Non-canonical DNA in bird telomere-to-telomere genomes

Linnéa Smeds[1], Simona Secomandi[2,3], Chul Lee[3], Francesca Chiaromonte[4,5], Erich D. Jarvis[2,3,6], 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 Statistics, Penn State University, University Park, PA 16802, USA
[5]School of Economics, Sant'Anna University, 56127 Pisa, Italy
[6]Howard Hughes Medical Institute, Chevy Chase, MD 20815, USA
*Corresponding author

Highlights

- We present the first analysis of sequences with the potential to adopt non-canonical (non-B) DNA conformations within a telomere-to-telomere (T2T) assembly of a bird genome, the zebra finch, and compare it to that in the near T2T assembly of chicken.
- Non-B DNA, particularly G-quadruplexes, is markedly enriched at regulatory regions such as promoters and 5'UTRs, suggesting its role in regulating gene expression in bird genomes.
- Z-DNA shows strong enrichment at centromeric regions, implying a contribution to centromere architecture and function in the zebra finch.
- The short, gene-rich, and highly recombining dot chromosomes have a strong overrepresentation of non-B DNA, which may act as a tunable regulator of euchromatin activity, but is also correlated with low sequencing depth.

Summary

Non-canonical (non-B) DNA motifs are genomic sequences capable of folding into three-dimensional structures distinct from the canonical right-handed helix. These structures regulate gene expression but also serve as mutation hotspots and are linked to cancer. Because non-B DNA is difficult to sequence, its annotations have been incomplete in most genome assemblies. Telomere-to-telomere (T2T) assemblies now overcome this limitation. Here, we provide a comprehensive analysis of eight types of non-B DNA motifs (e.g., G-quadruplexes and Z-DNA) in the zebra finch T2T genome. Motif content varied strongly by chromosome categories; gene-rich dot chromosomes showed the highest motif levels (22.8-40.5%), microchromosomes intermediate levels (9.8-24.8%), and macrochromosomes the lowest (9.1-10.1%). Within chromosomes, Z-DNA was enriched at centromeres, and G-quadruplexes were enriched at promoters and 5'UTRs. Low methylation at G-quadruplexes suggests they can form and contribute to gene regulation in these regions. Comparable patterns of non-B DNA distribution were observed in the near T2T chicken genome, except that A-phased repeats and not Z-DNA were enriched at chicken centromeres. Overall, our findings indicate that the non-B DNA distribution reflects the distinctive architecture of avian genomes, implicating non-canonical DNA in gene expression and centromere organization. The unusually high density on dot chromosomes is negatively correlated with PacBio sequencing depth, and thus helps explain why these chromosomes have posed exceptional challenges for sequencing.

Introduction

The recent improvements in genome assemblies, driven by advances in long-read sequencing and computational algorithms, have led to a common goal of sequencing *complete and gapless* genomes—generating so-called 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. The 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] In particular, 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 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 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 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 have been critically understudied.

Bird genomes have been included in several large-scale evolutionary studies of non-B DNA motifs.[17–19] However, to our knowledge, a detailed study of the non-B DNA motif landscape in bird genomes has been lacking so far. More than 1,500 bird genomes are now publicly available.[20] However, almost none of them have been assembled gap-free. 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] Yet, they have proven challenging to sequence in full, likely due to their unique genome organization, which consists of 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 chromosomes have been missing from the assemblies.[26] Long-read sequencing technologies do not have the same GC bias as short-read technologies, and have led to a vast increase in bird chromosome-level assemblies,[27] with more than 230 species with such assemblies available on NCBI at the time of writing (NCBI, July 2025). Despite this, sequencing and assembling small bird chromosomes have proven to be continuously difficult, even with the latest sequencing technologies.[28,29] Indeed, the first bird genome assembly with a complete set of assembled small chromosomes—that of a chicken—was published only recently[30] suggested a further classification 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 high enrichment for housekeeping genes.[30]

Here, we present a comprehensive analysis of the non-B DNA landscape in the first complete bird T2T genome of a zebra finch.[31] We study the distribution of non-B DNA in functionally important regions (e.g., different genic compartments), as well as in the sequences added to the T2T vs. previous assembly (e.g., centromeres and repeats). Using the lack of methylation as a proxy for non-B DNA formation,[32–34] we predict where in the genome such structures are likely to fold. This analysis has allowed us to formulate hypotheses about functions of non-B DNA in the bird genome. We compare our results to the non-B DNA motif content of another, near T2T bird genome, that of a chicken,[30] and provide a potential explanation as to why bird small chromosomes have been notoriously challenging to sequence.

Results

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

We annotated eight different types of motifs with the potential of forming non-canonical (non-B) DNA structures (Figure 1A) in the recently available zebra finch diploid T2T assembly.[31] These included A-phased repeats, direct repeats, G-quadruplexes (G4s), inverted repeats, mirror repeats, short tandem repeats, triplex motifs, and Z-DNA (see Methods). Taken together, the non-B DNA motif coverage across the genome was 11.4% (Table S1). The zebra finch genome consists of 11 macrochromosomes (including Z and W), 19 microchromosomes, and 11 dot chromosomes.[31] The chromosomal distribution of non-B DNA motifs in the zebra finch assembly was strikingly skewed towards high densities at small chromosomes, especially for the 11 dot chromosomes, with an overall motif coverage of 30% (Figure 1B, Figure S1, Table S1, see Table S2 for motif counts). Some dot chromosomes had non-B DNA motif coverage as high as 40.5% (Table S1). We expected to observe a particularly high coverage of G4s, due to the high GC content at microchromosomes and dot chromosomes (Figure S2). However, *all* non-B DNA motif types showed significant enrichment at dot chromosomes compared to macrochromosomes. This included A-phased repeats, which require recurring stretches of adenines. Microchromosomes sometimes had intermediate coverage when compared to macro- and dot chromosomes, e.g., for short tandem repeats and Z-DNA motifs, but often had coverage more similar to those on macrochromosomes (Figure 1B). The non-B DNA coverage on the two sex chromosomes—the Z and the W—was 9.8% and 10.1%, respectively, which was typical of the macrochromosomes (Table S1). Thus, we observed substantial differences in non-B DNA motif coverage among chromosome categories, with particular contrasts between macro- and dot chromosomes.

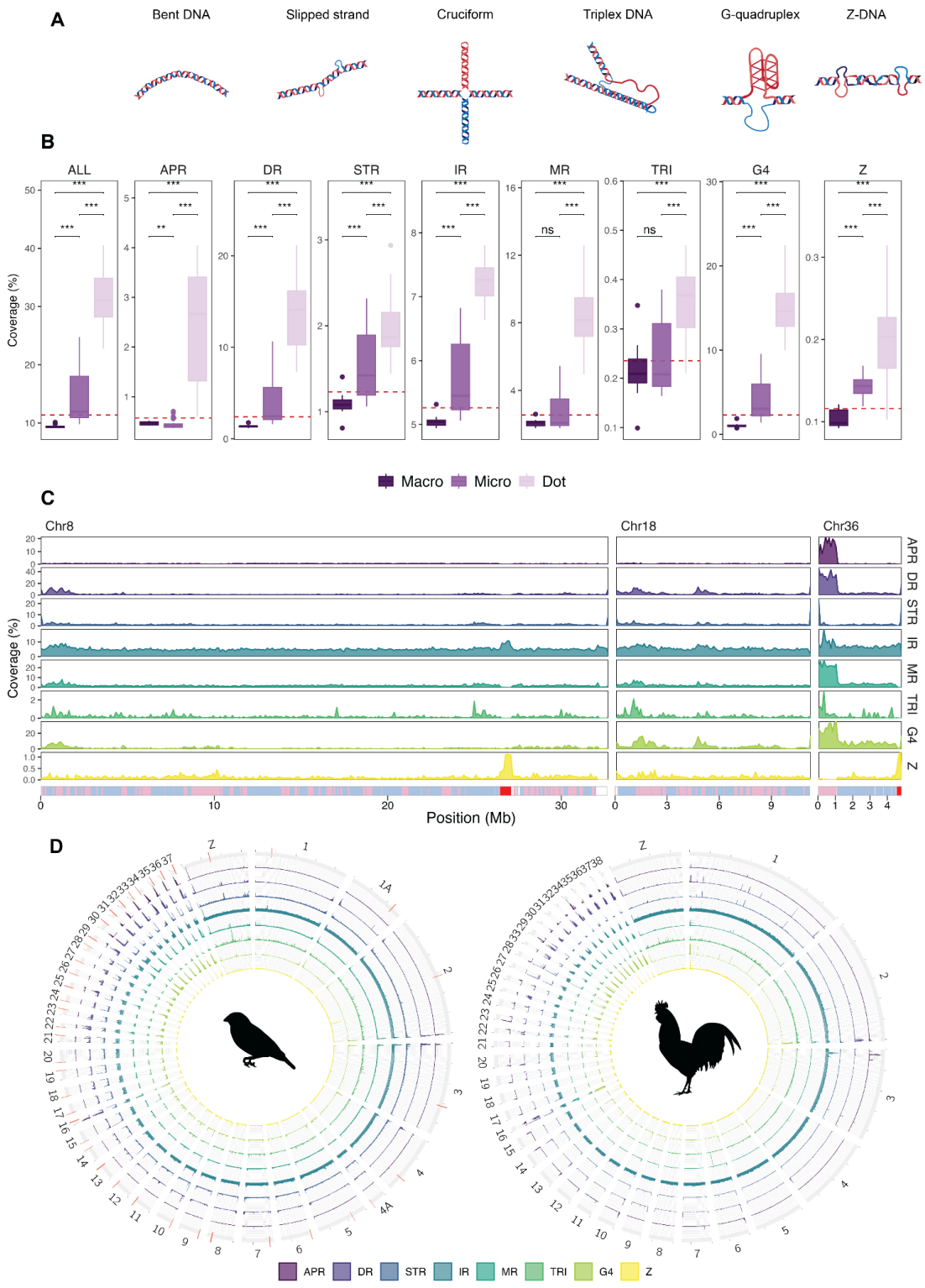

Figure 1. (**A**) Non-B DNA structures: bent DNA formed by A-phased repeats (APR), slipped-strand structure formed by direct repeats (DR) or short tandem repeats (STR), cruciform formed by inverted repeats (IR), triplex DNA formed by triplex motifs (TR), which are a subset of mirror repeats (MR), G-quadruplexes (G4) formed by $G_{3+}N_{1\text{-}12}G_{3+}N_{1\text{-}12}G_{3+}N_{1\text{-}12}G_{3+}$, and Z-DNA (Z) formed by purimidine-purine stretches. (**B**) Overall differences in non-B DNA motif coverage among macro-, micro-, and dot chromosomes. "ALL" means all non-B DNA motif types taken together. ** denotes *P* <0.01, *** denotes *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. **(C)** Examples of non-B DNA motif coverage in 100-kb windows along one macrochromosome (Chr8), one microchromosome (Chr18), and one dot chromosome (Chr36). Maternal haplotypes were used in each case. For each motif type, the y-axis is the same for the three chromosomes. The centromeres are marked in red below the panels, and compartments are shown in pink (A compartment) and blue (B compartment). **(D)** Circos plots with non-B motif coverage for zebra finch (left) and chicken (right). Centromeres in zebra finch are marked with red bars across the karyotype track. Chromosome W is excluded due to its incompleteness in the chicken assembly. Colors as in panel (D). Bird silhouettes are from https://www.phylopic.org.

Visual inspection indicated that the motif landscape was highly dynamic within chromosomes, with peak regions often colocalizing between motif types (Figure 1C). Intrachromosomal variation was larger and absolute peaks were higher on dot chromosomes as compared to macrochromosomes, with a tendency for more peaks close to chromosome ends and Z-DNA motif peaks at the centromeres (Figure 1C and Figure S1).

We also annotated non-B DNA motifs in the nearly complete chicken genome,[30] which showed a remarkably similar non-B DNA motif content to zebra finch, both genome-wide (11.0%, taking all motifs together) and per chromosome category, with the lowest non-B DNA motif coverage on macrochromosomes and the highest on dot chromosomes (Figure 1D, Table S3). Similar to the zebra finch genome, the non-B DNA coverage in the chicken genome displayed a negative correlation with the chromosome size and a positive one with GC-content (Figure S3). Interestingly, dot chromosomes displayed an even more extreme total non-B DNA content in chicken than in zebra finch, with some chromosomes approaching 70%. This is related to the fact that the smallest chicken dot chromosomes are shorter and display a higher GC-content than those of the zebra finch (FigureS3).

We also assessed actual overlaps between motif types (i.e., when the same position was annotated as part of two or more motif types) computationally (Figure S4). In zebra finch, 20% (48 Mb) of non-B annotated bases were annotated as at least two different motif types, and 7% (17 Mb) were annotated as more than two types. Genome-wide, the most common overlap was between G4s and direct repeats (7% of non-B annotated bases, or 16.6 Mb). Additionally, we found contrasting patterns of motif overlaps for the different chromosome categories. On macrochromosomes, overlaps between non-B motif types were sparse (Figure S4, top panels), with the most common overlaps occurring between mirror repeats, direct repeats, and short tandem repeats. On the contrary, dot chromosomes showed substantial overlaps between G4s and several other motif types, especially direct repeats and mirror repeats (Figure S4, bottom panels). Microchromosomes displayed an intermediate pattern, with more G4 overlaps than macrochromosomes, however, not as many as on the dot chromosomes (Figure S4, middle panels). The sex chromosomes showed a lower percentage of bases with overlapping annotations as compared to the autosomes as a group, but similar levels as compared to other macrochromosomes (Figure S5, Table S4).

Overlaps between different motif types in the chicken genome were largely similar to those of zebra finch. Dot chromosomes were an exception, where G4s had even more overlaps with other motif types in chicken than in zebra finch (Figure S6). W chromosome was another

exception (Figure S7). It is a macrochromosome in zebra finch and a (incompletely assembled) microchromosome in chicken, and in general it had more motif overlaps in the latter.

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

Because non-B DNA motifs were found to be enriched at functional regions in the primate genomes,[10] we examined whether the same was true in bird genomes. In zebra finch, macro- and microchromosomes showed moderate enrichment at putative promoter regions (1 kb from the transcription start site, TSS) and 5'UTRs for all non-B DNA motif types (between 1.3-5.9× compared to genome-wide densities) except for A-phased repeats (Figure 2A). The fold-enrichment was usually higher for microchromosomes compared to macrochromosomes, especially for G4s (5.9× and 4.3×, respectively). Dot chromosomes displayed a remarkably different pattern, with the highest fold enrichment observed at intronic regions, which contained more than 10 times the content of A-phased repeats, direct repeats, and G4s compared to genome-wide densities (Figure 2A, bottom panel). A-phased repeats, direct repeats, G4s, and mirror repeats also showed a higher enrichment at promoters and 5'UTRs on dot chromosomes than on macro- and microchromosomes (Figure 2A). G4s were enriched at all functional categories on dot chromosomes, reflecting the higher overall G4 coverage on these chromosomes compared to the genome-wide coverage. Thus, the coverage of non-B DNA motifs at functional regions also varied among the different chromosome categories.

We additionally assessed which strands the G4s preferentially occurred on in relation to the direction of gene transcription. For putative promoters, 5'UTRs, protein-coding regions (CDS), and 3'UTRs, there were more G4s annotated on the template than on the coding strand (Figure 2B). 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.

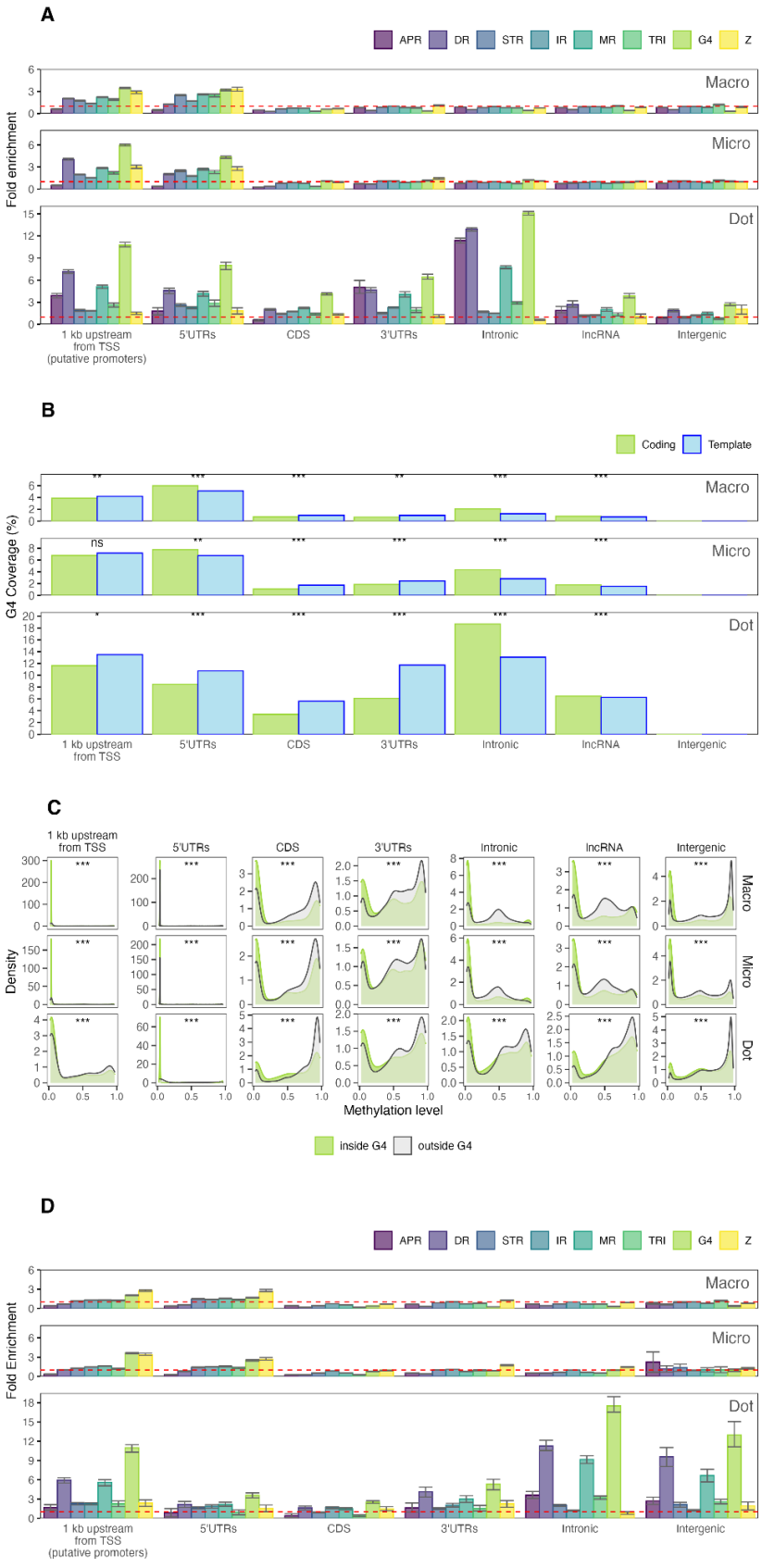

Figure 2. (**A)** Non-B DNA enrichment at functional regions on zebra finch macro-, micro-, and dot chromosomes, as compared to genome-wide non-B DNA motif content (red dashed line). For each bar, an interval was constructed by downsampling the data to 50% 100 times, and excluding the two highest and the two lowest enrichment values obtained; enrichment is considered non-significant if such interval overlaps the red dashed line. **(B)** Mean coverage of G4 motifs on the coding and template strand in zebra finch. Bars are compared using a paired Wilcoxon test corrected with FDR, *** P<0.001, ** P<0.01, * P<0.05, and 'ns' denotes non-significant. Note that intergenic regions are strand-ignorant and therefore excluded here. **(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 Figure S8 . *** P<0.001, Mann-Whitney U test**.** TSS: transcription start site; UTR: untranslated region, CDS: protein-coding region, lncRNA: long non-coding RNA. **(D)** Fold enrichment in functional regions of the chicken genome, separated by chromosome category.

To assess whether G4s fold at functional regions, we used blood methylation levels as a proxy, because G4 formation is inversely correlated with methylation.[32–34]. Comparing median methylation levels per gene for CpG sites inside and outside of G4s, we found that G4s overlapping with genes had significantly lower median methylation levels than the rest of the genic regions (methylation patterns for G4s located on the template vs. the coding strand were similar; Figure S8). This pattern was most prominent at putative promoter regions (1 kb upstream of transcription start sites, TSSs) and 5'UTRs, where the vast majority of CpG sites were unmethylated—especially if overlapping with a G4 motif—and suggests that G4s are likely to fold in these regions (Figure 2C). Interestingly, for dot chromosomes, there was less methylation in 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 (Figure S9). 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 enrichment patterns at functional regions in the chicken genome were remarkably similar to that of zebra finch, with putative promoters and 5'UTRs enriched in G4s for all chromosome categories, and in Z-DNA on macro- and microchromosomes (Figure 2D). Putative promoters on dot chromosomes were enriched in all motif types, as in zebra finch, while 5'UTRs showed a less prominent enrichment. Dot chromosome introns were enriched in almost all motif types, but especially direct repeats, mirror repeats and G4s—just as in zebra finch. Contrary to observations for zebra finch, the same motifs were also enriched in intergenic regions of these chromosomes in chicken. This could explain the even more extreme non-B DNA content for chicken dot chromosomes.

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

All 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 content. The enrichment patterns were mainly shared across the chromosome categories, with some exceptions (Figure 3A). 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 were entirely missing from these chromosomes. Both DNA transposons and miniature inverted-repeat transposable elements (MITEs) belonging to the Mutator superfamily (DTM) showed enrichment for essentially all non-B DNA motif types (Figure 3A). The most striking enrichment among TEs was observed for Ngaro elements, a distinct group of retrotransposons (Goodwin and Poulter 2004), which exhibited an ~288-fold enrichment of Z-DNA motifs compared to the genome-wide content. Ngaro elements are not abundant in zebra finch (they occupy only ~70 kb in the diploid genome), but are present 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 for non-B DNA motifs (Figure 3A), consistent with the pattern observed in great apes.[10]

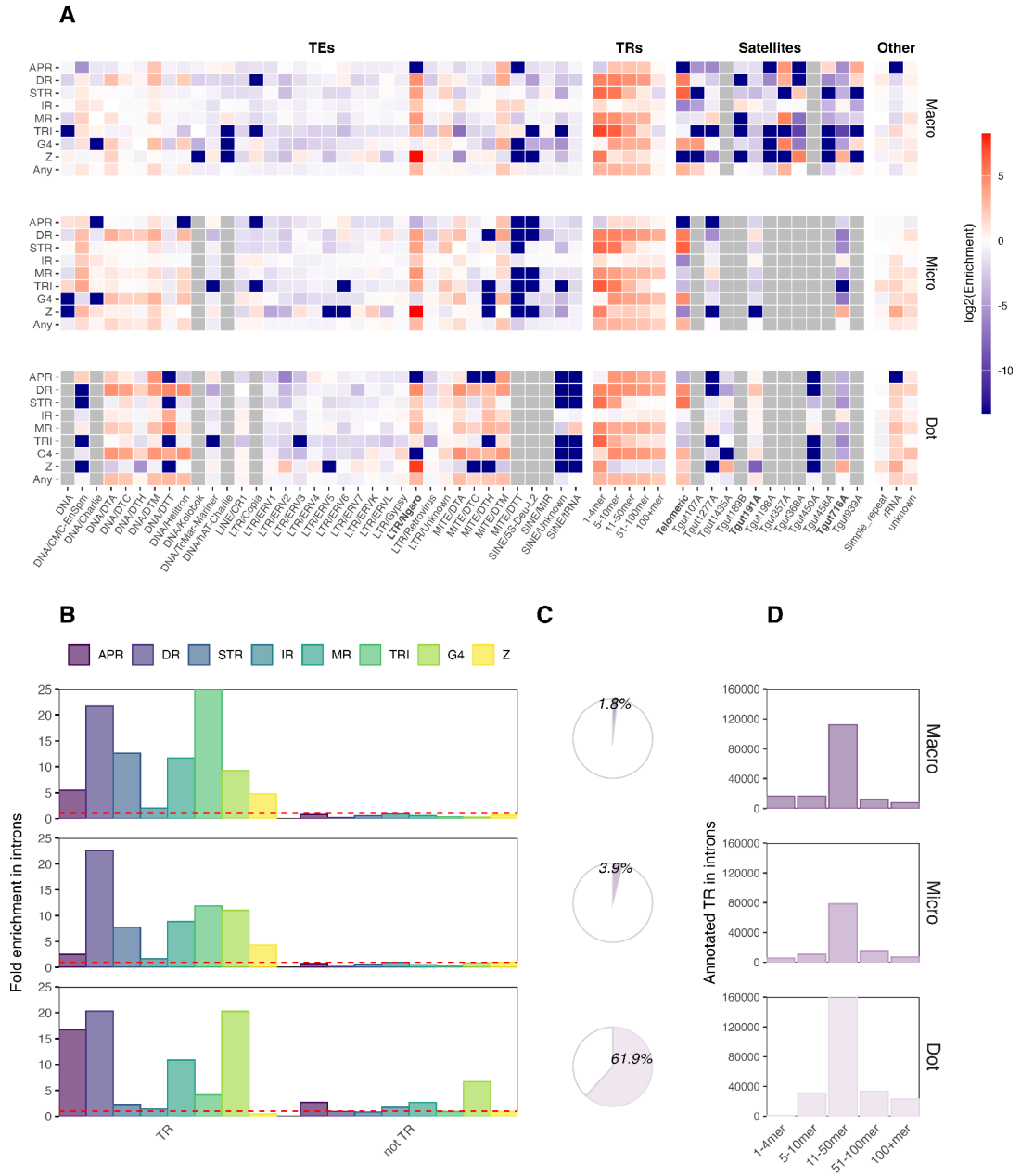

Figure 3. (A) Enrichment of non-B DNA motifs at different repeat classes for each chromosome category. Only repeat classes occupying at least 10 kb in the diploid genome are shown. TE: Transposable elements; TR: tandem repeats. White denotes no enrichment compared to the genome-wide average motif coverage; 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 satellites are marked in bold. **(B)** Enrichment of non-B DNA motifs within introns annotated as tandem repeats ("TR") vs. not annotated as tandem repeats ("not TR"). The dashed red line denotes genome-wide average. **(C)** Fraction of intronic base pairs overlapping with tandem repeats. **(D)** Number of annotated tandem repeats in introns, grouped by repeat unit length.

Tandem repeats of all lengths showed enrichment for essentially all non-B DNA motif types, and this enrichment was negatively correlated with the repeat unit length (Figure 3A, Figure S10). We further investigated overlaps between introns and tandem repeats. We found that tandemly repeated parts of introns exhibited notably higher enrichment for non-B DNA motifs compared to the rest of the introns (Figure 3B). This enrichment was high for all chromosome categories. However, we noticed that tandem repeats were rare on macro- and microchromosomes (they constituted 1.8% and 3.9% of introns, respectively) but highly abundant on dot chromosomes (where they constituted 61.9% of introns, Figure 3C). The most common tandem repeats on dot chromosome introns were minisatellites (Figure 3D), especially with repeated units of 10-11 bp and 20-21 bp (Figure S11).

The telomeres were enriched for direct repeats, short tandem repeats (because of their short unit size and high sequence similarity), and G4s (as expected, because each G4 stem corresponds to different repeat units of the telomeric repeat motif). 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 (Figure S12). One of the two most common satellites, Tgut191A, did not show any enrichment in non-B DNA elements on macrochromosomes and was only enriched in direct repeats on micro- and dot chromosomes. The other most common satellite, Tgut716A, was enriched for Z-DNA motifs in all chromosome categories (Figure 3A). Both of these satellites were previously suggested to be associated with centromeres in the zebra finch.[35] Tgut716A has been identified as the dominant centromeric satellite repeat in our companion study.[31] At the same time, Tgut191A appears to play a different role, potentially mediating ZW conjugation in the PAR region.[31] The 18S/28S rRNA, almost exclusively present on dot chromosome pair 37 in the zebra finch, was enriched in all motif types except for APRs. The 5S rRNA (in Figure 3A named Tgut368A), located in two clusters on macrochromosome pair 2,[31], was highly enriched in Z-DNA and STRs (36.2× and 4.9×, respectively, compared to the genome average).

Bird centromeres are strongly enriched in non-B DNA

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 at almost all (75 out of 80) chromosomes (2.83-17.02× compared to the genome-wide average, Figure 4A, Figure S13), consistent with Tgut716 being enriched in Z-DNA as indicated by our repeat analysis above (Figure 3A). An exception was the W chromosome (Figure 4A), which had a relatively short centromere (~51 kb compared to the average centromere length in the genome of 200 kb) and completely lacked Z-DNA motifs despite containing multiple copies of Tgut716A. A single repeat unit of Tgut716A generally contains two separate Z-DNA motifs, but a detailed investigation of the W centromere showed single-nucleotide changes that interrupted both motifs in each repeat unit. Interestingly, the W chromosome has another much larger cluster (879 kb) of Tgut716A located around 20 Mb, which contains Z-DNA motifs but lacks a centromeric function. Because the average non-B DNA motif coverage differed strikingly among chromosome categories (Figure 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 coverage (Figure S14). This reduced the fold enrichment of Z-DNA on dot chromosomes and microchromosomes only slightly (from 2.8-17.0× to 2.2-12.0×), while there was an increase on macrochromosomes (from 3.4-15.6× to 4.3-18.1×). We also noted that inverted repeats were overrepresented at 73 of 80 centromeres (1.01-2.47× compared to the genome-wide average), but the majority of these enrichments were not statistically significant (Figure 4A). Chicken also showed enrichment in non-B DNA motifs on almost all centromeres, but for other motif types, mainly A-phased repeats (Figure 4B). Indeed, A-phased repeats were significantly enriched for 25 chicken chromosomes.

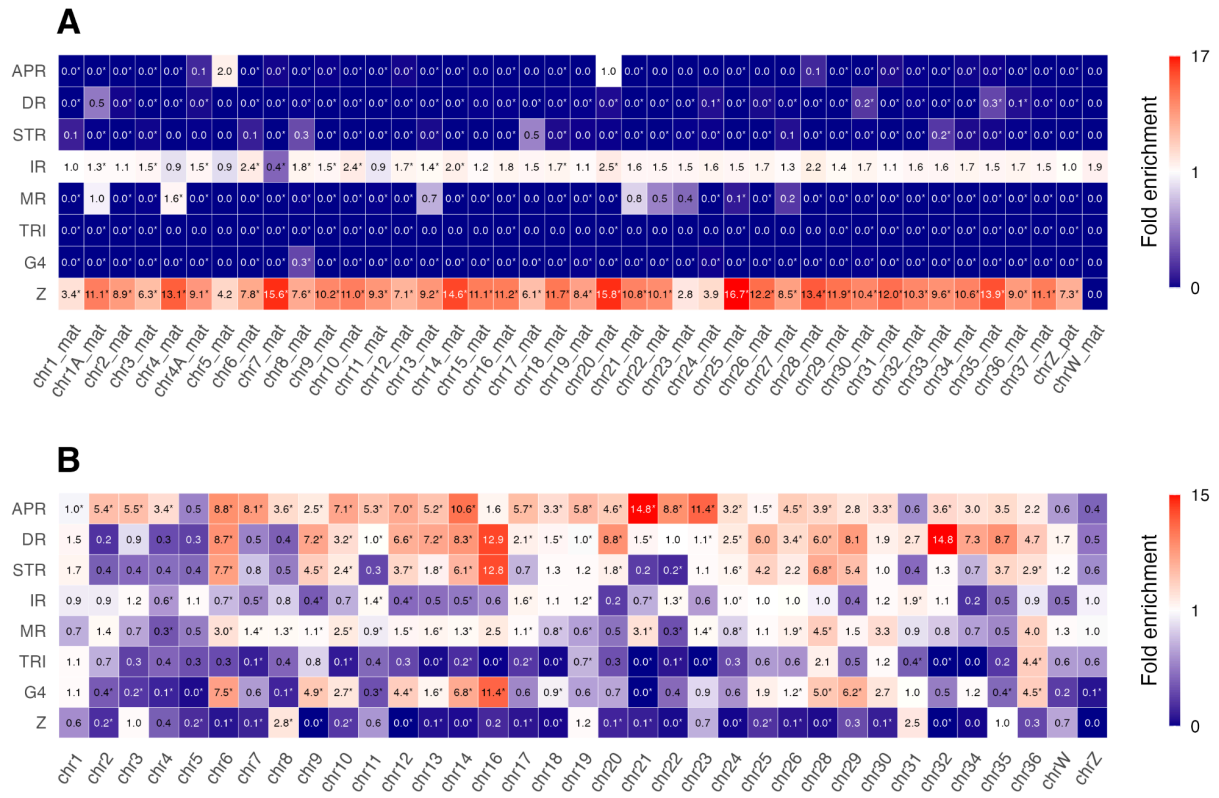

Figure 4. (A) Enrichment of non-B DNA motifs at centromeres in the haploid zebra finch genome (maternal + Z), compared to the genome-wide average. Red and blue denote, respectively, enrichment and depletion compared to the genome-wide non-B DNA content. Values marked with "*" are significantly enriched or depleted as compared to the genome-wide average (two-sided randomization test, $P < 0.05$). Abbreviations are as in Figure 1. See Figure S13 for a full version including all haplotypes, and Figure S14 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 greater detail in zebra finch. Non-B DNA motifs taken together were significantly more enriched in the euchromatic "A" compartment than the heterochromatic "B" compartment of dot chromosomes (Figure 5A).

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 (Figure 5A). 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, compared to the B compartment. Indeed, 2,533 and 1,513 protein-coding genes overlap the A and B compartments, respectively.[31] Moreover, the A compartment has many minisatellites in the introns, in which A-phased repeats, direct repeats, and G4s are highly enriched (Figure S15A). The same enrichment was also seen at intronic minisatellites in the B compartment, but they were much less prevalent (Figure S15B-C). 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.

There were a handful of visible outliers, namely chr29, chr33 and chr34, which had a similar non-B DNA enrichment between their A and B compartments. An additional inspection indicated that, compared to the more typical dot chromosomes, these outliers had a very different compartment organization–with small A and B regions intermingled and not well aligned with gene density, repeat structure, and sequence coverage patterns (see Figure S16 for an example). Outliers notwithstanding, we detected significant intrachromosomal differences in non-B DNA content between dot chromosomes' euchromatic (compartment A) and heterochromatic (compartment B) sequences.

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

To test a hypothesis about the link between non-B DNA and reduced sequencing coverage 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 done 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 54% of the variability in sequencing depth (Figure 5B). The content of direct repeats, G4s, and mirror repeats, when considered separately, explained 41%, 35%, and 25% of the variability in sequencing depth, respectively. Similar but much weaker patterns were found for micro- and macrochromosomes (Figure S17). We also found a negative relationship between non-B DNA and ONT sequencing coverage (Figure 5C, Figure S18), but this was much less pronounced. Indeed, non-B DNA content explained only 16% of the variability in ONT sequencing depth on dot chromosomes.

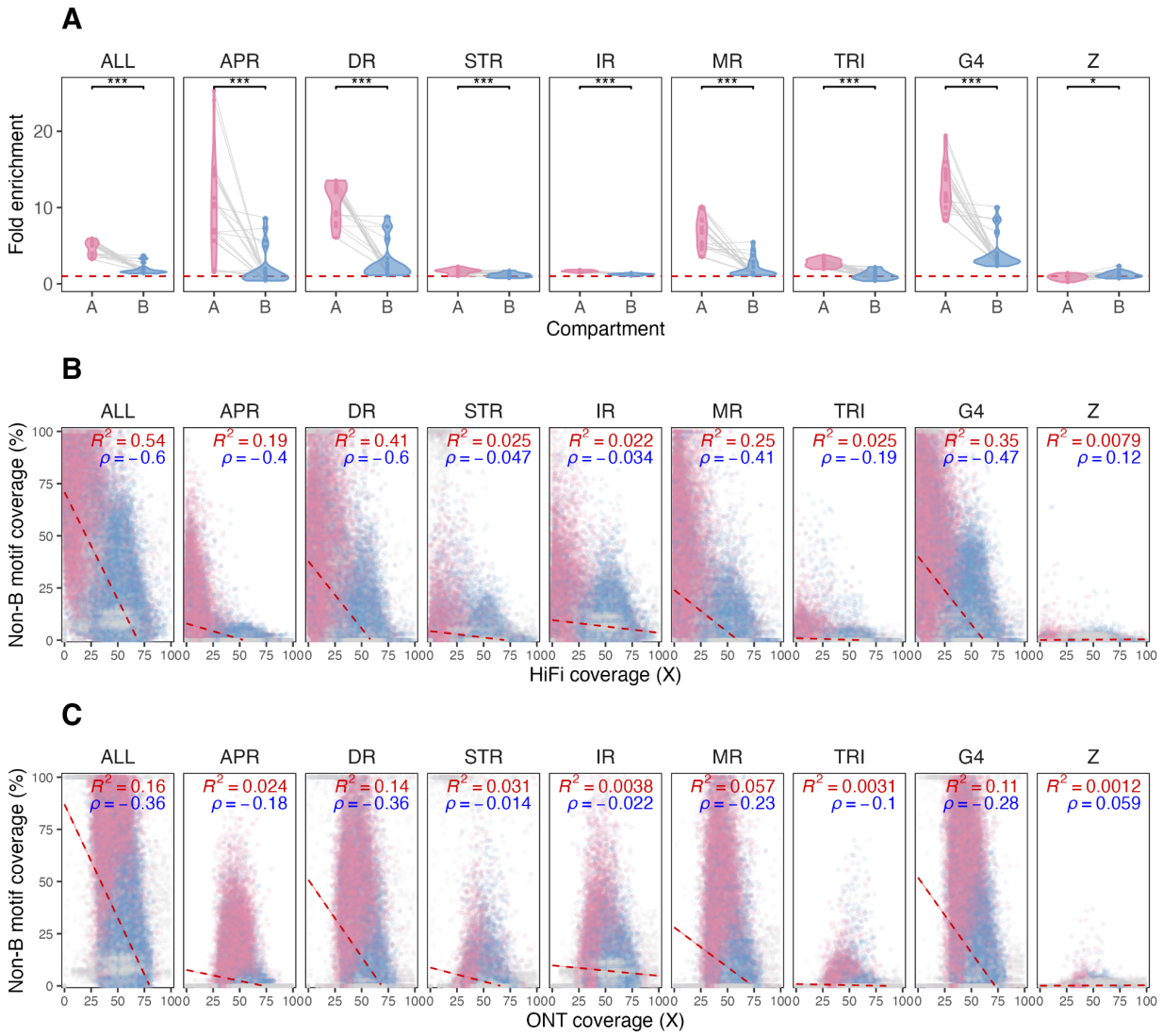

Figure 5. (A) Enrichment of non-B DNA at euchromatic (A) and heterochromatic (B) compartments of the 11 dot chromosomes (averaged across the compartments for each chromosome and haplotype separately) compared to the genome-wide average (red dashed line). Gray lines between violins connect the compartments from the same dot chromosome. * denotes *P* <0.05 and *** denotes *P* <0.001, Wilcoxon test adjusted for multiple testing using FDR. Note that not all parts of the chromosomes were assigned to A or B; 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. Windows are colored by compartment: windows in A are shown in pink, windows in B 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 (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.

Discussion

Here, we conducted a comprehensive analysis of the zebra finch and chicken T2T genomes and found that a total of 11.4% and 11%, respectively, contain sequences with the potential to form non-B DNA structures. This is remarkably similar despite approximately 100 million years (Mya) of divergence.[36] Moreover, these values are very similar to the non-B DNA motif content observed in ape T2T genomes (9.2-14.9%),[10] notwithstanding approximately 300 Mya of divergence from birds.[37]

Micro- and dot chromosomes

The non-B DNA motif content is highly skewed towards small (micro- and dot) chromosomes in both zebra finch and chicken genomes. Small chromosomes in bird genomes are known to have high GC-content, high gene density, elevated recombination rates, and high repetitive element content.[30] Thus, small bird chromosomes represent particularly active, functionally significant genomic units with several processes (e.g., transcription and recombination) requiring active regulation. Non-B DNA likely plays an active role in such regulation in birds, based on previous evidence about its participation in gene expression regulation in humans,[15] data about its facilitation of recombination (e.g.,[38]), and our findings about the high frequency of non-B DNA on the small gene-rich chromosomes in birds. A particularly strong enrichment in non-B DNA was found on the dot chromosomes. Whereas we expected dot chromosomes to be enriched in some non-B DNA motif types because they are G-rich (e.g., G4s), our data support that all non-B DNA motif categories, including A-rich ones (i.e. A-phased repeats), are overrepresented at such chromosomes.

Dot chromosomes portray a highly polarized distribution of non-B DNA motifs between the two compartments. Their euchromatic A compartment is highly enriched in several non-B DNA motif categories—G4s and direct repeats in particular—above the genome-wide levels. In contrast, their heterochromatic B compartment usually has non-B DNA coverage at genome-wide levels. Because the A compartment represents euchromatin, these patterns all point towards a high concentration and availability of non-B DNA for the regulation of functionally important regions of the bird genome.

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, 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[39] in the human genome. Another study, using 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 functionality 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 (see Fig. 5C in [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 birds than in great apes.

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 primates.[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 aligns with the results for human and ape G4s.[11]

Additionally, our results suggest that, compared to macro- and microchromosomes, dot chromosomes have evolved to have 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 is consistent 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]
We did not find a general enrichment of non-B DNA motifs in 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 bears similarities to long terminal repeats (LTRs) and is found in many animals and fungi.[40] LTRs that contain Z-DNA-forming sequences have previously been suggested to act as alternative promoters for genes,[41] and the Z-DNA motif containing Ngaro elements in zebra finch—present on 70% of 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 transposable elements, e.g., L1s (e.g., [42]), and Z-DNA at Ngaro might have a similar function.

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.[43] 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—the macrosatellite previously defined as centromeric[35]—was highly abundant in the pseudoautosomal region (PAR) on sex chromosomes and between the telomere and euchromatin of microchromosomes. However, it showed no overall enrichment for non-B DNA motifs (Figure 3A). The other abundant macrosatellite, Tgut716A, which is presently the strongest candidate for centromeric function,[31]

showed Z-DNA motif enrichment on all chromosomes except for the W chromosome centromere. However, we note that the W had a much longer cluster of Tgut716A with Z-DNA motif enrichment at a different location. Possibly, this is a remnant of an older centromere that has become inactivated. 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 share the Z-DNA enrichment in centromeres.

Non-B DNA structures have been suggested to contribute to defining centromeres[44] and to generating satellite copy number variation important for centromere drive.[45] Z-DNA motifs have previously been shown to be enriched at centromeres in plants[46] and at specific chromosomes in human, chimpanzee, and bonobo.[10] Future studies should investigate whether Z-DNA forms 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 assumed to evolve at high rates.[47] Huang et al.[30] found that most chicken micro- and dot chromosome centromeres had a 41-bp tandem repeat called CNM that often formed higher-order repeats (HORs), while macrochromosomes had different, chromosome-specific tandem repeats. In terms of non-B DNA motifs, we see no clear separation between macro- and micro/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 from 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 explains why some chicken centromeres are enriched in G4s. Further 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 for the different chromosome categories, the dot chromosomes had a 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.[48] Consistent with this, G4s were overrepresented on the coding strand of bird genes (Figure 2B). Alternatively or additionally, non-B DNA at introns might facilitate recombination.[38] Because dot chromosomes are so compact,[31] intronic minisatellites might represent recombination hotspots.

Interestingly, both zebra finch sex chromosomes displayed no enrichment in non-B DNA. This pattern was similar between the zebra finch Z and the ape X.[10] However, the fact that the highly

repetitive zebra finch W (with repeat content of 86%, Table S5) did not show enrichment in non-B DNA is surprising. This is in contrast to the also highly repetitive ape Y chromosome, which was found to have the highest density 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 S5), and TEs are not enriched in non-B DNA in the zebra finch. Additionally, unlike for the ape Y,[10] we did not observe many overlapping motifs of different types on the zebra finch W. The chicken W shows a higher non-B DNA motif enrichment, on a level between microchromosomes and dot chromosomes. This chromosome is not assembled to the T2T level, so the motif content might change depending on what sequences are still missing. However, the fact that the W chromosome is defined as a microchromosome in chicken but a macrochromosome in zebra finch suggests that the observed differences in non-B DNA enrichment are real.

High non-B DNA content on dot chromosomes as a potential reason for not sequencing them in previous 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 PacBio and even less so for the ONT platform.[49,50] 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[51] and the speed of going through pores for Oxford Nanopore Technology.[52] Moreover, non-B DNA motifs elevate sequencing errors, particularly for Illumina sequencing technology, but to a lesser extent for ONT and PacBio technologies.[53] Even in the recent sequencing effort of the zebra finch genome,[31] we observed lower sequencing depths for the euchromatic parts of dot chromosomes, particularly for the PacBio HiFi technology. These results are consistent with a small-scale study examining sequencing depth at five bird genes using Illumina and PacBio technologies and suggesting that non-canonical DNA structures may explain dropout in sequencing depth.[54] Examining this phenomenon genome-wide, 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, on the dot chromosomes. However, the high long-read sequencing depth used for generating T2T assemblies, and the combination of long-read sequencing technologies used—as was suggested[53]—alleviates some of these limitations of 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 or 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 have particularly high density at parts of the bird genome with a high concentration of genes—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 a hypothesis about whether non-B DNA also contributes to regulating other functionally important regions, such as enhancers and origins of replication, as was 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.

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 of housekeeping genes in the euchromatic regions of dot chromosomes. Since dot chromosomes are enriched in all eight 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 HiFi-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 in 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 chicken 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), mirror repeats (MR), short tandem repeats (STR), triplex motifs (TRI), and Z-DNA (Z), as previously described in.[31] In short, `gfa` (https://github.com/abcsFrederick/non-B_gfa,[55]), with default motif parameter settings and the flags `-skipGQ -skipWGET`, was used to annotate all motifs except G4s. This allows for spacer lengths up to 100 bp for MRs and IRs, and 10 bp for DRs. It is debated if motifs with long spacers actually form non-B DNA, but we choose to include them in this study as potentially forming, and we note that the majority of spacers are short, below 10bp. The output was converted to `BED` format using custom scripts (available on github), and triplex motifs were extracted from mirror repeats using `grep 'subset=1'`. G4 motifs were annotated with `Quadron`[56] using a dockerized version (docker://kxk302/quadron:1.0.0) and default settings for one chromosome at a time. The output was merged and converted to `BED` format using custom scripts. Quadron does not score motifs if the flanks are shorter than 50 bp, and any unscored

motif (in general 1-2 per chromosome) was removed from further analysis. All motifs were also merged into a single track (named ALL) using `bedtools merge v2.31.0`,[v2.31.0, 57]. 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) was processed in the same manner.

Coverage, enrichment in genes, repeats, and at 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. 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`[58].

Genome-wide coverage for all motifs was calculated as the sum of (unique) positions in each motif type divided by the diploid genome length (haploid genome length for chicken). This was also done for the three categories of chromosomes: macro-, micro-, and dot chromosomes, defined in, [30,31] as well as for each chromosome separately. These genome-, category-, and chromosome-wide coverages (as fractions of the regions) were used as baseline levels for different downstream enrichment calculations; see further below.

Different functional regions were extracted from the diploid zebra finch gene annotations (created with EGAPx and available at GenomeArk). We used the longest available isoform for each gene and extracted CDS directly from the gtf file. UTRs were not included in the annotations: we extracted them by subtracting CDS from exons using `bedtools subtract`, and defined them as 5' or 3'UTRs depending on their position relative to the annotated start and start codons using a custom script. Introns of protein-coding genes were defined as the region between adjacent exons on the same gene (using the longest isoform), and any overlaps to previously defined CDS and UTRs (in case of overlapping genes) were removed using `bedtools intersect`. Non-protein coding genes were extracted directly from the annotations file containing the longest isoform only (all of which were classified as long non-coding RNAs). Promoters were not annotated, so we defined them as 1 kb upstream of each TSS (using only the longest isoforms of protein-coding genes). Intergenic regions were defined as the regions between adjacent genes, excluding the ends before the first and after the last genes on each chromosome, and removing any potential overlap with previous gene categories. Gene annotations for chicken did not include information on gene type (protein-coding or long non-coding DNA), so genes with CDS were assumed to be protein-coding. 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 regions for each functional class and chromosome category 100 times and calculated the enrichment for each subsample. 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 G-quadruplexes, 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, converted into `BED` format, and labelled based on their classification using custom bash scripts. Tandem repeats were grouped into the following length categories: 1-4 mers, 5-10 mers, 11-50 mers, 51-100 mers, and >100mers based on the length of a single repeat unit. Note that the three different repeat annotation sets overlap significantly, as the same sequence can belong 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 chicken T2T genome.

Centromeric annotations for zebra finch (defined in[31] using previously published primer sequences[59]) was downloaded from GenomeArk and converted to `BED` format. Centromeres 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 if 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 test.

Methylation analysis

Methylation data for zebra finch downloaded from GenomeArk was converted from bigwig to bedgraph with the UCSC software `bigWigToWig`. 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 overlapped with 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 macro chromosomes using a z-test, correcting for multiple testing with FDR.

Detailed analysis of dot chromosomes

Euchromatic and heterochromatic regions (defined in[31]) in both zebra finch haplotypes of the 11 dot chromosomes were downloaded as `BED` format for 10-kb windows from GenomeArk, 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.

Coverage analysis in relation to non-B DNA motifs

PacBio HiFi and ONT coverage for zebra finch (in wig format, averaged for 1024bp windows) were downloaded from GenomeArk and converted to `BED` format using `wig2bed`. 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`[60]. Pearson's R^2 and Spearman's ρ were calculated in `R` using the package `ggpubr`. All figures were prepared in `R` unless otherwise indicated, using the `tidyverse` library[61].

Data availability

The zebra finch genome from our companion paper[31] is available on NCBI (accessions GCA_048771995.1 and GCA_048772025.1). The non-B DNA motifs in bed format are available for download through GenomeArk, and 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 thank Jacob Sieg for a template of the non-B DNA structures in Figure 1A. We are very grateful to Matthias Weissensteiner who provided thoughtful comments on the 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., McCarthy, S.A., Fedrigo, O., Damas, J., Formenti, G., Koren, S., Uliano-Silva, M., Chow, W., Fungtammasan, A., Kim, J., et al. (2021). Towards complete and error-free genome assemblies of all vertebrate species. Nature *592*, 737–746.

2. Nurk, S., Koren, S., Rhie, A., Rautiainen, M., Bzikadze, A.V., Mikheenko, A., Vollger, M.R., Altemose, N., Uralsky, L., Gershman, A., et al. (2022). The complete sequence of a human genome. Science *376*, 44–53.

3. Rhie, A., Nurk, S., Cechova, M., Hoyt, S.J., Taylor, D.J., Altemose, N., Hook, P.W., Koren, S., Rautiainen, M., Alexandrov, I.A., et al. (2023). The complete sequence of a human Y chromosome. Nature *621*, 344–354.

4. Makova, K.D., Pickett, B.D., Harris, R.S., Hartley, G.A., Cechova, M., Pal, K., Nurk, S., Yoo, D., Li, Q., Hebbar, P., et al. (2024). The complete sequence and comparative analysis of ape sex chromosomes. Nature *630*, 401–411.

5. Yoo, D., Rhie, A., Hebbar, P., Antonacci, F., Logsdon, G.A., Solar, S.J., Antipov, D., Pickett, B.D., Safonova, Y., Montinaro, F., et al. (2025). Complete sequencing of ape genomes. Nature *641*, 401–418.

6. Vollger, M.R., Guitart, X., Dishuck, P.C., Mercuri, L., Harvey, W.T., Gershman, A., Diekhans, M., Sulovari, A., Munson, K.M., Lewis, A.P., et al. (2022). Segmental duplications and their variation in a complete human genome. Science *376*, eabj6965.

7. Jeong, H., Dishuck, P.C., Yoo, D., Harvey, W.T., Munson, K.M., Lewis, A.P., Kordosky, J., Garcia, G.H., Human Genome Structural Variation Consortium (HGSVC), Yilmaz, F., et al. (2025). Structural polymorphism and diversity of human segmental duplications. Nat Genet *57*, 390–401.

8. Hoyt, S.J., Storer, J.M., Hartley, G.A., Grady, P.G.S., Gershman, A., de Lima, L.G., Limouse, C., Halabian, R., Wojenski, L., Rodriguez, M., et al. (2022). From telomere to telomere: The transcriptional and epigenetic state of human repeat elements. Science *376*, eabk3112.

9. Altemose, N., Logsdon, G.A., Bzikadze, A.V., Sidhwani, P., Langley, S.A., Caldas, G.V., Hoyt, S.J., Uralsky, L., Ryabov, F.D., Shew, C.J., et al. (2022). Complete genomic and epigenetic maps of human centromeres. Science *376*, eabl4178.

10. Smeds, L., Kamali, K., Kejnovská, I., Kejnovský, E., Chiaromonte, F., and Makova, K.D. (2025). Non-canonical DNA in human and other ape telomere-to-telomere genomes. Nucleic Acids Res *53*. https://doi.org/10.1093/nar/gkaf298.

11. Mohanty, S.K., Chiaromonte, F., and Makova, K.D. (2025). Evolutionary dynamics of predicted G-quadruplexes in human and other great apes. Genome Biol *26*, 161.

12. Guiblet, W.M., DeGiorgio, M., Cheng, X., Chiaromonte, F., Eckert, K.A., Huang, Y.-F., and Makova, K.D. (2021). Selection and thermostability suggest G-quadruplexes are novel functional elements of the human genome. Genome Res *31*, 1136–1149.

13. Makova, K.D., and Weissensteiner, M.H. (2023). Noncanonical DNA structures are drivers of genome evolution. Trends Genet *39*, 109–124.

14. Guiblet, W.M., Cremona, M.A., Harris, R.S., Chen, D., Eckert, K.A., Chiaromonte, F., Huang, Y.-F., and Makova, K.D. (2021). 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.

15. Wang, G., and Vasquez, K.M. (2023). Dynamic alternative DNA structures in biology and disease. Nat Rev Genet *24*, 211–234.

16. Esnault, C., Zine El Aabidine, A., Robert, M.-C., Cucchiarini, A., Magat, T., Pigeot, A., Bouchouika, S., Garcia-Oliver, E., Gawron, K., Basyuk, E., et al. (2025). G-quadruplexes are promoter elements controlling nucleosome exclusion and RNA polymerase II pausing. Nat Genet *57*, 1981–1993.

17. Gummadi, A.S.C., Muppa, D.K., and Yella, V.R. (2024). Dissecting non-B DNA structural motifs in untranslated regions of eukaryotic genomes. Genomics Inform *22*, 25.

18. Wang, Y.-R., Chang, S.-M., Lin, J.-J., Chen, H.-C., Lee, L.-T., Tsai, D.-Y., Lee, S.-D., Lan, C.-Y., Chang, C.-R., Chen, C.-F., et al. (2024). A comprehensive study of Z-DNA density and its evolutionary implications in birds. BMC Genomics *25*, 1123.

19. Yella, V.R., and Vanaja, A. (2023). Computational analysis on the dissemination of non-B DNA structural motifs in promoter regions of 1180 cellular genomes. Biochimie *214*, 101–111.

20. Genome NCBI. https://www.ncbi.nlm.nih.gov/datasets/genome/?taxon=8782.

21. International Chicken Genome Sequencing Consortium (2004). Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. Nature *432*, 695–716.

22. Warren, W.C., Clayton, D.F., Ellegren, H., Arnold, A.P., Hillier, L.W., Künstner, A., Searle, S., White, S., Vilella, A.J., Fairley, S., et al. (2010). The genome of a songbird. Nature *464*, 757–762.

23. Liao, X., Zhu, W., Zhou, J., Li, H., Xu, X., Zhang, B., and Gao, X. (2023). Repetitive DNA sequence detection and its role in the human genome. Commun Biol *6*, 954.

24. Ellegren, H. (2013). The evolutionary genomics of birds. Annual Review of Ecology, Evolution and Systematics *44*, 239-259.

25. Benjamini, Y., and Speed, T.P. (2012). Summarizing and correcting the GC content bias in high-throughput sequencing. Nucleic Acids Res *40*, e72.

26. Peona, V., Weissensteiner, M.H., and Suh, A. (2018). How complete are “complete” genome assemblies?-An avian perspective. Mol Ecol Resour *18*, 1188–1195.

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., Blom, M.P.K., Xu, L., Burri, R., Sullivan, S., Bunikis, I., Liachko, I., Haryoko, T.,

Jønsson, K.A., Zhou, Q., et al. (2021). Identifying the causes and consequences of assembly gaps using a multiplatform genome assembly of a bird-of-paradise. Mol Ecol Resour *21*, 263–286.

29. Barros, C.P., Derks, M.F.L., Mohr, J., Wood, B.J., Crooijmans, R.P.M.A., Megens, H.-J., Bink, M.C.A.M., and Groenen, M.A.M. (2022). A new haplotype-resolved turkey genome to enable turkey genetics and genomics research. Gigascience *12*. https://doi.org/10.1093/gigascience/giad051.

30. Huang, Z., Xu, Z., Bai, H., Huang, Y., Kang, N., Ding, X., Liu, J., Luo, H., Yang, C., Chen, W., et al. (2023). Evolutionary analysis of a complete chicken genome. Proc Natl Acad Sci U S A *120*, e2216641120.

31. Formenti G., Jain N., Medico J., Sollitto M., Antipov D., Barcellos S., Biegler M., Borges I., Chang JK, Chen Y., Cheng H., Conceição H., Davenport M., De Oliveira L. Duarte E. Durham G. Fenn J. Forde N. Galante P. A. Gerhardt K Giani A. Giunta S. Kim J. Komissarov A. Koo B. Koren S. Larkin D. Lee C. Li H. Makova K. Masterson P. Murphy T. McCaffrey K. Mercuri R. Na Y. O'Connell M. J. Ou S. Phillippy A. Popova M. Rhie A. Ruiz-Ruano F. J. Secomandi S. Smeds L. Suh A. Tilley T. Vontzou N. Waters P. Balacco J. Jarvis E. (2025). The complete genome of a songbird. BioRxiv, https://doi.org/10.1101/2025.10.14.682431

32. Mao, S.-Q., Ghanbarian, A.T., Spiegel, J., Martínez Cuesta, S., Beraldi, D., Di Antonio, M., Marsico, G., Hänsel-Hertsch, R., Tannahill, D., and Balasubramanian, S. (2018). DNA G-quadruplex structures mold the DNA methylome. Nat Struct Mol Biol *25*, 951–957.

33. Niu, K., Xiang, L., Zhang, X., Li, X., Yao, T., Li, J., Zhang, C., Liu, J., Peng, Y., Xu, G., et al. (2025). DNA 5mC methylation inhibits the formation of G-quadruplex structures in the genome. Genome Biol *26*, 202.

34. Halder, R., Halder, K., Sharma, P., Garg, G., Sengupta, S., and Chowdhury, S. (2010). Guanine quadruplex DNA structure restricts methylation of CpG dinucleotides genome-wide. Mol Biosyst *6*, 2439–2447.

35. Takki, O., Komissarov, A., Kulak, M., and Galkina, S. (2022). Identification of Centromere-Specific Repeats in the Zebra Finch Genome. Cytogenet Genome Res *162*, 55–63.

36. Hackett, S.J., Kimball, R.T., Reddy, S., Bowie, R.C.K., Braun, E.L., Braun, M.J., Chojnowski, J.L., Cox, W.A., Han, K.-L., Harshman, J., et al. (2008). A phylogenomic study of birds reveals their evolutionary history. Science *320*, 1763–1768.

37. Hedges, S.B., Parker, P.H., Sibley, C.G., and Kumar, S. (1996). Continental breakup and the ordinal diversification of birds and mammals. Nature *381*, 226–229.

38. Kshirsagar, R., Khan, K., Joshi, M.V., Hosur, R.V., and Muniyappa, K. (2017). Probing the Potential Role of Non-B DNA Structures at Yeast Meiosis-Specific DNA Double-Strand Breaks. Biophys. J. *112*, 2056–2074.

39. Lee, D.S.M., Ghanem, L.R., and Barash, Y. (2020). Integrative analysis reveals RNA G-quadruplexes in UTRs are selectively constrained and enriched for functional associations. Nat. Commun. *11*, 527.

40. Goodwin, T.J.D., and Poulter, R.T.M. (2004). A new group of tyrosine recombinase-encoding retrotransposons. Mol Biol Evol *21*, 746–759.

41. Lee, D.H., Bae, W.H., Ha, H., Park, E.G., Lee, Y.J., Kim, W.R., and Kim, H.-S. (2022). Z-DNA-Containing Long Terminal Repeats of Human Endogenous Retrovirus Families Provide Alternative Promoters for Human Functional Genes. Mol Cells *45*, 522–530.

42. Sahakyan, A.B., Murat, P., Mayer, C., and Balasubramanian, S. (2017). G-quadruplex structures within the 3' UTR of LINE-1 elements stimulate retrotransposition. Nat. Struct. Mol. Biol. *24*, 243–247.

43. Eckert, K.A., and Hile, S.E. (2009). Every microsatellite is different: Intrinsic DNA features dictate mutagenesis of common microsatellites present in the human genome. Mol Carcinog *48*, 379–388.

44. Kasinathan, S., and Henikoff, S. (2018). Non-B-Form DNA Is Enriched at Centromeres. Mol. Biol. Evol. *35*, 949–962.

45. Talbert, P.B., and Henikoff, S. (2025). Centromeres drive and take a break. Chromosome Res *33*, 17.

46. Liu, Q., Yi, C., Zhang, Z., Su, H., Liu, C., Huang, Y., Li, W., Hu, X., Liu, C., Birchler, J.A., et al. (2023). 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.

47. Henikoff, S., and Furuyama, T. (2010). Epigenetic inheritance of centromeres. Cold Spring Harb Symp Quant Biol *75*, 51–60.

48. Georgakopoulos-Soares, I., Parada, G.E., Wong, H.Y., Medhi, R., Furlan, G., Munita, R., Miska, E.A., Kwok, C.K., and Hemberg, M. (2022). Alternative splicing modulation by G-quadruplexes. Nat Commun *13*, 2404.

49. Ross, M.G., Russ, C., Costello, M., Hollinger, A., Lennon, N.J., Hegarty, R., Nusbaum, C., and Jaffe, D.B. (2013). Characterizing and measuring bias in sequence data. Genome Biology *14*, 1–20.

50. Browne, P.D., Nielsen, T.K., Kot, W., Aggerholm, A., Gilbert, M.T.P., Puetz, L., Rasmussen, M., Zervas, A., and Hansen, L.H. (2020). GC bias affects genomic and metagenomic reconstructions, underrepresenting GC-poor organisms. Gigascience *9*. https://doi.org/10.1093/gigascience/giaa008.

51. Guiblet, W.M., Cremona, M.A., Cechova, M., Harris, R.S., Kejnovská, I., Kejnovsky, E., Eckert, K., Chiaromonte, F., and Makova, K.D. (2018). Long-read sequencing technology indicates genome-wide effects of non-B DNA on polymerization speed and error rate. Genome Res *28*, 1767–1778.

52. Hosseini, M., Palmer, A., Manka, W., Grady, P.G.S., Patchigolla, V., Bi, J., O'Neill, R.J., Chi, Z., and Aguiar, D. (2023). Deep statistical modelling of nanopore sequencing translocation times reveals latent non-B DNA structures. Bioinformatics *39*, i242–i251.

53. Weissensteiner, M.H., Cremona, M.A., Guiblet, W.M., Stoler, N., Harris, R.S., Cechova, M., Eckert, K.A., Chiaromonte, F., Huang, Y.-F., and Makova, K.D. (2023). Accurate sequencing

of DNA motifs able to form alternative (non-B) structures. Genome Res *33*, 907–922.

54. Beauclair, L., Ramé, C., Arensburger, P., Piégu, B., Guillou, F., Dupont, J., and Bigot, Y. (2019). Sequence properties of certain GC rich avian genes, their origins and absence from genome assemblies: case studies. BMC Genomics *20*, 734.

55. Cer, R.Z., Bruce, K.H., Mudunuri, U.S., Yi, M., Volfovsky, N., Luke, B.T., Bacolla, A., Collins, J.R., and Stephens, R.M. (2011). Non-B DB: a database of predicted non-B DNA-forming motifs in mammalian genomes. Nucleic Acids Res *39*, D383–D391.

56. Sahakyan, A.B., Chambers, V.S., Marsico, G., Santner, T., Di Antonio, M., and Balasubramanian, S. (2017). Machine learning model for sequence-driven DNA G-quadruplex formation. Sci Rep *7*, 14535.

57. Quinlan, A.R., and Hall, I.M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics *26*, 841–842.

58. Krzywinski, M., Schein, J., Birol, I., Connors, J., Gascoyne, R., Horsman, D., Jones, S.J., and Marra, M.A. (2009). Circos: an information aesthetic for comparative genomics. Genome Res *19*, 1639–1645.

59. Knief, U., and Forstmeier, W. (2016). Mapping centromeres of microchromosomes in the zebra finch (Taeniopygia guttata) using half-tetrad analysis. Chromosoma *125*, 757–768.

60. R Core Team (2024). R: A Language and Environment for Statistical Computing. Preprint at R Foundation for Statistical Computing.

61. Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., et al. (2019). Welcome to the tidyverse. J. Open Source Softw. *4*, 1686.