**Non-B DNA Predictor: Identifying Multiple Potential Intramolecular Non-B DNAs**

***Authors(email):*** Hannah Ajoge (hajoge@uwo.ca), Hinissan P. Kohio (hkohio@uwo.ca), Henry He (hhe65@uwo.ca), Stephen D. Barr (stephen.barr@uwo.ca),
***Affiliation:*** Department of Microbiology and Immunology, Western University, Canada
**Keywords:** non-B DNA; cancer; replication; genome

Genome biology is not limited to the confines of the canonical B-forming DNA duplex, but includes over ten different types of other secondary structures, collectively termed non-B DNA structures. Non-B DNA structures such as A-motif, triplex, G-quadruplex and left-handed Z-DNA have garnered much attention recently because of their implication in DNA replication, telomere maintenance and gene expression in eukaryotic cells. They have also been shown to play important roles in the replication cycle of human viruses, immune evasion mechanisms, neurologic disorders and cancers [1-5]. Recognition of non-B DNAs has become essential to genome biology, and *in silico* approaches to studying these structures allows large-scale and detailed analysis of genes [6]. Presently, only one web resource is available for the identification of multiple non-B DNA motifs (non-B DNA Motif Search Tool (nBMST) ) [6]. Another webserver (QGRS Mapper) is available for identifying only G-quadruplexes [7]. Non-B DNA Predictor (NBDP) is capable of identifying seven motifs (A phased repeats, G Quadruplexes, Short tandem repeats, Z DNA, H DNA, Slipped motifs and Triplex forming oligonucleotides). NBDP has the added advantage of a backend R package 'gquad,' which is freely available as a standalone software package from the Comprehensive R Archive Network (http://cran.r-project.org). The standalone version can be run offline using Windows, Mac OS and Linux operating systems and can accommodate datasets in excess of 3 Gb (e.g. human genome datasets). Gquad is presently the only available R package and non-web tool for identifying multiple non-B DNAs. 'Triplex' is the only other available R package that identifies a non-B DNA motif and only predicts triplex DNA (H DNA) motifs.

NBDP accepts input as raw, fasta or GenBank accession numbers format and output result on the browser in a tabular format which can be copied or downloaded as a '.csv' file. NBDP improves upon current tools by identifying signature patterns (using regular expressions as with nBMST and QGRS Mapper [6, 7]) in nucleic acids that are predicted to form non-B DNA structures based on recent *in vitro*-validated whole genome data [8-10]. The validation datasets consisted of 616,828 positive sequences and 600,004 negative sequences[11]; NBDP excelled at a sensitivity of 97.70% (95% confidence interval: 97.66% - 97.74%), specificity of 100.00% (95% CI: 100.00% -100.00), precision of 100.00% (95% CI: 100.00% -100.00%), and accuracy of 98.83% (95% CI: 98.81% - 98.85%). Using a subset of the validation data, NBDP was compared to other web resources and shown to be the most sensitive and accurate at identifying non-B DNAs (Table 1). We also analyzed selected genomes of Viruses, Archaea, Prokaryotes and Eukaryotes using NBDP/gquad. We showed that non-B DNA densities are stable across kingdoms, and that G-quadruplex motifs are the most abundant non-B DNA motif. NBDP/gquad have been up and running since June 2017 and was recently presented as an oral abstract at the Rocky 2017 bioinformatics conference: (https://www.iscb.org/cms_addon/conferences/rocky2017/track/oral.php). NBDP testers outside our group included four bioinformaticians and two non-bioinformaticians.

The connection between non-B DNAs and human diseases is not yet well understood, with many unanswered questions [12]. Together NBDP/gquad offers a valuable online and offline toolset for the identification of non-B DNAs. NBDP will be useful to researchers with limited computational skills working on modestly-sized datasets and the backend gquad will be useful to researchers with basic computational skills to analyze high-throughput sequencing data files.

Table 1. Comparison of web tools for identifying potential intramolecular non-B DNAs

| | Input | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| **Web tools** | **Multiple Raw & Fasta Sequences** | **Accession No.** | **Offline standalone tool** | **Identify multiple non-B DNAs** | **Sensitivity (95% CI)** | **Specificity (95% CI)** | **Precision (95% CI)** | **Accuracy (95% CI)** |
| QGRS Mapper | No | Yes | No | No | 89.3(80.1-95.2) | 97.3(90.6-99.7) | 97.1(92.6-98.9) | 93.3(88.0-96.7) |
| nBMST | Yes | No | No | Yes | 50.7(38.9-62.4) | 100(95.1-100) | 100 (None) | 75.2(67.4-81.9) |
| NBDP | Yes | Yes | yes | Yes | 93.3(85.1-97.8) | 97.3(90.6-99.7) | 97.2(92.9-98.9) | 95.3(90.6-98.1) |

**References:**

1. Maizels, N. and L.T. Gray, *The G4 genome.* PLoS Genet, 2013. **9**(4): p. e1003468.
2. Metifiot, M., et al., *G-quadruplexes in viruses: function and potential therapeutic applications.* Nucleic Acids Res, 2014. **42**(20): p. 12352-66.
3. Shalaby, T., et al., *G-quadruplexes as potential therapeutic targets for embryonal tumors.* Molecules, 2013. **18**(10): p. 12500-37.
4. Simone, R., et al., *G-quadruplexes: Emerging roles in neurodegenerative diseases and the non-coding transcriptome.* FEBS Lett, 2015. **589**(14): p. 1653-1668.
5. Scheibye-Knudsen, M., et al., *Cockayne syndrome group A and B proteins converge on transcription-linked resolution of non-B DNA.* Proc Natl Acad Sci U S A, 2016. **113**(44): p. 12502-12507.
6. Cer, R.Z., et al., *Searching for non-B DNA-forming motifs using nBMST (non-B DNA motif search tool).* Curr Protoc Hum Genet, 2012. **Chapter 18**: p. Unit 18 7 1-22.
7. Kikin, O., L. D'Antonio, and P.S. Bagga, *QGRS Mapper: a web-based server for predicting G-quadruplexes in nucleotide sequences.* Nucleic Acids Res, 2006. **34**(Web Server issue): p. W676-82.
8. Chambers, V.S., et al., *High-throughput sequencing of DNA G-quadruplex structures in the human genome.* Nat Biotechnol, 2015. **33**(8): p. 877-81.
9. Stellwagen, E., Q. Dong, and N.C. Stellwagen, *Flanking A.T basepairs destabilize the B( *) conformation of DNA A-tracts.* Biophys J, 2015. **108**(9): p. 2291-9.
10. Boyer, A.S., et al., *The human specialized DNA polymerases and non-B DNA: vital relationships to preserve genome integrity.* J Mol Biol, 2013. **425**(23): p. 4767-81.
11. Villesen, P., *FaBox: an online toolbox for FASTA sequences.* Molecular Ecology Notes, 2007. **7**(6): p. 965-968.
12. Maizels, N., *G4-associated human diseases.* EMBO Rep, 2015. **16**(8): p. 910-22.