Journal Pre-proofs

Research paper

New telomere to telomere assembly of human chromosome 8 reveals a previous underestimation of G-quadruplex forming sequences and inverted repeats

Vaclav Brazda, Natalia Bohalova, Richard P. Bowater

PII: S0378-1119(21)00653-3
DOI: https://doi.org/10.1016/j.gene.2021.146058
Reference: GENE 146058

To appear in: *Gene Gene*

Received Date: 31 August 2021
Revised Date: 14 October 2021
Accepted Date: 29 October 2021

Please cite this article as: V. Brazda, N. Bohalova, R.P. Bowater, New telomere to telomere assembly of human chromosome 8 reveals a previous underestimation of G-quadruplex forming sequences and inverted repeats, *Gene Gene* (2021), doi: https://doi.org/10.1016/j.gene.2021.146058

This is a PDF file of an article that has undergone enhancements after acceptance, such as the addition of a cover page and metadata, and formatting for readability, but it is not yet the definitive version of record. This version will undergo additional copyediting, typesetting and review before it is published in its final form, but we are providing this version to give early visibility of the article. Please note that, during the production process, errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

© 2021 Elsevier B.V. All rights reserved.

New telomere to telomere assembly of human chromosome 8 reveals a previous underestimation of G-quadruplex forming sequences and inverted repeats

Vaclav Brazda[a]*, Natalia Bohalova [a,b], and Richard P. Bowater [c]*,

[a]Institute of Biophysics of the Czech Academy of Sciences, Královopolská 135, 612 65 Brno, Czech Republic;

[b]Department of Experimental Biology, Faculty of Science, Masaryk University, Kamenice 5, 62500, Brno, Czech Republic;

[c] School of Biological Sciences, University of East Anglia, Norwich Research Park, Norwich NR4 7TJ, United Kingdom

* Authors to whom correspondence should be addressed.

Email: vaclav@ibp.cz, R.Bowater@uea.ac.uk

Highlights

- Describes the first analyses of G-quadruplex forming sequences and inverted repeats in the complete telomere to telomere assembly of human chromosome 8
- The complete chromosome sequence contains significantly more inverted repeats and G-quadruplex forming sequences than was previously identified in the broadly used human genome reference assembly 38
- The prevalence of potential non-B DNA secondary structures in the human genome has not been fully appreciated and we anticipate that similar observations will be made as the full human genome sequence is completed

Keywords: G-quadruplex; Inverted repeat; Genome sequence of human chromosome 8; Non-B DNA structures

New telomere to telomere assembly of human chromosome 8 reveals a previous underestimation of G-quadruplex forming sequences and inverted repeats

Vaclav Brazda[a]*, Natalia Bohalova [a,b], and Richard P. Bowater [c]*,

[a]Institute of Biophysics of the Czech Academy of Sciences, Královopolská 135, 612 65 Brno, Czech Republic;

[b]Department of Experimental Biology, Faculty of Science, Masaryk University, Kamenice 5, 62500, Brno, Czech Republic;

[c] School of Biological Sciences, University of East Anglia, Norwich Research Park, Norwich NR4 7TJ, United Kingdom

* Authors to whom correspondence should be addressed.

Email: vaclav@ibp.cz, R.Bowater@uea.ac.uk

Highlights

- We describe the first analyses of G-quadruplex forming sequences and inverted repeats in the complete telomere to telomere assembly of human chromosome 8
- The complete chromosome sequence contains significantly more inverted repeats and G-quadruplex forming sequences than was previously identified in the broadly used human genome reference assembly 38
- The prevalence of potential non-B DNA secondary structures in the human genome has not been fully appreciated and we anticipate that similar observations will be made as the full human genome sequence is completed

Keywords: G-quadruplex; Inverted repeat; Genome sequence of human chromosome 8; Non-B DNA structures

New telomere to telomere assembly of human chromosome 8 reveals a previous underestimation of G-quadruplex forming sequences and inverted repeats

Vaclav Brazda[a*], Natalia Bohalova [a,b], and Richard P. Bowater [c*],

Abstract

Taking advantage of evolving and improving sequencing methods, human chromosome 8 is now available as a gapless, end-to-end assembly. Thanks to advances in long-read sequencing technologies, its centromere, telomeres, duplicated gene families and repeat-rich regions are now fully sequenced. We were interested to assess if the new assembly altered our understanding of the

Journal Pre-proofs

potential impact of non-B DNA structures within this completed chromosome sequence. It has been shown that non-B secondary structures, such as G-quadruplexes, hairpins and cruciforms, have important regulatory functions and potential as targeted therapeutics. Therefore, we analysed the presence of putative G-quadruplex forming sequences and inverted repeats in the current human reference genome (GRCh38) and in the new end-to-end assembly of chromosome 8. The comparison revealed that the new assembly contains significantly more inverted repeats and G-quadruplex forming sequences compared to the current reference sequence. This observation can be explained by improved accuracy of the new sequencing methods, particularly in regions that contain extensive repeats of bases, as is preferred by many non-B DNA structures. These results show a significant underestimation of the prevalence of non-B DNA secondary structure in previous assembly versions of the human genome and point to their importance being not fully appreciated. We anticipate that similar observations will occur as the improved sequencing technologies fill in gaps across the genomes of humans and other organisms.

New telomere to telomere assembly of human chromosome 8 reveals a previous underestimation of G-quadruplex forming sequences and inverted repeats

Vaclav Brazda[a]*, Natalia Bohalova [a,b], and Richard P. Bowater [c]*,

[a]Institute of Biophysics of the Czech Academy of Sciences, Královopolská 135, 612 65 Brno, Czech Republic;

[b]Department of Experimental Biology, Faculty of Science, Masaryk University, Kamenice 5, 62500, Brno, Czech Republic;

[c] School of Biological Sciences, University of East Anglia, Norwich Research Park, Norwich NR4 7TJ, United Kingdom

* Authors to whom correspondence should be addressed.

Email: vaclav@ibp.cz, R.Bowater@uea.ac.uk

Highlights

- We describe the first analyses of G-quadruplex forming sequences and inverted repeats in the complete telomere to telomere assembly of human chromosome 8

- The complete chromosome sequence contains significantly more inverted repeats and G-quadruplex forming sequences than was previously identified in the broadly used human genome reference assembly 38
- The prevalence of potential non-B DNA secondary structures in the human genome has not been fully appreciated and we anticipate that similar observations will be made as the full human genome sequence is completed

Keywords: G-quadruplex; Inverted repeat; Genome sequence of human chromosome 8; Non-B DNA structures

Journal Pre-proofs

Abstract

Taking advantage of evolving and improving sequencing methods, human chromosome 8 is now available as a gapless, end-to-end assembly. Thanks to advances in long-read sequencing technologies, its centromere, telomeres, duplicated gene families and repeat-rich regions are now fully sequenced. We were interested to assess if the new assembly altered our understanding of the potential impact of non-B DNA structures within this completed chromosome sequence. It has been shown that non-B secondary structures, such as G-quadruplexes, hairpins and cruciforms, have important regulatory functions and potential as targeted therapeutics. Therefore, we analysed the presence of putative G-quadruplex forming sequences and inverted repeats in the current human reference genome (GRCh38) and in the new end-to-end assembly of chromosome 8. The comparison revealed that the new assembly contains significantly more inverted repeats and G-quadruplex forming sequences compared to the current reference sequence. This observation can be explained by improved accuracy of the new sequencing methods, particularly in regions that contain extensive repeats of bases, as is preferred by many non-B DNA structures. These results show a significant underestimation of the prevalence of non-B DNA secondary structure in previous assembly versions of the human genome and point to their importance being not fully appreciated. We anticipate that similar observations will occur as the improved sequencing technologies fill in gaps across the genomes of humans and other organisms.

Journal Pre-proofs

1. Introduction

In the current and most complete reference assembly of the human genome (GRCh38), there are no chromosomes that are fully sequenced from telomere-to-telomere, meaning there are many gaps and missing regions that are not determined at the base sequence level. Recently, using ultra-long Oxford Nanopore Technologies (ONT) and Pacific Biosciences (PacBio) high-fidelity (HiFi) data, the telomere-to-telomere sequence of human chromosome 8 was published: it is 146,259,670 bases in length, including 3,334,256 bases that are missing from GRCh38, with an estimated base accuracy in excess of 99.99% [1]. The use of such advanced sequencing technologies mean that completion of the entire human genome sequence and the associated updated knowledge about complex genome-directed regulatory pathways is close.

Whereas DNA molecules mostly adopt double-stranded, right-handed B-helical conformations, a range of alternative (non-B) structures can also occur. It has been shown that such non-B DNA structures can play crucial roles in many basic cellular processes and can exist in vivo [2–5]. Alternative structures, such as G-quadruplexes (G4) and cruciforms, can be favoured under specific environmental (and cellular) conditions due to their improved thermodynamic stability compared to B-form DNA [6,7]. G4 structures regularly arise in G-rich regions of DNA sequence, mainly because the typical building block of a G4 is a G-quartet that emerges via Hoogsteen base pairing of four guanines. Two or more G-quartets stack on top of each other and are physiologically stabilized by monovalent cations [8]. Sites of inverted repeats are able to form hairpin stem-loop secondary structures in a single-stranded section of DNA or a cruciform structure if they occur on opposing strands of a double-stranded DNA [9].

We decided to analyze the new telomere to telomere assembly of human chromosome 8 to evaluate the presence of sequences that have potential to form well characterised local non-B DNA structures, focusing on G4-forming and inverted repeat sequences. Potential G4 structures were identified by G4Hunter software, which calculates the G4Hunter score of sequences depending on the asymmetrical distribution of G and C bases on each strand [10,11]. The presence of inverted repeats was assessed by Palindrome analyser [12]. We reveal that the number of potential G4 and inverted repeats sequences is significantly increased in the new telomere to telomere assembly of human chromosome 8 compared to the same analyses of GRCh38. As we go on to highlight, these results suggest that the number of these types of regulatory sequences is strongly underestimated in the previous broadly used reference assemblies of the human genome.

2. Results

To compare the human chromosome 8 assemblies, we downloaded full chromosome sequences (NCBI ID CM000670.2 of GRCh38 human genome assembly and CP061028.1 of the telomere to telomere chromosome 8 assembly) and analyzed them. The fully assembled chromosome 8 sequence (156,259,670 bp) is 1,121,034 bp longer than the older chromosome 8 assembly within GRCh38 (145,138,636 bp), with a 1.21% increase in GC content. While the changes in total number of bases are below 1%, analyses by G4Hunter software revealed a remarkable increase in the abundance of G4-forming sequences (Table 1). In general, the higher the G4Hunter score, the more likely it is that a stable G4 will be formed. Excitingly, although the new genome assembly contained higher G4Hunter scores at all levels, the increase is most dramatic at the higher scores, which are those where there is greatest confidence that they can adopt a stable G4. For example, for G4Hunter scores up to 2.0 there is an increase of up to 10% in the G4-prone sequences in the new assembly, but for higher G4Hunter scores of 2.5 and 3 the increase is 13% and 25%, respectively. For a G4Hunter score of 3.5 there are more than twice as many G4-prone sequences in the new gapless assembly of chromosome 8 compared to GRCh38 (Table 1).

Table 1. Comparison of the prevalence of G4-forming sequences in human chromosome 8. G4Hunter web applications with default parameters (window 25, G4Hunter score indicated in the first column) were used to identify G4-forming sequences in the current reference genome assembly ("GRCh 38") and the new telomere to telomere assembly of chromosome 8 ("chr8 new"). First column: G4Hunter score; second column: number of G4-forming sequences in chromosome 8 GRCh38 assembly; third column: number of G4-forming sequences in new chromosome 8 assembly; forth column: change in number of G4-forming sequences (new assembly-GRCh38), fifth column: difference in % (the number of G4-forming sequences in CRCh 38 was taken as 100%).

G4Hunter score	GRCh 38	new assembly	Δ	Δ%
1.2	232,871	237,055	+4,184	**+1.80**
1.4	121,355	124,004	+2,649	**+2.18**
1.6	64,699	66,638	+1,939	**+3.00**
1.8	34,514	36,008	+1,494	**+4.33**
2.0	18,368	19,526	+1,158	**+6.30**
2.5	3,746	4,259	+513	**+13.69**
3.0	520	650	+130	**+25**
3.5	22	47	25	**+113.63**

An even more obvious trend was observed with inverted repeats as analyzed by Palindrome finder (Table 2). The number of shorter inverted repeats (6-9 bases as the length of each half of the repeat) is similar in both assemblies (0.73% increase in new assembly) and corresponds well to the change of the corrected

chromosome 8 length (0,77% increase). However, longer inverted repeats (with a size of > 9 bases in each half of the repeat) are much more prevalent in the newly assembled chromosome 8 sequence than expected (1.90% increase for inverted repeats 10-14). Importantly, the length of the inverted repeats are connected to their proposed biological functions, with the highest abundance of inverted repeats immediately before transcription start sites being for repeats of 8 bp or longer [13]. The difference between the two assemblies increases with the length of the inverted repeat and there are almost twice as many inverted repeats with the repeat size of 30 in the previous newly assembled sequence compared to GRCh38 (Table 2). Such long inverted repeats are more prone to be a threat for genetic instability and recombination [14,15].

Table 2. Comparison of the inverted repeats presence in human chromosome 8. Palindrome finder with default parameters (Length of one repeat unit is indicated in the first column) were used to identify inverted repeats in the current reference genome assembly ("GRCh 38") and the new telomere to telomere assembly of chromosome 8 ("chr8 new"). First column: Length of repeat unit; second column: number of inverted repeats in chromosome 8 GRCh38 assembly; third column: number of inverted repeats in new chromosome 8 assembly; fourth column: change in number of inverted repeats (new assembly-GRCh38); fifth column: difference in % (with the number of inverted repeats in GRCh 38 taken as 100%).

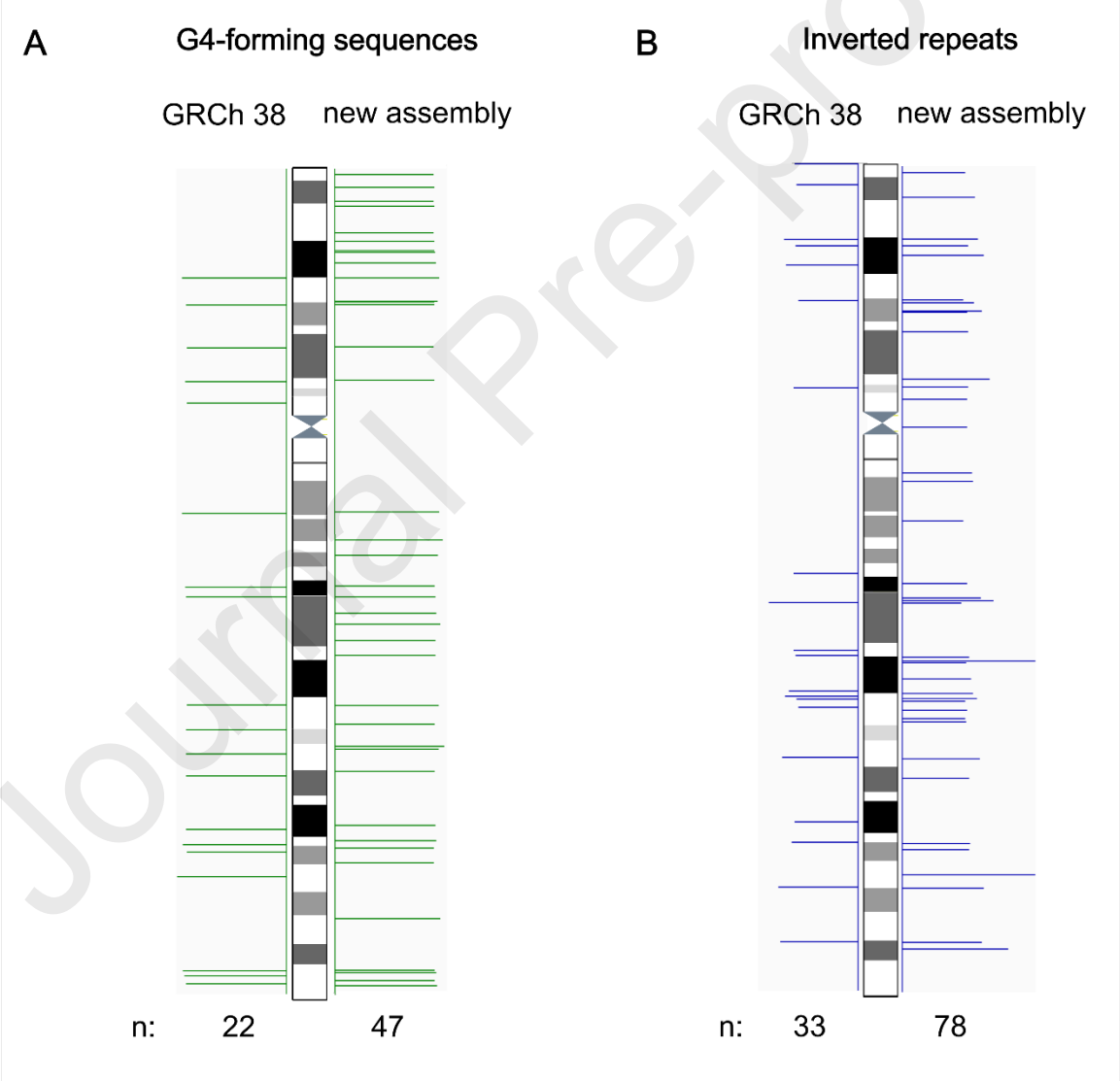
Inverted repeat size range	GRCh 38	new assembly	Δ	Δ%
6-9	5,023,663	5,060,277	+36,614	**+0.73**
10-14	114,281	116,455	+2,174	**+1.90**
15-19	5,914	6,501	+587	**+9.93**
20-24	1,670	2,081	+411	**+24.61**
25-29	526	668	+142	**+27.00**
30-44	364	528	+164	**+45.05**
Over 45	33	78	+45	**+136.36**

Then we visualize the localization of G4-prone sequences and inverted repeats according to Giemsa banding [16] of chromosome 8 (Figure 1). This highlights the non-random enrichment of these non-B DNA structures in the fully assembled human chromosome 8 and points to the previous underestimation of non-B DNA structures in the human genome. It is important to appreciate that the sequences that are highly prone to adopt G4 structures have always been present within the human genome but they were not identified in genome databases due to incomplete sequence information.

We further investigated the localization of the highest-scoring G4-prone sequences (with a G4Hunter score above 3.5) and the longest inverted repeats (over 45 bp in the repeat unit) in their genomic context (Supplementary Material 01). The nearest *gene* and *repeat* that were identified in the complete sequence

of chromosome 8 [1] were assigned to each G4-prone and inverted repeat sequence identified in this study. Interestingly, the majority (62%) of the G4-prone sequences are present upstream of *gene* regions compared to downstream of the transcription start site. By contrast, the inverted repeat sequences (longer than 45 bp in the repeat unit) are approximately equal in regions before and after the transcription start site. Moreover, all highest-scoring G4-prone sequences and the longest inverted repeats that are present within *genes* are located in introns.

Figure 1. Comparison of the localization of G4-forming sequences with a G4Hunter score threshold of 3.5 (A) and inverted repeats with an individual size for each half of the repeat to be 45 bases and higher (B) in human chromosome 8. The length of green dashes denotes the actual G4Hunter score, the blue dashes denotes the overall length of inverted repeat. For both parts, "n" reports the number of respective sequences found in each corresponding assembly.



Journal Pre-proofs

3. Discussion

Although the current reference human genome (GRCh38) is the latest example as part of the process that continually improves it compared to the initial sequencing [17], several problems still persist [18]. The new completed human chromosome 8 assembly sequence (published in April 2021) is 1.1 Mbp longer than GRCh38, but this comprehensive dataset is also more accurate about its number of open reading frames and regulatory regions [1]. Whilst it is not surprising that completion of a genome sequence will alter its overall size (either longer or shorter), the effects on its gene number and their expression are less predictable. For example, the completed genomes can have more or less genes (open reading frames), and even shorter genomic sequences may be able to express more genes due to reorganisation of the bases. Altered arrangement of the sequence can also lead to altered propensity to form local non-B DNA structures, which can have effects on the numbers of regulatory regions, such as promoters and terminators [9]. Contemporary sequencing methods with long-reads and combinations of several sequencing methods allow previously unimaginable progress in accuracy of genome assemblies. This has dramatic impacts on some types of sequences, particularly those that are repetitive in nature. Previously, many repetitive sequences were missed entirely because the sequencing methods did not allow their detection, which led to “gaps” in the genome assembly. Furthermore, even when repetitive sequences were detected by sequencing methods they were often deleted from assemblies due to the impossibility of determining their exact location in the DNA molecule.

Recent advances in DNA sequencing technologies allow determination of the complete information without any gaps and ambiguities [19]. GRCh38 is the most widely used reference assembly of the human genome, but it has 370,500 unknown (N) bases in the sequence of chromosome 8, which point to many uncertainties with its usage and it should be replaced by the fully assembled sequence in the near future. As we demonstrate, the new human chromosome 8 telomere-to-telomere assembly fixes obvious problems with the reference genome and it also brings many unexpected findings like the strong underestimation of non-B DNA structures in the human genome, as demonstrated here for chromosome 8. Local non-B DNA structures are involved in the regulation of basic molecular processes [20,21]. Bioinformatics studies have demonstrated that G4-forming sequences are enriched in promoters [20] and other regulatory regions of the human genome, such as untranslated regions (UTRs) and telomeres [22–24]. Inverted repeats have been detected in human promoters and represent targets for regulatory proteins [13,25,26] and the impact of G4 structural variations on gene activity was also shown [27]. As highlighted in Figure 2, it has been demonstrated that these non-B DNA structures play a role in replication, transcription, translation, genomic stability, and telomere

maintenance [14,28–30]. Data included in Supplementary Material 01 demonstrates that the most stable of the observed G4-prone sequences (with a G4Hunter score above 3.5) are located before genes or in introns. An important potential conclusion from this is that if such sequences form stable non-B DNA structures then they could impair transcription machinery or have a role in alternative splicing [31,32]. Moreover, the majority of these observed G4-prone and inverted repeat sequences are located within "*repeat*" regions, such as "*simple repeats*", which are known to be regulatory elements of gene expression [9]. Together, these observations warrant further investigation to the potential effects of non-B DNA structures on gene expression. An increasing number of studies have reported on the cellular significance of these non-B DNA structures, especially for G4 due to their possible utilization in antiviral and antiproliferative therapeutic targeting and because of their great stability under physiological conditions [33–35]. Diverse roles of G4 could be the result of a synergic effect with other structurally compatible non-B DNA structures, such as an R loop, which is formed by one DNA strand annealed to an RNA, thus forming a DNA:RNA hybrid [30].

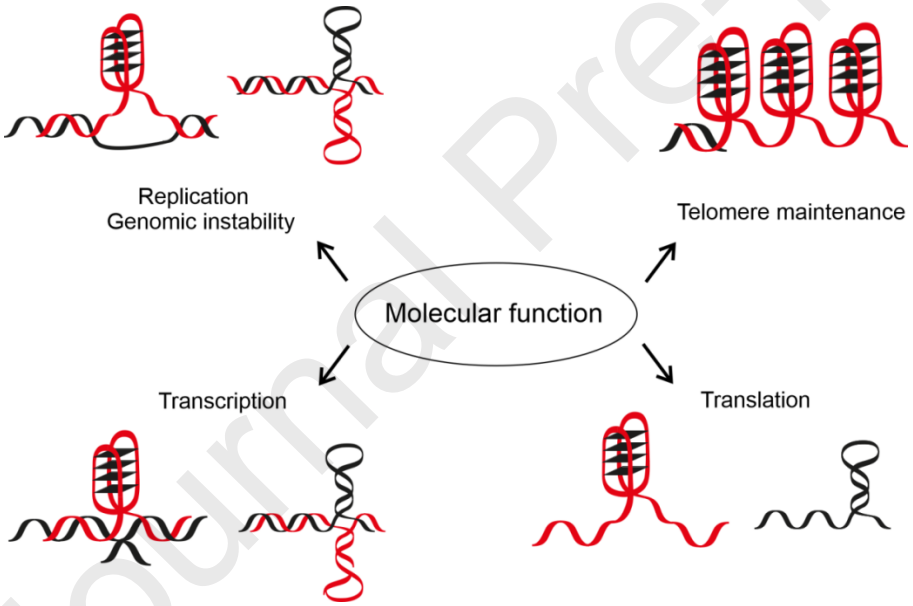


Figure 2. Non-B DNA structures in the regulation of basic molecular processes. Both G4 and cruciforms play a role in replication initiation, but could also impair replication machinery and thus be a source of genomic instability [36–39]. Stable G4s form in human telomere sequences and are involved in the protection of telomere ends [28,40]. G4 or cruciform formation in promoter regions could modulate the initiation of transcription [20]. G4 or hairpin formation in mRNA could affect mRNA stability and prematurely terminate translation [14,41].

In summary, the exciting findings from the recently published telomere to telomere assembly of chromosome 8 uncovered an increased abundance of G4-

forming sequences and inverted repeats compared to the current reference genome sequence (GRCh38). We found significantly more putative G4-forming sequences and inverted repeats in the newly assembled sequence, which strongly highlights the previous underestimation of the prevalence of potential non-B secondary structures in the human genome. Similar findings appear to be valid also for a new telomere to telomere assembly of chromosome X [42], as demonstrated for G4-prone sequences [43]. These potential regulatory sequences within newly-assembled genomes should be evaluated using methods that allow targeted detection of the non-B DNA structures, particularly within cells [44,45]. This new complete assembly of human chromosome 8 highlights the exciting findings that await as we approach the accurate reading and interpretation of the full human genome with its broad variety of regulatory sequences, including local non-B DNA structures, such as G4-forming sequences and inverted repeats.

4. Materials and Methods

To compare the human chromosome 8 assemblies, we downloaded full chromosome sequences (NCBI ID CM000670.2 of GRCh38 human genome assembly and CP061028.1 of the telomere to telomere chromosome 8 assembly) and analyzed them by Palindrome finder and G4Hunter web applications with default parameters [10,12]. *Gene* and *repeat* annotations were provided by the authors of the original paper that described the telomere to telomere assembly of chromosome 8 [1].

Acknowledgments

This work was supported by The Czech Science Foundation (18-15548S) to VB.

Competing Interest Statement

The authors declare no competing interest.

Author Contributions

V.B. designed research; N.B. and V.B. performed research; V.B., N.B. and R.P.B. analyzed data; V.B., N.B. and R.P.B. wrote the paper.

References

[1] G.A. Logsdon, M.R. Vollger, P. Hsieh, Y. Mao, M.A. Liskovykh, S. Koren, S. Nurk, L. Mercuri, P.C. Dishuck, A. Rhie, L.G. de Lima, T. Dvorkina, D. Porubsky, W.T. Harvey, A. Mikheenko, A.V. Bzikadze, M. Kremitzki, T.A. Graves-Lindsay, C. Jain, K. Hoekzema, S.C. Murali, K.M. Munson, C. Baker, M. Sorensen, A.M. Lewis, U. Surti, J.L. Gerton, V. Larionov, M. Ventura, K.H. Miga, A.M. Phillippy, E.E. Eichler, The structure, function and evolution of a complete human chromosome 8, Nature. 593 (2021) 101–107. https://doi.org/10.1038/s41586-021-03420-7.

[2] F. Kouzine, D. Wojtowicz, A. Yamane, R. Casellas, T.M. Przytycka, D.L. Levens, In Vivo Chemical Probing for G-Quadruplex Formation, Methods Mol Biol. 2035 (2019) 369–382. https://doi.org/10.1007/978-1-4939-9666-7_23.
[3] T. Nakamura, S. Okabe, H. Yoshida, K. Iida, Y. Ma, S. Sasaki, T. Yamori, K. Shin-ya, I. Nakano, K. Nagasawa, H. Seimiya, Targeting glioma stem cells in vivo by a G-quadruplex-stabilizing synthetic macrocyclic hexaoxazole, Sci Rep. 7 (2017) 3605. https://doi.org/10.1038/s41598-017-03785-8.
[4] R. Hänsel-Hertsch, A. Simeone, A. Shea, W.W.I. Hui, K.G. Zyner, G. Marsico, O.M. Rueda, A. Bruna, A. Martin, X. Zhang, S. Adhikari, D. Tannahill, C. Caldas, S. Balasubramanian, Landscape of G-quadruplex DNA structural regions in breast cancer, Nat Genet. 52 (2020) 878–883. https://doi.org/10.1038/s41588-020-0672-8.
[5] L. Poggi, G.-F. Richard, Alternative DNA Structures In Vivo: Molecular Evidence and Remaining Questions, Microbiol Mol Biol Rev. 85 (2021) e00110-20. https://doi.org/10.1128/MMBR.00110-20.
[6] R.P. Bowater, Z.A. Waller, DNA Structure, in: ELS, John Wiley & Sons, Ltd: Chichester, 2014. https://doi.org/10.1002/9780470015902.a0006002.pub2.
[7] S. Matsumoto, N. Sugimoto, New Insights into the Functions of Nucleic Acids Controlled by Cellular Microenvironments, Top Curr Chem (Cham). 379 (2021) 17. https://doi.org/10.1007/s41061-021-00329-7.
[8] R.W. Harkness, A.K. Mittermaier, G-quadruplex dynamics, Biochim Biophys Acta Proteins Proteom. 1865 (2017) 1544–1554. https://doi.org/10.1016/j.bbapap.2017.06.012.
[9] V. Brazda, M. Fojta, R.P. Bowater, Structures and stability of simple DNA repeats from bacteria, Biochem J. 477 (2020) 325–339. https://doi.org/10.1042/BCJ20190703.
[10] V. Brázda, J. Kolomazník, J. Lýsek, M. Bartas, M. Fojta, J. Šťastný, J.-L. Mergny, G4Hunter web application: a web server for G-quadruplex prediction, Bioinformatics. 35 (2019) 3493–3495. https://doi.org/10.1093/bioinformatics/btz087.
[11] A. Bedrat, L. Lacroix, J.-L. Mergny, Re-evaluation of G-quadruplex propensity with G4Hunter, Nucleic Acids Res. 44 (2016) 1746–1759. https://doi.org/10.1093/nar/gkw006.
[12] V. Brázda, J. Kolomazník, J. Lýsek, L. Hároníková, J. Coufal, J. Šťastný, Palindrome analyser - A new web-based server for predicting and evaluating inverted repeats in nucleotide sequences, Biochem. Biophys. Res. Commun. 478 (2016) 1739–1745. https://doi.org/10.1016/j.bbrc.2016.09.015.
[13] V. Brázda, M. Bartas, J. Lýsek, J. Coufal, M. Fojta, Global analysis of inverted repeat sequences in human gene promoters reveals their non-random distribution and association with specific biological pathways, Genomics. 112 (2020) 2772–2777. https://doi.org/10.1016/j.ygeno.2020.03.014.
[14] V. Brázda, R.C. Laister, E.B. Jagelská, C. Arrowsmith, Cruciform structures are a common DNA feature important for regulating biological processes, BMC Mol Biol. 12 (2011) 33. https://doi.org/10.1186/1471-2199-12-33.
[15] Y. Wang, F.C.C. Leung, Long inverted repeats in eukaryotic genomes: Recombinogenic motifs determine genomic plasticity, FEBS Letters. 580 (2006) 1277–1284. https://doi.org/10.1016/j.febslet.2006.01.045.
[16] M.R. Speicher, N.P. Carter, The new cytogenetics: blurring the boundaries with molecular biology, Nat Rev Genet. 6 (2005) 782–792. https://doi.org/10.1038/nrg1692.

[17] E.S. Lander, L.M. Linton, B. Birren, C. Nusbaum, M.C. Zody, J. Baldwin, K. Devon, K. Dewar, M. Doyle, W. FitzHugh, R. Funke, D. Gage, K. Harris, A. Heaford, J. Howland, L. Kann, J. Lehoczky, R. LeVine, P. McEwan, K. McKernan, J. Meldrim, J.P. Mesirov, C. Miranda, W. Morris, J. Naylor, C. Raymond, M. Rosetti, R. Santos, A. Sheridan, C. Sougnez, N. Stange-Thomann, N. Stojanovic, A. Subramanian, D. Wyman, J. Rogers, J. Sulston, R. Ainscough, S. Beck, D. Bentley, J. Burton, C. Clee, N. Carter, A. Coulson, R. Deadman, P. Deloukas, A. Dunham, I. Dunham, R. Durbin, L. French, D. Grafham, S. Gregory, T. Hubbard, S. Humphray, A. Hunt, M. Jones, C. Lloyd, A. McMurray, L. Matthews, S. Mercer, S. Milne, J.C. Mullikin, A. Mungall, R. Plumb, M. Ross, R. Shownkeen, S. Sims, R.H. Waterston, R.K. Wilson, L.W. Hillier, J.D. McPherson, M.A. Marra, E.R. Mardis, L.A. Fulton, A.T. Chinwalla, K.H. Pepin, W.R. Gish, S.L. Chissoe, M.C. Wendl, K.D. Delehaunty, T.L. Miner, A. Delehaunty, J.B. Kramer, L.L. Cook, R.S. Fulton, D.L. Johnson, P.J. Minx, S.W. Clifton, T. Hawkins, E. Branscomb, P. Predki, P. Richardson, S. Wenning, T. Slezak, N. Doggett, J.-F. Cheng, A. Olsen, S. Lucas, C. Elkin, E. Uberbacher, M. Frazier, R.A. Gibbs, D.M. Muzny, S.E. Scherer, J.B. Bouck, E.J. Sodergren, K.C. Worley, C.M. Rives, J.H. Gorrell, M.L. Metzker, S.L. Naylor, R.S. Kucherlapati, D.L. Nelson, G.M. Weinstock, Y. Sakaki, A. Fujiyama, M. Hattori, T. Yada, A. Toyoda, T. Itoh, C. Kawagoe, H. Watanabe, Y. Totoki, T. Taylor, J. Weissenbach, R. Heilig, W. Saurin, F. Artiguenave, P. Brottier, T. Bruls, E. Pelletier, C. Robert, P. Wincker, A. Rosenthal, M. Platzer, G. Nyakatura, S. Taudien, A. Rump, D.R. Smith, L. Doucette-Stamm, M. Rubenfield, K. Weinstock, H.M. Lee, J. Dubois, H. Yang, J. Yu, J. Wang, G. Huang, J. Gu, L. Hood, L. Rowen, A. Madan, S. Qin, R.W. Davis, N.A. Federspiel, A.P. Abola, M.J. Proctor, B.A. Roe, F. Chen, H. Pan, J. Ramser, H. Lehrach, R. Reinhardt, W.R. McCombie, M. de la Bastide, N. Dedhia, H. Blöcker, K. Hornischer, G. Nordsiek, R. Agarwala, L. Aravind, J.A. Bailey, A. Bateman, S. Batzoglou, E. Birney, P. Bork, D.G. Brown, C.B. Burge, L. Cerutti, H.-C. Chen, D. Church, M. Clamp, R.R. Copley, T. Doerks, S.R. Eddy, E.E. Eichler, T.S. Furey, J. Galagan, J.G.R. Gilbert, C. Harmon, Y. Hayashizaki, D. Haussler, H. Hermjakob, K. Hokamp, W. Jang, L.S. Johnson, T.A. Jones, S. Kasif, A. Kaspryzk, S. Kennedy, W.J. Kent, P. Kitts, E.V. Koonin, I. Korf, D. Kulp, D. Lancet, T.M. Lowe, A. McLysaght, T. Mikkelsen, J.V. Moran, N. Mulder, V.J. Pollara, C.P. Ponting, G. Schuler, J. Schultz, G. Slater, A.F.A. Smit, E. Stupka, J. Szustakowki, D. Thierry-Mieg, J. Thierry-Mieg, L. Wagner, J. Wallis, R. Wheeler, A. Williams, Y.I. Wolf, K.H. Wolfe, S.-P. Yang, R.-F. Yeh, F. Collins, M.S. Guyer, J. Peterson, A. Felsenfeld, K.A. Wetterstrand, R.M. Myers, J. Schmutz, M. Dickson, J. Grimwood, D.R. Cox, M.V. Olson, R. Kaul, C. Raymond, N. Shimizu, K. Kawasaki, S. Minoshima, G.A. Evans, M. Athanasiou, R. Schultz, A. Patrinos, M.J. Morgan, International Human Genome Sequencing Consortium, C. for G.R. Whitehead Institute for Biomedical Research, The Sanger Centre:, Washington University Genome Sequencing Center, US DOE Joint Genome Institute:, Baylor College of Medicine Human Genome Sequencing Center:, RIKEN Genomic Sciences Center:, Genoscope and CNRS UMR-8030:, I. of M.B. Department of Genome Analysis, GTC Sequencing Center:, Beijing Genomics Institute/Human Genome Center:, T.I. for S.B. Multimegabase Sequencing Center, Stanford Genome Technology Center:, University of Oklahoma's Advanced Center for Genome Technology:, Max Planck Institute for Molecular Genetics:, L.A.H.G.C. Cold Spring Harbor Laboratory, GBF—German Research Centre for Biotechnology:, also includes individuals listed under other headings): *Genome Analysis Group (listed in alphabetical order, U.N.I. of H.

Scientific management: National Human Genome Research Institute, Stanford Human Genome Center:, University of Washington Genome Center:, K.U.S. of M. Department of Molecular Biology, University of Texas Southwestern Medical Center at Dallas:, U.D. of E. Office of Science, The Wellcome Trust:, Initial sequencing and analysis of the human genome, Nature. 409 (2001) 860–921. https://doi.org/10.1038/35057062.

[18] S. Ballouz, A. Dobin, J.A. Gillis, Is it time to change the reference genome?, Genome Biology. 20 (2019) 159. https://doi.org/10.1186/s13059-019-1774-4.

[19] A.M. Phillippy, New advances in sequence assembly, Genome Res. 27 (2017) xi–xiii. https://doi.org/10.1101/gr.223057.117.

[20] V. Brázda, M. Bartas, R.P. Bowater, Evolution of Diverse Strategies for Promoter Regulation, Trends in Genetics. 37 (2021) 730–744. https://doi.org/10.1016/j.tig.2021.04.003.

[21] W.M. Guiblet, M.A. Cremona, R.S. Harris, D. Chen, K.A. Eckert, F. Chiaromonte, Y.-F. Huang, K.D. Makova, Non-B DNA: a major contributor to small- and large-scale variation in nucleotide substitution frequencies across the genome, Nucleic Acids Res. 49 (2021) 1497–1516. https://doi.org/10.1093/nar/gkaa1269.

[22] P. Kharel, G. Becker, V. Tsvetkov, P. Ivanov, Properties and biological impact of RNA G-quadruplexes: from order to turmoil and back, Nucleic Acids Res. 48 (2020) 12534–12555. https://doi.org/10.1093/nar/gkaa1126.

[23] F. Wu, K. Niu, Y. Cui, C. Li, M. Lyu, Y. Ren, Y. Chen, H. Deng, L. Huang, S. Zheng, L. Liu, J. Wang, Q. Song, H. Xiang, Q. Feng, Genome-wide analysis of DNA G-quadruplex motifs across 37 species provides insights into G4 evolution, Communications Biology. 4 (2021) 1–11. https://doi.org/10.1038/s42003-020-01643-4.

[24] C. Nakanishi, H. Seimiya, G-quadruplex in cancer biology and drug discovery, Biochem. Biophys. Res. Commun. 531 (2020) 45–50. https://doi.org/10.1016/j.bbrc.2020.03.178.

[25] V. Brázda, J. Coufal, Recognition of Local DNA Structures by p53 Protein, Int J Mol Sci. 18 (2017) 375. https://doi.org/10.3390/ijms18020375.

[26] A.M. Fleming, J. Zhu, M. Jara-Espejo, C.J. Burrows, Cruciform DNA Sequences in Gene Promoters Can Impact Transcription upon Oxidative Modification of 2′-Deoxyguanosine, Biochemistry. 59 (2020) 2616–2626. https://doi.org/10.1021/acs.biochem.0c00387.

[27] J.-Y. Gong, C.-J. Wen, M.-L. Tang, R.-F. Duan, J.-N. Chen, J.-Y. Zhang, K.-W. Zheng, Y. He, Y.-H. Hao, Q. Yu, S.-P. Ren, Z. Tan, G-quadruplex structural variations in human genome associated with single-nucleotide variations and their impact on gene activity, Proc Natl Acad Sci U S A. 118 (2021). https://doi.org/10.1073/pnas.2013230118.

[28] S. Ravichandran, J.-H. Ahn, K.K. Kim, Unraveling the Regulatory G-Quadruplex Puzzle: Lessons From Genome and Transcriptome-Wide Studies, Front Genet. 10 (2019) 1002. https://doi.org/10.3389/fgene.2019.01002.

[29] J. Spiegel, S. Adhikari, S. Balasubramanian, The Structure and Function of DNA G-Quadruplexes, TRECHEM. 2 (2020) 123–136. https://doi.org/10.1016/j.trechm.2019.07.002.

[30] A. De Magis, S.G. Manzo, M. Russo, J. Marinello, R. Morigi, O. Sordet, G. Capranico, DNA damage and genome instability by G-quadruplex ligands are mediated by R loops in human cancer cells, Proc Natl Acad Sci U S A. 116 (2019) 816–825. https://doi.org/10.1073/pnas.1810409116.

[31] D. Gomez, T. Lemarteleur, L. Lacroix, P. Mailliet, J.-L. Mergny, J.-F. Riou, Telomerase downregulation induced by the G-quadruplex ligand 12459 in A549 cells is mediated by hTERT RNA alternative splicing, Nucleic Acids Res. 32 (2004) 371–379. https://doi.org/10.1093/nar/gkh181.
[32] V. Marcel, P.L.T. Tran, C. Sagne, G. Martel-Planche, L. Vaslin, M.-P. Teulade-Fichou, J. Hall, J.-L. Mergny, P. Hainaut, E. Van Dyck, G-quadruplex structures in TP53 intron 3: role in alternative splicing and in production of p53 mRNA isoforms, Carcinogenesis. 32 (2011) 271–278. https://doi.org/10.1093/carcin/bgq253.
[33] E.P. Lombardi, Allyson Holmes, D. Verga, M.-P. Teulade-Fichou, A. Nicolas, A. Londoño-Vallejo, Thermodynamically stable and genetically unstable G-quadruplexes are depleted in genomes across species, Nucleic Acids Res. 47 (2019) 6098–6113. https://doi.org/10.1093/nar/gkz463.
[34] E. Ruggiero, S.N. Richter, Viral G-quadruplexes: New frontiers in virus pathogenesis and antiviral therapy, Annu Rep Med Chem. 54 (2020) 101–131. https://doi.org/10.1016/bs.armc.2020.04.001.
[35] S. Asamitsu, S. Obata, Z. Yu, T. Bando, H. Sugiyama, Recent Progress of Targeted G-Quadruplex-Preferred Ligands Toward Cancer Therapy, Molecules. 24 (2019). https://doi.org/10.3390/molecules24030429.
[36] M.-N. Prioleau, G-Quadruplexes and DNA Replication Origins, Adv Exp Med Biol. 1042 (2017) 273–286. https://doi.org/10.1007/978-981-10-6955-0_13.
[37] Y. Wang, J. Yang, A.T. Wild, W.H. Wu, R. Shah, C. Danussi, G.J. Riggins, K. Kannan, E.P. Sulman, T.A. Chan, J.T. Huse, G-quadruplex DNA drives genomic instability and represents a targetable molecular abnormality in ATRX-deficient malignant glioma, Nat Commun. 10 (2019) 943. https://doi.org/10.1038/s41467-019-08905-8.
[38] H. Inagaki, T. Ohye, H. Kogo, T. Kato, H. Bolor, M. Taniguchi, T.H. Shaikh, B.S. Emanuel, H. Kurahashi, Chromosomal instability mediated by non-B DNA: Cruciform conformation and not DNA sequence is responsible for recurrent translocation in humans, Genome Res. 19 (2009) 191–198. https://doi.org/10.1101/gr.079244.108.
[39] S. Takahashi, N. Sugimoto, Quantitative Analysis of Stall of Replicating DNA Polymerase by G-Quadruplex Formation, Methods Mol Biol. 2035 (2019) 257–274. https://doi.org/10.1007/978-1-4939-9666-7_15.
[40] L.I. Jansson, J. Hentschel, J.W. Parks, T.R. Chang, C. Lu, R. Baral, C.R. Bagshaw, M.D. Stone, Telomere DNA G-quadruplex folding within actively extending human telomerase, PNAS. 116 (2019) 9350–9359. https://doi.org/10.1073/pnas.1814777116.
[41] T. Kamura, Y. Katsuda, Y. Kitamura, T. Ihara, G-quadruplexes in mRNA: A key structure for biological function, Biochem. Biophys. Res. Commun. 526 (2020) 261–266. https://doi.org/10.1016/j.bbrc.2020.02.168.
[42] K.H. Miga, S. Koren, A. Rhie, M.R. Vollger, A. Gershman, A. Bzikadze, S. Brooks, E. Howe, D. Porubsky, G.A. Logsdon, V.A. Schneider, T. Potapova, J. Wood, W. Chow, J. Armstrong, J. Fredrickson, E. Pak, K. Tigyi, M. Kremitzki, C. Markovic, V. Maduro, A. Dutra, G.G. Bouffard, A.M. Chang, N.F. Hansen, A.B. Wilfert, F. Thibaud-Nissen, A.D. Schmitt, J.-M. Belton, S. Selvaraj, M.Y. Dennis, D.C. Soto, R. Sahasrabudhe, G. Kaya, J. Quick, N.J. Loman, N. Holmes, M. Loose, U. Surti, R. ana Risques, T.A. Graves Lindsay, R. Fulton, I. Hall, B. Paten, K. Howe, W. Timp, A. Young, J.C. Mullikin, P.A. Pevzner, J.L. Gerton, B.A. Sullivan, E.E. Eichler, A.M.

Phillippy, Telomere-to-telomere assembly of a complete human X chromosome, Nature. 585 (2020) 79–84. https://doi.org/10.1038/s41586-020-2547-7.
[43] N. Bohálová, J.-L. Mergny, V. Brázda, Novel G-quadruplex prone sequences emerge in the complete assembly of the human X chromosome, Biochimie. 191 (2021) 87–90. https://doi.org/10.1016/j.biochi.2021.09.004.
[44] M. Tassinari, M. Zuffo, M. Nadai, V. Pirota, A.C. Sevilla Montalvo, F. Doria, M. Freccero, S.N. Richter, Selective targeting of mutually exclusive DNA G-quadruplexes: HIV-1 LTR as paradigmatic model, Nucleic Acids Res. 48 (2020) 4627–4642. https://doi.org/10.1093/nar/gkaa186.
[45] M. Di Antonio, A. Ponjavic, A. Radzevičius, R.T. Ranasinghe, M. Catalano, X. Zhang, J. Shen, L.-M. Needham, S.F. Lee, D. Klenerman, S. Balasubramanian, Single-molecule visualization of DNA G-quadruplex formation in live cells, Nat. Chem. 12 (2020) 832–837. https://doi.org/10.1038/s41557-020-0506-4.

G4	G-quadruplexes
HiFi	high-fidelity
ONT	Oxford Nanopore Technologies
PacBio	Pacific Biosciences

Credit author statement

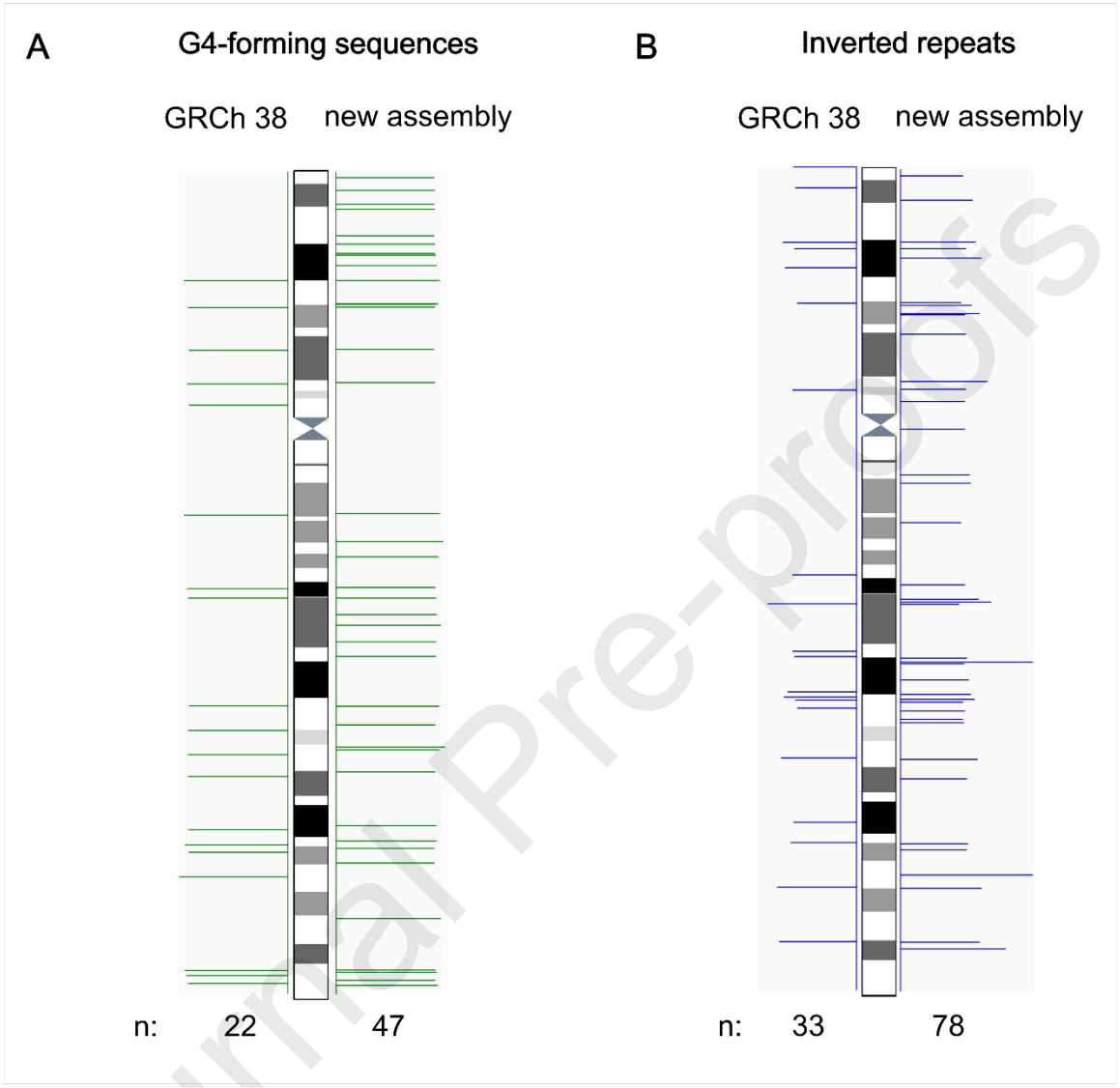
V.B. designed research; N.B. and V.B. performed research; V.B., N.B. and R.P.B. analyzed data; V.B., N.B. and R.P.B. wrote the paper.

Declaration of interests

☒ The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

☐ The authors declare the following financial interests/personal relationships which may be considered as potential competing interests:

Journal Pre-proofs



Journal Pre-proofs