Briefings in Bioinformatics, 2022, **23(6)**, 1–20

https://doi.org/10.1093/bib/bbac441

Advance access publication date 7 October 2022

Review

An updated overview of experimental and computational approaches to identify non-canonical DNA/RNA structures with emphasis on G-quadruplexes and R-loops

Xiaohui Shi, Huajing Teng and Zhongsheng Sun

Corresponding author: Zhongsheng Sun, Key Laboratory of Clinical Laboratory Diagnosis and Translational Research of Zhejiang Province, The 1st Affiliated Hospital of WMU, Nanbaixiang Wenyi Yiyuan Xinyuan District, Ouhai District, Wenzhou 325000, China. E-mail: sunzs@biols.ac.cn

Abstract

Multiple types of non-canonical nucleic acid structures play essential roles in DNA recombination and replication, transcription, and genomic instability and have been associated with several human diseases. Thus, an increasing number of experimental and bioinformatics methods have been developed to identify these structures. To date, most reviews have focused on the features of non-canonical DNA/RNA structure formation, experimental approaches to mapping these structures, and the association of these structures with diseases. In addition, two reviews of computational algorithms for the prediction of non-canonical nucleic acid structures have been published. One of these reviews focused only on computational approaches for G4 detection until 2020. The other mainly summarized the computational tools for predicting cruciform, H-DNA and Z-DNA, in which the algorithms discussed were published before 2012. Since then, several experimental and computational methods have been developed. However, a systematic review including the conformation, sequencing mapping methods and computational prediction strategies for these structures has not yet been published. The purpose of this review is to provide an updated overview of conformation, current sequencing technologies and computational identification methods for non-canonical nucleic acid structures, as well as their strengths and weaknesses. We expect that this review will aid in understanding how these structures are characterised and how they contribute to related biological processes and diseases.

Keywords: motif prediction, computational methods, sequencing techniques, conformation, non-canonical nucleic acid structure

Introduction

The canonical right-handed double-helical structure of B-DNA has been well recognized to play essential role in determining DNA function [1]. However, multiple types of motif sequences, such as direct repeats (DR), inverted repeats (IR), mirror repeats (MR) (H-DNA), tetraplexes and short tandem repeats (STR), are prone to form non-canonical nucleic acid structures at different loci in the genome [2, 3]. Since the late 1950s, many non-canonical nucleic acid structures, containing Z-DNA, hairpin/cruciform, DNA unwinding element (DUE)/base unpairing regions (BURs), triplex DNA, slipped strand DNA, sticky DNA, G-quadruplex DNA (G4), G-quadruplex RNA (rG4) and R-loop/DNA:RNA hybrids have been discovered, and their biological functions have been partially elucidated [4, 5]. The abundance and distribution of these structural sequences varies across the genomes of the species. Eukaryotes have more non-canonical nucleic acid structural motif sequences than prokaryotes [6]. It has been documented that these non-canonical structures play vital roles in DNA metabolism, DNA damage and repair, and genomic instability [7]. Furthermore, numerous studies have reported the association of these structures with various types of neurodevelopmental diseases and cancers, such as G4 in frontotemporal dementia/amyotrophic lateral sclerosis as well as R-loop in breast cancer [2, 8–13].

Experimental approaches for mapping two forms of non-canonical nucleic acid structures, G4/rG4 and R-loops, have been

Xiaohui Shi is a PhD candidate in Key Laboratory of Clinical Laboratory Diagnosis and Translational Research of Zhejiang Province, The first Affiliated Hospital of WMU; Beijing Institutes of Life Science, Chinese Academy of Sciences; University of Chinese Academy of Sciences. Her research interests include genomics and bioinformatics.

Huajing Teng is an associate professor in Department of Radiation Oncology, Key Laboratory of Carcinogenesis and Translational Research (Ministry of Education) at Peking University Cancer Hospital and Institute. His research is in the areas of bioinformatics and genomics.

Zhongsheng Sun is a professor in Key Laboratory of Clinical Laboratory Diagnosis and Translational Research of Zhejiang Province, The first Affiliated Hospital of WMU; Beijing Institutes of Life Science, Chinese Academy of Sciences; CAS Center for Excellence in Biotic Interactions and State Key Laboratory of Integrated Management of Pest Insects and Rodents, University of Chinese Academy of Sciences; Institute of Genomic Medicine, Wenzhou Medical University; IBMC-BGI Center, the Cancer Hospital of the University of Chinese Academy of Sciences (Zhejiang Cancer Hospital); Institute of Basic Medicine and Cancer (IBMC), Chinese Academy of Sciences. His research is in the areas of bioinformatics and genomics.

Received: April 21, 2022. **Revised:** August 22, 2022. **Accepted:** September 13, 2022

detailed in earlier reviews [8, 14–19]. With regard to sequencing strategies for G4/rG4 mapping, antibody-based pull-down and chain-extension stalling [8, 14–16] are the two major approaches. To date, two *in vitro* and two *in vivo* updated sequencing techniques to identify G4s/rG4s have not been reviewed [20–23]. L1H1-7OTD whole-genome amplification (WGA) sequencing and G4-RNA-specific precipitation (G4RP) with sequencing using G4-specific small-molecule ligand/probe BioTASQ are two methods to identify G4s and G4-RNAs *in vitro* [21, 22]. SHALiPE-seq and G4P-ChIP are *in vivo* genome-wide rG4s and G4s profiling methods [20, 23]. Thus, they have been included in the present review. Although three types of techniques for R-loop structure identification, including S9.6 antibody-based methods, RNase H1 IP methods and Single-Molecule R-loop Footprinting (SMRF-seq) have been discussed in previous reviews [17–19], antibody-based method of qDRIP-Seq [24] and a most recent approach of R-loop CUT&Tag published in 2021 [25] have not been included; therefore, we have comprehensively summarized these techniques in our review.

Dozens of computational programs have been developed to identify these structures, and two reviews of these bioinformatics approaches have been reported to predict putative non-canonical DNA/RNA structures on a whole-genome scale [26, 27]. One of these two reviews summarized strategies for identifying four types of non-B DNA motifs, including G4, Z-DNA, cruciform and H-DNA. Nevertheless, only algorithms for predicting non-canonical nucleic acid structures, except for G4 published before 2012, were provided, while the published methods since then were not included [26]. Another review focused on only four types of G4 prediction algorithms [27]. Given that five bioinformatic tools for G4/rG4 motif prediction reported recently and the related computational prediction strategies of the R-loop were also not included in these two reviews [28–35], we overviewed the updated computational methods related to non-canonical nucleic acid structures in this review. Overall, the present review includes not only databases and web servers containing non-canonical DNA/RNA-forming sequences but also the independent software for detecting non-canonical structures, which are mostly based on regular expression (regex), scoring and machine learning. This review provides a comprehensive overview of all updated experimental and computational methods related to non-canonical nucleic acid structures.

Herein, we describe the basic classification of non-canonical nucleic acid structures, high-throughput methods to identify these structures, and computational approaches to predict putative structures. The structures and discovery timelines of various non-canonical structures are shown in Figure 1A. In addition, the merits and demerits of sequencing and computational strategies to predict non-canonical nucleic acid structures are briefly discussed. A summary of the features of sequencing and computational methods to identify non-canonical nucleic acid structures can deepen our understanding of the formation and function of these structures, thereby facilitating further studies on the functions of these structures in different biological processes and related diseases, as well as the application of these structures in basic and applied biosciences.

The classification of non-canonical nucleic acid structures

DNA/RNA G-quadruplex (G4/rG4)

In 1910, Bang reported that concentrated solutions of guanylic acid could form a gel, indicating that guanine-rich DNA sequences could form 'secondary' structures (Figure 1B) [36]. Later, the G4 structure was analysed *in vitro* using X-ray crystal structure diffraction in 1962 [37]. Since then, a large number of biochemical and structural analyses have confirmed that guanine tetrad-forming sequence motifs, including intramolecular DNA and RNA, even intermolecular RNA, spontaneously fold into G4 and rG4 structure *in vitro* [37–40]. In 2001, scientists used antibodies in immunofluorescence experiments to prove the existence of G4 structures *in vivo* [41]. G4/rG4 are bound together by four guanines via Hoogsteen-type hydrogen bonds and are further stabilised by monovalent cations (usually K^+ or Na^+). G-quadruplexes are located in gene promoters, the borders between introns and exons, immunoglobulin heavy chain switch regions, transcription start sites and chromosomal telomeres [9, 42–44]. The role of G4/rG4 in biological functions has been well comprehensively summarized, including telomere maintenance, DNA replication, genome rearrangements, DNA damage response, chromatin structure formation, RNA processing and transcriptional or translational regulation [27]. In addition, the development of sequencing methods and bioinformatics algorithms for G4 has been extensively reviewed [45]. Thus, two sequencing techniques and three computational methods for G4/rG4 developed most recently are evaluated in our review.

R-loop/DNA:RNA hybrid

The R-loop is a non-canonical nucleic acid structure that is formed during transcription. When the nascent RNA transcript segment was annealed to the template DNA strand, a fragment of the displaced non-template DNA was formed. The nascent RNA and its DNA template form DNA:RNA hybrids, whereas the single-strand non-template DNA is displaced (Figure 1C). The R-loop was first reported in 1976 in *in vitro* studies [46]. Twenty years later, this structure was discovered in *in vivo* studies [47]. Genome-wide analysis showed that the R-loops in normal cells occupied more regions than expected [48]. The proportion of this structure in the genome regions of different species shows as follows: about 8% in yeast, 10% in arabidopsis and 5% in human cells [49–51]. A recent review reported that the R-loop structure was more frequently located in highly transcribed genes as well as some centromeres and telomeres, transposable elements and antisense-RNAs or ncRNA regions [48]. It also illustrated that the length of gene regions identified as R-loop-prone was distributed around 100 bp to 2 kb, depending on experimental methods [52–54]. Moreover, the R-loop plays different regulatory roles in different gene regions in the human genome. For example, R-loops at promoters can prevent DNA methylation and promote TF binding, whereas R-loops located in transcription termination sites are required for transcription termination. In addition, R-loops located at gene bodies can block transcriptional elongation [11]. Furthermore, R-loops in genomic regions regulate downstream cellular processes in a sequence-specific manner including telomere maintenance, DNA replication, MMR (DNA mismatch repair), DNA double strand break repair such as Non Homologous End Joining, transcriptional regulation etc. [13].

Regarding the relationship of G4s and R-loops, it has been found that there is a great overlap between the genome-wide map of R-loops in human embryonic kidney cells with G4-forming sequences identified by G4-seq [55]. It has been suggested that the formation of G4s and R-loops is beneficial to similar aspects of DNA structure through stabilizing each other's structures, such as G abundance of displacement chain and negative torsional tension [56].. For example, a study by Yang Zhao *et al.* [57] reported that the R-loop structure could retard the unfolding of G4 by preventing the annealing of double strands of DNA. Meanwhile,

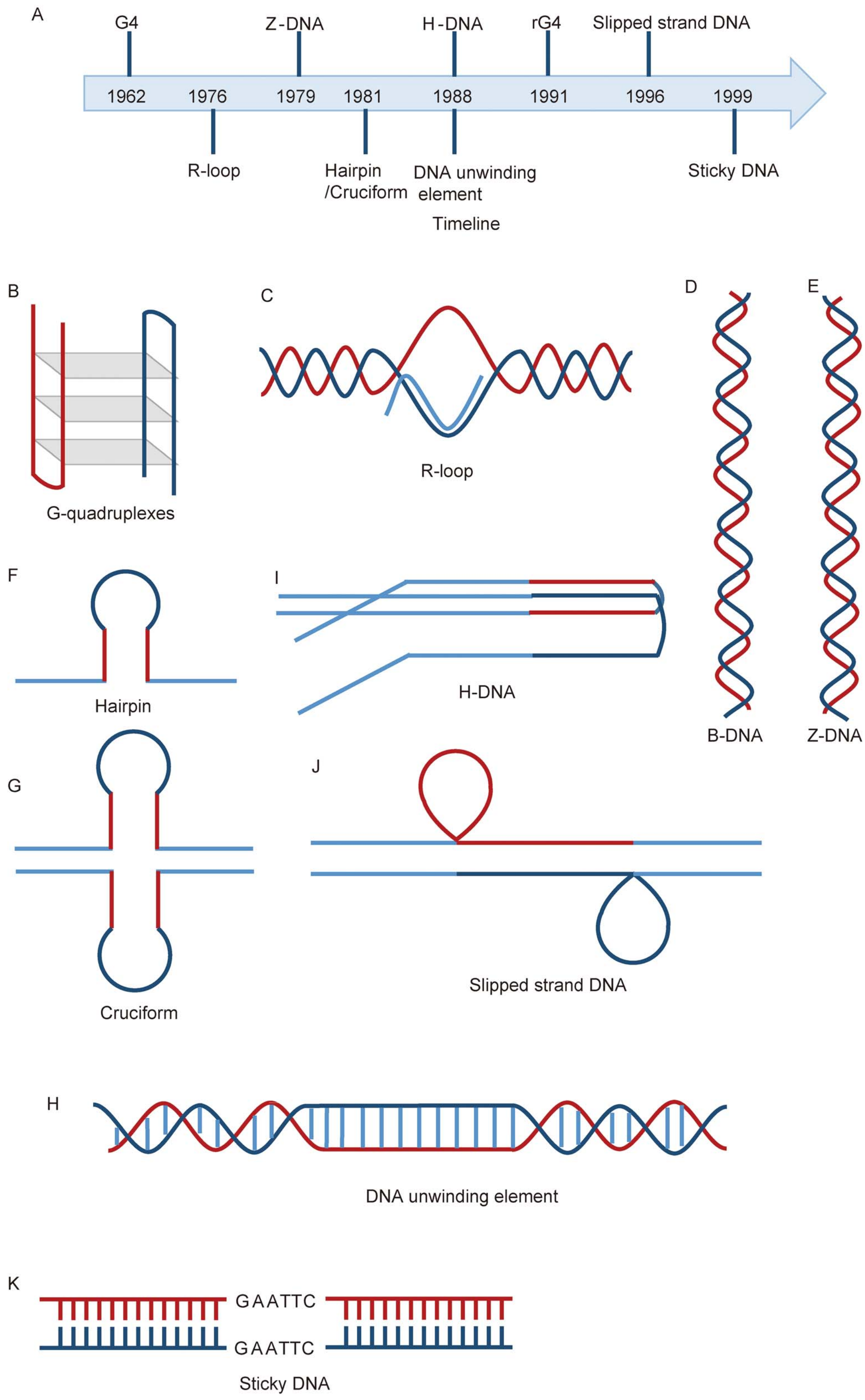

Figure 1. The structures and discovery timelines of various non-canonical structures. (**A**) The discovery timelines of non-canonical DNA/RNA structures. (**B**, **C**, **F**–**K**) Schematic of various non-canonical structures (G-quadruplexes (**B**), R-loop (**C**), hairpin (**F**), H-DNA (**I**), cruciform (**G**), slipped strand DNA (**J**), DNA unwinding element (**H**) and sticky DNA (**K**)). (**D**) The right-handed helical configuration of human DNA (B-DNA). (**E**) The left-handed helix structure of human DNA (Z-DNA).

the R-loop formation and/or stability *in vivo* may also be affected by the G content or G-quartets in displaced single stranded DNA (ssDNA), as observed by electron microscopy [58, 59]. Furthermore, G4 stabilisation by pyridostatin could be prevented by R-loop overexpression, indicating that the R-loop is necessary for G4 ligand-induced DNA damage [56].

Other non-canonical nucleic acid structures except G4/rG4 and R-loop

Z-DNA

Z-DNA was first identified in dsDNA using circular dichroism in 1979 [60]. In 2005, the crystal structure of the B-to-Z-DNA junction was analysed (Figure 1D and E) [61]. The Z-DNA contains extruded base pair on each side of the DNA duplex, which is susceptible to DNA modification. When topoisomerase activity increases, negative relaxation of supercoiled DNA is induced, which promotes the transition from Z to B [62]. Recently, a review of Z-DNA suggested that Z-DNA-binding proteins that interact with B-DNA could directly induce and stabilise Z-DNA [12]. Z-DNA plays an essential role in several biological processes. For example, they are formed during transcription in prokaryotic systems or human cells [63, 64]. The importance of Z-DNA in cellular activities, including regulating gene expression, eliciting immunogenic responses, recruitment of specific proteins/transcription activators or repressors and control of genome instability, has also been mentioned [12]. For instance, the presence of Z-DNA-forming sequences at the breakpoints of human tumour chromosomes suggests that Z-DNA may cause genomic instability by inducing double-strand breaks and large deletions [65].

Hairpin/cruciform (inverted repeats)

Hairpin and cruciform structures were hypothesised to exist soon after Watson and Crick's discovery (Figure 1F and G) [66]. Their existence was experimentally (for example, two-dimensional gel electrophoresis [67]) assessed *in vitro* under natural superhelical densities in the 1980s [68]. Later, a technology for revealing the cruciform *in vivo* was developed, and the biological functions involved in this type of DNA secondary structure were discovered [69–71]. Hairpin/cruciform structures can form by IR [72], also termed palindromes. In somatic and germ cells, the distribution of hairpin/cruciform motifs often overlaps with chromosome regions that are prone to produce large rearrangements [73]. In cell replication and transcription, double-stranded DNA often unrolls into a single strand, and the repeated sequences may be allowed to fold back or form alternative base pairs on the same DNA strand [3]. The expansion and deletion of repeats may be caused by the hairpin structure owing to the stability of the intermediates [3]. A review of hairpins in prokaryotes summarized their functions in many biological processes, including replication origins, transcription, conjugation and recombination [74].

DNA unwinding element (DUE)/base unpairing region (BUR)

The DUE structure was first reported in 1988 (Figure 1H) [75]. When there is a high A + T content feature of DNA sequences that will form unpaired structures under superhelix tension—for instance, $(ATTCT)_n \bullet (AGAAT)_n$ unique repeats among expanding repeat sequences—and is unavailable for these sequences to form other non-B DNAs such as hairpins, intramolecular triplexes, quadruplexes or slipped-strand DNA, the A + T-rich sequences usually form DUEs or BURs [75–77], which are usually related to the origin of replication and the attachment sites of the chromosome matrix [78]. Recent studies have indicated an association between DUEs and aberrant DNA replication [79], suggesting a potential role of DUEs in sequence instability [3].

Triplex DNA (H-DNA, mirror repeats)

Triple-helical nucleic acids were first reported in 1957 [80], in which a stable complex of polyuridylic acids and polyadenylic acid strands with a ratio of 2:1 was demonstrated. In 1986, a short triplex-forming oligonucleotide formed a stable specific triple-helical DNA complex was described (Figure 1I) [81]. In 1988, triplex DNA was reported to include intermolecular triplex and naturally intramolecular triplex [82]. For intramolecular triplex, according to whether the third chain interacts with the purine chain in the double bond through Hoogsteen or reverse Hoogsteen base pairs, triplex DNA can be divided into two types: Hoogsteen and reverse Hoogsteen [83], where the reverse-Hoogsteen structure referring to 'H-DNA'. In the H-palindrome, H-DNA and canonical B-DNA structure are in a state of dynamic equilibrium, which is prone to change into H-DNA with an increase in the negative superhelix. A previous study found that motif sequences prone to form H-DNA occurred more frequently in the promoters of genes than any other random distribution of bases in eukaryotic genomes [84]. A review reported that H-DNAs are important in regulating DNA metabolism, including transcription regulation and gene function. They are inherently mutagenic and have a potential role in gene targeting strategies [10].

Slipped strand DNA (direct repeats)

Slipped-strand DNA structures are composed of DR of complementary strands along the DNA helix axis in a misaligned or slipped manner (Figure 1J). Most slipped chains are found in short DR, termed unstable DNA repeats. After knowing that slipped strand DNA structures exist, $(CTG)_n \bullet (CAG)_n$ and $(CGG)_n \bullet (CCG)_n$ repeats were reported as the first characterization of these structures in 1996 [85], which were associated with myotonic dystrophy (DM) and fragile X syndrome, respectively. The DM2 (myotonic dystrophy types 2) $(CCTG)_n \bullet (CAGG)_n$ repeats can also form slipped strand DNA structures, as illustrated by experiments [86]. There are two possible isomers of slipped DNA strands. Under these two isomers, loops are located at the 5′ and 3′ ends of the DR. This alternative DNA structure or similar intermediate DNA formed during DNA replication or repair may be one of the reasons for the instability of the DNA sequences that form these structures [86]. Slipped strand DNA structures play potential roles in several biological processes, including DNA repair, DNA replication and genomic instability [86].

Sticky DNA (direct repeats)

Sticky DNA is a novel non-B DNA structure formed from GAA•TTC repeats, which is associated with Friedreich ataxia [87] (Figure 1K). It is an intramolecular structure of two long repeat chains of GAA•TTC in the DNA molecule, forming a dumbbell-like conformation in the bacterial plasmid. The formation of sticky DNA has special requirements, including two DR of the GAA•TTC chain in the same molecule, a neutral pH environment, a negative superhelix state and Mg^{2+} [3]. Meanwhile, sticky DNA can only be formed by intramolecular reactions and does not easily dissociate even when heated to 80°C for 60 min [88]. Moreover, researchers have found that in *E. coli*, yeast and eukaryotic cells, long fragments of the GAA•TTC sequence inhibit replication and transcription, suggesting that sticky DNA might regulate gene expression and DNA metabolism [89, 90]. Furthermore, a study showed that the decrease in homologous recombination (HR) was related to an increase in the length of GAA•TTC, indicating that the formation of sticky DNA may contribute to the decrease in HR frequency [91].

High-throughput experimental methods to map non-canonical nucleic acid structure

To the best of our knowledge, only a few high-throughput sequencing approach for other non-canonical DNA/RNA structures except for G4/rG4 and R-loop has been reported; therefore, we mainly focused on sequencing methods for identifying G4/rG4 and R-loop structures (Table 1).

DNA/RNA G-quadruplex (G4/rG4)

For sequencing technologies of G4/rG4 mapping, two approaches have been designed to identify DNA and RNA G4s summarized in published reviews [14–16]. These were antibody-based pull-down and chain-extension stalling methods, respectively. For antibody-based pull-down methods, DNA G4s were first predicted by indirect chromatin immunoprecipitation and next-generation sequencing (ChIP-seq) of phosphorylation of histone H2AX on Ser-139 (γH2AX) in 2012. γH2AX is a marker indicating the activation of DNA damage response (DDR) after treatment with pyridostatin, a highly selective G4 binding small molecule [92]. Then, ChIP-seq methods based on specific single-chain variable fragment antibodies of hf2, D1 and BG4 have been performed to map G4s in the genome [93–96]. In solution, hf2 could pull down the DNA G-quadruplex structure formed by synthetic DNA oligonucleotides [96]. D1 antibody had a high affinity for parallel G4 DNA but not antiparallel G4 DNA and hybrid G4 DNA based on the high level of conformational polymorphisms of G4 [95]. Antibody-based methods used G4 antibodies (e.g. the single-chain variable fragment antibody BG4) together with fluorescently bound secondary or tertiary antibodies to pull down G4s in the nucleus and cytoplasm by ChIP-seq [16]. Moreover, the location of G4 binding proteins (XPB and XPD, transcription-associated helicases) by ChIP-seq could indirectly infer the formation of DNA G4 [97]. G4 clusters could be identified by high-throughput sequencing of genomic DNA amplified via whole-genome amplification in the presence of telomestatin derivative L1H1-7OTD ligand of G4 [21]. However, all of the mentioned approaches rely on antibody specificity, targeted accessibility and cell population averaging. Given that the formation extent of rG4s in human cells is controversial, G4-RNA-specific precipitation (G4RP) with sequencing (G4RP-seq) was designed for capturing transient rG4s in the human transcriptome. It requires chemical crosslinking step followed by affinity capture with G4 ligand of BioTASQ [22].

In terms of chain-extension stalling methods, as G4/rG4 plays a role in the process of stalling DNA polymerase or reverse transcriptase in DNA or RNA [98, 99], the 5′ end of G4/rG4 can be detected *in vitro* by comparing the polymerase pause sites under stable and unstable conditions of G4/rG4. Based on this principle, researchers have proposed technologies to predict G4/rG4 structure, followed by high-throughput sequencing. For experimental procedure of chain-extension stalling methods, G4-seq first provides a reference under non-G4-forming conditions (Read 1) and then determines the locations of G4-dependent DNA polymerase stalling under G4-stabilizing conditions (e.g. in the presence of K^+ or the G4-stabilizing ligand pyridostatin (PDS; Read 2)). In rG4-seq, poly(A)-enriched RNA was used as a reference in the presence of Li^+ under non-G4-forming condition and reverse transcribed with K^+ or PDS to make the RNA G4-dependent reverse transcriptase stalling. Overall, more than 700 000 DNA G4s in human genome as well as thousands of RNA G4s in human HeLa cells have been detected by G4-seq and rG4-seq, respectively [100–103]. The chain-extension stalling approach can map the G4-forming sequences in the genome (G4-seq) or transcriptome (rG4-seq) *in vitro*. However, it should be noted that the effect of the cellular environment (for instance, protein or chromatin structure) on G4 was not considered because of DNA or RNA extraction.

Mapping G4/rG4s from living cells can solve the problem mentioned above, in which the cellular environment was ignored. However, there are still few methods for detecting G4/rG4s in living cells for two major reasons. First, the reducing environment in living cells is not conducive to forming disulfide bonds of antibodies. Second, in the process of formaldehyde fixation, permeabilisation or lysis of cells for detection of G4/rG4s by antibodies, the intracellular environment that maintains the G-quadruplex is usually destroyed, and very few G-quadruplexes can be bound and detected by antibodies as formaldehyde crosslinks the G-quadruplex with endogenous G-quadruplex binding protein. The lack of methods to quantitatively detect the structure of G4 and RNA G4 in living cells greatly limits the in-depth study of the biological function of G4 and RNA G4. To overcome these difficulties, in September 2020, taking advantage of the unique chemical labelling characteristics of RNA G4, Yang *et al.* proposed the first *in vivo* genome-wide RNA G4s profiling method, Selective 2′-hydroxyl acylation with lithium ion-based primer extension (SHALiPE-seq). They designed a set of standard procedures for RNA G4s *in vivo* labelling and quantitative calculation, to realize the quantitative evaluation of the RNA G4s in the whole transcriptome. Hundreds of RNA G4s with strong folding were detected in rice and *Arabidopsis thaliana* by SHALiPE-seq [23]. Meanwhile, in October 2020, Tan Zheng *et al.* screened G-quadruplex-binding peptides from natural G-quadruplex binding proteins and composed an artificial protein (G4P) with high affinity and specificity, which only contains 64 amino acids to recognise the G-quadruplex. G4P contains two short peptides derived from the natural protein RHAU, which can bind to the upper and lower guanine planes of the G-quadruplex, respectively. In their study, G-quadruplex maps on chromosomes of living human, mouse and chicken cells were drawn using G4P with the ChIP-Seq technique, and the existence of G-quadruplexes in cells was further confirmed [20]. Although G4P can detect G-quadruplexes in living cells, some atypical G4 structures, which G4P cannot recognise, may not be detected using this small artificial protein.

For the above methods of identifying G4/rG4 *in vivo*, the antibody/ligand/probe approaches may induce structure formation by perturbing the G4 fold balance in cells or binding G-rich sequence motifs [8, 104]. Therefore, it is important to effectively rule out the possibility of G4 structure being recognized during cell fixation, permeation or staining. Then, because the folding of G4 structure is different in various growth environments and organisms [102], how to accurately identify the diverse folded G4 structures *in vivo* and distinguish the typical and atypical G4 structures are also tremendous challenge. In addition, hundreds of RNA G4s have been found in plant species [23]. How to accurately depict G4 profiling *in vivo* in animal cells at low cost as *in vitro* sequencing methods is also the development trend of identifying G4 structure *in vivo*.

R-loop/DNA:RNA hybrid

For R-loop structure sequencing technologies, three categories of approaches to characterise and map R-loop-forming sequences

Table 1. List of sequencing methods for G4 and R-loop prediction

Structure	Method	Year	Advantages and disadvantages	Reference
Reviewed				
G4	*Antibody-based pull-down methods*			
	Indirect ChIP: histone variant γH2AX	2012	The first method to predict DNA G4s in human cancer cells. The wide distribution of histone markers in damage sites leads to large sequence domains enriched in quadruplex-forming sequences. The location of G4 binding proteins by ChIP-seq can indirectly infer the formation of DNA G4.	[92]
	scFv hf2 G4 pull-down sequencing	2013	G4 ChIP-seq can be used to identify genome-wide DNA G4s. However, it relies on antibody specificity, targeted accessibility and cell population averaging.	[96]
	scFv D1 ChIP-seq	2016		[95]
	scFv BG4 ChIP-seq	2016		[93]
		2018		[94]
	G4BPs ChIP-seq	2014		[97]
	Chain-extension stalling methods			
	G4-seq	2015	The 5′ end of G4/rG4 can be detected in vitro by comparing the polymerase pause sites under stable and unstable conditions of G4/rG4. G4-seq has been applied to the human genome (more than 700 000 DNA G4s have been identified) and other model organisms. RNA G4s can be predicted in human HeLa cells using rG4-seq (thousands of RNA G4s have been detected). The chain-extension stalling approach can map the G4-forming sequences in the genome (G4-seq) or transcriptome (rG4-seq) in vitro. In contrast, the effect of the cellular environment (for instance, protein or chromatin structure) on G4 was not considered because of DNA or RNA extraction.	[100, 101]
	rG4-seq	2016		[102, 103]
R-loop	S9.6 *antibody-based methods*			
	DRIP-seq	2012	Commonly used on the population-average scale. The resolution of R-loop mapping has been improved to near-nucleotides in bisDRIP-seq. S9.6 antibody-based methods are highly antibody-dependent and whether these sequencing approaches can accurately map and quantify the R-loop has been challenged because the binding affinity of the S9.6 antibody varies greatly among different sequencing methods.	[105]
	DRIPc-seq	2016		[51]
	S1-DRIP-seq	2016		[49]
	ssDRIP-seq	2017		[50]
	bisDRIP-seq	2017		[106]
	RNase H1 IP method			
	R-ChIP-seq	2017	Commonly used on the population-average scale. It relies on RNase H1, which can recognise but not process DNA:RNA hybrids. The bias in chromatin fragmentation and digestion efficiency of restriction enzymes may influence the resolution of R-loop mapping.	[55]
	MapR	2019		[108]
	SMRF-seq (Single-Molecule R-loop Footprinting)	2020	It is suitable for R-loop characterisation of single-molecule amplicons with multi-kilobase size at ultra-deep coverage in human genome, but the genome-wide profiling of R-loop was not supported.	[109]
Updated				
G4	L1H1-7OTD whole-genome amplification (WGA) sequencing	2018	G4 clusters could be identified in human genome based on genomic DNA via whole-genome amplification in the presence of telomestatin derivative L1H1-7OTD ligand of G4.	[21]
	G4RP-seq (G4-RNA-specific precipitation (G4RP) with sequencing)	2018	It can capture transient rG4s in the human transcriptome. Sequencing using G4-specific small-molecule ligand/probe BioTASQ to identify G4-RNAs.	[22]
	SHALiPE-seq	2020	The first in vivo genome-wide RNA G4s profiling method, Selective 2′-hydroxyl acylation with lithium ion-based primer extension (SHALiPE-seq).	[23]
	G4P-ChIP	2020	This method screened G-quadruplex-binding peptides from natural G-quadruplex binding proteins and composed a small artificial protein (G4P) with high affinity and specificity, which only contains 64 amino acids to recognise the G-quadruplex. Although it can detect G-quadruplexes in living human, mouse and chicken cells, some atypical G4 structures that G4P cannot recognise may not be detected.	[20]

(continue)

Table 1. Continued.

Structure	Method	Year	Advantages and disadvantages	Reference
R-loop	qDRIP-seq	2020	It combines internal standards human cells for synthetic RNA–DNA hybridisation with high-resolution, strand-specific sequencing. It allows accurate standardisation in the case of interference with the R-loops and allows quantitative measurements.	[24]
	R-loop CUT&Tag	2021	Using Tn5 to tag DNA and DNA:RNA hybrids provides the possibility of avoiding fragmentation bias caused by RNase H1. It only requires fewer cells (half of a million) and is easier and more straightforward, so it can save time in library preparation. Furthermore, it can more specifically identify the R-loop structure in the promoter region and can more sensitively detect the transient R-loop in the gene body and enhancer regions in human genome.	[25]
Other non-canonical nucleic acid structures	ChIP-seq to detect Z-DNA	2016	The ChIP-seq experiment for detection of Z-DNA-forming sites genome-wide in HeLa cells using Zaa probe for the first time. There are 391 Z-DNA-forming sites were found totally, which have been functionally examined in vivo. In addition, Z-DNA-forming sites were related to H3K4me3 and H3K9ac, indicating that the Z-DNA site was associated with active transcription.	[64]
	Potassium permanganate footprinting combined with high-throughput sequencing	2017	The first method is based on Potassium permanganate footprinting. The second method is based on kethoxal-assisted single-stranded DNA sequencing. The first method found abundant formation of Z-DNA, G-quadruplex and H-DNA. The second method found G-quadruplex, Z-DNA, hairpin, H-DNA and cruciform structures. They generate single stranded DNA (ssDNA) first and then find the computational predicted non-B DNA structures near the ssDNA regions. They revealed the existence of multitude of non-B DNA structures in human genome, but the limitation lies in their dependence on non-B DNA structure prediction algorithms.	[113]
	Kethoxal-assisted single-stranded DNA sequencing	2020	[114]	
RNA	SHAPE-seq	2011/2012/2014	It can obtain the quantitative and single nucleotide analysis of secondary and tertiary structure information of hundreds of arbitrary sequence RNA molecules. It has higher sensitivity with approximately 0.1 pmol of RNA needed. It can resolve RNA structural changes due to point mutations with bar coding. It is not limited by the environment required for enzyme function and is usually used for variable buffer and temperature conditions. An automated and rigorous pipeline was designed due to the digital characteristic of direct cDNA sequencing.	[115–117]

in vitro have been reviewed so far: S9.6 antibody-based methods, RNase H1-based methods and SMRF-seq [18, 19]. However, another independent R-loop CUT&Tag method that has been recently published has not been reviewed until now [25]. S9.6 antibody-based and RNase H1-based methods are the two main types of strategies commonly used to provide the distribution and abundance of R-loops on the population-average scale. The first approach is mostly carried out through DNA:RNA immunoprecipitation (DRIP) to predict the R-loop structure [105]. The R-loop structure can be profiled by sequencing the DNA strand (DRIP-seq) [105] or RNA strand of the R-loop after cDNA synthesis (DRIPc-seq) [45], as well as by degrading the displaced ssDNA of an R-loop using S1-nuclease (S1-DRIP-Seq) [49] and identifying single strands of DNA:RNA hybrids by strand-specific R-loop maps [50]. Meanwhile, unpaired cytosines in the displaced DNA strand could be deaminated by combining the S9.6 antibody with sodium bisulfite treatment using bisDRIP-seq, which improved the resolution of R-loop mapping to near-nucleotides [106]. Moreover, qDRIP combines internal standards for synthetic RNA–DNA hybridisation with high-resolution, strand-specific sequencing. It allows accurate standardisation in the case of interference with the R-loops and allows quantitative measurements, thus providing biological insights that were unattainable before [24]. The second approach relies on RNase H1, which can recognise but not process DNA:RNA hybrids [107]. For example, R-ChIP uses an RNase H1 mutant with catalytic death for immunoprecipitation [55]. Meanwhile, defective RNase H1 (dRNase H1) has also been applied in a CUT&RUN-based method termed 'MapR' [108]. The third technology, SMRF-seq, is suitable for R-loop characterisation of single-molecule amplicons with multi-kilobase size at ultra-deep coverage, but the genome-wide profiling of R-loop was not supported [109].

The two types of R-loop mapping methods mentioned above revealed that the R-loop is mainly formed at various sites. For example, using the S9.6 antibody-based method, the main formation sites of the R-loop structure included transcribed gene bodies, GC-skewed CpG island promoters and terminal genic regions. However, the RNase H1-based method showed that the R-loop was mainly formed in G-rich genes related to the promoter-proximal pausing of RNA polymerase II. Even in the loci identified by these two strategies, there is a significant disparity in the location of the signal, which may be attributed to the difference in the formation position of the R-loop structure under different mapping methods [19]. Therefore, improving the rigor and reproducibility of these methods and how to evaluate the inconsistency of sequencing quality effectively is one of the research trends in high-throughput sequencing methods for R-loop identification.

Although these techniques have been widely discussed, they have limitations that need to be addressed. For RNase H1 IP-based methods, the bias in chromatin fragmentation and digestion efficiency of restriction enzymes may influence the resolution of R-loop mapping [110]. In addition, sequencing strategies based on the S9.6 antibody are highly antibody-dependent, and the binding affinity of the S9.6 antibody varies greatly among different sequencing methods, so whether these sequencing approaches can accurately map and quantify the R-loop has been challenged [111, 112]. Given these issues, the R-loop Tn5-based cleavage under targets and tagmentation (CUT&Tag) method, which combines CUT&Tag and glutathione S-transferase–hexahistidine–2× hybrid-binding domain (GST-His6–2 × HBD), was proposed, in which HBD can specifically recognise DNA:RNA hybrids [25]. Based on the fact that Tn5 transposase can combine with DNA:RNA hybrids randomly and transpose the adaptor to two strands of DNA:RNA hybrids, and transposed products can displace the strand, researchers have used Tn5 to tag DNA and DNA:RNA hybrids, which provides the possibility of avoiding fragmentation bias caused by RNase H1. Compared with MapR, R-ChIP or DRIPc-seq, the R-loop CUT&Tag method only requires fewer cells (half of a million) and is easier and more straightforward, so it can save time in library preparation. Furthermore, it can more specifically identify the R-loop structure in the promoter region and can more sensitively detect the transient R-loop in the gene body and enhancer regions.

Other non-canonical nucleic acid structures

For other non-canonical nucleic acid structures, three high-throughput sequencing approaches to map Z-DNA, hairpin, H-DNA or cruciform have been reported [64, 113, 114]. In 2016, the ChIP-seq experiment for detection of Z-DNAs genome-wide in HeLa cells using Zaa probe (containing two Za copies) was published for the first time [64]. The Zaa can be generated by replacing Zβ with Zα that has a much higher affinity and specificity to Z-DNA, where Zα and Zβ are the Z-DNA-binding domain motifs of the N-terminal region of human RNA adenosine deaminase (hADAR1). A total of 391 Z-DNAs have been identified *in vivo*. Most of 10 randomly selected Z-DNAs were verified by Zaa ChIP-qPCR and *in vitro* Z-DNA cleavage assay. Most of detected Z-DNAs were located in the promoter region. In addition, Z-DNAs were related to H3K4me3 and H3K9ac, indicating that the Z-DNA site was associated with active transcription.

Since then, two similar techniques of detecting non-B DNAs were reported. The first one was published in 2017 by combining potassium permanganate footprinting with high-throughput sequencing [113], whereas another approach based on kethoxal-assisted single-stranded DNA sequencing was published in 2020 [114]. Both techniques generate ssDNA first, and then find the computational predicted non-B DNA structures near the ssDNA regions. Both of these two approaches revealed the existence of multitude of non-B DNA structures in human genome, but the limitation of these methods lies in their dependence on non-B DNA structure prediction algorithms. The difference between these two methods is that in the first method, Z-DNA, G-quadruplex and H-DNA could be identified, whereas the second technique could found hairpin and cruciform structures in addition to the structures detected in the first one.

Regarding that there is no specific high-throughput sequencing method for other non-canonical RNA structures, a method named SHAPE-seq, which could obtain the quantitative and single nucleotide analysis of secondary and tertiary structure information of hundreds of arbitrary sequence RNA molecules, was developed [115–117]. SHAPE-seq combines selective 2′-hydroxyl acylation analysed by primer extension (SHAPE) chemistry and deep sequencing of primer extension products with multiple paired ends, resulting higher sensitivity with approximately 0.1 pmol of RNA needed. It can resolve RNA structural changes due to point mutations with bar coding. In addition, SHAPE-seq is not limited by the environment required for enzyme function, and is usually used for variable buffer and temperature conditions. Furthermore, an automated and rigorous pipeline was designed due to the digital characteristic of direct cDNA sequencing.

In general, it is difficult to detect the non-canonical nucleic acid structures in genome-wide by experimental sequencing methods. Because these structures are in the process of dynamic formation, which usually perform specific functions and then decompose. The existing experimental methods are only limited to identifying these structures that are still active at the time of the experiment.

Computational methods to predict the non-canonical nucleic acid structure

The computational methods for analysing other non-canonical nucleic acid structures, except for G4/rG4 and the R-loop, reviewed [24], are listed in Table 2. Herein, we focused on recently published computational methods for the identification of G4/rG4, R-loop and other non-canonical DNA/RNA structures (Table 3). We divided these bioinformatics methods into two classes, including the integration resources of databases and web servers as well as the standalone tools for the identification of putative non-canonical nucleic acid structures, which are mostly based on regex, scoring and machine learning.

DNA/RNA G-quadruplex (G4/rG4)

In January 2020, a review of open-source computational methods for the prediction and implementation of G4 was published. The description of the architecture as well as the merits and drawbacks of three types of models (regex-based, scoring-based and machine learning-based methods) have been discussed [27]. The classical method of regular expression matching algorithms initially use consensus sequence, which usually has a strict pattern to identify potential non-canonical DNA/RNA structures from the primary sequence. Nevertheless, the variable structures of regular expression-based approaches often ignore the nonstandard motif of non-canonical structures. The principle of the scoring-based method is to score and rank each possible sequence to predict the sequence most likely to form non-canonical DNA/RNA structures when there are multiple alternatives. Compared with regular expression-based algorithms, scoring-based methods have less strict criteria, which provide more possibilities for predicting the results. Regular expression and scoring-based methods are based on biophysical and biochemical data or known observed structures, which may not be suitable for predicting novel conformations or sequences purely through computational research. Therefore, algorithms based on machine learning, data-driven predictions have been developed. The main drawback of machine-learning-based algorithms is that they rely on the quality and

Table 2. List of computational methods for the prediction of H-DNA, Z-DNA and cruciform reviewed in current bioinformatics

Method	Structure	URL	Year	Reference
_	cruciform/H-DNA	not provided	1995	[155]
_	cruciform	not provided	2012	[156]
UNAFold	cruciform	http://www.unafold.org/ (accessible)	2008	[157]
TRACTS	H-DNA	http://bioportal.weizmann.ac.il/tracts/tracts.html (no update/not accessible)	2003	[158]
TFO	H-DNA	spi.mdanderson.org/tfo (no update/not accessible)	2006	[159]
TTS mapping	H-DNA/G4	http://ggeda.bii.a-star.edu.sg/~piroonj/TTS_mapping/TTS_mapping.php (under maintenance)	2009	[160]
_	H-DNA	http://www.fi.muni.cz/lexa/triplex (no update/not accessible)	2011	[161]
Z-hunt	Z-DNA	not provided	1986	[162]
Z-Hunt-II	Z-DNA	not provided	1992	[163]

quantity of available training datasets. After this review, five updated programs of PENGUINN, G4detector, DeepG4, rG4-seeker and G4-miner for G4 and rG4 exploration were published [28, 29, 33–35].

In 2020, a machine learning-based software, PENGUINN, was reported by using convolutional neural network (CNN) to detect G4 forming sequences in the nucleus [28]. PENGUINN is a CNN-based approach that provides an independent tool for implementing the training model and a web source that can measure the potential of a sequence to form G4. It identifies G4s from raw DNA sequence data. The input data of PENGUINN could be a single sequence, FASTA format or multiple sequences in multiple lines. In terms of the length of the input sequence, the web application can accept a sequence from 20 to 200 nt. After evaluated the sequence, the score and threshold evaluation were be outputted finally. Using human G4-seq data as a training set, they proved that PENGUINN is superior to state-of-the-art approaches [118–121] in simulating high background testing sets with high genomic variation. G4s in other species could also be predicted by human trained model of PENGUINN, such as the nematode *Cenorhabditis elegans* and the plant *A. thaliana*, but the performance of this model in these species is lower than that expected by human or mouse. In addition, when intersecting PENGUINN detections against G4detector and the regular expression method predictions. PENGUINN detected 3818 'unique' G4s with a threshold of 0.5, which was more than the G4detector (440 G4s) [34] and regular expression (24 G4s). Compared to the regular expression method, PENGUINN offers a scoring system for priority ranking sequences, although the former method is widely used.

Similarly, G4detector is also a machine learning-based software to predict G4 based on CNN. G4detector improves the detection accuracy by adding RNA secondary structure information to the sequence information. It shows excellent performance when benchmarked against novel G4-seq datasets of multiple species genomes. It can predict the whole genome G4s with high accuracy, and can extrapolate the measurement results of human-trained to various non-human species. Moreover, to better interpreting the 'black boxes' prediction, G4detector visualizes the most important features in a given input sequence by integrating gradient and mutation maps. However, it does not integrate RNA structure information into G4 prediction task well [34].

DeepG4 is also a CNN-based method to map DNA motifs predicting G4 region activity and cell-type specific active G4 regions accurately at 201-bp resolution (with AUC (area under the receiver operating characteristic curve) > 0.98). G4 activity can be used for assessing the ability of G4 sequences to form *in vivo*. The combination dataset of G4-seq, G4 ChIP-seq and ATAC-seq has been used for training set. DeepG4 focuses on predicting specific motifs in the active G4 region rather than G4 sequences with flexible patterns. In addition, it has been used to identify active G4 regions in a tremendous amount of tissues and cancers, thus representing a resource for the G4 community. However, DeepG4 requires several hundred bases to map G4, thereby limiting the resolution of G4 mapping. Moreover, the dependence on training set of machine learning-based methods will restrict their performance. Using human data as training set, G4 detection of DeepG4 on non-mammalian genomes seems to be less accurate [35].

In addition to directly using the DNA sequence data to predict the motif that can form G4, Chow *et al.* proposed an rG4-seeker to identify rG4 motifs from rG4-seq experimental data to improve the experimental result and false-positive identification [29]. In order to infer the biological function of rG4 *in vivo* based on association analysis, the relationship between the preparation chemistry of rG4-seq library and the attributes of sequencing data was established. rG4-seeker was then employed to mitigate local sampling errors and background noise in rG4 sequences. In order to screen rG4 candidates with high confidence in functional research, the replication independence of rG4-seeker would significantly reduce the cost of rG4 screening in rG4-seq dataset. Compared with previous method [103], rG4-seeker could better discriminate false-positive, and its sensitivity to non-canonical rG4s is also improved. rG4-seeker could also be incorporated into the bioinformatic pipeline of other RNA-seq technologies, which can improve analysis results. Based on the approach, some novel features missed in the HeLa rG4-seq dataset were recognised using the rG4-seeker with experimental validation.

In contrast to existing methods that need to change the sequencing environment, G4-miner, reported in 2021, use the conventional sequencing process and sequencing data based on the often-neglected sequencing quality information in the sequencing data [33]. G4-miner is a user-friendly and portable genome-wide DNA G-quadruplex (G4) map analysis method, which can identify G4 structures from ordinary genome-wide resequencing data by determining slight fluctuations in sequencing quality. 736 689 G4 structures were identified in the human genome using this approach. More than 89% of the detected typical G4 were identified by polymerase termination analysis and next-generation sequencing. The predicted detection rates of the typical tetrads for different species (*Homo sapiens, Mus musculus, Drosophila melanogaster, A. thaliana, Caenorhabditis elegans and Saccharomyces cerevisiae*) ranged from 32% to 58%,

Table 3. The methods and databases for non-canonical nucleic acid structures identification

Method	URL	Description	Advantages and disadvantages/greater detail explanation of databases	Year	Reference
G4 (updated)					
PENGUINN	https://github.com/ML-Bioinfo-CEITEC/penguinn(accessible)	A machine learning method based on Convolutional neural networks to explore nuclear G-Quadruplexes.	It is a CNN based approach, which implements the training model and a web source, and can measure the potential of a sequence to form G4. The input data of PENGUINN could be raw DNA sequence data including a single sequence, FASTA format or multiple sequences in multiple lines with the length from 20 to 200 nt. The final output is the sequence score. It is superior to state-of-the-art approaches [114–117] in simulating high background testing sets with high genomic variation. It supports the nematode *Cenorhabditis elegan* and the plant *Arabidopsis thaliana*, but the performance of this model in these species is lower than that expected by human or mouse.	2020	[28]
G4detector	https://github.com/OrensteinLab/G4detector(accessible)	A machine learning method based on Convolutional neural networks to explore genome-wide G-Quadruplexes.	G4detector uses a multi-kernel CNN to classify DNA sequences that tend to form G4. It allows the use of the raw DNA sequence as input. It outperforms pqsfinder, G4hunter, and Quadron in predicting G4s. It can predict the whole genome G4s with high accuracy, and can extrapolate the measurement results of human-trained to various non-human species. However, it does not integrate RNA structure information into G4 prediction task well. When intersecting PENGUINN detections against G4detector and the regular expression method predictions. PENGUINN detected 3818 'unique' G4s with a threshold of 0.5, which was more than the G4detector (440 G4s) and regular expression (24 G4s).	2022	[34]
DeepG4	https://github.com/raphaelmourad/DeepG4(accessible)	A deep learning approach to predict cell-type specific active G-quadruplex regions.	DeepG4 is also a CNN-based method to map DNA motifs predicting G4 region activity and active G4 regions in the cell at 201-bp resolution from the DNA sequence and chromatin accessibility accurately. It is firstly designed to assess the ability of G4 sequences to form in vivo. It focuses on predicting specific motifs in the active G4 region rather than G4 sequences with flexible patterns. In addition, it has been used to identify active G4 regions in a tremendous amount of tissues and cancers. However, the resolution of G4 mapping for DeepG4 is limited. Using human data as training set, G4 detection of DeepG4 on non-mammalian genomes seems to be less accurate.	2021	[35]
rG4-seeker	https://github.com/TF-Chan-Lab/rG4-seeker (accessible)	A pipeline used tailored noise models to predict non-canonical rG4s from rG4-seq data.	rG4-seeker mitigates local sampling errors and background noise in rG4 sequences. It screens rG4 candidates with high confidence. Compared with previous method [100], rG4-seeker could better discriminate false-positive, with improved sensitivity. rG4-seeker could also be incorporated into the bioinformatic pipeline of other RNA-seq technologies such as SHAPE-seq and eCLIP, which can improve analysis results. It recognised some novel features missed in the HeLa rG4-seq.	2020	[29]
G4-miner	https://github.com/tulabcode/G4-miner (accessible)	Direct genome-wide identification of G-quadruplex structures by whole-genome resequencing.	It supports G4 structure identification in the human genome and other species including *Mus musculus, Drosophila melanogaster, A. thaliana, Caenorhabditis elegans and Saccharomyces cerevisiae*, which demonstrated that G4-miner is widely applicable. It can be used to identify and characterize genome-wide G4s of specific individuals. G4-seq tends to detect bulges, while G4-miner tends to detect two quartets.	2021	[33]
R-loop					
R-loopDB	http://rloop.bii.a-star.edu.sg/ (accessible)	An integrated database; The first bioinformatics tool for RLFS search and visualization; All RLFS identified in more than half of human genes were collected; A variety of bioinformatics sources were integrated.	The first R-loop related database that contain R-loop information. It contains 245 181 RLFSs in more than half of human genes. A computational tool for RLFS search and visualization has been developed. Many oncogenes, tumour suppressor genes and neurodegenerative disease-related genes have been suggested to prone to form R-loop. Users can enter the name of the coding gene and the website will predict the possibility of R-loop formation based on the sequence of the gene. The RLFS in human genes could be visualised. As the first edition of the R-loop-related database, R-loopDB provides researchers with the first comprehensive RLFS catalogue and inspires to predict the selective R-loop structure using the RLFS model. However, it only integrates the gene sequence from UCSC, so the quantity and quality of R-loop structures on different genes in the genome are limited. In addition, it only supports detection in human genome.	2012	[30]

(continue)

Table 3. Continued.

Method	URL	Description	Advantages and disadvantages/greater detail explanation of databases	Year	Reference
QmRLFS-finder	http://rloop.bii.a-star.edu.sg/?pg=qmrlfs-finder (accessible)	A web server or a tool support command line; The first open source software for R-loop prediction; RLFS coordinates identification and visualization; Any DNA or RNA sequence is supported.	It is used for detecting and analysing the structure and sequence coordinates of RLFS. It supports the input data of DNA/RNA sequence as R-loopDB of 2012. All RLFS tables, FASTA, BED and CUSTOM TRACK can be obtained as outputs. The overlapped RLFS regions from DRIP-seq and QmRLFS-finder confirmed its sensitivity of 79.2%. QmRLFS-finder is the first open-sourced R-loop prediction tool.	2015	[32]
R-loopDB	http://rloop.bii.a-star.edu.sg (accessible)	The updated R-loopDB; Support accessing to experimental data; genome-scale prediction of RLFSs for humans, mouse, rat, chimpanzee, chicken, frog, fruit fly and yeast.	It is the renewed R-loopDB combined RLFS strand, GC skew value and experimental R-loop detection data based on the predicted RLFSs by QmRLFS-finder. It includes experimental data compared with R-loopDB in 2012. It supports searching for RLFSs in the 2 kb upstream and downstream flanking sequences of the entire gene and any gene. Chromosome coordinates, sequences and genomic data of 1 565 795 RLFSs across 121 056 genes from eight species (human, chimp, mouse, rat, chicken, frog, fruit fly and yeast) were collected. It provides the scientific community with a tool that integrates RLFS query and prediction analysis on one platform with strong interactivity.	2017	[31]
R-loopBase	https://rloopbase.nju.edu.cn (accessible)	A database systematic integrating R-loop regulators in human, mouse, yeast and *Escherichia coli*; the functional relationship between individual R-loops and their putative regulators were deduce.	It is the first database integrating R-loop distribution and regulation. To systematically combing and annotating R-loop regulatory proteins, corresponding information integration and visual presentation have been made. First, the author collected 107 sets of high-quality genome-wide R-loop detection data from 11 R-loop detection technologies and 26 human tissues and cells that have been published so far. After strict quality control and standardized analysis, the basic supporting data of R-loopBase was finally formed. Secondly, the authors integrated all the known 1293 R-loop regulatory proteins in human, mouse, yeast and *E. coli* and annotated their molecular functions and gene expression profiles in detail. Finally, in order to facilitate R-loop researchers to make full use of the above information, the author further integrates rich multi-omics data resources and constructs an interactive R-loop expert database interface R-loopBase. Users can retrieve the R-loop formation of their region of interest by gene name, location coordinates and sequence; The interested proteins can be retrieved to understand their regulatory information and regulatory regions related to R-loop; In addition, users can visualize the above data through the genome browser of R-loopBase.	2021	[127]
Other non-canonical nucleic acid structures					
Databases and web servers					
Non-B DB v2.0	https://nonb-abcc.ncifcrf.gov/apps/site/default(accessible)	A database integrating complete categories of motifs prone to form non-B DNAs.	It integrates Z-DNA, G4, DR, IR, MR, STR and a phased repeat. Non-B DB has established search criteria for each type of motif according to the sequence characteristics. It supports non-B DNA motif identification in humans, chimps, dogs, macaques and mice. As a curation database, Non-B DB provides data on known non-B DNA motifs across multiple species but lacks tools for the identification of novel non-B DNA. In 2013, Non-B DB was updated into Non-B DB v2.0, which deepened the non-B DNA forming motif coverage, added visualisation tools and increased seven organisms, including orangutans, rats, cows, pigs, horses, platypus and *A. thaliana*.	2011/2013	[128–129]

(continue)

Table 3. Continued.

Method	URL	Description	Advantages and disadvantages/greater detail explanation of databases	Year	Reference
nBMST	http://nonb.abcc.ncifcrf.gov/apps/nBMST/ (noupdate/notaccessible)	A Web server that allows customizing analyse and detect non-B DNA motifs.	It supports graphical user interface, batch processing capability, dynamic visualization, result storage for up to 6 months, various downloadable file formats for further analysis and FAQs content. A method for adjusting the trade-off between false negatives and false positives to meet the needs of users is provided. It supports various types of genome-wide analyses. Its PolyBrowse function includes the possible association between predicted non-B DNA forming motifs and pathogenic effects. It is widely used without bioinformatics skills. It can be applied to any type of DNA sequence, including viral and bacterial genomes, up to 20 MB. It can be used combined with other bioinformatics tools for non-B structure motif prediction.	2012	[130]
DNA structure search	http://www.utexas.edu/pharmacy/dnastructure/ (no update/not accessible)	A new web-based search engine to detect H-DNA-forming and Z-DNA-forming sequences in whole genomes or at selected sequences of interest.	The search parameters and the identified potential H-DNA or Z-DNA forming sequences and their positions in the genes will be displayed on the search results page. In the result, each sequence will also get a score to indicate the possibility/stability of each non-B DNA conformation.	2013	[131]
Palindrome analyser	http://bioinformatics.ibp.cz(accessible)	A web-based server for predicting and evaluating inverted repeats.	It supports the circular genomes as input data. In this web-based server, it allows easy sorting according to the characteristics of IRs and compares all IRs through similarity analysis in the sequence. It supports different genomes. The distribution and localization of IRs can be visualized and the types and frequencies of different IRs from various species can be compared. The comparison of the probability of cruciform formation is supported.	2016	[132]
3D-NuS	http://iith.ac.in/3dnus/ (accessible)	A web server for automated modelling and visualization of nucleic acid structures. Overall, 80 types of triplexes including 6 homomers and 4 hetermers as well as 64 types of G4s are generated.	The constructed 3D structures included: RNA–DNA hybrid duplexes, intra/intermolecular DNA/RNA duplexes, triplexes, Z-DNAs and G-quadruplexes. 3D-NuS does not support mismatch base triplets. After obtaining the provided sequence information, 3D-NuS can display whether there is a mismatch and the location of the mismatch.	2017	[133]
Software					
Emboss	http://www.sanger.ac.uk/Software/EMBOSS/ (noupdate/notaccessible)	Cruciform detection.	Because there is no any visualization function in Emboss, it requires a higher level of computing skills.	2000	[134]
MFOLD	http://www.unafold.org/ (accessible)	Cruciform detection.	MFOLD was developed to detect secondary structures from RNA or ssDNA. The MFOLD server is limited to predict a secondary structure under specific conditions, which can be useful for cruciform identification within the input sequence of up to 9000 bases, but it could be used together with other tools or integrated in pipelines.	2003	[135]
IRF (Inverted Repeats Finder)	http://tandem.bu.edu (accessible)	Inverted repeats prediction.	A command line algorithm used to predict inverted repeat structure. It needs a higher level of computer skills to achieve applications.	2004	[136]
BioPHP - Find Palindromic sequences	http://www.biophp.org/minitools/find_palindromes/ (accessible)	Find Palindromic sequences.	It is a webpage that could search the sequence to find palindromic subsequences. It allows selection of minimum and maximum size of palindromic subsequences.	2011	/
SHAPE-seq	not provided	Nucleotide-resolution RNA structure prediction.	The output of this pipeline can be immediately used in RNA folding tools to predict the structure of each RNA molecule.	2011/2012/2014	[115–117]

(continue)

Table 3. Continued.

Method	URL	Description	Advantages and disadvantages/greater detail explanation of databases	Year	Reference
findIR	http://bioinfolab.miamioh.edu/bioinfolab/palindrome.php (accessible)	A MATLAB-based program for perfect inverted repeats prediction.	It allows genome-scale input data while only supports perfect inverted repeat identification. In comparison with existing IR detection tools including EMBOSS [126] and BioPHP (http://www.biophp.org/minitools/find_palindromes/), findIR demonstrates a high accuracy in detecting nested and overlapping IRs. It is available for download but it used the commercial software package, MATLAB.	2014	[137]
detectIR	https://sourceforge.net/projects/detectir (accessible)	A MATLAB-based program for the perfect and imperfect inverted repeat prediction.	It allows genome-scale input data and supports the prediction of both perfect and imperfect IR. detectIR has been proven to have higher accuracy and efficiency. It is available for download but it used the commercial software package, MATLAB.	2014	[138]
NeSSie	https://github.com/B3rse/nessie (accessible)	An algorithm for imperfection-tolerant search of DNA palindromes, mirrors, potential triplex forming patterns and symmetrical DNA sequence patterns.	It does not support the prediction of many non-B DNA types in databases and web servers. The results of NeSSie need to be transformed using Python tools to make it more readable. It needs higher level of computational skills.	2018	[141]
DeepZ	https://github.com/Nazar1997/DeepZ (accessible)	Z-DNA prediction.	DeepZ is not only used to verify the Z-DNA in the experiment, but also used to annotate the entire genome and found some new Z-DNA that has not been found in the experiment.	2020	[143]

which demonstrated that G4-miner is widely applicable. Because single-nucleotide variations (SNVs) affects the formation of G4 structures and has individual differences, this method can be used to identify and characterize genome-wide G4s of specific individuals. By comparing the detected G4 sequences of optimized G4-seq [101] and G4-miner, they are reliable to detect canonical G4, but have a preference when detecting noncanonical G4s. G4-seq tends to detect bulges, while G4-miner tends to detect two quartets.

R-loop/DNA:RNA hybrid

Unlike computational algorithms for identifying G4, to date, only a few bioinformatics strategies to identify R-loop structures in the genome have been reported. In 2012, Kuznetsov *et al.* constructed a database of R-loopDB [30]. Using 66 803 USCS reference genes in FASTA format from the UCSC Genome Browser [122], they defined the R-loop forming sequence (RLFS) as the configuration of three segments: the R-loop initiation zone (RIZ), linker and R-loop elongation zone (REZ). The R-loopDB database contains 245 181 RLFSs from 39 720 known UCSC gene sequences. The RLFS for a specific gene can be visualised for the database interface. As the first edition of the R-loop-related database, R-loopDB provides researchers with the first comprehensive RLFS catalogue and inspires to predict the selective R-loop structure using the RLFS model. However, it only integrates the gene sequence from UCSC, so the quantity and quality of R-loop structures on different genes in the genome are limited.

In 2015, based on the R-loopDB of 2012, the Quantitative Model of RLFS finder (QmRLFS-finder) for detecting and analysing the structure and sequence coordinates of RLFS was reported [32]. The RIZ-Linker-REZ model is also implemented in the QmRLFS-finder, which is based on regular expression matching. Moreover, the RIZ in the QmRLFS-finder supports the emergence of two linked G-clusters and increases the number of adjacent G-clusters to four guanines, based on an empirical R-loop sequence model [58]. Using the same input DNA/RNA sequence as R-loopDB of 2012, all RLFS tables, FASTA, BED and CUSTOM TRACK can be obtained as outputs. To evaluate the performance of the QmRLFS-finder, the authors compared the experimentally verified R-loop and DRIP-seq data [123] with RLFS detected by the QmRLFS-finder. The 3311 RLFS regions overlapped from 4181 DRIP-seq defined regions and QmRLFS-finder predicted regions, with a sensitivity of 79.2%. The QmRLFS-finder introduces the first R-loop prediction tool open to the public. Compared with R-loopDB 2012, QmRLFS-finder has been confirmed to possess a higher confidence [124].

With the development of DNA:RNA immunoprecipitation coupled with high-throughput sequencing (DRIP-seq) [105, 123], Kuznetsov *et al.* integrated the experimental data [51, 123, 125, 126] into the updated R-loopDB in 2017 [31]. In addition to humans, the database supports genome-scale detection of RLFS in seven additional organisms, including mouse, rat, chimpanzee, chicken, frog, fruit fly and yeast using QmRLFS-finder. Chromosome coordinates, sequences and genomic data of 1 565 795 RLFSs across 121 056 genes from eight species were collected in R-loopDB 2017. The renewed R-loopDB provides the scientific community with a tool that integrates RLFS query and prediction analysis on one platform with strong interactivity.

To deepen the understanding R-loop genome localisation and its regulatory network, the database of R-loopBase integrating R-loop distribution and regulation was developed in 2021 [127]. Through the integration of genomics and literature data, the following were included in the database: 107 high-quality genome-wide R-loop mapping datasets generated based on 11 different technologies. To date, the most comprehensive R-loop regulatory proteins and their targeted R-loops in multiple species, billion functional genome annotations and interactive interfaces have been developed to search, visualise, download and analyse R-loops and R-loop regulators in the context of well-annotated genomes. R-loopBase was used to identify R-loop-forming regions with high confidence by unifying the commonness of different detection technologies. In addition, the database integrates and visualises information on the systematic combination and annotation of R-loop regulatory proteins. This provides convenient data resources for scholars and researchers in the field.

It is very important to consider the dynamic changes in the cell environment in R-loop formation when developing the algorithm because the formation of the R-loop is spatiotemporally specific. Meanwhile, the published R-loop map data have obvious technical preferences due to the significant differences in the principles and experimental processes of different R-loop identification technologies. The effective integration of published experimental data of the R-loop into the software or database for RLFS detection is an urgent problem to be solved. Furthermore, the innovation of the RLFS identification algorithm also provides the possibility for a more accurate prediction of the R-loop structure.

Other non-canonical nucleic acid structures (except for G4/rG4 and R-loop)

Databases and web servers for identification of other non-canonical nucleic acid structures

The databases and web servers of non-canonical nucleic acid structure-forming sequences curated most categories of non-canonical structures. For example, non-B DB is one of the most comprehensive databases for non-B DNA prediction in mammalian genomes, which was reported in 2011 by Stephens *et al.* [128]. It integrates a relatively complete type of non-B DNA motif, including Z-DNA motifs, G4 motifs, DR, IR, MR and phased repeats, which are prone to form their corresponding non-canonical nucleic acid structures, including Z-DNA, G4, sticky DNA and slipped strand DNA, hairpin and cruciform, triplex DNA and H-DNA and static bending, respectively. Meanwhile, the non-B DB also collects motifs that tend to form STR, which are related to disease. For each type of motif prone to form non-B DNA, the Non-B DB has established search criteria according to their sequence characteristics. Furthermore, five organisms, humans, chimps, dogs, macaques and mice, were included to support non-B DNA motif identification. The database also lists the number of non-B DNA-forming motifs across five species. For humans, each type of motif is between 0.1 and 1 million, even more than one million for G4 and MR. As a curation database, Non-B DB provides data on known non-B DNA motifs across multiple species but lacks tools for the identification of novel non-B DNA. In 2013, Non-B DB was updated into Non-B DB v2.0, which deepened the non-B DNA forming motif coverage, added visualisation tools and increased seven organisms, including orangutans, rats, cows, pigs, horses, platypus and *A. thaliana* [129]. By using search criteria for perfect repeats, imperfect repeats have been ignored, but these should be considered in the future.

Similar studies were conducted by the team of *Robert M. Stephens*. In 2012, nBMST (non-B DNA motif search tool), a web server used for searching non-B DNA motifs, was published [130]. Similar to Non-B DB, the program of nBMST can recognize several types of non-B DNA motifs. However, unlike the Non-B DB, which only supports queries, it allows customising analysis and detecting unknown non-B DNA motifs using user-submitted nucleotide sequences.

In 2013, a web-based engine titled ‘DNA structure search’ was reported to predict H-DNA formation and Z-DNA forming sequences in whole genomes or at selected sequences of interest. [131]. Using nucleotide sequence or gene name information as input files, they developed algorithms to identify H-DNA and Z-DNA motifs based on their sequence features. This search engine supports users in customising parameters to specify the length range of mirrored arms and spacers and allows a mismatch in the number of arms. At the same time, each sequence is scored according to the possibility and stability of non-B DNA conformation. Compared to nBMST [130], this search engine contains fewer types of non-B DNA motifs for prediction.

In 2016, the Palindrome analyser, a web-based server for detecting and evaluating IR only and especially for longer nucleotide sequences, was published [132]. It supports the input data of genome sequences and oligonucleotides, and the results of the Palindrome analyser provide information on the length, sequence, coordinates and energy required for cruciform formation. Moreover, the Palindrome analyser can display inverted repeat sequences both interactively and graphically and supports a variety of filtering options. In 2017, 3D-NuS, a web server for automated modelling and visualisation of non-canonical 3-Dimensional Nucleic acid Structures, was reported. This can support the simultaneous prediction of triplexes, G4s, Z-DNA/RNA and DNA–RNA hybrid double strands [133].

Software for identification of other non-canonical nucleic acid structures

In addition to databases and web servers, several standalone software packages have been developed for methods that can be used to detect other types of non-B DNA. In 2000, Emboss software used for cruciform detection was reported [134]. Because there is no any visualization function in Emboss, it requires a higher level of computing skills. After that, MFOLD was developed to detect secondary structures (such as stem-loop structures) from RNA or ssDNA in 2003 [135]. The MFOLD server is limited to predict a secondary structure under specific conditions, which can be useful for cruciform identification within the input sequence of up to 9000 bases, but it could be used together with other tools or integrated in pipelines. Furthermore, IRF (Inverted Repeats Finder), a command line algorithm used to predict inverted repeat structure was developed in 2004 [136]. It also needs a higher level of computer skills to achieve applications. BioPHP - Find Palindromic sequences is a webpage that can search the sequence to find palindromic subsequences. It allows selection of minimum and maximum size of palindromic subsequences (http://www.biophp.org/minitools/find_palindromes/). Sequentially, based on published high-throughput sequencing method of SHAPE-seq, a pipeline for analysing its data was developed [115–117]. The output of this pipeline can be immediately used in RNA folding tools to predict the structure of each RNA molecule.

For the prediction of IR that are prone to form hairpins and cruciform repeats, two similar MATLAB-based programs, findIR and detectIR, which allow genome-scale input data, were reported in 2014 [137, 138]. In terms of the algorithm, both transform the sequence search into a numerical calculation and operate with complex numbers. The two software have a common inconvenience, that is, although they are open source and downloadable, they both used the commercial software package, MATLAB. As previous studies have suggested, IR can be classified into perfect and imperfect based on whether the two halves are perfectly complementary [72, 139, 140]. The difference between these two tools is that the former only supports perfect inverted repeat identification, whereas the latter supports the prediction of both perfect and imperfect IR. In addition, in comparison with existing IR detection tools including EMBOSS [134] and BioPHP (http://www.

biophp.org/minitools/find_palindromes/), findIR demonstrates a high accuracy in detecting nested and overlapping IRs. Meanwhile, detectIR has been proven to have higher accuracy and efficiency.

Several softwares are available that can detect more than one specific non-canonical structure. Nucleic acid elements of Sequence Symmetry identification (NeSSie), a dynamic programming C/C++ 64-bit library for imperfection-tolerant searches of DNA palindromes, mirrors and symmetrical DNA sequence patterns, was published in 2018 [141]. NeSSie was inspired by the phenomenon that although there are diverse basic characteristics in the primary sequences that form different types of non-B DNA, experimental evidence suggests that the features of non-B DNA-forming sequences may be provided with high polymorphism and instability [142]. For example, the genome of *Mycobacterium bovis* was analysed by NeSSie as a case study. However, it does not support the prediction of many non-B DNA types in databases and web servers. Moreover, the results of NeSSie need to be transformed using Python tools to make it more readable.

Given that the deep learning method could analyse and extract information from a large number of molecular biology data, a machine learning strategy of DeepZ was designed for Z-DNA detection [143]. This approach aggregated information from epigenetic markers, transcription factor and RNA polymerase binding sites and chromosome accessibility maps. The authors of DeepZ not only used the model to verify the Z-DNA in the experiment, but also annotated the whole genome and found some new Z-DNA regions that have not yet been found in the experiments. These regions may arouse the interest of researchers in various fields.

In general, for software to detect non-canonical DNA/RNA structures, memory consumption is a key factor that needs to be considered for computation and time saving. Moreover, as data from high-throughput sequencing can be used for the development of machine learning-based algorithms, one of the trends in this field is the prediction of other potential non-canonical DNA/RNA structures, except for G4/rG4 and R-loop, with machine learning algorithms on the premise that there are relevant high-throughput sequencing methods to map them.

Summary and outlook

Nucleic acid structure is an essential element in determining the function of nucleic acids. Because non-canonical DNA and RNA structures are dynamic in the human genome, their formation is influenced by several molecular players. Biological reactions, such as replication, transcription and reverse transcription controlled by non-canonical DNA/RNA structures, provide therapeutic opportunities for targeting non-canonical nucleic acid structures [8, 10, 44, 144–146]. Although there have been many studies on the structural characteristics and biological functions of non-B DNA and RNA structures, our understanding is still in its infancy and further research is warranted.

Dozens of experimental approaches have been designed to predict these structures, including nuclease cleavage, polyacrylamide gel electrophoresis, NMR, chemical probing, electron microscopy, circular dichroism, atomic force microscopy, ultraviolet absorption and crystallography (reviewed in [46]). For sequencing strategies, to the best of our knowledge, the majority of approaches are for G4/rG4 and R-loop mapping (Table 1) (reviewed in [8, 14–19], with several updated methods reviewed in this overview). To date, there is only a few sequencing technology for other non-canonical nucleic acid structures (other than G4/rG4 and R-loop). This may be attributed to the fact that other non-canonical nucleic acid structures account for a relatively small proportion of the genome [140, 147, 148] and they are often dynamic in biological processes. Therefore, the development of sequencing technologies to identify other non-canonical nucleic acid structures, especially non-canonical RNA structures, is a potential research hotspot in the future. In addition, studies have found that RNA binding proteins (RBPs) are key factors in regulating gene expression. In order to study the mechanism of RBPs regulating RNA, a variety of research technologies have been developed, such as RIP-seq, MeRIP-seq, iCLIP-seq and eCLIP-seq [149–152]. However, the approach to detect the relationship of non-canonical RNAs and proteins has not been designed. This may be a potential interest in this area.

Although an increasing number of computational prediction approaches for non-B DNA/RNA structures have been developed, accurate detection of the formation, presence, genomic locations and target of non-canonical nucleic acid structures is still one of the obstacles that need to be overcome in the field of non-canonical DNA/RNA structures. Because most algorithms were designed based on known knowledge of the formation of DNA/RNA secondary structures, the dynamic conditions that affect the formation of non-B DNA/RNA structures (such as negative superhelices, binding proteins, chromatin structures, epigenetic modifications and DNA transactions) are usually not included in the search criteria. Therefore, some software packages may yield inaccurate results [153, 154]. Thus, only a known structure can provide preliminary prediction guidance based on non-canonical structure identification approaches. Furthermore, the biological effects of complex and dynamic regions in the genome must be carefully considered before developing a direct and conclusive method to detect specific DNA/RNA structures within multiple regions of non-B DNA/RNA formation sequences. Moreover, no computational method that can simultaneously identify multiple types of non-canonical nucleic acid structures has been developed. It is imperative to integrate the software for identifying various non-canonical nucleic acid structures into a comprehensive tool box for analysing these structures simultaneously. Most notably, although we cover the updated computational methods of non-canonical nucleic acid structures, our review lacks an evaluation of their performance. On one hand, the inconsistency between different sequencing technologies leads to the missing of a ‘gold standard’ for non-B DNAs, on the other, the unavailability of consistent input DNA data increases the difficulty of evaluating these software, which makes it challenging to obtain the test data set for evaluation.

In general, more G4s/rG4s were screened out by the computational approaches compared with those by experimental methods. Thus, the experimental prediction was considered more reliable. This is mainly because the screening criteria of the computational methods are not strict enough. The computational methods are difficult to evaluate, which makes these computational approaches more complicated and difficult to choose. Using the experimental data set as the benchmark set to evaluate similar algorithms can improve the accuracy of the computational methods. To a certain extent, this can improve the problem of large divergence between computational prediction and experimental prediction of G4/rG4. In short, effective filtering of the noise in the experimental process can obtain high confidence experimental verification results, and then matching the corresponding computational methods and doing experimental verification can better solve the problem of huge divergence between experimental techniques and computational methods.

Given that non-canonical DNA/RNA structures play an important role in genomic instability, analysing what genes/pathways/proteins participate in the non-canonical structures related to genome instability or repair is important for understanding the life activities mediated by the central dogma, studying the mechanisms involved in genome instability, and developing potential methods to reduce or stimulate non-B DNA/RNA-induced replication-transcription conflicts in healthy cells or diseased cells. Moreover, the contributors of non-B DNA/RNA-induced mutations, epigenetic modifications and gene fusion to diseases warrant further study. Furthermore, an important direction of future efforts for non-canonical nucleic acid structures includes identifying the proteins binding to these non-canonical structures, stabilising or unwinding structures and cleaving or repairing structures. Environmental conditions that can affect the formation of these structures are also essential.

Key Points

- We firstly comprehensively review all updated experimental and computational methods related to non-canonical nucleic acid structures.
- The updated experimental methods related to non-canonical nucleic acid structures including L1H1-7OTD WGA sequencing and G4P-ChIP for G4s identification, G4RP-seq and SHALiPE-seq for identifying rG4s, as well as an antibody-based method of qDRIP-Seq and the most recent approach of R-loop CUT&Tag for R-loops published in 2021.
- The updated computational methods of non-canonical nucleic acid structures, including databases and web servers to integrate non-canonical DNA/RNA-forming sequences as well as the independent software for detecting non-canonical structures.

Data availability

All data relevant to the study are included in the article.

Funding

Key Laboratory of Clinical Laboratory Diagnosis and Translational Research of Zhejiang Province (Grant No: 2022E10022); National Natural Science Foundation of China [32170650]; Guangzhou and Guangdong Key Project [202007030002 and 2018B030335001].

Author contributions

X.S.: conceptualization, writing, review and editing; H.T.: review and editing; Z.S.: conceptualization, supervision, review and editing.

References

1. JD WATSON, FHC CRICK. Molecular structure of nucleic acids: a structure for deoxyribose nucleic acid. *Nature* 1953;**171**:737–8.
2. Georgakopoulos-Soares I, Morganella S, Jain N, *et al.* Noncanonical secondary structures arising from non-B DNA motifs are determinants of mutagenesis. *Genome Res* 2018;**28**:1264–71.
3. Wells RD. Advances in mechanisms of genetic instability related to hereditary neurological diseases. *Nucleic Acids Res* 2005;**33**:3785–98.
4. Ghosh A, Bansal M. A glossary of DNA structures from A to Z. *Acta Crystallogr Sect D Biol Crystallogr* 2003;**59**:620–6.
5. Mirkin SM. Discovery of alternative DNA structures: a heroic decade (1979–1989). *Front Biosci* 2008;**13**:1064.
6. Cox R, Mirkin SM. Characteristic enrichment of DNA repeats in different genomes. *Proc Natl Acad Sci* 1997;**94**:5237–42.
7. Wang G, Vasquez KM. Impact of alternative DNA structures on DNA damage, DNA repair, and genetic instability. *DNA Repair (Amst)* 2014;**19**:143–51.
8. Kharel P, Balaratnam S, Beals N, *et al.* The role of RNA G-quadruplexes in human diseases and therapeutic strategies. *Wiley Interdiscip Rev RNA* 2020;**11**:e1568.
9. Rhodes D, Lipps HJ. G-quadruplexes and their regulatory roles in biology. *Nucleic Acids Res* 2015;**43**:8627–37.
10. Jain A, Wang G, Vasquez KM. DNA triple helices: biological consequences and therapeutic potential. *Biochimie* 2008;**90**: 1117–30.
11. Santos-Pereira JM, Aguilera A. R loops: new modulators of genome dynamics and function. *Nat Rev Genet* 2015;**16**: 583–97.
12. Ravichandran S, Subramani VK, Kim KK. Z-DNA in the genome: from structure to disease. *Biophys Rev* 2019;**11**:383–7.
13. Niehrs C, Luke B. Regulatory R-loops as facilitators of gene expression and genome stability. *Nat Rev Mol Cell Biol* 2020;**21**: 167–78.
14. Hänsel-Hertsch R, Di Antonio M, Balasubramanian S. DNA G-quadruplexes in the human genome: detection, functions and therapeutic potential. *Nat Rev Mol Cell Biol* 2017;**18**: 279–84.
15. Kwok CK, Merrick CJ. G-quadruplexes: prediction, characterization, and biological application. *Trends Biotechnol* 2017;**35**: 997–1013.
16. Varshney D, Spiegel J, Zyner K, *et al.* The regulation and functions of DNA and RNA G-quadruplexes. *Nat Rev Mol Cell Biol* 2020;**21**:459–74.
17. Vanoosthuyse V. Strengths and weaknesses of the current strategies to map and characterize R-loops. *Non-Coding RNA* 2018;**4**:9.
18. Hegazy YA, Fernando CM, Tran EJ. The balancing act of R-loop biology: the good, the bad, and the ugly. *J Biol Chem* 2020;**295**: 905–13.
19. Chédin F, Hartono SR, Sanz LA, *et al.* Best practices for the visualization, mapping, and manipulation of R-loops. *EMBO J* 2021;**40**:e106394.
20. Zheng K, Zhang J, He Y, *et al.* Detection of genomic G-quadruplexes in living cells using a small artificial protein. *Nucleic Acids Res* 2020;**48**:11706–20.
21. Yoshida W, Saikyo H, Nakabayashi K, *et al.* Identification of G-quadruplex clusters by high-throughput sequencing of whole-genome amplified products with a G-quadruplex ligand. *Sci Rep* 2018;**8**:3116.
22. Yang SY, Lejault P, Chevrier S, *et al.* Transcriptome-wide identification of transient RNA G-quadruplexes in human cells. *Nat Commun* 2018;**9**:4730.
23. Yang X, Cheema J, Zhang Y, *et al.* RNA G-quadruplex structures exist and function in vivo in plants. *Genome Biol* 2020;**21**:226.
24. Crossley MP, Bocek MJ, Hamperl S, *et al.* qDRIP: a method to quantitatively assess RNA–DNA hybrid formation genome-wide. *Nucleic Acids Res* 2020;**48**:e84–4.

25. Wang K, Wang H, Li C, *et al.* Genomic profiling of native R loops with a DNA-RNA hybrid recognition sensor. *Sci Adv* 2021;**7**:eabe3516.
26. Parveen N, Shamim A, Cho S, *et al.* Computational approaches to predict the non-canonical DNAs. *Curr Bioinform* 2019;**14**: 470–9.
27. Puig Lombardi E, Londoño-Vallejo A. A guide to computational methods for G-quadruplex prediction. *Nucleic Acids Res* 2020;**48**: 1–15.
28. Klimentova E, Polacek J, Simecek P, *et al.* PENGUINN: Precise exploration of nuclear G-quadruplexes using interpretable neural networks. *Front Genet* 2020;**11**:568546.
29. Chow EY-C, Lyu K, Kwok CK, *et al.* rG4-seeker enables high-confidence identification of novel and non-canonical rG4 motifs from rG4-seq experiments. *RNA Biol* 2020;**17**:903–17.
30. Wongsurawat T, Jenjaroenpun P, Kwoh CK, *et al.* Quantitative model of R-loop forming structures reveals a novel level of RNA-DNA interactome complexity. *Nucleic Acids Res* 2012;**40**:e16–6.
31. Jenjaroenpun P, Wongsurawat T, Sutheeworapong S, *et al.* R-loopDB: a database for R-loop forming sequences (RLFS) and R-loops. *Nucleic Acids Res* 2017;**45**:D119–27.
32. Jenjaroenpun P, Wongsurawat T, Yenamandra SP, *et al.* QmRLFS-finder: a model, web server and stand-alone tool for prediction and analysis of R-loop forming sequences. *Nucleic Acids Res* 2015;**43**:W527–34.
33. Tu J, Duan M, Liu W, *et al.* Direct genome-wide identification of G-quadruplex structures by whole-genome resequencing. *Nat Commun* 2021;**12**:6014.
34. Barshai M, Aubert A, Orenstein Y. G4detector: convolutional neural network to predict DNA G-quadruplexes. *IEEE/ACM Trans Comput Biol Bioinforma* 2022;**19**:1946–55.
35. Rocher V, Genais M, Nassereddine E, *et al.* DeepG4: a deep learning approach to predict cell-type specific active G-quadruplex regions. *PLoS Comput Biol* 2021;**17**:e1009308.
36. Bang I. Untersuchungen über die Guanylsäure. *Biochem Z* 1910;**26**:293–311.
37. Gellert M, Lipsett MN, Davies DR. Helix formation by guanylic acid. *Proc Natl Acad Sci* 1962;**48**:2013–8.
38. Sen D, Gilbert W. A sodium-potassium switch in the formation of four-stranded G4-DNA. *Nature* 1990;**344**:410–4.
39. Sen D, Gilbert W. Formation of parallel four-stranded complexes by guanine-rich motifs in DNA and its implications for meiosis. *Nature* 1988;**334**:364–6.
40. Kim J, Cheong C, Moore PB. Tetramerization of an RNA oligonucleotide containing a GGGG sequence. *Nature* 1991;**351**:331–2.
41. Schaffitzel C, Berger I, Postberg J, *et al.* In vitro generated antibodies specific for telomeric guanine-quadruplex DNA react with Stylonychia lemnae macronuclei. *Proc Natl Acad Sci* 2001;**98**:8572–7.
42. Schiavone D, Guilbaud G, Murat P, *et al.* Determinants of G quadruplex-induced epigenetic instability in REV$_1$-deficient cells. *EMBO J* 2014;**33**:2507–20.
43. Lipps HJ, Gruissem W, Prescott DM. Higher order DNA structure in macronuclear chromatin of the hypotrichous ciliate Oxytricha nova. *Proc Natl Acad Sci* 1982;**79**:2495–9.
44. Patel DJ, Phan AT, Kuryavyi V. Human telomere, oncogenic promoter and 5′-UTR G-quadruplexes: diverse higher order DNA and RNA targets for cancer therapeutics. *Nucleic Acids Res* 2007;**35**:7429–55.
45. Mendoza O, Bourdoncle A, Boulé J-B, *et al.* G-quadruplexes and helicases. *Nucleic Acids Res* 2016;**44**:1989–2006.
46. Thomas M, White RL, Davis RW. Hybridization of RNA to double-stranded DNA: formation of R-loops. *Proc Natl Acad Sci* 1976;**73**:2294–8.
47. Drolet M, Phoenix P, Menzel R, *et al.* Overexpression of RNase H partially complements the growth defect of an Escherichia coli delta topA mutant: R-loop formation is a major problem in the absence of DNA topoisomerase I. *Proc Natl Acad Sci* 1995;**92**: 3526–30.
48. García-Muse T, Aguilera A. R loops: from physiological to pathological roles. *Cell* 2019;**179**:604–18.
49. Wahba L, Costantino L, Tan FJ, *et al.* S1-DRIP-seq identifies high expression and polyA tracts as major contributors to R-loop formation. *Genes Dev* 2016;**30**:1327–38.
50. Xu W, Xu H, Li K, *et al.* The R-loop is a common chromatin feature of the Arabidopsis genome. *Nat Plants* 2017;**3**:704–14.
51. Sanz LA, Hartono SR, Lim YW, *et al.* Prevalent, dynamic, and conserved R-loop structures associate with specific epigenomic signatures in mammals. *Mol Cell* 2016;**63**:167–78.
52. García-Pichardo D, Cañas JC, García-Rubio ML, *et al.* Histone mutants separate R loop formation from genome instability induction. *Mol Cell* 2017;**66**:597–609.e5.
53. Li X, Manley JL. Inactivation of the SR protein splicing factor ASF/SF2 results in genomic instability. *Cell* 2005;**122**:365–78.
54. Yu K, Chedin F, Hsieh C-L, *et al.* R-loops at immunoglobulin class switch regions in the chromosomes of stimulated B cells. *Nat Immunol* 2003;**4**:442–51.
55. Chen L, Chen J-Y, Zhang X, *et al.* R-ChIP using inactive RNase H reveals dynamic coupling of R-loops with transcriptional pausing at gene promoters. *Mol Cell* 2017;**68**:745–757.e5.
56. De Magis A, Manzo SG, Russo M, *et al.* DNA damage and genome instability by G-quadruplex ligands are mediated by R loops in human cancer cells. *Proc Natl Acad Sci* 2019;**116**:816–25.
57. Zhao Y, Zhang J, Zhang Z, *et al.* Real-time detection reveals responsive cotranscriptional formation of persistent intramolecular DNA and intermolecular DNA:RNA hybrid G-quadruplexes stabilized by R-loop. *Anal Chem* 2017;**89**: 6036–42.
58. Roy D, Lieber MR. G clustering is important for the initiation of transcription-induced R-loops in vitro, whereas high G density without clustering is sufficient thereafter. *Mol Cell Biol* 2009;**29**: 3124–33.
59. Duquette ML, Handa P, Vincent JA, *et al.* Intracellular transcription of G-rich DNAs induces formation of G-loops, novel structures containing G4 DNA. *Genes Dev* 2004;**18**: 1618–29.
60. Wang AHJ, Quigley GJ, Kolpak FJ, *et al.* Molecular structure of a left-handed double helical DNA fragment at atomic resolution. *Nature* 1979;**282**:680–6.
61. Ha SC, Lowenhaupt K, Rich A, *et al.* Crystal structure of a junction between B-DNA and Z-DNA reveals two extruded bases. *Nature* 2005;**437**:1183–6.
62. Wang G. Z-DNA, an active element in the genome. *Front Biosci* 2007;**12**:4424.
63. Peck LJ, Wang JC. Transcriptional block caused by a negative supercoiling induced structural change in an alternating CG sequence. *Cell* 1985;**40**:129–37.
64. Shin S-I, Ham S, Park J, *et al.* Z-DNA-forming sites identified by ChIP-Seq are associated with actively transcribed regions in the human genome. *DNA Res* 2016;**23**:477–86.
65. Wang G, Christensen LA, Vasquez KM. Z-DNA-forming sequences generate large-scale deletions in mammalian cells. *Proc Natl Acad Sci* 2006;**103**:2677–82.

66. Platt JR. POSSIBLE SEPARATION OF INTERTWINED NUCLEIC ACID CHAINS BY TRANSFER-TWIST. *Proc Natl Acad Sci* 1955;**41**: 181–3.
67. Lyamichev VI, Panyutin IG, Frank-Kamenetskii MD. Evidence of cruciform structures in superhelical DNA provided by two-dimensional gel electrophoresis. *FEBS Lett* 1983;**153**:298–302.
68. Panayotatos N, Wells RD. Cruciform structures in supercoiled DNA. *Nature* 1981;**289**:466–70.
69. Masai H, Arai K. Frpo: a novel single-stranded DNA promoter for transcription and for primer RNA synthesis of DNA replication. *Cell* 1997;**89**:897–907.
70. Horwitz MSZ, Loeb LA. An *E. coli* promoter that regulates transcription by DNA superhelix-induced cruciform extrusion. *Science (80-)* 1988;**241**:703–5.
71. White JH, Bauer WR. Superhelical DNA with local substructures. *J Mol Biol* 1987;**195**:205–13.
72. Smith GR. Meeting DNA palindromes head-to-head. *Genes Dev* 2008;**22**:2612–20.
73. Repping S, Skaletsky H, Lange J, *et al.* Recombination between palindromes P5 and P1 on the human Y chromosome causes massive deletions and spermatogenic failure. *Am J Hum Genet* 2002;**71**:906–22.
74. Bikard D, Loot C, Baharoglu Z, *et al.* Folded DNA in action: Hairpin formation and biological functions in prokaryotes. *Microbiol Mol Biol Rev* 2010;**74**:570–88.
75. YuL L, Shlyakhtenko LS. Early melting of supercoiled DNA. *Nucleic Acids Res* 1988;**16**:3269–81.
76. Kowalski D, Natale DA, Eddy MJ. Stable DNA unwinding, not ‘breathing,’ accounts for single-strand-specific nuclease hypersensitivity of specific A+T-rich sequences. *Proc Natl Acad Sci* 1988;**85**:9464–8.
77. Kohwi-Shigematsu T, Kohwi Y. Torsional stress stabilizes extended base unpairing in suppressor sites flanking immunoglobulin heavy chain enhancer. *Biochemistry* 1990;**29**: 9551–60.
78. Drakesmith H, Townsend A. *DNA Structure and Function*, 1994.
79. Potaman VN, Bissler JJ, Hashem VI, *et al.* Unpaired structures in SCA10 (ATTCT)*n*·(AGAAT)*n* repeats. *J Mol Biol* 2003;**326**: 1095–111.
80. Felsenfeld G, Rich A. Studies on the formation of two- and three-stranded polyribonucleotides. *Biochim Biophys Acta* 1957;**26**:457–68.
81. Dervan PB. Design of sequence-specific DNA-binding molecules. *Science (80-)* 1986;**232**:464–71.
82. Htun H, Dahlberg JE. Single strands, triple strands, and kinks in H-DNA. *Science (80-)* 1988;**241**:1791–6.
83. Mirkin SM, Frank-Kamenetskii MD. H-DNA and related structures. *Annu Rev Biophys Biomol Struct* 1994;**23**:541–76.
84. Praseuth D, Guieysse AL, Hélène C. Triple helix formation and the antigene strategy for sequence-specific control of gene expression. *Biochim Biophys Acta Gene Struct Expr* 1999;**1489**: 181–206.
85. Pearson CE, Sinden RR. Alternative structures in duplex DNA formed within the trinucleotide repeats of the myotonic dystrophy and fragile X loci. *Biochemistry* 1996;**35**:5041–53.
86. Sinden RR. Slipped strand DNA structures. *Front Biosci* 2007;**12**:4788.
87. Sakamoto N, Chastain PD, Parniewski P, *et al.* Sticky DNA. *Mol Cell* 1999;**3**:465–75.
88. Vetcher AA, Napierala M, Iyer RR, *et al.* Sticky DNA, a long GAA·GAA·TTC triplex that is formed intramolecularly, in the sequence of Intron 1 of the frataxin gene. *J Biol Chem* 2002;**277**: 39217–27.
89. Krasilnikova MM, Mirkin SM. Replication stalling at Friedreich’s ataxia (GAA) *n* repeats in vivo. *Mol Cell Biol* 2004;**24**:2286–95.
90. Ohshima K, Montermini L, Wells RD, *et al.* Inhibitory effects of expanded GAA·TTC triplet repeats from Intron I of the Friedreich ataxia gene on transcription and replicationin vivo. *J Biol Chem* 1998;**273**:14588–95.
91. Napierala M, Dere R, Vetcher A, *et al.* Structure-dependent recombination hot spot activity of GAA·TTC sequences from Intron 1 of the Friedreich’s ataxia gene. *J Biol Chem* 2004;**279**: 6444–54.
92. Rodriguez R, Miller KM, Forment JV, *et al.* Small-molecule–induced DNA damage identifies alternative DNA structures in human genes. *Nat Chem Biol* 2012;**8**:301–10.
93. Hänsel-Hertsch R, Beraldi D, Lensing SV, *et al.* G-quadruplex structures mark human regulatory chromatin. *Nat Genet* 2016;**48**:1267–72.
94. Hänsel-Hertsch R, Spiegel J, Marsico G, *et al.* Genome-wide mapping of endogenous G-quadruplex DNA structures by chromatin immunoprecipitation and high-throughput sequencing. *Nat Protoc* 2018;**13**:551–64.
95. Liu H-Y, Zhao Q, Zhang T-P, *et al.* Conformation selective antibody enables genome profiling and leads to discovery of parallel G-quadruplex in human telomeres. *Cell Chem Biol* 2016;**23**: 1261–70.
96. Lam EYN, Beraldi D, Tannahill D, *et al.* G-quadruplex structures are stable and detectable in human genomic DNA. *Nat Commun* 2013;**4**:1796.
97. Gray LT, Vallur AC, Eddy J, *et al.* G quadruplexes are genomewide targets of transcriptional helicases XPB and XPD. *Nat Chem Biol* 2014;**10**:313–8.
98. Woodford KJ, Howell RM, Usdin K. A novel K(+)-dependent DNA synthesis arrest site in a commonly occurring sequence motif in eukaryotes. *J Biol Chem* 1994;**269**:27029.
99. Han H, Hurley LH, Salazar M. A DNA polymerase stop assay for G-quadruplex-interactive compounds. *Nucleic Acids Res* 1999;**27**:537–42.
100. Chambers VS, Marsico G, Boutell JM, *et al.* High-throughput sequencing of DNA G-quadruplex structures in the human genome. *Nat Biotechnol* 2015;**33**:877–81.
101. Marsico G, Chambers VS, Sahakyan AB, *et al.* Whole genome experimental maps of DNA G-quadruplexes in multiple species. *Nucleic Acids Res* 2019;**47**:3862–74.
102. Guo JU, Bartel DP. RNA G-quadruplexes are globally unfolded in eukaryotic cells and depleted in bacteria. *Science (80-)* 2016;**353**:aaf5371–1.
103. Kwok CK, Marsico G, Sahakyan AB, *et al.* rG4-seq reveals widespread formation of G-quadruplex structures in the human transcriptome. *Nat Methods* 2016;**13**:841–4.
104. Guo JU, Bartel DP. RNA G-quadruplexes are globally unfolded in eukaryotic cells and depleted in bacteria. *Science (80-)* 2016;**353**:aaf5371–aaf5371.
105. Ginno PA, Lott PL, Christensen HC, *et al.* R-loop formation is a distinctive characteristic of unmethylated human CpG island promoters. *Mol Cell* 2012;**45**:814–25.
106. Dumelie JG, Jaffrey SR. Defining the location of promoter-associated R-loops at near-nucleotide resolution using bisDRIP-seq. *Elife* 2017;**6**:e28306.
107. Wu H, Lima WF, Crooke ST. Investigating the structure of human RNase H1 by site-directed mutagenesis. *J Biol Chem* 2001;**276**:23547–53.
108. Yan Q, Shields EJ, Bonasio R, *et al.* Mapping native R-loops genome-wide using a targeted nuclease approach. *Cell Rep* 2019;**29**:1369–1380.e5.

109. Malig M, Hartono SR, Giafaglione JM, *et al.* Ultra-deep coverage single-molecule R-loop footprinting reveals principles of R-loop formation. *J Mol Biol* 2020;**432**:2271–88.
110. Halász L, Karányi Z, Boros-Oláh B, *et al.* RNA-DNA hybrid (R-loop) immunoprecipitation mapping: an analytical workflow to evaluate inherent biases. *Genome Res* 2017;**27**:1063–73.
111. Hartono SR, Malapert A, Legros P, *et al.* The affinity of the S9.6 antibody for double-stranded RNAs impacts the accurate mapping of R-loops in fission yeast. *J Mol Biol* 2018;**430**: 272–84.
112. König F, Schubert T, Längst G. The monoclonal S9.6 antibody exhibits highly variable binding affinities towards different R-loop sequences. *PLoS One* 2017;**12**:e0178875.
113. Kouzine F, Wojtowicz D, Baranello L, *et al.* Permanganate/S1 nuclease footprinting reveals non-B DNA structures with regulatory potential across a mammalian genome. *Cell Syst* 2017;**4**:344–356.e7.
114. Wu T, Lyu R, You Q, *et al.* Kethoxal-assisted single-stranded DNA sequencing captures global transcription dynamics and enhancer activity in situ. *Nat Methods* 2020;**17**:515–23.
115. Lucks JB, Mortimer SA, Trapnell C, *et al.* Multiplexed RNA structure characterization with selective 2′-hydroxyl acylation analyzed by primer extension sequencing (SHAPE-Seq). *Proc Natl Acad Sci* 2011;**108**:11063–8.
116. Mortimer SA, Trapnell C, Aviran S, *et al.* SHAPE–seq: high-throughput RNA structure analysis. *Curr Protoc Chem Biol* 2012;**4**:275–97.
117. Loughrey D, Watters KE, Settle AH, *et al.* SHAPE-Seq 2.0: systematic optimization and extension of high-throughput chemical probing of RNA secondary structure with next generation sequencing. *Nucleic Acids Res* 2014;**42**:e165–5.
118. Bedrat A, Lacroix L, Mergny J-L. Re-evaluation of G-quadruplex propensity with G4Hunter. *Nucleic Acids Res* 2016;**44**:1746–59.
119. Sahakyan AB, Chambers VS, Marsico G, *et al.* Machine learning model for sequence-driven DNA G-quadruplex formation. *Sci Rep* 2017;**7**:14535.
120. Hon J, Martínek T, Zendulka J, *et al.* Pqsfinder: an exhaustive and imperfection-tolerant search tool for potential quadruplex-forming sequences in R. *Bioinformatics* 2017;**33**: 3373–9.
121. Barshai M, Orenstein Y. Predicting G-quadruplexes from DNA sequences using multi-kernel convolutional neural networks. *Proc. 10th ACM Int. Conf. Bioinformatics, Comput. Biol. Heal. Informatics* 2019; 357–65.
122. Rhead B, Karolchik D, Kuhn RM, *et al.* The UCSC genome browser database: update 2010. *Nucleic Acids Res* 2010;**38**: D613–9.
123. Ginno PA, Lim YW, Lott PL, *et al.* GC skew at the 5′ and 3′ ends of human genes links R-loop formation to epigenetic regulation and transcription termination. *Genome Res* 2013;**23**: 1590–600.
124. Kuznetsov VA, Bondarenko V, Wongsurawat T, *et al.* Toward predictive R-loop computational biology: genome-scale prediction of R-loops reveals their association with complex promoter structures, G-quadruplexes and transcriptionally active enhancers. *Nucleic Acids Res* 2018;**46**:7566–85.
125. Lim YW, Sanz LA, Xu X, *et al.* Genome-wide DNA hypomethylation and RNA:DNA hybrid accumulation in Aicardi–Goutières syndrome. *Elife* 2015;**4**:e08007.
126. Nadel J, Athanasiadou R, Lemetre C, *et al.* RNA:DNA hybrids in the human genome have distinctive nucleotide characteristics, chromatin composition, and transcriptional relationships. *Epigenetics Chromatin* 2015;**8**:46.
127. Lin R, Zhong X, Zhou Y, *et al.* R-loopBase: a knowledgebase for genome-wide R-loop formation and regulation. *Nucleic Acids Res* 2022;**50**:D303–15.
128. Cer RZ, Bruce KH, Mudunuri US, *et al.* Non-B DB: a database of predicted non-B DNA-forming motifs in mammalian genomes. *Nucleic Acids Res* 2011;**39**:D383–91.
129. Cer RZ, Donohue DE, Mudunuri US, *et al.* Non-B DB v2.0: a database of predicted non-B DNA-forming motifs and its associated tools. *Nucleic Acids Res* 2012;**41**:D94–100.
130. Cer RZ, Bruce KH, Donohue DE, *et al.* Searching for non-B DNA-forming motifs using nBMST (non-B DNA motif search tool). *Curr. Protoc. Hum. Genet.* 2012;**Chapter 18**:Unit 18.7.1–22.
131. Wang G, Gaddis S, Vasquez KM. Methods to detect replication-dependent and replication-independent DNA structure-induced genetic instability. *Methods* 2013;**64**:67–72.
132. Brázda V, Kolomazník J, Lýsek J, *et al.* Palindrome analyser—a new web-based server for predicting and evaluating inverted repeats in nucleotide sequences. *Biochem Biophys Res Commun* 2016;**478**:1739–45.
133. Patro LPP, Kumar A, Kolimi N, *et al.* 3D-NuS: a web server for automated modeling and visualization of non-canonical 3-D imensional nucleic acid structures. *J Mol Biol* 2017;**429**: 2438–48.
134. Rice P, Longden I, Bleasby A. EMBOSS: the European molecular biology open software suite. *Trends Genet* 2000;**16**:276–7.
135. Zuker M. Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res* 2003;**31**:3406–15.
136. Warburton PE, Giordano J, Cheung F, *et al.* Inverted repeat structure of the human genome: the X-chromosome contains a preponderance of large, highly homologous inverted repeats that contain testes genes. *Genome Res* 2004;**14**:1861–9.
137. Sreeskandarajan S, Flowers MM, Karro JE, *et al.* A MATLAB-based tool for accurate detection of perfect overlapping and nested inverted repeats in DNA sequences. *Bioinformatics* 2014;**30**:887–8.
138. Ye C, Ji G, Li L, *et al.* detectIR: a novel program for detecting perfect and imperfect inverted repeats using complex numbers and vector calculation. *PLoS One* 2014;**9**:e113349.
139. Lilley DM. The inverted repeat as a recognizable structural feature in supercoiled DNA molecules. *Proc Natl Acad Sci* 1980;**77**: 6468–72.
140. Strawbridge EM, Benson G, Gelfand Y, *et al.* The distribution of inverted repeat sequences in the *Saccharomyces cerevisiae* genome. *Curr Genet* 2010;**56**:321–40.
141. Berselli M, Lavezzo E, Toppo S. NeSSie: a tool for the identification of approximate DNA sequence symmetries. *Bioinformatics* 2018;**34**:2503–5.
142. Kaushik M. Structural polymorphism exhibited by a quasi-palindrome present in the locus control region (LCR) of the human -globin gene cluster. *Nucleic Acids Res* 2006;**34**:3511–22.
143. Beknazarov N, Jin S, Poptsova M. Deep learning approach for predicting functional Z-DNA regions using omics data. *Sci Rep* 2020;**10**:19134.
144. Tateishi-Karimata H, Sugimoto N. Chemical biology of non-canonical structures of nucleic acids for therapeutic applications. *Chem Commun* 2020;**56**:2379–90.
145. Lefebvre J, Guetta C, Poyer F, *et al.* Copper-alkyne complexation responsible for the nucleolar localization of quadruplex nucleic acid drugs labeled by click reactions. *Angew Chem Int Ed* 2017;**56**:11365–9.
146. Asamitsu S, Takeuchi M, Ikenoshita S, *et al.* Perspectives for applying G-quadruplex structures in neurobiology and neuropharmacology. *Int J Mol Sci* 2019;**20**:2884.

147. Lu L, Jia H, Dröge P, *et al.* The human genome-wide distribution of DNA palindromes. *Funct Integr Genomics* 2007;**7**:221–7.
148. Zhao J, Bacolla A, Wang G, *et al.* Non-B DNA structure-induced genetic instability and evolution. *Cell Mol Life Sci* 2010;**67**:43–62.
149. Nicholson CO, Friedersdorf M, Keene JD. Quantifying RNA binding sites transcriptome-wide using DO-RIP-seq. *RNA* 2017;**23**: 32–46.
150. Meyer KD, Saletore Y, Zumbo P, *et al.* Comprehensive analysis of mRNA methylation reveals enrichment in 3′ UTRs and near stop codons. *Cell* 2012;**149**:1635–46.
151. König J, Zarnack K, Rot G, *et al.* iCLIP reveals the function of hnRNP particles in splicing at individual nucleotide resolution. *Nat Struct Mol Biol* 2010;**17**:909–15.
152. Van Nostrand EL, Pratt GA, Shishkin AA, *et al.* Robust transcriptome-wide discovery of RNA-binding protein binding sites with enhanced CLIP (eCLIP). *Nat Methods* 2016;**13**:508–14.
153. Feigon J, Wang AHJ, van der Marel GA, *et al.* Z-DNA forms without an alternating purine-pyrimidine sequence in solution. *Science (80-)* 1985;**230**:82–4.
154. Eichman BF, Schroth GP, Basham BE, *et al.* The intrinsic structure and stability of out-of-alternation base pairs in Z-DNA. *Nucleic Acids Res* 1999;**27**:543–50.
155. Schroth GP, Ho PS. Occurrence of potential cruciform and H-DNA forming sequences in genomic DNA. *Nucleic Acids Res* 1995;**23**:1977–83.
156. Lexa M, Navrátilová L, Brázdová M. Prediction of significant cruciform structures from sequence in topologically constrained DNA. *Bioinformatics* 2012;124–30.
157. Markham NR, Zuker M. UNAFold: software for nucleic acid folding and hybridization. *Methods Mol Biol* 2008;3–31.
158. Gal M. TRACTS: a program to map oligopurine.oligopyrimidine and other binary DNA tracts. *Nucleic Acids Res* 2003;**31**: 3682–5.
159. Gaddis SS, Wu Q, Thames HD, *et al.* A Web-Based Search Engine for Triplex-forming Oligonucleotide Target Sequences. *Oligonucleotides* 2006;**16**:196–201.
160. Jenjaroenpun P, Kuznetsov VA. TTS Mapping: integrative WEB tool for analysis of triplex formation target DNA Sequences, G-quadruplets and non-protein coding regulatory DNA elements in the human genome. *BMC Genomics* 2009;**10**:S9.
161. Lexa M, Martínek T, Burgetová I, *et al.* A dynamic programming algorithm for identification of triplex-forming sequences. *Bioinformatics* 2011;**27**:2510–7.
162. Ho PS, Ellison MJ, Quigley GJ, *et al.* A computer aided thermodynamic approach for predicting the formation of Z-DNA in naturally occurring sequences. *EMBO J* 1986;**5**:2737–44.
163. Schroth GP, Chou PJ, Ho PS. Mapping Z-DNA in the human genome. Computer-aided mapping reveals a nonrandom distribution of potential Z-DNA-forming sequences in human genes. *J Biol Chem* 1992;**267**:11846–55.