Current Genetics (2019) 65:575–590
https://doi.org/10.1007/s00294-018-0907-8

ORIGINAL ARTICLE

A strong structural correlation between short inverted repeat sequences and the polyadenylation signal in yeast and nucleosome exclusion by these inverted repeats

Osamu Miura[1] · Toshihiro Ogake[2] · Hiroki Yoneyama[2] · Yo Kikuchi[2] · Takashi Ohyama[1,2]

Received: 18 September 2018 / Revised: 14 November 2018 / Accepted: 15 November 2018 / Published online: 29 November 2018

Abstract

DNA sequences that read the same from 5′ to 3′ in either strand are called inverted repeat sequences or simply IRs. They are found throughout a wide variety of genomes, from prokaryotes to eukaryotes. Despite extensive research, their in vivo functions, if any, remain unclear. Using *Saccharomyces cerevisiae*, we performed genome-wide analyses for the distribution, occurrence frequency, sequence characteristics and relevance to chromatin structure, for the IRs that reportedly have a cruciform-forming potential. Here, we provide the first comprehensive map of these IRs in the *S. cerevisiae* genome. The statistically significant enrichment of the IRs was found in the close vicinity of the DNA positions corresponding to polyadenylation [poly(A)] sites and ~30 to ~60 bp downstream of start codon-coding sites (referred to as 'start codons'). In the former, ApT- or TpA-rich IRs and A-tract- or T-tract-rich IRs are enriched, while in the latter, different IRs are enriched. Furthermore, we found a strong structural correlation between the former IRs and the poly(A) signal. In the chromatin formed on the gene end regions, the majority of the IRs causes low nucleosome occupancy. The IRs in the region ~30 to ~60 bp downstream of start codons are located in the +1 nucleosomes. In contrast, fewer IRs are present in the adjacent region downstream of start codons. The current study suggests that the IRs play similar roles in *Escherichia coli* and *S. cerevisiae* to regulate or complete transcription at the RNA level.

Keywords Inverted repeat (IR) · Yeast genome · IR map · Nucleosome exclusion · 3′-Untranslated region

Introduction

The multifarious structures and physical properties of DNA are thought to provide additional structural and functional dimensions to chromatin organization and gene expression (Schroth et al. 1992; Herbert et al. 1998; Liu et al. 2001; Ohyama 2001; Fukue et al. 2004, 2005; Paeschke et al. 2005; Sumida et al. 2006; Kamiya et al. 2007; Jain et al. 2008; Qin and Hurley 2008; Strawbridge et al. 2010; Du et al. 2013; Kimura et al. 2013; Nishikawa and Ohyama 2013). The occurrence of diverse DNA structures usually requires special sequence characteristics or defined symmetry elements, which are frequently found in the genomes of both prokaryotes and eukaryotes. For example, alternating purine–pyrimidine sequences, periodically occurring A-tracts, inverted repeat (IR) sequences, homopurine/homopyrimidine sequences and guanine-rich sequences lead to the formation of left-handed Z-DNAs, curved DNAs, cruciforms, triple-stranded H-DNAs (triplexes) and four-stranded G-quadruplexes, respectively (Sinden 1994). Except for curved DNAs, however, the other structures additionally require local DNA underwinding for their occurrence (Paleček 1991; Van Holde and Zlatanova 1994; Krasilnikov et al. 1999; Kouzine and Levens 2007; Sun and

Communicated by M. Kupiec.

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s00294-018-0907-8) contains supplementary material, which is available to authorized users.

✉ Takashi Ohyama
ohyama@waseda.jp

[1] Department of Biology, Faculty of Education and Integrated Arts and Sciences, Waseda University, 2-2 Wakamatsu-cho, Shinjuku-ku, Tokyo 162-8480, Japan

[2] Major in Integrative Bioscience and Biomedical Engineering, Graduate School of Science and Engineering, Waseda University, 2-2 Wakamatsu-cho, Shinjuku-ku, Tokyo 162-8480, Japan

Hurley 2009). The dynamic processes of DNA replication and transcription generate the local DNA underwinding.

A DNA sequence that reads the same from 5′ to 3′ in each strand is known as an IR or a palindrome. IR sequences are commonly found in a wide variety of genomes, from prokaryotes to eukaryotes (Warburton et al. 2004; Wang and Leung 2009; Strawbridge et al. 2010; Cer et al. 2013; Du et al. 2013). Some of these sequences can form cruciforms with the aid of energy from negative supercoiling of DNA and in turn, cruciforms can reduce the negative superhelicity in that region (Lilley 1980; Lilley and Markham 1983; Courey and Wang 1988; Paleček 1991; Van Holde and Zlatanova 1994; Shlyakhtenko et al. 1998; Krasilnikov et al. 1999; Oussatcheva et al. 2004; Kouzine and Levens 2007). Thus, cruciforms have the potential to influence nucleosome formation and/or positioning and the local chromatin structure in eukaryotes. Numerous studies have sought to clarify the biological functions of IR sequences or cruciform structures, and suggested their participation in DNA replication (Pearson et al. 1996; Zannis-Hadjopoulos et al. 2008; Brázda et al. 2011), transcription (Dai et al. 1997; Dai and Rothman-Denes 1998; Jagelská et al. 2010; Brázda et al. 2012; Coufal et al. 2013; Miura et al. 2018), recombination (Lin et al. 1997; Shlyakhtenko et al. 2000; Lobachev et al. 2002; Wang and Leung 2006) and genome or chromosome instability (Wang and Leung 2006; Inagaki et al. 2013; Javadekar and Raghavan 2015). Furthermore, a recent study showed that short IRs with cruciform-forming potential are hotspots for genome instability in human cancer cells (Lu et al. 2015; Bacolla et al. 2016). Many reports have also suggested the presence of cruciform-binding proteins (for review, Brázda et al. 2011; Qian and Adhya 2017). However, determining the presence of cruciforms and identifying their biological role have generally been difficult, particularly in eukaryotic systems (Gentry and Hennig 2016).

With the availability of genome sequence databases, we can now easily search for IR sequences in genomic DNA. Thus, genome-wide analyses of IR sequences would provide a powerful means to assess their biological significance. Recently, genome-wide computational analyses for the distribution of IR sequences have been performed for the proteobacterium *Escherichia coli* and the budding yeast *Saccharomyces cerevisiae* (Strawbridge et al. 2010; Du et al. 2013; Miura et al. 2018). In *E. coli*, a strong enrichment of IRs with cruciform-forming potential was found in the adjacent regions downstream of the stop codon-coding sites (referred to as 'stop codons') and on and around the positions corresponding to mRNA ends (referred to as 'gene ends'). Furthermore, most of the IRs with a repeat unit length of ≥ 8 bp and a spacer size of ≤ 8 bp were parts of the intrinsic terminators (Miura et al. 2018). For the *S. cerevisiae* genome, Strawbridge et al. reported that the IRs were significantly enriched and highly clustered in the intergenic regions (in this study, the genome was partitioned into coding and non-coding regions, referred to as 'genic' and 'intergenic' regions, respectively), especially in the 3′-flanking regions of the genic regions, while their occurrence in coding sequences was random (2010). These studies revealed the somewhat similar features for the occurrence of IRs or cruciform motifs between prokaryotes and eukaryotes. However, many unanswered questions still remain for the IRs in the yeast genome, including where they are located in the 3′-flanking regions of the genic regions, what primary structures they adopt, whether there is some relationship between their primary structures and positions in the genome, how these sequences influence chromatin structure in vivo, and so forth. Addressing these questions would provide clues toward clarifying the biological significance of IRs or cruciforms.

In the current study, we constructed the first *S. cerevisiae* genome-wide comprehensive map of the IRs that reportedly have a cruciform-forming potential. Furthermore, by introducing the information about the DNA positions corresponding to polyadenylation [poly(A)] sites [referred to as 'poly(A) sites'] (i.e., gene ends), we could perform more accurate analyses than previously possible for the biological relevance of the focused IRs. We found that the IRs occur frequently in the close vicinity of poly(A) sites and ~ 30 to ~ 60 bp downstream of start codon-coding sites (referred to as 'start codons'), and these enrichments are statistically significant. However, the effects of these IRs on the chromatin structure are different: the majority in the former regions excludes nucleosomes, while the IRs in the latter regions are incorporated into the + 1 nucleosomes. The DNA sequence analysis revealed that the enriched IRs comprise three different types: two types are in the close vicinity of poly(A) sites and another type is in the open reading frame (ORF) region. Furthermore, we found a strong structural correlation between the former two types and the poly(A) signal. Moreover, our analyses provided clues about the functions of the IRs conserved between *E. coli* and *S. cerevisiae*.

Materials and methods

Genome sequence and gene annotation

We obtained the full genome sequence of *S. cerevisiae* from the *Saccharomyces* Genome Database (SGD, https://www.yeastgenome.org). Gene annotations for *S. cerevisiae* were from SGD (R64) and Park et al. (2014).

Partitioning of the genome

We defined the 'genic' and 'intergenic' regions as follows: genic: ORF, 5′- and 3′-UTRs (untranslated regions) and

OUR-1, -2, and -3 (OUR: overlapping untranslated region; OUR-1, the 5′-UTR of one gene partially or completely overlaps that of another gene; OUR-2, the 3′-UTR of one gene partially or completely overlaps the 5′-UTR of another gene; OUR-3, the 3′-UTR of one gene partially or completely overlaps that of another gene); and intergenic: 'TAN' (the region between tandem genes), 'DIV' (that between divergent genes) and 'CON' (that between convergent genes). The information about the transcription start sites (TSSs) and the poly(A) sites for protein-coding genes was obtained from Park et al. (2014) and that about the start codons and the stop codons was obtained from the SGD. The terms 'tandem', 'divergent' and 'convergent' refer to the directions of transcription for the abutting genes. For intergenic regions, only those that had two clear ends, such as two TSSs, a poly(A) site and a TSS, or two poly(A) sites, were analyzed. In the cases where two protein-coding genes contain a pseudogene, tRNA gene, rRNA gene or these genes in between, the entire region between the two protein-coding genes was not subjected to further analyses.

IR identifier

We used the computer program 'CIRI', which judges a given sequence as a target IR when the repeat unit length is longer than or equal to 5 bp, the spacer length is 0–8 bp and the entire IR length is longer than or equal to 13 bp (Miura et al. 2018). The CIRI program was run against the *S. cerevisiae* genome.

Genome-wide distribution map of IR sites

The method was recently reported (Miura et al. 2018). Briefly, the location of each IR was mapped by the position of the central base pair. When an IR is located inside a larger IR, only the outer IR was used for the analyses. To construct the genome-wide distribution map of the IR sites, the Circos software (Krzywinski et al. 2009) was used. Furthermore, we developed a web-based server, 'Cruciform-formable IRs in the *S. cerevisiae* genome (CFIRs-Sc)' (http://www.waseda.jp/sem-ohyama/CFIRs-Sc), which is an application for browsing the map interactively.

Regional distribution profiles of IRs

The regional distribution profiles of IRs were drawn using two homemade scripts. One sorts the IRs into the partitioned regions (ORF, 5′- and 3′-UTRs, etc.). The other measures the distance between a given IR and each end of the relevant region.

Randomized control sequences and statistical analysis

The *S. cerevisiae* genome was partitioned into coding (ORF) and noncoding (non ORF) regions, according to its SGD annotations. The sequence randomization was performed by the method of Strawbridge et al. (2010) and Miura et al. (2018). Using 100 randomized genomes as the "control genomes", we obtained control data. Using the test datum and the corresponding 100 control data for each bin of 10 bp, the Grubbs test was performed to examine whether the former was a significant outlier.

Sorting of the IR sequences

Based on the AT content, the occupancy of the longest A (or T)-tract (greater than or equal to three runs of A or T) and the occupancy of the longest $(ApT)_n$ [or $(TpA)_n$] ($n \geq 1$) in a repeat unit, the IR sequences were sorted into seven types (types I–VII).

Nucleosome occupancy

The MNase-seq data were downloaded from the NCBI SRA database under the accession number SRR2045610, and processed to generate the BED files of the paired-end read data corresponding to 16 chromosomes (Ocampo et al. 2016). Using the files and the iNPS algorithm (Chen et al. 2014), the nucleosome positions in each chromosome were determined. When a given region was incorporated into a nucleosome, the nucleosome occupancy of the region was defined as 1.0 and when it was not, the value was defined as 0. The nucleosome occupancy data based on the chemical cleavage were obtained from Chereji et al. (2018).

The IRs were collected independently (IR by IR) and aligned with their center positioned at 0. Subsequently, the per-position nucleosome occupancy values were calculated and averaged from the upstream position to the downstream position. The averaged values were normalized to the average nucleosome occupancy of each chromosome that was defined as 1.0. The resulting values were abbreviated as average *nNuOcs*.

Results

The current analyses excluded the IRs that seemed to have no potential for transition into cruciforms. To our knowledge, the shortest stem in a cruciform heretofore reported is 5 bp (Sheflin and Kowalski 1985; Iacono-Connors and Kowalski 1986; Müller and Wilson 1987; McMurray et al. 1991; Dai et al. 1997; Dai and Rothman-Denes 1998; Jagelská et al. 2010; Nuñez et al. 2015), and the typical number of

nucleotides in a loop has been suggested to be 3–6 (Hilbers et al. 1985; Furlong and Lilley 1986; Gough et al. 1986; Nag and Petes 1991; Sinden 1994; Potaman and Sinden 2005). However, larger loops can also be formed in some cases, and even motifs with no spacer can form loops in the resulting cruciform (Furlong and Lilley 1986; Gough et al. 1986; Scholten and Nordheim 1986; Müller and Wilson 1987; Damas et al. 2012). Thus, we focused on the IRs with repeat unit lengths greater than or equal to 5 bp, spacer lengths between 0 and 8 bp and an entire IR length longer than or equal to 13 bp. The IRs are named and grouped in the following manner; e.g., R8S4 (the IR with repeat unit length of 8 bp and spacer length of 4 bp), for convenience. Imperfect IRs were excluded from the screening. The reasons were as follows: they occur less frequently than perfect IRs, undergo spontaneous mutations to form more perfect IRs and require higher energies for cruciform formation (Benham et al. 2002; Van Noort et al. 2003).

Distribution of IR sequences with cruciform-forming potential

At first, we constructed a comprehensive map for the $R \geq 5S \leq 8$ ($2R + S \geq 13$) IRs with the following information: their positions and structures, genes with annotations, and positions of TSSs and poly(A) sites (Fig. 1, http://www.waseda.jp/sem-ohyama/CFIRs-Sc). Although the loci of pseudogenes and rRNA and tRNA genes are shown in the map, these were not subjected to further analyses. This is because pseudogenes generally have incomplete information for the TSS and poly(A) site, and most of the IRs detected in rRNA and tRNA gene loci are used to form the secondary structures of the corresponding RNA molecules. Thus, the analyses described below focus on protein-coding genes and their flanking regions.

The distribution profile of the IRs in the yeast genome shows that the IRs with a repeat unit of ≥ 10 are rare in the genome (Fig. 1, http://www.waseda.jp/sem-ohyama/CFIRs-Sc). In contrast, the IRs belonging to the $R5S \leq 8$ ($2R + S \geq 13$) seem to be abundant. Subsequently, we examined whether any regional characteristics are associated with the IR occurrence. For this analysis, the yeast

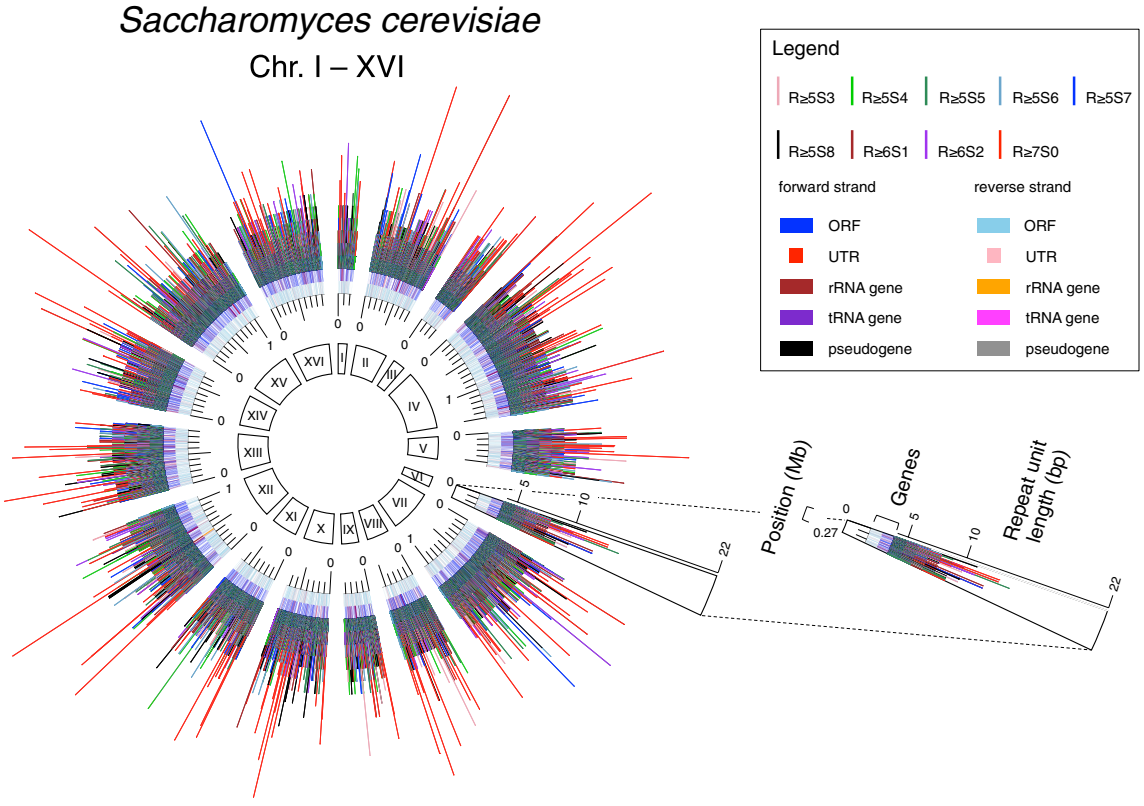


Fig. 1 Distribution of IRs in the *S. cerevisiae* genome. The position coordinates of the $R \geq 5S \leq 8$ ($2R + S \geq 13$) IRs are overlaid on the map of genes with annotations, with their repeat lengths shown as line heights. The map can also be browsed interactively in the CFIRs-Sc (http://www.waseda.jp/sem-ohyama/CFIRs-Sc)

genome was partitioned into six genic and three intergenic regions, as shown in Fig. 2. Furthermore, 100 randomized sequences were generated for each of the genic and intergenic regions ("Materials and methods") to determine whether the apparent enrichment or deficiency of the IRs in a given region is statistically significant. The analysis showed that the IRs with cruciform-forming potential are enriched in 3′-UTRs and ~30 to ~60 bp downstream of start codons and the enrichments are statistically significant (Fig. 2). In 3′-UTRs, the regions of enrichment were ~20 to ~90 bp upstream of poly(A) sites and ~100 to ~130 bp downstream of stop codons. The data suggested that the IRs are located closer to poly(A) sites than stop codons. To confirm this, 3′-UTRs were sorted by width and the same analysis was performed, which clearly showed that the IRs are located closer to poly(A) sites (Fig. 3). Finally, we note that the distribution analysis shown in Fig. 2 also revealed that fewer IRs were present in the adjacent regions downstream of start codons and around ~15 bp downstream of TSSs.

Sequence characteristics of the IRs

The sequence characteristics of the IRs located in 3′-UTRs and ~30 to ~60 bp downstream of start codons were subsequently examined. In this analysis, the sequences of the $R \geq 5 S \leq 8$ $(2R + S \geq 13)$ IRs were sorted into seven types, according to AT content, A- or T-tract occupancy and $(ApT)_n$ or $(TpA)_n$ occupancy in a repeat unit (Fig. 4). Regarding the AT content, the value of 0.6 comes from that of the *S. cerevisiae* genome of 0.62. The sequence type III occurred in 3′-UTRs most frequently, and was especially eminent in the ~−30 to ~−60 region relative to poly(A) sites. The sequence type II was the second most frequent in 3′-UTRs and ~−10 to ~−20 relative to poly(A) sites were more eminent for this phenomenon. Thus, the type III IRs are generally located slightly upstream of the type II IRs. The sequence types III and II are both AT-rich (AT content ≥ 0.6), but they differ in that the former is $(ApT)_n$ or $(TpA)_n$-rich (≥ 0.5) in a repeat unit while the latter is A- or T-tract-rich (≥ 0.5). The sequence type I, which is somewhat similar to both types III and II, is also enriched in restricted small regions that are located within the type III and/or type II-enriched regions, although it occurs much less frequently than these types.

In the region ~30 to ~60 bp downstream of start codons, the sequence type VII, which is neither AT-rich, A- or T-tract-rich, nor $(ApT)_n$ or $(TpA)_n$-rich, was enriched. The sequence type V [neither AT-rich, $(ApT)_n$-rich nor $(TpA)_n$-rich] is also enriched in a restricted small region within the type VII-enriched regions, although its occurrence frequency is much lower than that of type VII.

Localizations of the IRs in chromatin

Cruciform structures are incompatible with nucleosome structures (Nickol and Martin 1983; Nobile et al. 1986; Battistoni et al. 1988; Van Holde and Zlatanova 1994; Pearson et al. 1996). Accordingly, we can roughly speculate on the potential of a given IR to transition into a cruciform in vivo, by examining where it is located in the chromatin. Several groups have reported genome-wide nucleosome maps for budding yeast (Kaplan et al. 2009; Brogaard et al. 2012; Henikoff et al. 2014; Hu et al. 2014; Ramachandran et al. 2015; Ocampo et al. 2016; Chereji et al. 2018). At first, we used the MNase-seq-based map of Ocampo et al. (2016) for this purpose. This map is based on the paired-end sequencing, which provides more accurate nucleosome positions than single-read data (Cole et al. 2012; Ocampo et al. 2016). For the chromatin of 3′-UTRs, the positions of the types III and II IRs were examined and for those formed ~30 to ~60 bp downstream of start codons, the type VII IRs were examined. As shown in Fig. 5, a clear difference was found between the two results. In the chromatin of 3′-UTRs, the types III and II IRs are generally located at the bottom or very close to it in each profile, indicating that these types are more preferentially located in the linker DNA regions than the other DNA sequences in 3′-UTRs. Furthermore, the profile in each panel is asymmetric and the peak appearing on the upstream side is generally higher than that on the downstream side, indicating that the nucleosome occupancies generally differ between upstream and downstream of the IRs. In contrast, for the chromatin formed on the ~30 to ~60 bp downstream region of start codons, the majority of the type VII IRs is located within nucleosomes, which are most certainly the +1 nucleosomes (Tirosh et al. 2010; Tsui et al. 2011).

As an alternative to drawing nucleosome maps, chemical cleavage-based methods are known, and they can reportedly avoid the cleavage bias caused by the preference of MNase for A/T-rich regions and be thought to provide more accurate data on nucleosome positions (Brogaard et al. 2012; Henikoff et al. 2014; Chereji et al. 2018). Thus, using the chemical cleavage-based nucleosome map of Chereji et al. (2018), which was based on the H3Q85C cleavage method, we also performed the same analysis. The profiles were generally similar to those obtained based on the MNase-seq-based map. In this analysis, however, the asymmetry in the 3′-UTR profiles was more pronounced, confirming that the nucleosome occupancies change between upstream and downstream regions of the IRs in 3′-UTRs, from high to low. For the focused region in ORFs, the majority of the type VII IRs was also found within nucleosomes.

Finally, we examined the relationship between the IR structure and the nucleosome occupancy for the IRs found in 3′-UTRs (Fig. 6). This analysis revealed several interesting

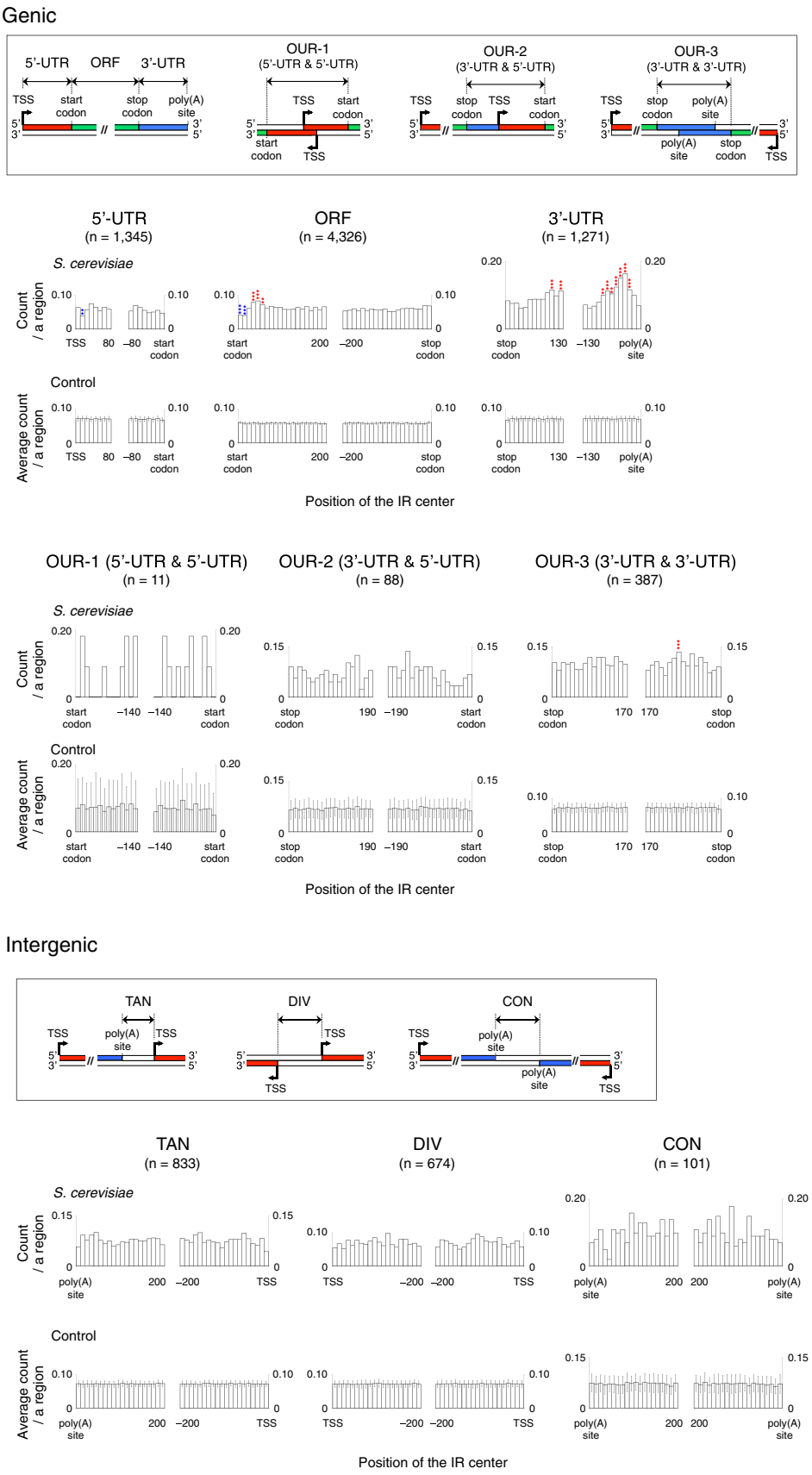


Fig. 2 Regional occurrence frequencies of the IRs. The regional occurrence frequencies of the $R \geq 5S \leq 8$ ($2R + S \geq 13$) IRs were analyzed. Genic and intergenic regions were subdivided, as schematically shown in the insets. The positions of the IRs are represented by their center positions. A TSS, the first nucleotide of a start codon, the third nucleotide of a stop codon, and a poly(A) site were each defined as position 0. In each panel, the span of the *x*-axis indicates the average length of a given region, except for the ORF, TAN, DIV and CON panels (Supplementary Table S1). The samples with lengths larger than the average length were subjected to the analysis, to obtain the information about the region that all samples have in common ('*n*' indicates the number of samples). The average lengths of ORFs, TANs, DIVs and CONs are 1536 bp, 305 bp, 420 bp and 209 bp, respectively, and thus only 200 bp regions from the relevant two positions were analyzed. The control data were obtained using 100 control genomes ("Materials and methods") and the statistical significance levels were calculated based on the Grubbs test. The bin size is 10 bp. $^{**}P < 0.01$, $^{***}P < 0.001$ (red, enrichment; blue, deficiency)

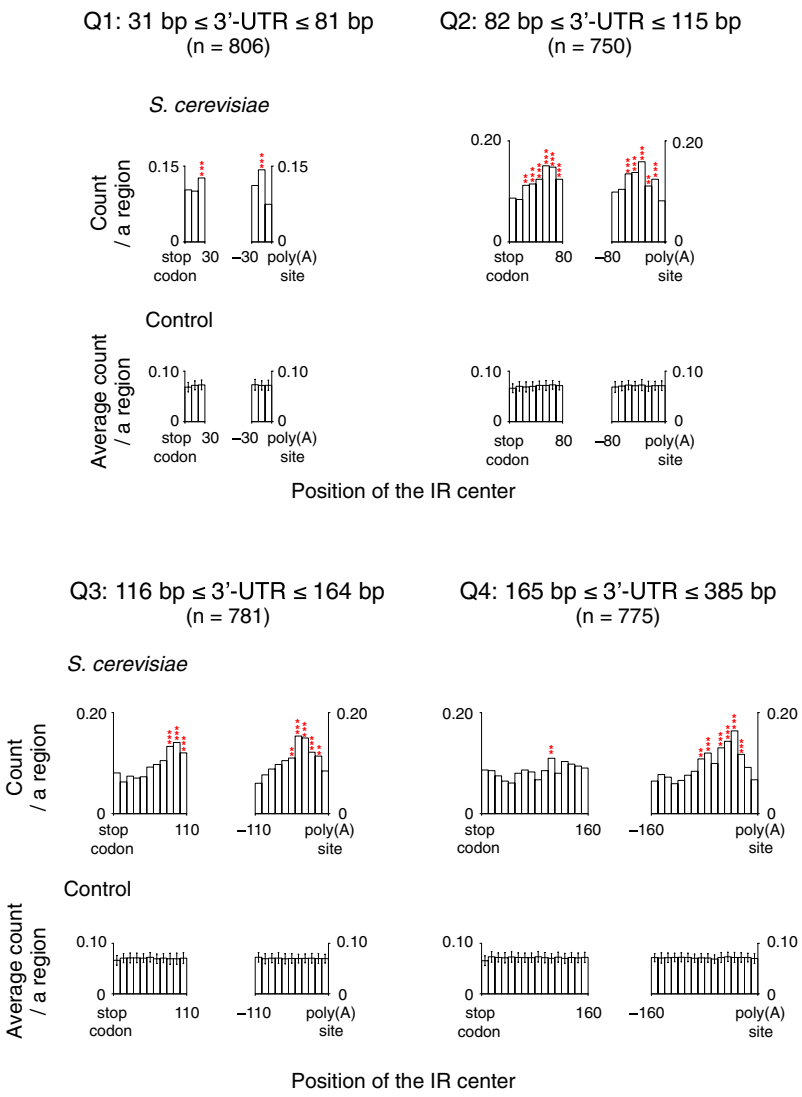


Fig. 3 Position of the IRs in 3′-UTRs. According to the length, 3′-UTRs were sorted into five groups, and the four groups named Q1–Q4 were subjected to the analysis: Q1, 31 bp ≤ 3′-UTR ≤ 81 bp; Q2, 82 bp ≤ 3′-UTR ≤ 115 bp; Q3, 116 bp ≤ 3′-UTR ≤ 164 bp; Q4, 165 bp ≤ 3′-UTR ≤ 385 bp. In each group, the position histogram of the R ≥ 5S ≤ 8 (2R + S ≥ 13) IRs is shown. The span of the *x*-axis corresponds to the region range common among a given group. '*n*' indicates the number of 3′-UTR samples. The bin size is 10 bp. $^{**}P < 0.01$, $^{***}P < 0.001$ (red, enrichment)

points. Firstly, as the unit lengths of the type III IRs increased, the average values of the normalized nucleosome occupancies (*nNuOcs*; "Materials and methods") decreased. Second, in most cases, when the length of the type III IRs becomes ≥ 9, the average *nNuOc* value becomes ~ 0 on the IRs or very close to them. In the latter cases, another IR or IRs or A/T-rich tracts were often found to be the sites of the ~ 0 value (Supplementary Fig. S1). Third, the type II IRs that showed values ~ 0 are rare (this may be caused by the lengths of their repeat units: those with R ≥ 10 were not found and 92% of them had a repeat unit length of 5–6 bp).

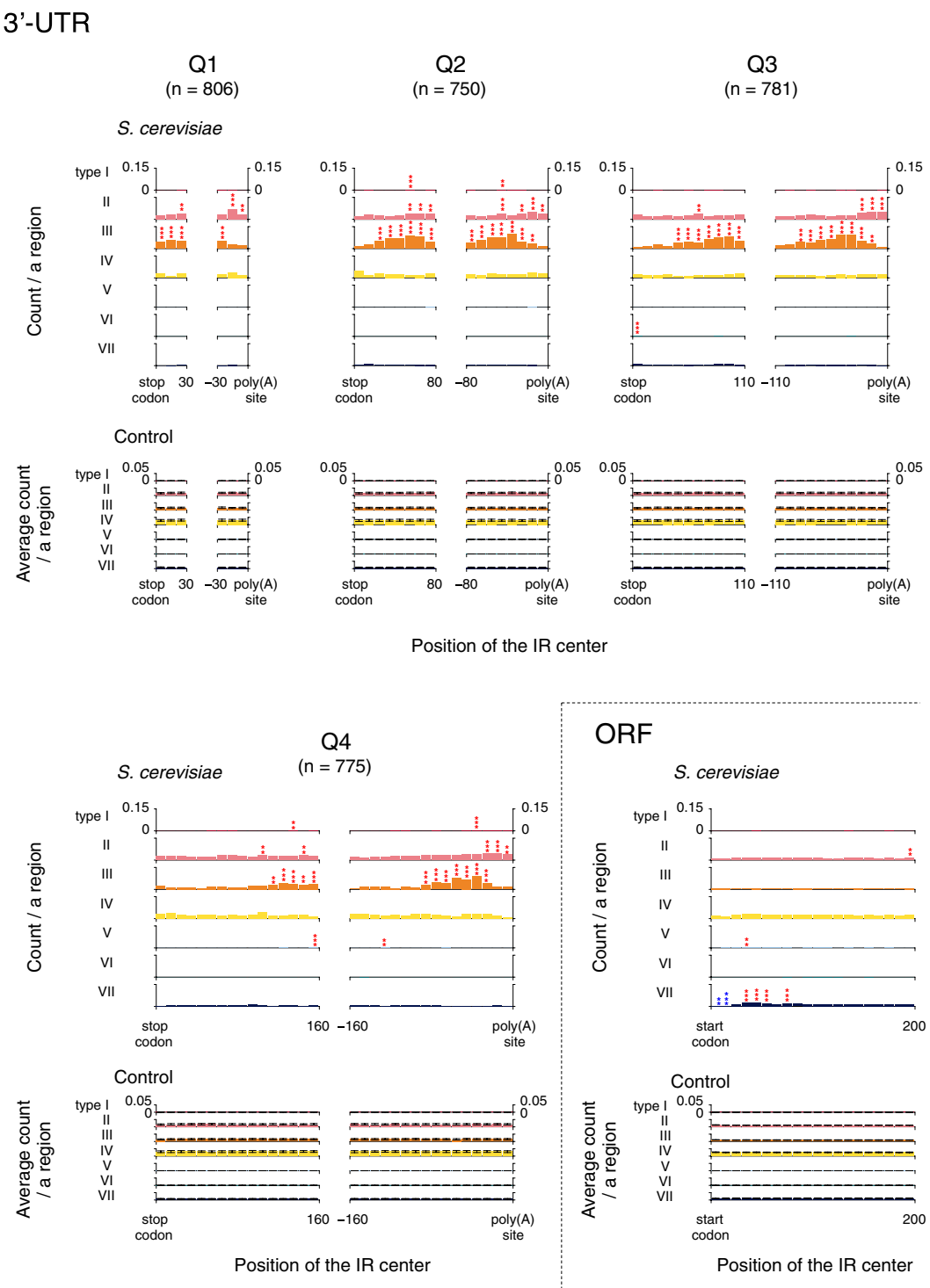
Discussion

We performed genome-wide analyses for the distribution, occurrence frequency, sequence characteristics and relevance to chromatin structure of the IRs that reportedly have a cruciform-forming potential. The IRs are widely distributed in the yeast genome. The ApT- or TpA-rich type III IRs and A-tract- or T-tract-rich type II IRs are enriched in 3′-UTRs, especially in the close vicinity of poly(A) sites. The majority of these types is located in linker DNA regions. In the region ~ 30 to ~ 60 bp downstream of start codons, the type VII IRs, which are neither AT-rich, A- or T-tract-rich, nor $(ApT)_n$ or $(TpA)_n$-rich, are enriched and located within the + 1 nucleosome. In contrast, fewer IRs are present in the adjacent region downstream of start codons and around ~ 15 bp downstream of TSSs. Here, we discuss what these phenomena suggest with regard to the genetic events.

What the positions and the types of IRs suggest

The types III and II IRs are enriched in 3′-UTRs. They seem to correspond to the important elements in RNA that are used as the poly(A) signal, PAS. Furthermore, the

Repeat unit \ Type	I	II	III	IV	V	VI	VII
AT content	1	≥0.6	≥0.6	≥0.6	<0.6	<0.6	<0.6
Occupancy of the longest A (or T)-tract in the unit	0.5	≥0.5	<0.5	<0.5	≥0.5	<0.5	<0.5
Occupancy of the longest $(ApT)_n$ [or $(TpA)_n$] in the unit	0.5	<0.5	≥0.5	<0.5	<0.5	≥0.5	<0.5



◂**Fig. 4** Sequence characteristics of the IRs in 3′-UTRs and ORFs. The $R \geq 5S \leq 8$ ($2R + S \geq 13$) IRs were classified into seven types according to AT content, A- or T-tract occupancy and $(ApT)_n$ or $(TpA)_n$ occupancy in a repeat unit. The occurrence profiles of these types in 3′-UTRs and ORFs are shown. For the Q1–Q4 groups, see Fig. 3. The bin size is 10 bp. $^{**}P < 0.01$, $^{***}P < 0.001$ (red, enrichment; blue, deficiency)

nucleosome occupancy changes within the 3′-UTR from high (upstream) to low (downstream), and these IRs are located at the border (Fig. 5). Although the PAS of *S. cerevisiae* is reportedly very degenerate and thus recognizing the PAS in a given gene is sometimes difficult, the current study provides a new perspective on this issue. Generally, from upstream to downstream, a PAS consists of an AU-rich efficiency element 'EE' (UAYRUA: Y = U or C, R = A or G), an A-rich positioning element 'PE' (AAWAAA: W = A or U) that is typically located ~ 10 to ~ 30 nucleotides upstream of the cleavage position, and a U-rich element spanning the cleavage position and the site of poly(A) addition (Guo and Sherman 1996; Zhao et al. 1999; Proudfoot 2011; Mischo and Proudfoot 2013). The EE and PE sequences seem to correspond to the types III and II IR sequences of the DNA, respectively. Furthermore, the mutual positional relationships among the EE, the PE and the site of poly(A) addition are very similar to those among the type III IR, the type II IR and the poly(A) site. The type II IRs occur slightly closer to the poly(A) sites than the type III IRs, in general. Thus, these analyses indicated that in a certain population of genes, the EE-coding DNA region and/or the PE-coding DNA region presumably constitute(s) the repeat units of the type III IRs and/or that of the type II IRs, respectively. Viewed in this light, the types III and II IRs seem to function at the RNA level, rather than the DNA level.

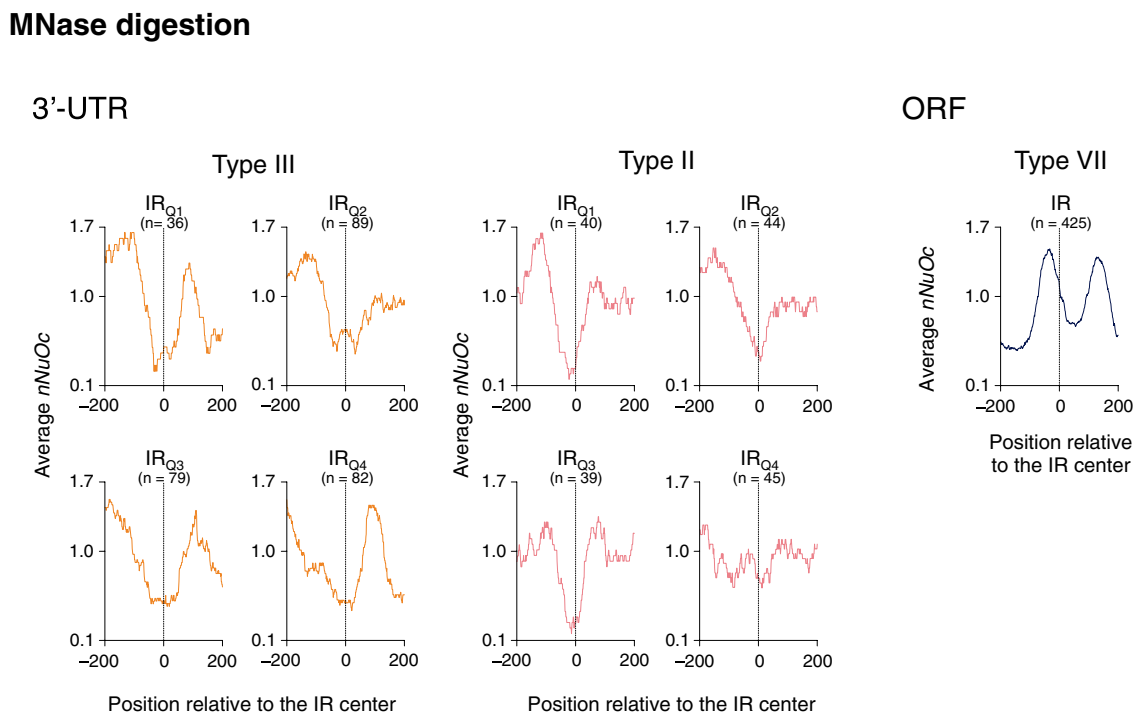
The type VII IRs are enriched in the regions ~ 30 to ~ 60 bp downstream of start codons. They are not AT-rich (the average AT content in the repeat units of the type VII IRs located in this region is ~ 40%) and lack the sequence advantage for cruciform formation, and are actually located within nucleosomes (Fig. 5). Thus, if they have some biological function, it would presumably be at the RNA level. The function may be some "riboregulator"-like one found in bacteria (Merino and Yanofsky 2005; Wachter 2014; Millman et al. 2017). The riboregulators can assume two mutually exclusive RNA structures in the primary transcripts: one forms a terminator and results in premature transcription termination, and the other forms an antiterminator that allows the production of a full-length mRNA by read-through into the coding sequence (Millman et al. 2017). Although riboregulator-related IRs usually occur in the 5′-UTR in *E. coli*, we suggested that such IRs may also occur in the region ~ 25 to ~ 60 bp downstream of the start codons in this organism (Miura et al. 2018). Furthermore, it must be noted that conditional transcriptional terminator-like structures, which have an IR followed by a U-rich tract, are sometimes found in the focused regions (data not shown). Thus, in *S. cerevisiae*, the IRs in the regions ~ 30 to ~ 60 bp downstream of start codons may play some riboregulator-like role.

We also found an IR-deficient region adjacent downstream of start codons (Fig. 2). Since a stem-loop RNA structure formed near a start codon would negatively influence translation initiation, this situation may be diminished in yeast. The region around ~ 15 bp downstream of TSSs was another site of low IR occurrence. For this phenomenon, we presently cannot give any plausible explanation.

Possible causes of low nucleosome occupancy on the types III and II IRs

The region around ~ 100 bp downstream of a stop codon is known to have relatively low nucleosome occupancy in yeast (Kaplan et al. 2009; Pan et al. 2011). To explain this phenomenon, a hypothesis was raised that PASs disfavor nucleosome formation (Kaplan et al. 2009). This putative propensity of PASs may be caused by the types III and II IRs in a certain population of genes. For the type III IRs, the cruciform formation is the first issue to discuss as a possible cause. Dayn et al. (1991, 1992) reported that all detected in vivo cruciforms are formed by AT-rich inverted repeats, particularly $(ApT)_n$ sequences. Other groups also arrived at similar conclusions (McClellan et al. 1986, 1990; Panayotatos and Fontaine 1987; Wells and Harvey 1987; Horwitz and Loeb 1988; Calladine et al. 2004). Mechanistically, the very small contribution of the stacking forces of the $(ApT)_n$ sequences to stabilize the B-form is likely to be the cause of the transition into cruciforms (Panayotatos and Fontaine 1987). However, the hypothesis of "B to cruciform transition" for the type III IRs has a "size-problem".

The size of a cruciform is a debated issue. Vologodskii et al. suggested that cruciform extrusion in short palindromes with low supercoiling is highly improbable (Vologodskaia and Vologodskii 1999; Vologodskii 2015), and a theoretical study by Zhabinskaya and Benham (2013) was in accordance with this suggestion. In the latter study, the cruciforms with stem lengths of < ~ 15 bp seemed improbable (however, DNA melting seemed possible even for the IRs with ~ 3 bp repeats). In the current study, short IRs with repeat units of < ~ 15 bp were found to be the majority, including the type III IRs, in the yeast genome (Figs. 1, 6, http://www.waseda.jp/sem-ohyama/CFIRs-Sc). Thus, based on the studies by Vologodskii et al. and Zhabinskaya and Benham, the in vivo transition of the type III IRs into cruciforms may be "highly improbable" (but melting or deformation seems possible). However, we must also note that numerous reports have shown or proposed the presence of cruciforms with short stems of 5–7 bp (Sheflin and Kowalski 1985; Iacono-Connors and Kowalski 1986; Müller and



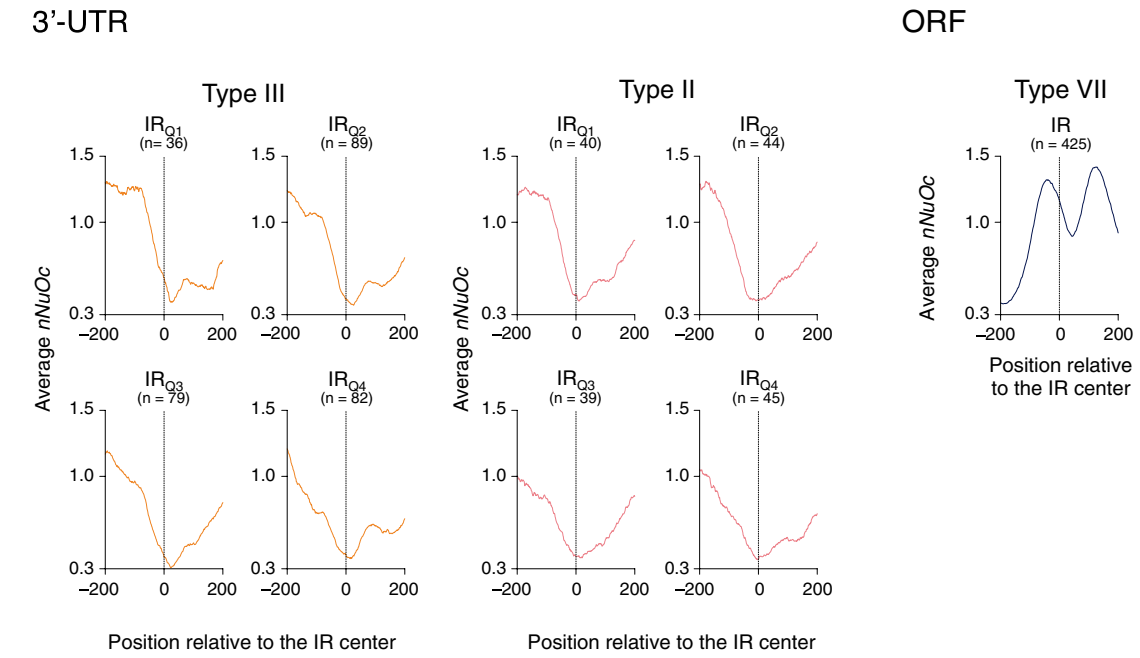


Fig. 5 Nucleosome occupancy on and around the IRs. The IRs (types III and II for 3′-UTRs and type VII for ORFs) that showed statistically significant scores for the occurrence (Fig. 4) were subjected to the analysis. IR_{Q1}–IR_{Q4} mean the IRs located in the 3′-UTR length groups Q1–Q4 (Fig. 3), respectively. The average *nNuOc* value ("Materials and methods") for each base pair located from −200 to +200 relative to the IR center, indicated as 0, was calculated and plotted. In the case of tandem genes, the low nucleosome occupancy on the promoter of the downstream gene may affect the total profile. Thus, only convergent genes were used in this analysis. The data of nucleosome positions were obtained from Ocampo et al. (2016) (based on MNase digestion) and Chereji et al. (2018) (based on chemical cleavage)

Chemical cleavage

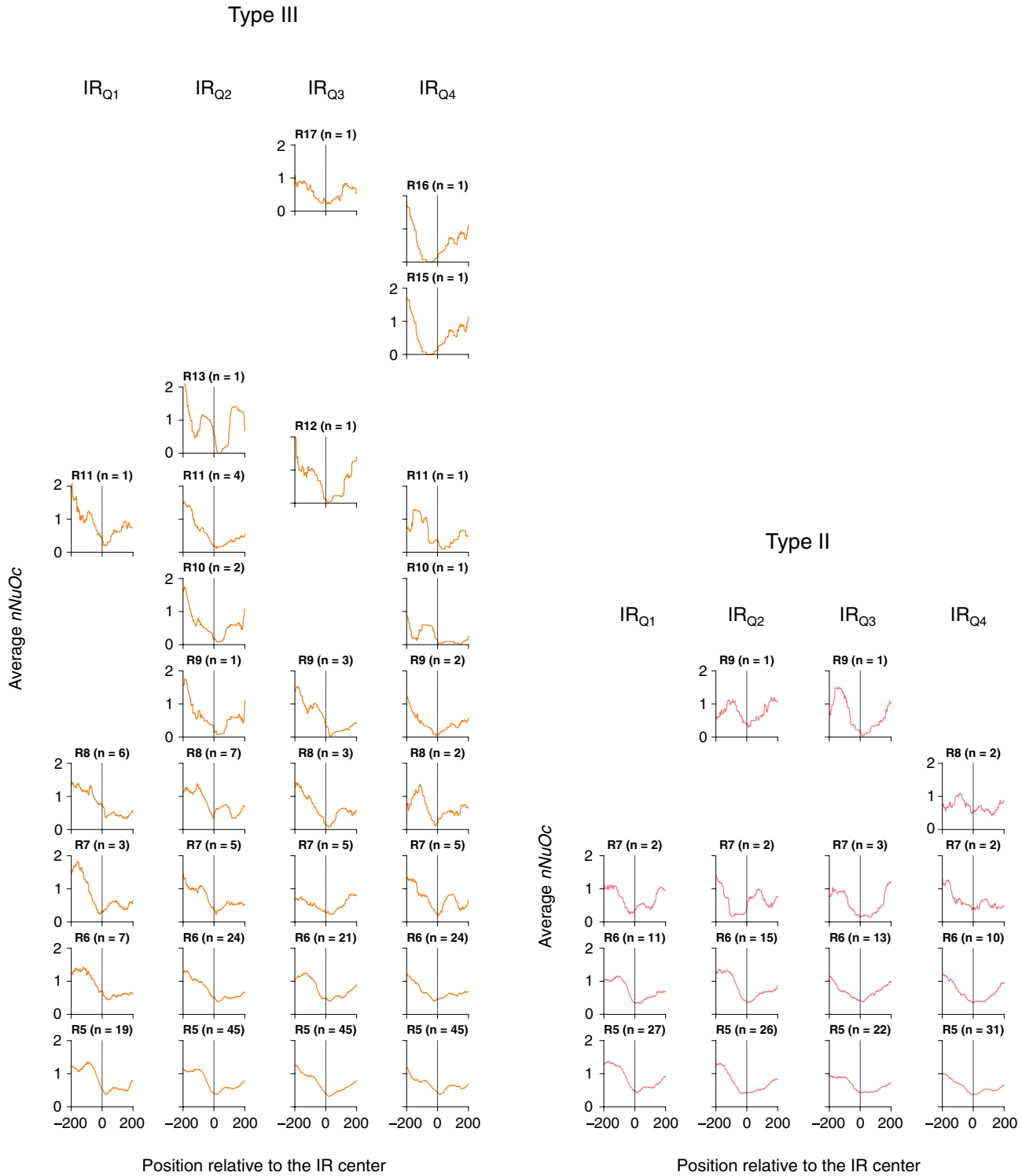


Fig. 6 Relationship between the IR structure and the nucleosome occupancy. The IRs located in 3′-UTRs were sorted according to the repeat unit length, and the same analysis as in Fig. 5 was performed. For the data on nucleosome positions, only the chemical cleavage-based data (Chereji et al. 2018) were used in this analysis. 'R5'–'R17', repeat unit lengths of 5–17 bp

Wilson 1987; McMurray et al. 1991; Dai et al. 1997; Dai and Rothman-Denes 1998; Jagelská et al. 2010; Nuñez et al. 2015). Based mainly on the latter reports, the current study regarded the $R \geq 5 S \leq 8$ ($2R + S \geq 13$) IRs as those that have a "potential" for transition into cruciforms. Importantly, this does not mean that they are actually forming cruciforms in vivo or have "high" potential for cruciform formation. The level of the potential was not the point in the current study.

The focus here is what causes the low nucleosome occupancy on the type III IRs. We found that the average *nNuOc* values decrease, even to 0, according to the increase of the repeat unit length of the type III IRs (Fig. 6). This phenomenon seems to be explained in terms of the increase of deformed B-form structures or cruciform occurrence. Although the occurrence of these non-B structures may be transient, even for the larger type III IRs, it may be sufficient to exclude nucleosomes. The presence of multiple IRs or A/T-rich tracts in a small region may increase these probabilities overall (Supplementary Fig. S1). However, at present, we cannot still deny the formation of "stable" cruciforms in some cases. Some unknown effect only seen in vivo, including dynamic genetic processes that can locally generate a high density of negative supercoiling temporarily or even a simple loss of nucleosomes may be able to generate the cruciforms with short stems. Finally, we must also discuss the possibility for the formation of alternative structures. The $(ApT)_n$ tracts can also form Z-DNA structures. However, this seems to be less probable. The propensity for forming Z-DNA is in the following order: $(GpC)_n > (CpA)_n > (CGGG)_n > (ApT)_n$ (Wang et al. 1984; Shin et al. 2016). Furthermore, it is known that the $(ApT)_n$ tracts more prefer cruciform formation than Z-DNA formation (Wang et al. 1984, 2013; Sinden 1994). In summary, (transient) deformation or cruciform formation is raised as a possible mechanism underlying the low nucleosome occupancy on or around the type III IRs.

For the A-tract- or T-tract-rich type II IRs, the low nucleosome occupancy may be caused by different mechanisms. The A/T-tracts and oligo(A/T) sequences are reportedly rigid (Nelson et al. 1987; Packer et al. 2000; Suter et al. 2000), and seem to resist bending around the histone core (Iyer and Struhl 1995; Segal and Widom 2009; Struhl and Segal 2013). Indeed, a genome-scale analysis for nucleosome positions showed that these sequences are usually not incorporated into nucleosomes (Yuan et al. 2005). Furthermore, intrinsically bent DNA structures, which can either inhibit or facilitate nucleosome formation due to the 3D structure (Ohyama 2001), may also be partly relevant. These structures are formed under the following conditions: an A- or T-tract is present within the spacer region in a given IR in phase with the tracts within the two repeat regions, or A- or T-tracts accidentally occur in the flanking regions of a given IR in phase with the tracts inside the IR. In the case where the periodicity of the tract is ≥ 11 bp, an unfavorable 3D structure for nucleosome formation is generated. Indeed, such cases are sometimes found in the type II IRs focused upon here (data not shown). However, we should also note the report by Kornberg's group. They found that the nucleosome-free regions are formed and maintained by an active mechanism involving chromatin remodeling, with RSC (the most abundant member of the SWI/SNF family) recognition of T-tract-rich sequences, rather than the DNA rigidity- or conformation-based mechanism described above (Lorch et al. 2014). Considering these possibilities, several A-tract- or T-tract-originated mechanisms other than cruciform formation are likely to cause the low nucleosome occupancy on the type II IRs. Thus, the mechanistic cause for the low nucleosome occupancy seems to be essentially different between the types III and II IR sequences.

In addition to the types III and II sequences and the putative action of RSC, the dynamic migration of RNA polymerase II (pol II) may also contribute to the low nucleosome occupancy. The rapid removal of pol II reportedly causes increased nucleosome occupancy around poly(A) sites (Fan et al. 2010). Thus, the dynamic changes in the superhelical state caused by transcription, pol II migration itself, some action by the RSC, and the intrinsic properties and/or conformations of the type III and II IRs may collaborate with one another and induce the nucleosome depletion.

Similarity in the IR occurrence between *E. coli* and *S. cerevisiae*

The genomes of *E. coli* and *S. cerevisiae* have two common regions with statistically significant enrichment of IRs: one is in the close vicinity of the positions corresponding to mRNA ends (*E. coli*; Miura et al. 2018) or poly(A) sites (*S. cerevisiae*) and the other is ~25 to ~60 bp (*E. coli*; Miura et al. 2018) or ~30 to ~60 bp (*S. cerevisiae*) downstream of the start codons (Fig. 7). For the former, most of the IRs in *E. coli* seem to be used as parts of intrinsic terminators and they are GC-rich (Miura et al. 2018). In contrast, the IRs in *S. cerevisiae* seem to function as parts of the PAS signal and they are AT-rich, as described above. Thus, the *E. coli* and *S. cerevisiae* IRs both seem to function at the RNA level in each transcription termination system, although their nucleotide compositions are quite different. The differences in the DNA sequences may originate from the absence or presence of chromatin structure. In the case of *S. cerevisiae*, the IRs are also used to decrease nucleosome occupancy at the DNA level and for this purpose, A- or T-tract-rich, or $(ApT)_n$ or $(TpA)_n$-rich IRs are favorable, as described above.

For the regions in ORFs, the similarity between the two organisms also alludes to the presence of some common role of the IRs, which is presumably played at the RNA level. Furthermore, it is notable that the IRs with

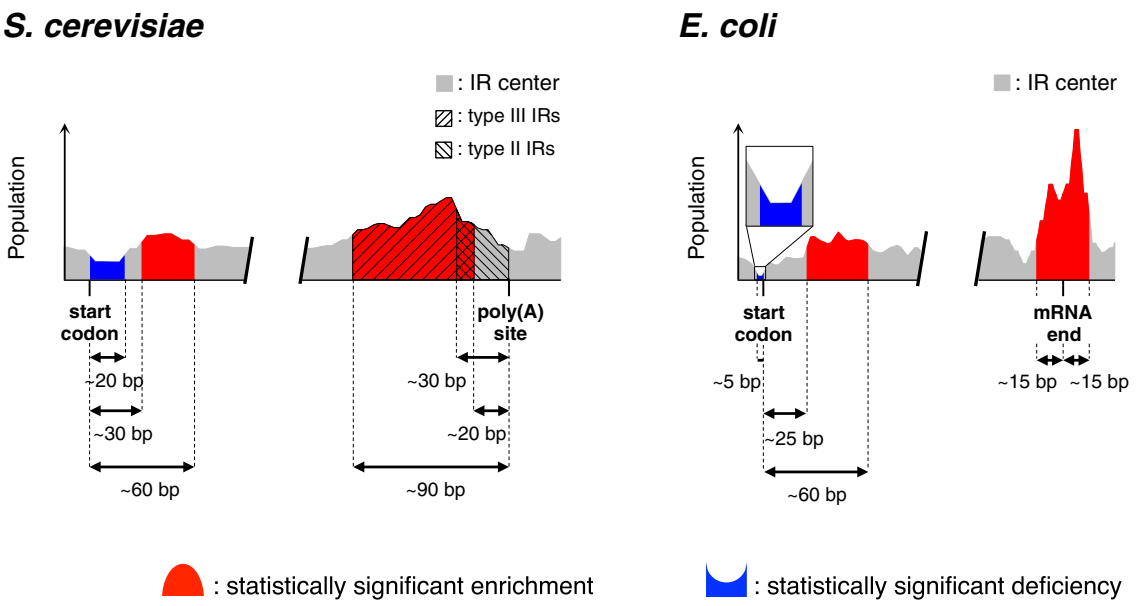


Fig. 7 Similarity in the IR occurrence between *E. coli* and *S. cerevisiae*. The illustration for *E. coli* is based on Miura et al. (2018) and that for yeast is based on the data shown in Fig. 2 ('3′-UTR' and 'ORF' panels). The 'mRNA end' in the *E. coli* illustration indicates the experimentally determined position, and the actual end position seems to be located farther downstream relative to the shown IR peak (Miura et al. 2018)

cruciform-forming potential are actively excluded in the translation initiation regions, not only in *S. cerevisiae* but also in *E. coli* (Miura et al. 2018). From this viewpoint, we can safely conclude that the IRs presumably play similar roles in the prokaryote *E. coli* and the lower eukaryote *S. cerevisiae* to regulate or complete transcription at the RNA level.

Acknowledgements This work was supported in part by JSPS KAKENHI Grant number 26440010 to T.O.

References

Bacolla A, Tainer JA, Vasquez KM, Cooper DN (2016) Translocation and deletion breakpoints in cancer genomes are associated with potential non-B DNA-forming sequences. Nucleic Acids Res 44:5673–5688. https://doi.org/10.1093/nar/gkw261

Battistoni A, Leoni L, Sampaolese B, Savino M (1988) Kinetic persistence of cruciform structures in reconstituted minichromosomes. Biochim Biophys Acta 950:161–171. https://doi.org/10.1016/0167-4781(88)90008-5

Benham CJ, Savitt AG, Bauer WR (2002) Extrusion of an imperfect palindrome to a cruciform in superhelical DNA: complete determination of energetics using a statistical mechanical model. J Mol Biol 316:563–581. https://doi.org/10.1006/jmbi.2001.5361

Brázda V, Laister RC, Jagelská EB, Arrowsmith C (2011) Cruciform structures are a common DNA feature important for regulating biological processes. BMC Mol Biol 12:33. https://doi.org/10.1186/1471-2199-12-33

Brázda V, Coufal J, Liao JC, Arrowsmith CH (2012) Preferential binding of IFI16 protein to cruciform structure and superhelical DNA. Biochem Biophys Res Commun 422:716–720. https://doi.org/10.1016/j.bbrc.2012.05.065

Brogaard K, Xi L, Wang JP, Widom J (2012) A map of nucleosome positions in yeast at base-pair resolution. Nature 486:496–501. https://doi.org/10.1038/nature11142

Calladine C, Drew H, Luisi B, Travers A (2004) Understanding DNA: the molecule and how it works, 3rd edn. Academic Press, San Diego

Cer RZ, Donohue DE, Mudunuri US, Temiz NA, Loss MA, Starner NJ, Halusa GN, Volfovsky N, Yi M, Luke BT et al (2013) Non-B DB v2.0: a database of predicted non-B DNA-forming motifs and its associated tools. Nucleic Acids Res 41:D94–D100. https://doi.org/10.1093/nar/gks955

Chen W, Liu Y, Zhu S, Green CD, Wei G, Han JD (2014) Improved nucleosome-positioning algorithm iNPS for accurate nucleosome positioning from sequencing data. Nat Commun 5:4909. https://doi.org/10.1038/ncomms5909

Chereji RV, Ramachandran S, Bryson TD, Henikoff S (2018) Precise genome-wide mapping of single nucleosomes and linkers in vivo. Genome Biol 19:19. https://doi.org/10.1186/s13059-018-1398-0

Cole HA, Howard BH, Clark DJ (2012) Genome-wide mapping of nucleosomes in yeast using paired-end sequencing. Methods

Enzymol 513:145–168. https://doi.org/10.1016/B978-0-12-391938-0.00006-9
Coufal J, Jagelská EB, Liao JC, Brázda V (2013) Preferential binding of p53 tumor suppressor to p21 promoter sites that contain inverted repeats capable of forming cruciform structure. Biochem Biophys Res Commun 441:83–88. https://doi.org/10.1016/j.bbrc.2013.10.015
Courey AJ, Wang JC (1988) Influence of DNA sequence and supercoiling on the process of cruciform formation. J Mol Biol 202:35–43. https://doi.org/10.1016/0022-2836(88)90516-5
Dai X, Rothman-Denes LB (1998) Sequence and DNA structural determinants of N4 virion RNA polymerase-promoter recognition. Genes Dev 12:2782–2790. https://doi.org/10.1101/gad.12.17.2782
Dai X, Greizerstein MB, Nadas-Chinni K, Rothman-Denes LB (1997) Supercoil-induced extrusion of a regulatory DNA hairpin. Proc Natl Acad Sci USA 94:2174–2179. https://doi.org/10.1073/pnas.94.6.2174
Damas J, Carneiro J, Gonçalves J, Stewart JB, Samuels DC, Amorim A, Pereira F (2012) Mitochondrial DNA deletions are associated with non-B DNA conformations. Nucleic Acids Res 40:7606–7621. https://doi.org/10.1093/nar/gks500
Dayn A, Malkhosyan S, Duzhy D, Lyamichev V, Panchenko Y, Mirkin SM (1991) Formation of $(dA\text{-}dT)_n$ cruciforms in *Escherichia coli* cells under different environmental conditions. J Bacteriol 173:2658–2664. https://doi.org/10.1128/jb.173.8.2658-2664.1991
Dayn A, Malkhosyan S, Mirkin SM (1992) Transcriptionally driven cruciform formation in vivo. Nucleic Acids Res 20:5991–5997. https://doi.org/10.1093/nar/20.22.5991
Du X, Wojtowicz D, Bowers AA, Levens D, Benham CJ, Przytycka TM (2013) The genome-wide distribution of non-B DNA motifs is shaped by operon structure and suggests the transcriptional importance of non-B DNA structures in *Escherichia coli*. Nucleic Acids Res 41:5965–5977. https://doi.org/10.1093/nar/gkt308
Fan X, Moqtaderi Z, Jin Y, Zhang Y, Liu XS, Struhl K (2010) Nucleosome depletion at yeast terminators is not intrinsic and can occur by a transcriptional mechanism linked to 3′-end formation. Proc Natl Acad Sci USA 107:17945–17950. https://doi.org/10.1073/pnas.1012674107
Fukue Y, Sumida N, Nishikawa J, Ohyama T (2004) Core promoter elements of eukaryotic genes have a highly distinctive mechanical property. Nucleic Acids Res 32:5834–5840. https://doi.org/10.1093/nar/gkh905
Fukue Y, Sumida N, Tanase J, Ohyama T (2005) A highly distinctive mechanical property found in the majority of human promoters and its transcriptional relevance. Nucleic Acids Res 33:3821–3827. https://doi.org/10.1093/nar/gki700
Furlong JC, Lilley DMJ (1986) Highly selective chemical modification of cruciform loops by diethyl pyrocarbonate. Nucleic Acids Res 14:3995–4007. https://doi.org/10.1093/nar/14.10.3995
Gentry M, Hennig L (2016) A structural bisulfite assay to identify DNA cruciforms. Mol Plant 9:1328–1336. https://doi.org/10.1016/j.molp.2016.06.003
Gough GW, Sullivan KM, Lilley DMJ (1986) The structure of cruciforms in supercoiled DNA: probing the single-stranded character of nucleotide bases with bisulphite. EMBO J 5:191–196. https://doi.org/10.1002/j.1460-2075.1986.tb04195.x
Guo Z, Sherman F (1996) 3′-end-forming signals of yeast mRNA. Trends Biochem Sci 21:477–481. https://doi.org/10.1016/S0968-0004(96)10057-8
Henikoff S, Ramachandran S, Krassovsky K, Bryson TD, Codomo CA, Brogaard K, Widom J, Wang J-P, Henikoff JG (2014) The budding yeast Centromere DNA Element II wraps a stable Cse4 hemisome in either orientation in vivo. Elife 3:e01861. https://doi.org/10.7554/eLife.01861
Herbert A, Schade M, Lowenhaupt K, Alfken J, Schwartz T, Shlyakhtenko LS, Lyubchenko YL, Rich A (1998) The Zα domain from human ADAR1 binds to the Z-DNA conformer of many different sequences. Nucleic Acids Res 26:3486–3493. https://doi.org/10.1093/nar/26.15.3486
Hilbers CW, Haasnoot CA, de Bruin SH, Joordens JJ, van der Marel GA, van Boom JH (1985) Hairpin formation in synthetic oligonucleotides. Biochimie 67:685–695. https://doi.org/10.1016/S0300-9084(85)80156-5
Horwitz MS, Loeb LA (1988) An *E. coli* promoter that regulates transcription by DNA superhelix-induced cruciform extrusion. Science 241:703–705. https://doi.org/10.1126/science.2456617
Hu Z, Chen K, Xia Z, Chavez M, Pal S, Seol JH, Chen CC, Li W, Tyler JK (2014) Nucleosome loss leads to global transcriptional up-regulation and genomic instability during yeast aging. Genes Dev 28:396–408. https://doi.org/10.1101/gad.233221.113
Iacono-Connors L, Kowalski D (1986) Altered DNA conformations in the gene regulatory region of torsionally-stressed SV40 DNA. Nucleic Acids Res 14:8949–8962. https://doi.org/10.1093/nar/14.22.8949
Inagaki H, Ohye T, Kogo H, Tsutsumi M, Kato T, Tong M, Emanuel BS, Kurahashi H (2013) Two sequential cleavage reactions on cruciform DNA structures cause palindrome-mediated chromosomal translocations. Nat Commun 4:1592. https://doi.org/10.1038/ncomms2595
Iyer V, Struhl K (1995) Poly(dA:dT), a ubiquitous promoter element that stimulates transcription via its intrinsic DNA structure. EMBO J 14:2570–2579. https://doi.org/10.1002/j.1460-2075.1995.tb07255.x
Jagelská EB, Pivoňková H, Fojta M, Brázda V (2010) The potential of the cruciform structure formation as an important factor influencing p53 sequence-specific binding to natural DNA targets. Biochem Biophys Res Commun 391:1409–1414. https://doi.org/10.1016/j.bbrc.2009.12.076
Jain A, Wang G, Vasquez KM (2008) DNA triple helices: biological consequences and therapeutic potential. Biochimie 90:1117–1130. https://doi.org/10.1016/j.biochi.2008.02.011
Javadekar SM, Raghavan SC (2015) Snaps and mends: DNA breaks and chromosomal translocations. FEBS J 282:2627–2645. https://doi.org/10.1111/febs.13311
Kamiya H, Fukunaga S, Ohyama T, Harashima H (2007) The location of the left-handedly curved DNA sequence affects exogenous DNA expression in vivo. Arch Biochem Biophys 461:7–12. https://doi.org/10.1016/j.abb.2007.02.012
Kaplan N, Moore IK, Fondufe-Mittendorf Y, Gossett AJ, Tillo D, Field Y, LeProust EM, Hughes TR, Lieb JD, Widom J et al (2009) The DNA-encoded nucleosome organization of a eukaryotic genome. Nature 458:362–366. https://doi.org/10.1038/nature07667
Kimura H, Shimooka Y, Nishikawa J, Miura O, Sugiyama S, Yamada S, Ohyama T (2013) The genome folding mechanism in yeast. J Biochem 154:137–147. https://doi.org/10.1093/jb/mvt033
Kouzine F, Levens D (2007) Supercoil-driven DNA structures regulate genetic transactions. Front Biosci 12:4409–4423. https://doi.org/10.2741/2398
Krasilnikov AS, Podtelezhnikov A, Vologodskii A, Mirkin SM (1999) Large-scale effects of transcriptional DNA supercoiling in vivo. J Mol Biol 292:1149–1160. https://doi.org/10.1006/jmbi.1999.3117
Krzywinski M, Schein J, Birol I, Connors J, Gascoyne R, Horsman D, Jones SJ, Marra MA (2009) Circos: an information aesthetic for comparative genomics. Genome Res 19:1639–1645. https://doi.org/10.1101/gr.092759.109
Lilley DMJ (1980) The inverted repeat as a recognizable structural feature in supercoiled DNA molecules. Proc Natl Acad Sci USA 77:6468–6472. https://doi.org/10.1073/pnas.77.11.6468

Lilley DMJ, Markham AF (1983) Dynamics of cruciform extrusion in supercoiled DNA: use of a synthetic inverted repeat to study conformational populations. EMBO J 2:527–533. https://doi.org/10.1002/j.1460-2075.1983.tb01458.x

Lin CT, Lyu YL, Liu LF (1997) A cruciform-dumbbell model for inverted dimer formation mediated by inverted repeats. Nucleic Acids Res 25:3009–3016. https://doi.org/10.1093/nar/25.15.3009

Liu R, Liu H, Chen X, Kirby M, Brown PO, Zhao K (2001) Regulation of CSF1 promoter by the SWI/SNF-like BAF complex. Cell 106:309–318. https://doi.org/10.1016/S0092-8674(01)00446-9

Lobachev KS, Gordenin DA, Resnick MA (2002) The Mre11 complex is required for repair of hairpin-capped double-strand breaks and prevention of chromosome rearrangements. Cell 108:183–193. https://doi.org/10.1016/S0092-8674(02)00614-1

Lorch Y, Maier-Davis B, Kornberg RD (2014) Role of DNA sequence in chromatin remodeling and the formation of nucleosome-free regions. Genes Dev 28:2492–2497. https://doi.org/10.1101/gad.250704.114

Lu S, Wang G, Bacolla A, Zhao J, Spitser S, Vasquez KM (2015) Short inverted repeats are hotspots for genetic instability: relevance to cancer genomes. Cell Rep 10:1674–1680. https://doi.org/10.1016/j.celrep.2015.02.039

McClellan JA, Palecek E, Lilley DM (1986) $(A\text{-}T)_n$ tracts embedded in random sequence DNA—formation of a structure which is chemically reactive and torsionally deformable. Nucleic Acids Res 14:9291–9309. https://doi.org/10.1093/nar/14.23.9291

McClellan JA, Boublíková P, Paleček E, Lilley DMJ (1990) Superhelical torsion in cellular DNA responds directly to environmental and genetic factors. Proc Natl Acad Sci USA 87:8373–8377. https://doi.org/10.1073/pnas.87.21.8373

McMurray CT, Wilson WD, Douglass JO (1991) Hairpin formation within the enhancer region of the human enkephalin gene. Proc Natl Acad Sci USA 88:666–670. https://doi.org/10.1073/pnas.88.2.666

Merino E, Yanofsky C (2005) Transcription attenuation: a highly conserved regulatory strategy used by bacteria. Trends Genet 21:260–264. https://doi.org/10.1016/j.tig.2005.03.002

Millman A, Dar D, Shamir M, Sorek R (2017) Computational prediction of regulatory, premature transcription termination in bacteria. Nucleic Acids Res 45:886–893. https://doi.org/10.1093/nar/gkw749

Mischo HE, Proudfoot NJ (2013) Disengaging polymerase: terminating RNA polymerase II transcription in budding yeast. Biochim Biophys Acta 1829:174–185. https://doi.org/10.1016/j.bbagrm.2012.10.003

Miura O, Ogake T, Ohyama T (2018) Requirement or exclusion of inverted repeat sequences with cruciform-forming potential in *Escherichia coli* revealed by genome-wide analyses. Curr Genet 64:945–958. https://doi.org/10.1007/s00294-018-0815-y

Müller UR, Wilson CL (1987) The effect of supercoil and temperature on the recognition of palindromic and non-palindromic regions in phi X174 replicative form DNA by S1 and *Bal*31. J Biol Chem 262:3730–3738

Nag DK, Petes TD (1991) Seven-base-pair inverted repeats in DNA form stable hairpins in vivo in *Saccharomyces cerevisiae*. Genetics 129:669–673. https://doi.org/10.1016/0168-9525(92)90341-Z

Nelson HC, Finch JT, Luisi BF, Klug A (1987) The structure of an oligo(dA)·oligo(dT) tract and its biological implications. Nature 330:221–226. https://doi.org/10.1038/330221a0

Nickol J, Martin RG (1983) DNA stem-loop structures bind poorly to histone octamer cores. Proc Natl Acad Sci USA 80:4669–4673. https://doi.org/10.1073/pnas.80.15.4669

Nishikawa J, Ohyama T (2013) Selective association between nucleosomes with identical DNA sequences. Nucleic Acids Res 41:1544–1554. https://doi.org/10.1093/nar/gks1269

Nobile C, Nickol J, Martin RG (1986) Nucleosome phasing on a DNA fragment from the replication origin of simian virus 40 and rephasing upon cruciform formation of the DNA. Mol Cell Biol 6:2916–2922. https://doi.org/10.1128/MCB.6.8.2916

Nuñez JK, Lee AS, Engelman A, Doudna JA (2015) Integrase-mediated spacer acquisition during CRISPR-Cas adaptive immunity. Nature 519:193–198. https://doi.org/10.1038/nature14237

Ocampo J, Chereji RV, Eriksson PR, Clark DJ (2016) The ISW1 and CHD1 ATP-dependent chromatin remodelers compete to set nucleosome spacing in vivo. Nucleic Acids Res 44:4625–4635. https://doi.org/10.1093/nar/gkw068

Ohyama T (2001) Intrinsic DNA bends: an organizer of local chromatin structure for transcription. Bioessays 23:708–715. https://doi.org/10.1002/bies.1100

Oussatcheva EA, Pavlicek J, Sankey OF, Sinden RR, Lyubchenko YL, Potaman VN (2004) Influence of global DNA topology on cruciform formation in supercoiled DNA. J Mol Biol 338:735–743. https://doi.org/10.1016/j.jmb.2004.02.075

Packer MJ, Dauncey MP, Hunter CA (2000) Sequence-dependent DNA structure: dinucleotide conformational maps. J Mol Biol 295:71–83. https://doi.org/10.1006/jmbi.1999.3236

Paeschke K, Simonsson T, Postberg J, Rhodes D, Lipps HJ (2005) Telomere end-binding proteins control the formation of G-quadruplex DNA structures in vivo. Nat Struct Mol Biol 12:847–854. https://doi.org/10.1038/nsmb982

Paleček E (1991) Local supercoil-stabilized DNA structures. Crit Rev Biochem Mol Biol 26:151–226. https://doi.org/10.3109/10409239109081126

Pan J, Sasaki M, Kniewel R, Murakami H, Blitzblau HG, Tischfield SE, Zhu X, Neale MJ, Jasin M, Socci ND et al (2011) A hierarchical combination of factors shapes the genome-wide topography of yeast meiotic recombination initiation. Cell 144:719–731. https://doi.org/10.1016/j.cell.2011.02.009

Panayotatos N, Fontaine A (1987) A native cruciform DNA structure probed in bacteria by recombinant T7 endonuclease. J Biol Chem 262:11364–11368

Park D, Morris AR, Battenhouse A, Iyer VR (2014) Simultaneous mapping of transcript ends at single-nucleotide resolution and identification of widespread promoter-associated non-coding RNA governed by TATA elements. Nucleic Acids Res 42:3736–3749. https://doi.org/10.1093/nar/gkt1366

Pearson CE, Zorbas H, Price GB, Zannis-Hadjopoulos M (1996) Inverted repeats, stem-loops, and cruciforms: significance for initiation of DNA replication. J Cell Biochem 63:1–22. https://doi.org/10.1002/(SICI)1097-4644(19961)63:1%3C1::AID-JCB1%3E3.0.CO;2-3

Potaman VN, Sinden RR (2005) DNA: alternative conformations and biology. In: Ohyama T (ed) DNA conformation and transcription. Springer, New York, pp 3–17. https://doi.org/10.1007/0-387-29148-2_1

Proudfoot NJ (2011) Ending the message: poly(A) signals then and now. Genes Dev 25:1770–1782. https://doi.org/10.1101/gad.17268411

Qian Z, Adhya S (2017) DNA repeat sequences: diversity and versatility of functions. Curr Genet 63:411–416. https://doi.org/10.1007/s00294-016-0654-7

Qin Y, Hurley LH (2008) Structures, folding patterns, and functions of intramolecular DNA G-quadruplexes found in eukaryotic promoter regions. Biochimie 90:1149–1171. https://doi.org/10.1016/j.biochi.2008.02.020

Ramachandran S, Zentner GE, Henikoff S (2015) Asymmetric nucleosomes flank promoters in the budding yeast genome. Genome Res 25:381–390. https://doi.org/10.1101/gr.182618.114

Scholten PM, Nordheim A (1986) Diethyl pyrocarbonate: a chemical probe for DNA cruciforms. Nucleic Acids Res 14:3981–3993. https://doi.org/10.1093/nar/14.10.3981

Schroth GP, Chou PJ, Ho PS (1992) 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 267:11846–11855

Segal E, Widom J (2009) Poly(dA:dT) tracts: major determinants of nucleosome organization. Curr Opin Struct Biol 19:65–71. https://doi.org/10.1016/j.sbi.2009.01.004

Sheflin LG, Kowalski D (1985) Altered DNA conformations detected by mung bean nuclease occur in promoter and terminator regions of supercoiled pBR322 DNA. Nucleic Acids Res 13:6137–6154. https://doi.org/10.1093/nar/13.17.6137

Shin SI, Ham S, Park J, Seo SH, Lim CH, Jeon H, Huh J, Roh TY (2016) Z-DNA-forming sites identified by ChIP-Seq are associated with actively transcribed regions in the human genome. DNA Res 23:477–486. https://doi.org/10.1093/dnares/dsw031

Shlyakhtenko LS, Potaman VN, Sinden RR, Lyubchenko YL (1998) Structure and dynamics of supercoil-stabilized DNA cruciforms. J Mol Biol 280:61–72. https://doi.org/10.1006/jmbi.1998.1855

Shlyakhtenko LS, Hsieh P, Grigoriev M, Potaman VN, Sinden RR, Lyubchenko YL (2000) A cruciform structural transition provides a molecular switch for chromosome structure and dynamics. J Mol Biol 296:1169–1173. https://doi.org/10.1006/jmbi.2000.3542

Sinden RR (1994) DNA structure and function. Academic Press, San Diego. https://doi.org/10.1016/C2009-0-02451-9

Strawbridge EM, Benson G, Gelfand Y, Benham CJ (2010) The distribution of inverted repeat sequences in the *Saccharomyces cerevisiae* genome. Curr Genet 56:321–340. https://doi.org/10.1007/s00294-010-0302-6

Struhl K, Segal E (2013) Determinants of nucleosome positioning. Nat Struct Mol Biol 20:267–273. https://doi.org/10.1038/nsmb.2506

Sumida N, Nishikawa J, Kishi H, Amano M, Furuya T, Sonobe H, Ohyama T (2006) A designed curved DNA segment that is a remarkable activator of eukaryotic transcription. FEBS J 273:5691–5702. https://doi.org/10.1111/j.1742-4658.2006.05557.x

Sun D, Hurley LH (2009) The importance of negative superhelicity in inducing the formation of G-quadruplex and i-motif structures in the c-Myc promoter: implications for drug targeting and control of gene expression. J Med Chem 52:2863–2874. https://doi.org/10.1021/jm900055s

Suter B, Schnappauf G, Thoma F (2000) Poly(dA·dT) sequences exist as rigid DNA structures in nucleosome-free yeast promoters in vivo. Nucleic Acids Res 28:4083–4089. https://doi.org/10.1093/nar/28.21.4083

Tirosh I, Sigal N, Barkai N (2010) Divergence of nucleosome positioning between two closely related yeast species: genetic basis and functional consequences. Mol Syst Biol 6:365. https://doi.org/10.1038/msb.2010.20

Tsui K, Dubuis S, Gebbia M, Morse RH, Barkai N, Tirosh I, Nislow C (2011) Evolution of nucleosome occupancy: conservation of global properties and divergence of gene-specific patterns. Mol Cell Biol 31:4348–4355. https://doi.org/10.1128/MCB.05276-11

Van Holde K, Zlatanova J (1994) Unusual DNA structures, chromatin and transcription. Bioessays 16:59–68. https://doi.org/10.1002/bies.950160110

Van Noort V, Worning P, Ussery DW, Rosche WA, Sinden RR (2003) Strand misalignments lead to quasipalindrome correction. Trends Genet 19:365–369. https://doi.org/10.1016/S0168-9525(03)00136-7

Vologodskaia MY, Vologodskii AV (1999) Effect of magnesium on cruciform extrusion in supercoiled DNA. J Mol Biol 289:851–859. https://doi.org/10.1006/jmbi.1999.2811

Vologodskii A (2015) Biophysics of DNA. Cambridge University Press, Cambridge. https://doi.org/10.1017/CBO9781139542371

Wachter A (2014) Gene regulation by structured mRNA elements. Trends Genet 30:172–181. https://doi.org/10.1016/j.tig.2014.03.001

Wang Y, Leung FC (2006) Long inverted repeats in eukaryotic genomes: recombinogenic motifs determine genomic plasticity. FEBS Lett 580:1277–1284. https://doi.org/10.1016/j.febslet.2006.01.045

Wang Y, Leung FC (2009) A study on genomic distribution and sequence features of human long inverted repeats reveals species-specific intronic inverted repeats. FEBS J 276:1986–1998. https://doi.org/10.1111/j.1742-4658.2009.06930.x

Wang AH, Hakoshima T, van der Marel G, van Boom JH, Rich A (1984) AT base pairs are less stable than GC base pairs in Z-DNA: the crystal structure of d(m^5CGTAm^5CG). Cell 37:321–331. https://doi.org/10.1016/0092-8674(84)90328-3

Wang G, Gaddis S, Vasquez KM (2013) Methods to detect replication-dependent and replication-independent DNA structure-induced genetic instability. Methods 64:67–72. https://doi.org/10.1016/j.ymeth.2013.08.004

Warburton PE, Giordano J, Cheung F, Gelfand Y, Benson G (2004) 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 14:1861–1869. https://doi.org/10.1101/gr.2542904

Wells RD, Harvey SC (eds) (1987) Unusual DNA structures. Springer, New York. https://doi.org/10.1007/978-1-4612-3800-3

Yuan GC, Liu YJ, Dion MF, Slack MD, Wu LF, Altschuler SJ, Rando OJ (2005) Genome-scale identification of nucleosome positions in *S. cerevisiae*. Science 309:626–630. https://doi.org/10.1126/science.1112178

Zannis-Hadjopoulos M, Yahyaoui W, Callejo M (2008) 14-3-3 cruciform-binding proteins as regulators of eukaryotic DNA replication. Trends Biochem Sci 33:44–50. https://doi.org/10.1016/j.tibs.2007.09.012

Zhabinskaya D, Benham CJ (2013) Competitive superhelical transitions involving cruciform extrusion. Nucleic Acids Res 41:9610–9621. https://doi.org/10.1093/nar/gkt733

Zhao J, Hyman L, Moore C (1999) Formation of mRNA 3′ ends in eukaryotes: mechanism, regulation, and interrelationships with other steps in mRNA synthesis. Microbiol Mol Biol Rev 63:405–445