- and minisatellites). Short tandem repeats fold into slipped-strands, which can lead to copy-number changes in repeat units at microsatellite sequences[61]. Satellites (i.e., macrosatellites), the class of repeats with the highest number of copies in the bird genome, had the lowest non-B DNA motif content among all repeat classes analyzed. Tgut191A was highly abundant in the PAR on sex chromosomes and in the region between the telomere and euchromatin of microchromosomes. However, it showed no overall enrichment in non-B DNA motifs (Fig. 4A). The other abundant macrosatellite, Tgut716A, which is presently the strongest candidate for centromeric function, showed only some Z-DNA motif enrichment on certain chromosomes with the annotation method used in this study, but significant enrichment on almost all chromosomes when annotated with gfa, the software used for the non-B DB[62]. Zebra finch centromeres have a high abundance of purine-pyrimidine repeats of length 10 bp (the actual lower threshold in gfa), but it is unclear whether such motifs can form Z-DNA (whose conformation has 12 bp per turn in its left-handed helix). Two newly published Z-DNA finding algorithms, Z-DNA Hunter[36] and ZSeeker[63] have both chosen a minimum threshold of 12 bp, and hence we adopt this threshold

in this study. We acknowledge that, apart from centromeres, the enrichment analyses did not change when we switched from a lenient to a more strict annotation method (Table S1). Formenti et al.[31] found that Tgut716A had sequence similarity to satellites in several other passerines; however, centromere positions are lacking from most bird assemblies, and it remains to be seen whether other passerines show non-B DNA enrichment at centromeres.

Non-B DNA structures have been suggested to contribute to defining centromeres[64,65], to help maintain centromere position and integrity[66], and to generate satellite copy number variation important for centromere drive[67]. Z-DNA motifs have previously been shown to be enriched at centromeres in plants[68] and at specific chromosomes in human, chimpanzee, and bonobo[10], but depleted at Drosophila centromeres[69]. Future experimental studies, for example using PDAL-seq[70], should investigate whether Z-DNA forms, especially for the very short motifs seen here, and what potential function it may have for the centromeres in zebra finch.

The finding that chicken centromeres were enriched for other types of non-B DNA than zebra finch might not be surprising, as no sequence similarity was found between chicken and zebra finch centromeres[31], and centromeres are known to evolve at high rates[71]. Huang et al.[30] found that most chicken micro- and dot chromosome centromeres harbor a 41-bp tandem repeat called CNM, which often forms higher-order repeats (HORs), whereas macrochromosomes harbor different, chromosome-specific tandem repeats. In terms of non-B DNA motifs, we see no clear separation between macro-, micro-, and dot chromosome centromeres, and no clear pattern that connects all centromeres with CNM repeats. However, we note that defining the exact centromere positions is difficult and somewhat arbitrary. The more precise CENP-A binding regions in chicken were not available for our study; however, Huang et al.[30] describe that these do not occur precisely on the CNM cluster, but rather overlap with a short tandem repeat, sometimes identical to the telomeric sequence $(TTAGGG)_n$. This might explain why some chicken centromeres are enriched in G4s. Further, more detailed analyses of the centromeric components and the distribution of non-B DNA motifs among them, as well as more complete bird genomes, are necessary to draw more general conclusions about the role of non-B DNA in centromeric function in birds.

Additionally, we note that even though the repeat enrichment pattern was—with a few exceptions—remarkably similar across the different chromosome categories, the dot chromosomes had higher repeat content than the other chromosome categories. This pattern was primarily driven by a specific minisatellite sequence located in introns and intergenic regions of euchromatin. Our finding that introns on dot chromosomes are enriched in non-B DNA, particularly within minisatellites, requires further investigation. For instance, non-B DNA might be involved in alternative splicing[72]. Consistent with this, G4s were overrepresented on the coding strand of bird introns (Fig. 3B). Alternatively or additionally, non-B DNA at introns might facilitate recombination[52]. Because dot chromosomes are so compact[31], intronic minisatellites might represent recombination hotspots.

The fact that the highly repetitive zebra finch W (with repeat content of 86%, Table S14) did not show enrichment in non-B DNA is surprising, and is in contrast to the highly repetitive ape Y,

which has the highest coverage of non-B DNA among all ape chromosomes[10]. The lack of non-B DNA enrichment on the zebra finch W can be explained by the fact that most of its repeats consist of TEs (Table S14), and TEs are in general not enriched in non-B DNA in the zebra finch. The notable exception—a peak enriched in all non-B DNA motifs except triplexes in the PAR region—had very similar features to the peaks on the dot chromosomes, with minisatellites in the introns. Interestingly, this region also shows a drop in PacBio sequencing coverage, and was missing in previous zebra finch assemblies[31]. The chicken W shows a higher non-B DNA motif enrichment than zebra finch W. This chromosome has not been assembled to the T2T level yet, so its non-B DNA motif coverage might still change. Regarding the other bird species considered, four have incompletely assembled W chromosomes and two lack W altogether, but the non-B motif coverage of the assembled sequence (4.4-10.5%) is more similar to that for zebra finch than that for chicken.

High non-B DNA content on dot chromosomes as a potential reason for their underrepresentation in older assemblies

The smallest chromosomes of bird genomes have been notoriously challenging to sequence and assemble[31]. Dot chromosomes have a high GC content—a known problem for short-read Illumina sequencing, but a smaller problem for long-read PacBio and even less so for the ONT sequencing[73,74]. Therefore, the fact that dot chromosomes have been largely missing from bird assemblies until very recently[30,31] could be in part explained by a shift from Illumina to long-read technologies. Additionally, non-B DNA structures, which we hypothesize form during sequencing, increase polymerase stalling for PacBio technology[75] and the speed of going through pores for ONT[76]. Moreover, non-B DNA motifs increase sequencing errors, particularly for Illumina sequencing technology, but to a lesser extent for ONT and PacBio technology[77]. Even in the recent zebra finch genome sequencing effort[31], we observed lower sequencing depths in euchromatic parts of the dot chromosomes, particularly for the PacBio technology. These results are consistent with a small-scale study examining sequencing depth at five bird genes using Illumina and PacBio technologies, which suggested that non-canonical DNA structures may explain dropout in sequencing depth[78]. Examining this phenomenon genome-wide and for a complete T2T genome, we found that non-B DNA content explains a large proportion of dropout in PacBio HiFi sequencing depth, and a smaller proportion in ONT sequencing depth, particularly on the dot chromosomes. PacBio HiFi targets double-stranded fragments for sequencing, which become single-stranded (circular) after attachment of the SMRT bell adapters to them, and reads them repeatedly in a circular manner to create high-confidence reads. The extreme density of non-B DNA motifs (especially G4s) in the euchromatic parts of dot chromosomes may result in many partial short single-stranded reads that cannot be captured by this technique. However, the high long-read sequencing depth used to generate T2T assemblies, and the combination of long-read sequencing technologies utilized—as was suggested[77]—alleviates some of the difficulties sequencing through non-B DNA, and results in more accurate non-B DNA motif sequences and gapless genome assemblies, rescuing dot chromosomes.

General conclusions and perspectives

Here we present the first extensive study of non-B DNA motifs in completely and nearly completely sequenced bird genomes. Even though the overall motif content was very similar to that in primates, the inter- and intrachromosomal variation in non-B DNA motif content in birds was substantial, mainly due to the unique genome organization of birds into macro-, micro-, and dot chromosomes. We found that non-B DNA motifs concentrate in portions of the bird genome where genes also concentrate—the euchromatic compartments of dot chromosomes, where they are likely implicated in regulating gene expression. Additionally, non-B DNA might play a role in defining centromeres and contributing to centromere drive in passerines.

Additional annotations of bird genomes will enable testing the hypothesis that non-B DNA also contributes to regulating other functionally important regions, such as enhancers and origins of replication, as suggested by the analysis of primate genomes[10]. Future wet-lab studies are needed to resolve this and determine when and where non-B DNA forms in bird genomes *in vivo*.

We hypothesize that bird dot chromosomes have been challenging to sequence and assemble in part because of their high non-B DNA content. This can be primarily attributed to G4 structures that likely form in the promoters and introns of housekeeping genes in the euchromatic regions of dot chromosomes. Since dot chromosomes are enriched in four non-B motif types investigated in this study, other structures apart from G4s may also form and impede sequencing. This poses a challenge for PacBio-only reference genome projects moving forward, since this sequencing technology appears to be substantially affected by the high density of non-B DNA in the sequence. The PacBio HiFi sequencing dropout at dot chromosomes may result in many genes being entirely missing from the assemblies. Until this is resolved, a combination of PacBio HiFi and ONT technologies should be preferred to guarantee a good representation of dot chromosome sequences in bird reference genomes.

Materials and Methods

Annotation of non-B DNA in the zebra finch and other bird genomes

The maternal and the paternal haplotypes of the zebra finch T2T genome (GCA_048771995.1 and GCA_048772025.1[31]) were annotated for the following non-B DNA motifs: A-phased repeats (APR), direct repeats (DR), G-quadruplexes (G4), inverted repeats (IR), short tandem repeats (STR), mirror triplex motifs (TRI), and Z-DNA (Z). G4 motifs were annotated with the `G4DISCOVERY` pipeline[11], a hybrid approach that utilizes both `PQSFINDER`[37] and `G4HUNTER`[45] and keeps only motifs that are likely to form, based on experimental data. Z-DNA motifs were annotated using the `Z-DNA HUNTER` webserver[36], downloaded in a bedgraph format and converted to chromosome positions using a python script (https://github.com/makovalab-psu/T2T_bird_nonB/blob/main/python/remap_zdna.py) . All other motifs were annotated with `GFA` (https://github.com/abcsFrederick/non-B_gfa[62]), with default motif parameter settings, which allow for spacer lengths up to 100 bp for mirror repeats and IRs,

and 10 bp for DRs. It is debated whether motifs with long spacers actually form non-B DNA. After inspecting the spacer and repeat length distributions, which showed that most spacers in fact were shorter than 10 bp (Fig. S24), we filtered out motifs with spacers longer than 10 bp. For inverted repeats, we allowed the repeat arm length to be from 6-30 bp (threshold taken from PALINDROME ANALYZER[38]), and for triplex motifs we only considered mirror repeats with 100% purine or pyrimidine content, as these are most likely to form triplex structures[39]. The output was converted to BED format using custom scripts (available on github). As an alternative approach, we assessed G4 with QUADRON[79] (using a dockerized version available at docker://kxk302/quadron:1.0.0) and Z-DNA with GFA and ZSEEKER[63]. QUADRON and GFA are more lenient and find more motifs than G4DISCOVERY and Z-DNA HUNTER, but they overall showed a very similar motif distribution across the genome (Fig. S25). ZSEEKER found more motifs than Z-DNA hunter, but fewer than gfa. Conservatively, we decided to proceed with only the strict set of motifs. For some analyses all motifs were merged into a single track (named All) using BEDTOOLS MERGE v2.31.0[80]. The latest version of the near complete chicken genome[30] (downloaded from https://www.dropbox.com/scl/fo/plq2tm2w9lzlk0ua1rzph/h?rlkey=l6z3rgmjs7ec9azun8nundnzl&e=1&dl=0, including gene annotations), and genome assemblies for Ural owl[40], band-tailed pigeon, Anna's hummingbird[1], great bustard[41], peking duck and emu (accession numbers GCF_047716275.1, GCF_037038585.1, GCF_003957555.1, GCA_026413225.1, GCF_047663525.1 and GCF_036370855.1, respectively, downloaded from NCBI) were processed in the same manner.

Chromosome synteny and assignment

Protein-coding gene sequences for all eight species were extracted from the genome assemblies using GFFREAD v0.12.7[81] and aligned to each other with ORTHOFINDER v2.5.5[82]. For each species pair, homologous genes were summarized based on chromosome location, and chromosomes were assigned to macro-, micro-, or dot chromosome categories based on their gene homology to the zebra finch genome (maternal +Z), for which chromosome group assignments were taken from [31]. This grouping in zebra finch was based on structure rather than just size, so sometimes dot chromosomes were larger than the smallest microchromosomes. For chicken and Pekin duck, the W chromosomes are much smaller than that for zebra finch, and the W chromosome is defined as a microchromosome for chicken in[30], an assignment we kept for our analysis as well. The other species' W chromosomes were assigned to macrochromosomes just as in zebra finch. The chromosome synteny across species is shown in Table S13.

Non-B DNA coverage and enrichment at genes, repeats, and centromeres

For coverage plots, chromosomes were divided into non-overlapping 100-kb windows using BEDTOOLS MAKEWINDOWS, and overlap to each motif type was calculated using BEDTOOLS INTERSECT and custom scripts (https://github.com/makovalab-psu/T2T_bird_nonB/tree/main/python). Positions that were annotated for more than one motif (where different motifs overlapped each other) were only

counted once per motif category. Hence, coverage is defined as the fraction of each window that is annotated as non-B DNA motifs. Circos plots were generated using `CIRCOS v0.69`[83].

Per-chromosome coverage for all motifs was calculated as the sum of positions annotated as each motif type divided by the chromosome length. Coverage was also aggregated into three categories of chromosomes: macro-, micro-, and dot chromosomes, as well as genome-wide (diploid for zebra finch and haploid for the other species). These chromosome-, category-, and genome-wide coverages (as fractions of the regions) were used as baseline levels for different downstream enrichment calculations. For the 11 zebra finch dot chromosomes, euchromatic (A compartment) and heterochromatic (B compartment) regions (defined in[31]) in both haplotypes were downloaded as `BED` format for 200-kb windows from GenomeArk (https://genomeark.s3.amazonaws.com/species/Taeniopygia_guttata/bTaeGut7/manuscript/annotations/3D/bTaeGut7v0.4_MT_rDNA.Cooltools.E1.200kbp.flipped.dip.collated.v0.1.bed) and intersected with the non-B DNA motif annotations. Fold enrichment for A and B compartments compared to genome-wide coverage was calculated separately for each chromosome and haplotype, and the compartments were compared to each other using a Wilcoxon test, corrected for multiple testing with FDR.

Different functional regions were extracted from the diploid zebra finch gene annotations (created with `EGAPX` (https://github.com/ncbi/egapx) and available at GenomeArk https://genomeark.s3.amazonaws.com/species/Taeniopygia_guttata/bTaeGut7/manuscript/annotations/genes/bTaeGut7v0.4_MT_rDNA.EGAPx.v0.1.gtf.gz) and haploid gene annotations for the other species (downloaded from NCBI for all species except chicken, that was downloaded from https://www.dropbox.com/scl/fo/plq2tm2w9lzlk0ua1rzph/h?rlkey=l6z3rgmjs7ec9azun8nundnzl&e=1&dl=0). We used only the longest available isoform for each gene. UTRs were not included in the annotations and we extracted them by subtracting CDSs from exons using `BEDTOOLS SUBTRACT`, and defined them as 5' or 3'UTRs depending on their position relative to the CDS and gene direction using a custom script (https://github.com/makovalab-psu/T2T_bird_nonB/blob/main/bash/05_functional_enrichment.sh). Promoters were not annotated, so we defined them as 1 kb upstream of each TSS (using only the longest isoforms of protein-coding genes). Introns were defined as the region between adjacent exons on the same gene (using the longest isoform), and any overlaps to CDS, UTRs, and promoters (in case of overlapping genes) were removed using `BEDTOOLS INTERSECT`. Intergenic regions were defined as the regions between adjacent genes (including promoters), excluding the ends before the first and after the last gene on each chromosome. All regions were converted into `BED` format before downstream analysis. Each functional region was overlapped with each motif `BED` file using `BEDTOOLS INTERSECT`, and the sum of the unique overlapping positions was divided by the length of the region to obtain the non-B DNA motif coverage. Enrichment was calculated as region motif coverage divided by the genome-wide motif coverage. To test the robustness of the data, we randomly subsampled half of the region length for each functional class and chromosome category and calculated the enrichment for each subsample. This was repeated 100 times. The range of obtained enrichment values, excluding the two lowest and two highest values, was used to obtain error bars for each class

and motif type. For G4s, enrichment was also calculated per gene based on whether they occurred on the coding or template strand, in relation to gene annotations. The content of coding and template G4s were compared in a paired Wilcoxon test, adjusted for multiple testing with FDR.

We used three sets of repeat annotations for zebra finch described in[31]—which contain transposable elements, tandem repeats, and satellites, respectively. All annotations were downloaded from GenomeArk (https://genomeark.s3.amazonaws.com/index.html?prefix=species/Taeniopygia_guttata/bTaeGut7/manuscript/annotations/repeats/), converted into `BED` format, and labelled based on their classification using custom bash scripts (https://github.com/makovalab-psu/T2T_bird_nonB/blob/main/bash/04_repeat_enrichment.sh). Tandem repeats were grouped into the following length categories: 1-4-mers, 5-10-mers, 11-50-mers, 51-100-mers, and >100-mers, based on the length of a single repeat unit. Note that the three different repeat annotation sets overlapped significantly, as the same sequence frequently belonged to more than one annotated repeat type. Coverage and enrichment were calculated for each repeat category in the same manner as for genes, using the genome-wide coverage as the denominator. No repeat annotations were publicly available for the other bird genomes.

Centromeric annotations for zebra finch (defined in[31]) was downloaded from GenomeArk (https://genomeark.s3.amazonaws.com/species/Taeniopygia_guttata/bTaeGut7/manuscript/annotations/centromeres/bTaeGut7v0.4_MT_rDNA.centromere_detector.v0.1.gff) and converted to `BED` format. Centromere locations for most chicken chromosomes were provided by the authors[30]. Enrichment in centromeres was calculated both in relation to the genome-wide coverage as well as to chromosome-wide coverage, the latter to account for inter-chromosomal differences and to test whether each centromere was enriched or depleted for non-B DNA motifs as compared to the rest of the chromosome. To test for significance, we divided the non-centromeric parts of each chromosome into 100 windows of the same size as the corresponding centromere. If there were more than 100 non-overlapping windows, 100 were selected randomly from the total set; if there were fewer, the windows were allowed to overlap. For each window, the coverage and genome- and chromosome-wide enrichment were calculated, and these were used as a background distribution per chromosome to which the centromere enrichment could be compared in a two-sided randomization test. If the centromere enrichment was in the top or bottom 2.5% of the background distribution, it was considered significantly different from the chromosome average. No centromere information was available for the other species.

Methylation analysis

Methylation data for zebra finch downloaded from GenomeArk was converted from bigwig to bedgraph with the UCSC software `BIGWIGTOWIG`[84]. This file contains the percentage of 5-methyl-cytosine (5mC) methylated PacBio HiFi reads for each CpG site in the zebra finch genome. We extracted relevant sites in genes from each chromosome category using `BEDTOOLS INTERSECT` and separated them based on whether they occurred within G4 motifs

or not. For each gene region, we calculated the median methylation level within or outside of G4s. Distributions of all medians were compared with a Mann-Whitney U test. We also investigated what fraction of genes had CpG sites (that could be methylated), for each chromosome category and gene class (within and outside of G4 motifs), and compared fractions from micro- and dot chromosomes to those of macrochromosomes using a Z-test, correcting for multiple testing with FDR.

Experimental *in vitro* validation of zebra finch G4 sequences

We obtained single-strand DNA oligos from four common zebra finch G4 motifs from Integrated DNA Technologies (IDT) with high pressure liquid chromatography purification. We investigated structure using circular dichroism (CD) spectra, UV-absorbance thermal melting spectra ,and native gel electrophoresis, as described in[48]. Briefly, oligos were resuspended in nuclease-free water, and buffer exchanged into 140 mM LiCl, 20 mM LiMOPS pH 7.2 buffer, to remove trace small-molecule impurities. Concentrations were determined with Beer-Lambert law[85] using the UV-absorbance at 260 nm and extinction coefficients provided by IDT. Samples were further diluted to a final 10 μM concentration in either 140 mM LiCl, 20 mM LiMOPS pH 7.2 or the same buffer with 100 mM KCl. Samples were denatured at 95°C for 1.5 min and slowly annealed to 4°C over 90 min. Samples were stored at 4°C for subsequent analysis. CD spectra were collected at 20°C in 0.1 cm cuvettes using a Jasco J-1500 spectrometer. Background CD from the buffer was subtracted from every wavelength for each sample, and molar ellipticity was calculated using the formula E = CD/(33000 × molarity × pathlength in cm). UV-absorbance thermal melting spectra were collected on an OLIS HP 8425 diode array spectrophotometer from 5 to 95℃ at a ramp rate of 0.5℃/min, collecting the absorbance spectra every 0.5℃. Background absorbance from the buffer was subtracted from each sample and absorptivity was calculated using Beer's law. Conformational heterogeneity was assessed by native gel electrophoresis, where samples were fractionated on a 15% polyacrylamide gel containing 10 mM KCl and 1× Tris Borate EDTA (TBE) buffer for about 3 hours at 150 V and 4℃. Bands were visualized using SYBR gold stain on a Bio-Rad Gel Doc Go imager. Molecular weights were compared to a New England Biolabs Low Molecular Weight dsDNA ladder (catalog number N3233).

Coverage analysis in relation to non-B DNA motifs

PacBio HiFi and ONT coverage for zebra finch (in wig format, averaged for 1024-bp windows) were downloaded from GenomeArk (https://genomeark.s3.amazonaws.com/index.html?prefix=species/Taeniopygia_guttata/bTaeGut7/assembly_verkko_0.1/manual_curation/bTaeGut7v0.4/mapping/) and converted to `BED` format using `WIG2BED`[84]. For each window, the non-B DNA motif content was calculated for each motif type, and the correlation between sequence coverage and non-B content was assessed for each chromosome category using a linear model in `R`[86]. Pearson's R^2 and Spearman's ρ^2 were calculated in `R`[86] using the package `GGPUBR`[87]. All figures were prepared in `R` unless otherwise indicated, using the libraries `TIDYVERSE`[88], `GGPLOT2`[89], `GGUPSET`[90] , `PATCHWORK`[91], `COWPLOT`[92], `VIRIDIS`[93], and `GGH4X`[94] .

Data availability

The zebra finch genome is available on NCBI (accessions GCA_048771995.1 and GCA_048772025.1)[31]. The non-B DNA motifs in bed format are available for download from Zenodo, doi 10.5281/zenodo.19225047, and all the code for this manuscript is available on GitHub: https://github.com/makovalab-psu/T2T_bird_nonB.

Declaration of interests

The authors declare no competing interests.

Acknowledgements

We thank Kaivan Kamali for the dockerized version of Quadron and Luohao Xu for centromere coordinates in the chicken genome assembly. We are grateful to Matthias Weissensteiner for providing thoughtful comments on an earlier version of this manuscript. This research was supported by the grant R35GM151945 and by the Willaman Chair Endowment Fund from the Eberly College of Science to KDM. Computations were performed at the Penn State Institute of Computational Data Sciences (RRID:SCR_025154), which provided access to computational research infrastructure within the Roar Core Facility (RRID:SCR_026424).

References

1. Rhie, A. *et al.* Towards complete and error-free genome assemblies of all vertebrate species. *Nature* **592**, 737–746 (2021).
2. Nurk, S. *et al.* The complete sequence of a human genome. *Science* **376**, 44–53 (2022).
3. Rhie, A. *et al.* The complete sequence of a human Y chromosome. *Nature* **621**, 344–354 (2023).
4. Makova, K. D. *et al.* The complete sequence and comparative analysis of ape sex chromosomes. *Nature* **630**, 401–411 (2024).
5. Yoo, D. *et al.* Complete sequencing of ape genomes. *Nature* **641**, 401–418 (2025).
6. Vollger, M. R. *et al.* Segmental duplications and their variation in a complete human genome. *Science* **376**, eabj6965 (2022).
7. Jeong, H. *et al.* Structural polymorphism and diversity of human segmental duplications. *Nat Genet* **57**, 390–401 (2025).
8. Hoyt, S. J. *et al.* From telomere to telomere: The transcriptional and epigenetic state of human repeat elements. *Science* **376**, eabk3112 (2022).
9. Altemose, N. *et al.* Complete genomic and epigenetic maps of human centromeres. *Science* **376**, eabl4178 (2022).
10. Smeds, L. *et al.* Non-canonical DNA in human and other ape telomere-to-telomere genomes. *Nucleic Acids Res* **53**, (2025).

11. Mohanty, S. K., Chiaromonte, F. & Makova, K. D. Evolutionary dynamics of predicted G-quadruplexes in human and other great apes. *Genome Biol.* **26**, 161 (2025).
12. Guiblet, W. M. *et al.* Selection and thermostability suggest G-quadruplexes are novel functional elements of the human genome. *Genome Res* **31**, 1136–1149 (2021).
13. Makova, K. D. & Weissensteiner, M. H. Noncanonical DNA structures are drivers of genome evolution. *Trends Genet* **39**, 109–124 (2023).
14. Guiblet, W. M. *et al.* Non-B DNA: a major contributor to small- and large-scale variation in nucleotide substitution frequencies across the genome. *Nucleic Acids Res* **49**, 1497–1516 (2021).
15. Wang, G. & Vasquez, K. M. Dynamic alternative DNA structures in biology and disease. *Nat Rev Genet* **24**, 211–234 (2023).
16. Esnault, C. *et al.* G-quadruplexes are promoter elements controlling nucleosome exclusion and RNA polymerase II pausing. *Nat Genet* **57**, 1981–1993 (2025).
17. Gummadi, A. S. C., Muppa, D. K. & Yella, V. R. Dissecting non-B DNA structural motifs in untranslated regions of eukaryotic genomes. *Genomics Inform* **22**, 25 (2024).
18. Wang, Y.-R. *et al.* A comprehensive study of Z-DNA density and its evolutionary implications in birds. *BMC Genomics* **25**, 1123 (2024).
19. Yella, V. R. & Vanaja, A. Computational analysis on the dissemination of non-B DNA structural motifs in promoter regions of 1180 cellular genomes. *Biochimie* **214**, 101–111 (2023).
20. Genome. *NCBI* https://www.ncbi.nlm.nih.gov/datasets/genome/?taxon=8782. (March 2026)
21. International Chicken Genome Sequencing Consortium. Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. *Nature* **432**, 695–716 (2004).
22. Warren, W. C. *et al.* The genome of a songbird. *Nature* **464**, 757–762 (2010).
23. Liao, X. *et al.* Repetitive DNA sequence detection and its role in the human genome. *Commun Biol* **6**, 954 (2023).
24. Ellegren, H. (2013). The evolutionary genomics of birds. Annual Review of Ecology, Evolution and Systematics *44*, 239-259.
25. Benjamini, Y. & Speed, T. P. Summarizing and correcting the GC content bias in high-throughput sequencing. *Nucleic Acids Res* **40**, e72 (2012).
26. Peona, V., Weissensteiner, M. H. & Suh, A. How complete are ‘complete’ genome assemblies?-An avian perspective. *Mol Ecol Resour* **18**, 1188–1195 (2018).
27. Bravo, G.A., Schmitt, C.J., and Edwards, S.V. (2021). What have we learned from the first 500 avian genomes? Annual Review of Ecology, Evolution and Systematics *52*, 611-639.
28. Peona, V. *et al.* Identifying the causes and consequences of assembly gaps using a multiplatform genome assembly of a bird-of-paradise. *Mol Ecol Resour* **21**, 263–286 (2021).
29. Barros, C. P. *et al.* A new haplotype-resolved turkey genome to enable turkey genetics and genomics research. *Gigascience* **12**, (2022).
30. Huang, Z. *et al.* Evolutionary analysis of a complete chicken genome. *Proc Natl Acad Sci U S A* **120**, e2216641120 (2023).
31. Formenti, G. *et al.* The complete genome of a songbird. *bioRxiv* 2025.10.14.682431 (2025) doi:10.1101/2025.10.14.682431.
32. Mao, S.-Q. *et al.* DNA G-quadruplex structures mold the DNA methylome. *Nat Struct Mol*

Biol **25**, 951–957 (2018).
33. Niu, K. *et al.* DNA 5mC methylation inhibits the formation of G-quadruplex structures in the genome. *Genome Biol* **26**, 202 (2025).
34. Halder, R. *et al.* Guanine quadruplex DNA structure restricts methylation of CpG dinucleotides genome-wide. *Mol Biosyst* **6**, 2439–2447 (2010).
35. Kumar, S., Stecher, G., Suleski, M. & Hedges, S. B. TimeTree: A Resource for Timelines, Timetrees, and Divergence Times. *Mol Biol Evol* **34**, 1812–1819 (2017).
36. Petrovič, M. *et al.* Z-DNA Hunter tool for straightforward detection of Z-DNA forming regions and a case study in Drosophila. *NAR Genom Bioinform* **7**, lqaf166 (2025).
37. Hon, J., Martínek, T., Zendulka, J. & Lexa, M. pqsfinder: an exhaustive and imperfection-tolerant search tool for potential quadruplex-forming sequences in R. *Bioinformatics* **33**, 3373–3379 (2017).
38. Brázda, V. *et al.* Palindrome analyser - A new web-based server for predicting and evaluating inverted repeats in nucleotide sequences. *Biochem. Biophys. Res. Commun.* **478**, 1739–1745 (2016).
39. Belotserkovskii, B. P. *et al.* Formation of intramolecular triplex in homopurine-homopyrimidine mirror repeats with point substitutions. *Nucleic Acids Res.* **18**, 6621–6624 (1990).
40. Chrysostomakis, I. *et al.* A high-quality reference genome for the Ural Owl (Strix uralensis) enables investigations of cell cultures as a genomic resource for endangered species. *Gigascience* **14**, (2025).
41. Luo, H. *et al.* A high-quality genome assembly highlights the evolutionary history of the great bustard (Otis tarda, Otidiformes). *Commun. Biol.* **6**, 746 (2023).
42. Goodwin, T. J. D. & Poulter, R. T. M. A new group of tyrosine recombinase-encoding retrotransposons. *Mol Biol Evol* **21**, 746–759 (2004).
43. Takki, O., Komissarov, A., Kulak, M. & Galkina, S. Identification of Centromere-Specific Repeats in the Zebra Finch Genome. *Cytogenet Genome Res* **162**, 55–63 (2022).
44. Esnault, C. *et al.* G4access identifies G-quadruplexes and their associations with open chromatin and imprinting control regions. *Nat Genet* **55**, 1359–1369 (2023).
45. Bedrat, A., Lacroix, L. & Mergny, J.-L. Re-evaluation of G-quadruplex propensity with G4Hunter. *Nucleic Acids Res.* **44**, 1746–1759 (2016).
46. Del Villar-Guerra, R., Trent, J. O. & Chaires, J. B. G-Quadruplex Secondary Structure Obtained from Circular Dichroism Spectroscopy. *Angew Chem Int Ed Engl* **57**, 7171–7175 (2018).
47. Mergny, J.-L., Phan, A.-T. & Lacroix, L. Following G-quartet formation by UV-spectroscopy. *FEBS Letters* **435**, 74–78 (1998).
48. Kejnovská, I. *et al.* Clustered abasic lesions profoundly change the structure and stability of human telomeric G-quadruplexes. *Nucleic Acids Res* **45**, 4294–4305 (2017).
49. Huang, Z. *et al.* Evolutionary analysis of a complete chicken genome. *Proc Natl Acad Sci U S A* **120**, e2216641120 (2023).
50. Hackett, S. J. *et al.* A phylogenomic study of birds reveals their evolutionary history. *Science* **320**, 1763–1768 (2008).
51. Hedges, S. B., Parker, P. H., Sibley, C. G. & Kumar, S. Continental breakup and the ordinal diversification of birds and mammals. *Nature* **381**, 226–229 (1996).

52. Kshirsagar, R., Khan, K., Joshi, M. V., Hosur, R. V. & Muniyappa, K. Probing the Potential Role of Non-B DNA Structures at Yeast Meiosis-Specific DNA Double-Strand Breaks. *Biophys. J.* **112**, 2056–2074 (2017).
53. Hron, T. *et al.* Decoding the avian missing gene mystery: Dot chromosomes unmask extensive gene loss and novel genetic instability. *Genome Biol. Evol.* **18**, (2026).
54. Waters, P. D. *et al.* Microchromosomes are building blocks of bird, reptile, and mammal chromosomes. *Proc Natl Acad Sci U S A* **118**, (2021).
55. Huang, Z. *et al.* Evolutionary analysis of a complete chicken genome. *Proc Natl Acad Sci U S A* **120**, e2216641120 (2023).
56. Habermann, F. A. *et al.* Arrangements of macro- and microchromosomes in chicken cells. *Chromosome Res.* **9**, 569–584 (2001).
57. Tamon, L., Ashford, J., Nicholls, M., de Bruijn, M. F. T. R. & Sahakyan, A. B. The emerging sequence grammar of 3D genome organisation. *Hum Genet* **144**, 917–928 (2025).
58. Lee, D. S. M., Ghanem, L. R. & Barash, Y. Integrative analysis reveals RNA G-quadruplexes in UTRs are selectively constrained and enriched for functional associations. *Nat. Commun.* **11**, 527 (2020).
59. Lee, D. H. *et al.* Z-DNA-Containing Long Terminal Repeats of Human Endogenous Retrovirus Families Provide Alternative Promoters for Human Functional Genes. *Mol Cells* **45**, 522–530 (2022).
60. Sahakyan, A. B., Murat, P., Mayer, C. & Balasubramanian, S. G-quadruplex structures within the 3' UTR of LINE-1 elements stimulate retrotransposition. *Nat. Struct. Mol. Biol.* **24**, 243–247 (2017).
61. Eckert, K. A. & Hile, S. E. Every microsatellite is different: Intrinsic DNA features dictate mutagenesis of common microsatellites present in the human genome. *Mol Carcinog* **48**, 379–388 (2009).
62. Cer, R. Z. *et al.* Non-B DB: a database of predicted non-B DNA-forming motifs in mammalian genomes. *Nucleic Acids Res* **39**, D383–91 (2011).
63. Wang, G. *et al.* ZSeeker: an optimized algorithm for Z-DNA detection in genomic sequences. *Brief Bioinform* **26**, (2025).
64. Kasinathan, S. & Henikoff, S. Non-B-Form DNA Is Enriched at Centromeres. *Mol. Biol. Evol.* **35**, 949–962 (2018).
65. Chittoor, S. S. & Giunta, S. Comparative analysis of predicted DNA secondary structures infers complex human centromere topology. *Am J Hum Genet* **111**, 2707–2719 (2024).
66. Chardon, F. *et al.* CENP-B-mediated DNA loops regulate activity and stability of human centromeres. *Mol Cell* **82**, 1751–1767.e8 (2022).
67. Talbert, P. B. & Henikoff, S. Centromeres drive and take a break. *Chromosome Res* **33**, 17 (2025).
68. Liu, Q. *et al.* Non-B-form DNA tends to form in centromeric regions and has undergone changes in polyploid oat subgenomes. *Proc Natl Acad Sci U S A* **120**, e2211683120 (2023).
69. Patchigolla, V. S. P. & Mellone, B. G. Enrichment of Non-B-Form DNA at D. melanogaster Centromeres. *Genome Biol Evol* **14**, (2022).
70. Lahnsteiner, A. *et al.* In vivo detection of DNA secondary structures using permanganate/S1 footprinting with direct adapter ligation and sequencing (PDAL-Seq). *Methods Enzymol* **695**, 159–191 (2024).

71. Henikoff, S. & Furuyama, T. Epigenetic inheritance of centromeres. *Cold Spring Harb Symp Quant Biol* **75**, 51–60 (2010).
72. Georgakopoulos-Soares, I. *et al.* Alternative splicing modulation by G-quadruplexes. *Nat Commun* **13**, 2404 (2022).
73. Ross, M. G. *et al.* Characterizing and measuring bias in sequence data. *Genome Biology* **14**, 1–20 (2013).
74. Browne, P. D. *et al.* GC bias affects genomic and metagenomic reconstructions, underrepresenting GC-poor organisms. *Gigascience* **9**, (2020).
75. Guiblet, W. M. *et al.* Long-read sequencing technology indicates genome-wide effects of non-B DNA on polymerization speed and error rate. *Genome Res* **28**, 1767–1778 (2018).
76. Hosseini, M. *et al.* Deep statistical modelling of nanopore sequencing translocation times reveals latent non-B DNA structures. *Bioinformatics* **39**, i242–i251 (2023).
77. Weissensteiner, M. H. *et al.* Accurate sequencing of DNA motifs able to form alternative (non-B) structures. *Genome Res* **33**, 907–922 (2023).
78. Beauclair, L. *et al.* Sequence properties of certain GC rich avian genes, their origins and absence from genome assemblies: case studies. *BMC Genomics* **20**, 734 (2019).
79. Sahakyan, A. B. *et al.* Machine learning model for sequence-driven DNA G-quadruplex formation. *Sci Rep* **7**, 14535 (2017).
80. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
81. Pertea, G. & Pertea, M. GFF utilities: GffRead and GffCompare. *F1000Res.* **9**, 304 (2020).
82. Emms, D. M. & Kelly, S. OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biol.* **20**, 238 (2019).
83. Krzywinski, M. *et al.* Circos: an information aesthetic for comparative genomics. *Genome Res* **19**, 1639–1645 (2009).
84. Kent, W. J., Zweig, A. S., Barber, G., Hinrichs, A. S. & Karolchik, D. BigWig and BigBed: enabling browsing of large distributed datasets. *Bioinformatics* **26**, 2204–2207 (2010).
85. Kejnovská, I., Renčiuk, D., Palacký, J. & Vorlíčková, M. CD Study of the G-Quadruplex Conformation. *Methods Mol Biol* **2035**, 25–44 (2019).
86. R Core Team. R: A Language and Environment for Statistical Computing. Preprint at https://www.R-project.org/ (2024).
87. Kassambara A (2026). *ggpubr: 'ggplot2' Based Publication Ready Plots*. R package version 0.6.3.
88. Wickham, H. *et al.* Welcome to the tidyverse. *J. Open Source Softw.* **4**, 1686 (2019).
89. Wickham, H. *ggplot2: Elegant Graphics for Data Analysis*. (Springer, 2016).
90. Ahlmann-Eltze C (2026). *ggupset: Combination Matrix Axis for 'ggplot2' to Create 'UpSet' Plots*. R package version 0.4.1.
91. Pedersen T (2025). *patchwork: The Composer of Plots.* R package version 1.3.2.9000.
92. Wilke C (2025). *cowplot: Streamlined Plot Theme and Plot Annotations for 'ggplot2'*. R package version 1.2.0.
93. Garnier, S., et al. (2024). *viridis(Lite) - Colorblind-Friendly Color Maps for R*. (Version 0.6.5).
94. van den Brand T (2025). *ggh4x: Hacks for 'ggplot2'*. R package version 0.3.1.9000.