Published in final edited form as:

Curr Protoc Hum Genet. 2012 April ; CHAPTER: Unit-18.722. doi:10.1002/0471142905.hg1807s73.

Searching for non-B DNA-forming motifs using nBMST (non-B DNA Motif Search Tool)

RZ Cer¹, KH Bruce¹, DE Donohue¹, NA Temiz¹, US Mudunuri¹, M Yi¹, N Volfovsky¹, A Bacolla^{1,2}, BT Luke¹, Collins JR¹, and RM Stephens^{1,*}

¹Advanced Biomedical Computing Center, Information Systems Program, SAIC-Frederick, Inc., National Cancer Institute-Frederick, Frederick MD 21702, USA

²The Dell Pediatric Research Institute, Division of Toxicology and Pharmacology, The University of Texas at Austin, Austin TX 78723, USA

Abstract

This unit describes basic protocols on using the non-B DNA Motif Search Tool (nBMST) to search for sequence motifs predicted to form alternative DNA conformations that differ from the canonical right-handed Watson-Crick double-helix, collectively known as non-B DNA and on using the associated PolyBrowse, a GBrowse (Stein et al., 2002) based genomic browser. The nBMST is a web-based resource that allows users to submit one or more DNA sequences to search for inverted repeats (cruciform DNA), mirror repeats (triplex DNA), direct/tandem repeats (slipped/hairpin structures), G4 motifs (tetraplex, G-quadruplex DNA), alternating purine-pyrimidine tracts (left-handed Z-DNA), and Aphased repeats (static bending). Basic protocol 1 illustrates different ways of submitting sequences, the required file input format, results comprising downloadable Generic Feature Format (GFF) files, static Portable Network Graphics (PNG) images, dynamic PolyBrowse link, and accessing documentation through the Help and Frequently Asked Questions (FAQs) pages. Basic Protocol 2 illustrates a brief overview of some of the PolyBrowse functionalities, particularly with reference to possible associations between predicted non-B DNA forming motifs and disease causing effects. The nBMST is versatile, simple to use, does not require bioinformatics skills, and can be applied to any type of DNA sequences, including viral and bacterial genomes, up to 20 megabytes (MB).

Keywords

nBMST; non-B DNA; nucleotide sequence analysis; G-quadruplex; triplex; cruciform; Z-DNA; hairpin; slipped DNA; alternative DNA structure; tandem repeats; PolyBrowse

introduction

Algorithms to identify a wide variety of non-B DNA forming motifs were developed at the Advanced Biomedical Computing Center (ABCC), National Cancer Institute-Frederick (Table 1). Several mammalian reference genomes including human (Build 37.1), mouse (Build 37.1), chimpanzee (Build 2.1), macaque (Build 1.1), cow (Build 5.2), dog (Build 2.1),

*Corresponding Author: stephensr@mail.nih.gov.

Conflict of Interest: none declared.

Disclaimer

The content of this publication does not necessarily reflect the views or policies of the Department of Health and Human Services, nor does mention of trade names, commercial products, or organizations imply endorsement by the U.S. Government.

rat (Build 4.2), and platypus (Build 1.1) were completely annotated and the data was made publicly available in non-B DB (<http://nonb.abcc.ncifcrf.gov>) (Cer et al., 2011).

Increasing public interest in the non-B DB inspired us to develop a systematic tool that can dynamically annotate user-submitted nucleotide sequences with all the available non-B DNA motifs. In addition to providing a comprehensive tool, the non-B DNA Motif Search Tool (nBMST) was intended to combine and improve the functionalities of existing motif prediction tools. It is freely available through a simple web interface at <http://nonb.abcc.ncifcrf.gov/apps/nBMST/> and is currently the first and only web resource that searches for six alternative DNA structure forming sequences, including inverted (palindromic) repeats (cruciforms), direct and short tandem repeats (STR) (slipped DNA), mirror repeats and their subset containing purine pyrimidine tracts (triplexes), alternating purine-pyrimidine motifs (Z-DNA), G quadruplexes (GQ), and A-phased repeats (static bending).

BASIC PROTOCOL 1: Using the nBMST server

The Basic Protocol 1 includes step-by-step procedures on utilizing nBMST and is demonstrated with several screen shots using the bacterial genome of *Bacillus anthracis* strain ‘Ames Ancestor’ as an example. The same procedures can be applied to any type of DNA sequence from any organism. The Commentary section discusses how nBMST may help researchers to identify any potential associations between non-B DNA structure formation and diseases, using the genomic sequences from the bacteriococcus *Neisseria gonorrhoeae* responsible for the sexually transmitted infection gonorrhea (Cahoon and Seifert 2009), and forty-three sequences from the brain tumor tissues of pediatric cancer patients (Lawson et al., 2011).

Necessary Resources

Computer with Internet Access—An up-to-date web browser, such as Firefox (Windows, Mac OS X, and Linux; <http://www.mozilla.org/firefox>); Safari (Windows, Mac OS X; <http://www.apple.com/safari>); or Internet Explorer (Windows; <http://www.microsoft.com/ie>)

File Size and Format Requirement—A text file up to 20 MB with one or more DNA sequences in FASTA format

A FASTA file begins with a greater than sign, “>” character in the header followed without any spaces by a description and on a new line, the DNA sequence. The DNA sequences may contain only letters A, C, G, T or N and uppercase, lowercase letters and spaces are allowed. If there is more than one DNA sequence, each sequence must be separated by a description line.

Below are two examples of FASTA sequences. Only short sequences are shown for simplicity, >seq1

```
TTTATAATTTTATAATTATAAAATTTTATAATTTTATAATTTTATAATTTT
ATAATTATTTATAAT >seq2 gggtgggtgggtgggg
```

Accessing nBMST

- 1 Start at the non-B DB home page <http://nonb.abcc.ncifcrf.gov/>. Follow the link to the non-B DNA Motif Search Tool (nBMST) on the left sidebar.

Alternatively, navigate directly to
<http://nonb.abcc.ncifcrf.gov/apps/nBMST>.

Registration

- 2 Users are highly encouraged to register to the non-B DB resource by visiting <http://nonb.abcc.ncifcrf.gov/apps/account/register>. Registered users can login to the resource, submit nBMST jobs without the need to enter captcha characters (Figure 1) which are included for anti-spambot and security purposes and access the results for a longer storage time than non-registered users.

Submitting a custom nBMST search

- 3 In the “Email address” text box, enter a valid email address (Figure 1).

Providing an email address is optional but is recommended for notification purposes. The server sends notification emails containing URL links to the results when the submitted jobs are completed, circumventing user’s potential browser or computer system conflicts and/or connection issues.
- 4 In the “Non-B DNA Motif(s)” drop down list, select the non-B DNA motifs of interest using the Ctrl (Windows) or Command (Mac) key.

To select all the motifs, use the Shift key. Often, two types of motifs are paired together since the algorithms search for biologically relevant subsets within the main motifs. The “main motifs” are identified using more relaxed search criteria than the “subset motifs” (Table 1). For example, when the main motif “Mirror Repeats” is selected, users will also get results for the subset of “Triplex Motifs”, which are Mirror Repeats comprising 90% purines or pyrimidines on the same strand of DNA. The same applies for the main motif “Inverted Repeats” and its subset of “Cruciform Motifs” and the main motif “Direct Repeats” and the subset of “Slipped Motifs”, both of which are characterized by shorter spacer sizes than their parental “main motifs”. Aphased repeats, which form static DNA bends, left-handed Z-DNA and G-Quadruplex forming-motifs are also available in the dropdown list but are not associated with subset motifs.

- 5 Choose one of the four methods of submitting FASTA-formatted DNA sequence(s).

Option 1: Click on “Use the single sequence sample”.

This will use a sample sequence containing only one FASTA sequence.

Option 2:

Click on “Use the multi sequence sample”.

This will use a sample sequence containing multiple FASTA sequence.

Option 3:

In the text box for the “Enter sequence(s)” field, copy and paste sequence(s).

Option 4:

Click on “Upload sequence(s)” and select a file sequence(s) (*shown in Figure 1*).

The input text file may contain multiple FASTA sequences, which must be separated by a new line beginning with the “>” character followed by a description of the sequence. Sequences containing letters other than ‘A’, ‘C’, ‘G’, ‘T’, or ‘N’ will produce an error. Currently, the maximum file size allowed is 20 MB. For files larger than the allowed size, users are encouraged to contact nonb@nih.gov for further assistance.

In the example here, the 5.2 mega base pairs (Mbp) genome sequence of Bacillus anthracis was downloaded from [http://www.ncbi.nlm.nih.gov/nuccore/NC_007530.2?report=fasta&log\\$=seqview](http://www.ncbi.nlm.nih.gov/nuccore/NC_007530.2?report=fasta&log$=seqview) and uploaded using Option 4.

- 6 Enter captcha information and hit the “Submit” button to process the job.

Turnaround time for results will vary depending on the size of the sequence(s), the number and types of the non-B DNA motifs selected and the computing resources available at the ABCC server at the time of job submission. In cases where input sequences are very large and/or multiple motifs are selected, an email address is recommended to avoid waiting on-line. A notification email will be sent when the job is completed.

It should be noted that the algorithms for mirror and inverted repeats are the most computationally intensive (Table 4), and therefore take more time to complete than the rest of the motifs. Thus, in cases where quick results are desired for large sequences, it is recommended that separate jobs be submitted for these two motif types.

Accessing and understanding the different sections of the result page—There are two major sections in the result page (Figure 2).

- 7 The upper section displays the major statistics of the nBMST run and links to the aggregated results.
 - a. *Job ID* unique to each nBMST request.
 - b. *Total non-B motifs found* gives the total number of non-B DNA motifs predicted in all queried sequences combined.
 - c. *Results will be stored until* displays the date and time the results will be removed from the server. The result files are stored on the server for six months for a registered user and for three days for a non-registered user.
 - d. *Link* provides a URL that will be valid till the time as specified in (c) for accessing results.
 - e. *Dynamic visualization* provides a link for viewing the results in PolyBrowse.
 - f. *Download all files for this job*, a clickable button for downloading all the result files for the current job.
- 8 The lower section displays the individual results for each motif type.
 - a. *Expand all results* is a link visible between the upper and lower sections that can be used for viewing all the motifs at the same time.

Conversely, the up arrow or the “Hide all results” link can be used to hide results for each individual motif or for all the motifs, respectively.

- b. The down arrow (▼), on the left of each motif, can be clicked to view the results for individual motifs (Figure 3). The total numbers of hits found for each motif in the sequence submitted is also included.
- c. *Download all files for this motif* is a clickable button for downloading all the result files for one type of motif.

Visualization of the nBMST results—To provide better visualization of the results, both static Portable Network Graphic (PNG) images and a dynamic PolyBrowse link are provided.

Static visualization

- 9 Click on *Expand all results* on the lower section of the result page to view all the motifs (Figure 3).

- a. The top panel is a graphical representation of the position of the direct repeats across the submitted nucleotide sequence. The red arrow shows the size of the sequence submitted and the direct repeats which are depicted as vertical ticks below the red line are shown across the genome.

In this *Bacillus anthracis* example, the direct repeats are seen to span across the 5 Mbp. Each direct repeat is labeled with the reference name followed by start and stop coordinates and the motif type (e.g. BA_Ames_40529_40551_Direct_Repeat).

- b. The bottom panel contains detailed information for each motif, including the start and stop coordinates, strand, length, structure, sequence composition, and a note if a sequence can also be categorized as a short tandem repeat (STR).

For example, in the following motif entry, the structure 10-3-10 indicates the composition of repeat-spacer-repeat (i.e. ATTCGAGGAA-TTT-ATTCGAGGAA).

Reference	Type	Start	Stop	Strand	Length	Structure
BA Ames	Direct Repeat	40529	40551	+	23	10-3-10

Notice that because slipped motifs are the subset of direct repeats with a spacer size of zero base pair, the sequence with coordinates 104120 to 104141 is categorized as both a direct repeat and a slipped motif.

Reference	Type	Start	Stop	Strand	Length	Structure
BA_Ames	Direct_Repeat	104120	104141	+	22	11-0-1
BA_Ames	Slipped_Motif	104120	104141	+.+	22	11-0-1

Dynamic Visualization

- 10 Click on *PolyBrowse* on the upper section of the result page. Viewing the non-B DNA-forming motifs interactively on PolyBrowse is particularly useful when there are too many entries to be viewed in a static image.

PolyBrowse view of the Bacillus anthracis genome (Figure 4) shows the inverted repeats, cruciform motifs, and mirror repeats as a nearly solid line due to their abundance. Direct repeats, slipped motifs, triplex motifs and Z-DNA motifs are seen more sparsely as bars. G-quadruplex forming repeats are very rare and, therefore, are seen labeled with exact locations. The location specific to any motif can be seen by moving the cursor over the bar representing the motif. Clicking on the glyph representing the motif will take users to view more detailed information.

Customizing PolyBrowse

- 11 The default display setting of PolyBrowse is 5 Mbp which provides a good overview for large amounts of data. Users can navigate by clicking on the *Scroll/Zoom minus* or *plus* signs as well as by selecting the drop down list in between *minus* and *plus* signs. Selecting smaller regions reduces the scale size allowing users to view tracks in greater details. Figure 5 shows a smaller region of Figure 4 after zooming-in. For easy visualization, each motif is represented in a different color. Acronyms for non-B DNA motifs used in PolyBrowse are listed in Table 2.

In cases where more than one DNA sequence is submitted, the multiple sequences will be displayed in PolyBrowse as shown in Figure 6.

Downloading and saving results

- 12 The results can be downloaded in two different ways. The files include text results in the Generic Feature Format (.gff), tab-delimited format (.txt) and Portable Network Graphic (.png) images for each motif.
 - a. Click on “*Download all files for this motif*” to download individual files for only one type of non-B DNA motif.
 - b. Click on “*Download all files for this job*” to download all the files for all the motifs found for the entire job.

In both cases, all the files are compressed into a zipped file which may be opened by double clicking on it and by using software such as WinZip, WinRAR, and 7-Zip.

Additional help

- a. Frequently Asked Questions (FAQs)
[http://nonb.abcc.ncifcrf.gov/FAQs/#cat_non-B_Motif_Search_Tool_\(nBMST\)](http://nonb.abcc.ncifcrf.gov/FAQs/#cat_non-B_Motif_Search_Tool_(nBMST))
If there is any question not addressed in the FAQs section, it may be submitted via
http://nonb.abcc.ncifcrf.gov/FAQs/submit_a_question/.
- b. Help - http://nonb.abcc.ncifcrf.gov/Help/help_nBMST.php
Any questions, comments, suggestions or concerns should be submitted to
nonb@nih.gov *through a direct email or by using the appropriate form at*
<http://nonb.abcc.ncifcrf.gov/Site-Information/Contact-Us/>.

BASIC PROTOCOL 2: Using the PolyBrowse viewer

The ability to visualize predicted non-B DNA forming motifs across several mammalian genomes in their genomic context is also available through the parent non-B DB web site (Cer et al., 2011). The following protocol is intended to provide an overview of some of the PolyBrowse functionalities, particularly with reference to possible associations between predicted non-B DNA forming motifs and disease causing effects. The screenshots illustrate how to turn on selected tracks. We discuss the visualization they produce and how to best interpret them in the context of integrating all available biological information. The PolyBrowse visualization tool is derived from GBrowse (Stein et al., 2002) and was first introduced in our earlier publications describing the characterization of variant retrotransposons across different mouse strains (Akagi et al., 2008, Akagi et al., 2009). Similar functionality is made available through the nBMST tool, with the distinction that in nBMST, rather than viewing the pre-computed mammalian genomic data from non-B DB, the dynamically and uniquely created PolyBrowse view pertains to the user-submitted sequence(s).

Another example of the utility of PolyBrowse is the ability to see overlaps between the various predicted non-B DNA motif positions relative to various cancer data such as breakpoints, somatic mutations and copy number variations (data not shown). These data are all also available and are frequently updated within the PolyBrowse environment.

Necessary Resources

Computer with Internet Access—An up-to-date web browser, such as Firefox (Windows, Mac OS X, and Linux; <http://www.mozilla.org/firefox>); Safari (Windows, Mac OS X; <http://www.apple.com/safari>); or Internet Explorer (Windows; <http://www.microsoft.com/ie>)

Experimental paradigm—For demonstration purposes, here we will use Human v37 PolyBrowse viewer to evaluate a hypothesis that there may be G-quadruplex motifs present in the promoter or first exon of the c-MYC gene and that this alternative structure may impact its expression or splicing

Accessing PolyBrowse

- 1 Navigate directly to <http://pbrowse3.abcc.ncifcrf.gov/cgi-bin/gb2/gbrowse/>
Alternatively, start at the non-B DB home page <http://nonb.abcc.ncifcrf.gov/>. From the scrollbar displaying “View in PolyBrowse”, select the species of interest to use as a starting point from the list. A new page showing the PolyBrowse interface will appear with the selected species in context. By default, the page will load with the myc proto-oncogene in view.

Navigating PolyBrowse

General layout

- 2 PolyBrowse contains *Overview*, *Region*, *Details*, and *Tracks* panels as other genome browsers do. Click on “*Help*” at the top left of the PolyBrowse panel. Select “*Help with this browser*” item from *Help* menu to learn in more details about navigational tips on moving to a specific landmark or genomic position and other operational considerations for the environment. Alternatively, go directly to http://pbrowse3.abcc.ncifcrf.gov/gbrowse2/general_help.html.

Species

- 3 Click on the “*Data Source*” selection menu and select “*Human v37*”.

This allows switching between different species. The naming convention is species name followed by version number. In this example, Human v37 is Homo sapiens version 37.1.

Landmark or region

- 4 In the “*Landmark or region*” search box, enter “chr8:128747620..128749619” for chromosome coordinates.

Moving the browser position to your region of interest operates similarly as in other genome browsers. Either a specific gene name or chromosome coordinates can be input to the “Landmark or Region” search box. In cases where multiple matches are found, they will be displayed in a list to permit the user to select the exact region of interest. A wildcard (for example, myc%) may be used to broaden the search capability. Once at the desired location, the user may zoom in or out using the controls under the scroll/zoom area at the upper right of the panel.

Tracks

- 5 In order to lay the groundwork for the experimental scenario, turn on all the following tracks.

- a. Overview

Turn on “*ideogram*” under “*Overview*” section.

- b. Genes

Turn on “*Refseq Genes (NCBI)*”, “*Refseq mRNAs (NCBI)*” and “*RefSeq Promoters (NCBI/ABCC)*” tracks under the “*Genes*” section.

- c. Non-B DNA motifs

Turn on “*G Quadruplex Forming Repeat*” under the “*Non-B DNA motifs*” section (Figure 7a) located at the bottom of the PolyBrowse page.

As seen in Figure 7, the PolyBrowse shows that there are G-quadruplex repeat motifs predicted within both the promoter and first exon of the MYC gene.

- d. Polymorphism

- i. Turn on the “*dbSNP(NCBI)*” track under the “*Polymorphism*” section.

Since we have confirmed that there are G-Quadruplex repeat motifs within the region of interest, it would be of interest to determine whether there have been polymorphisms identified within these locations. As shown in Figure 8, it does appear that there may be at least one SNP within one of these motif regions.

- ii. Turn on the “*trace GPlaxes clusters*” under the “*GPlaxes*” of “*Polymorphism*” section.

To assess the impact of this and other SNPs on the motifs more specifically, we have provided available Sanger trace reads against the human genome reference and scored the differences that occurred within the predicted G-quadruplex motifs. Please note that this capability only exists for the G-quadruplex repeat motifs at this time and these results are preliminary.

Using the various different options under this class of polymorphisms, it can be seen that some variations that would disrupt the presence of the predicted motif in some individual exist. In other individuals, there are additional variants that have additional predicted G-quadruplex motifs not present in the reference sequence. Both of these sources of polymorphism could contribute to the differential effects observed in different individuals at this location (Figure 9).

e. Synteny

Turn on the “*1kb LiftOver Blocks*” under the “*Synteny*” section.

Within-species conservation of a specific motif has often been used to indicate its likeliness to play some important functional role (neutral selection). An extension to this, observing the same motifs conserved in position in other species can also be used to support this model. In order to facilitate this type of comparison, we have adopted PolyBrowse to support synteny-sensitive context switching. This leverages the UCSC cross-species mapping information.

Once this track has been added, by mouse-ing over the selected species, an additional page will be displayed that shows the syntenic region within that other species.

6 Click on “*Go to Chimp*”.

Figure 11 shows that the three G-quadruplex forming repeat motifs are all conserved in the chimpanzee genome sequence. In this case, the motifs were not similarly conserved in the macaque (not shown) suggesting some possible divergence between these organisms.

COMMENTARY

Background information

In addition to the right-handed Watson-Crick B-form, DNA can adopt alternative (non-canonical, non-B) structures (Cer et al., 2011; Phan et al., 2006; Wells, 2007; Zhao et al., 2010), which have been associated with genetic instability and disease (Bacolla and Wells, 2009; Wells, 2007; Zhao et al., 2010)(Bacolla and Wells, 2004; Cooper et al., 2011). Microsatellite expansion diseases (MEDs) encompass approximately thirty neuromuscular and developmental human pathologic conditions, including myotonic dystrophy types 1 and 2, Friedreich ataxia, fragile X syndrome, fragile X-associated ataxia and tremor, FRAXE associated mental retardation, Huntington’s disease, spinal and bulbar muscular atrophy, oculopharyngeal muscular dystrophy, dentatorubral-pallidoluysian atrophy and at least nine types of spinocerebellar ataxias. MEDs are caused by the expansion of microsatellite repeats

within the affected genes (Brouwer et al., 2009; Lopez Castel et al., 2010; McMurray, 2010; Messaied and Rouleau, 2009; Orr and Zoghbi, 2007)).

The number of these repeats (often trinucleotide) is polymorphic in the general population and varies among repeat types (CAG-CTG, CGG-CCG, GAA-TTC, CCTG-CAGG and ATTCT-AGAAT). However, it is generally confined to below 50 repeats. In the disease state, the number of repeat units increases dramatically, up to several thousand copies in the UTRs and introns of genes, and up to a few hundred copies in coding exons (Brouwer et al., 2009). Both the timing and the mechanisms of repeat expansion remain open questions. However, expansion is thought to occur early during embryogenesis/gametogenesis from “pre-mutation” (generally longer than normal) alleles (Brouwer et al., 2009) by mechanisms involving the formation of non-canonical DNA structures, such as hairpins (Mirkin, 2007; Pearson et al., 2005; Wells et al., 2005), and their processing by DNA repair enzymes, leading to expansion.

Further somatic instability is also often observed with age in the affected tissues. The involvement of DNA repair pathways in repeat expansion is supported by a number of studies in transgenic mice, which show decreased repeat instability in DNA repair deficient strains (Dragileva et al., 2009; Entezam and Usdin, 2008; Fleming et al., 2003; Foirey et al., 2006; Gomes-Pereira et al., 2004; Lin and Wilson, 2011; Manley et al., 1999; Shelbourne et al., 2007; van den Broek et al., 2002; Wheeler et al., 2003). Expanded trinucleotide repeats mediate the process of pathogenesis through various mechanisms, including altered protein structure, heterochromatin formation, RNA gain-of-function and aberrant initiation of translation (Zu et al., 2011). However, in the case of Friedreich ataxia, triplexes or other DNA conformations (Wells, 2008) of the expanded (GAA-TTC) repeat in intron 1 of the frataxin (*FXN*) gene are also believed to contribute to pathogenesis, by repressing the process of transcription elongation and leading to gene silencing along with histones tail modification (Punga and Buhler, 2010).

It is estimated that one in every 600 individuals carries a balanced constitutional translocation (Vandyke et al., 1983). The supernumerary der(22)t(11;22) syndrome or Emanuel syndrome (MIM 609209) is characterized by severe mental retardation and morphologic abnormalities, and results from meiotic malsegregation of chromosomes carrying a t(11;22)(q23;q11), the most common non-Robertsonian constitutional translocation (Emanuel, 2008). Sequence analyses of several unrelated cases of Emanuel syndrome have revealed the clustering of breakpoints at the centers of ~450 and ~590 bp-long palindromic A+T-rich repeats on both chromosomes 11 and 22, termed PATRR11 and PATRR22 for palindromic A+T rich repeats (Inagaki et al., 2009; Kurahashi et al., 2007; Kurahashi et al., 2010). Despite their A+T-richness, PATRR11 and PATRR22 share limited homology; their sequences have been shown to form cruciform structures when cloned in plasmids *in vitro* and to undergo recombination in cell culture systems, reconstituting junction fragments that mimic those observed *in vivo* (Inagaki et al., 2009).

Analyses in sperm from healthy individuals have revealed the occurrence of *de novo* t(11;22) events at frequencies that correlate with PATTR sequence polymorphisms in the population, with carriers of partially deleted sequences, resulting in reduced inverted repeat symmetry, undergoing fewer *de novo* events than the full-length PATTRs (Tong et al., 2010). The PATRR22 is also embedded in the low copy repeat LCR-B, one of the most recombination prone regions in the human genome (Emanuel, 2008) and an unclonable gap in the current reference genome. PATRR22 has been observed to undergo recurrent constitutional translocations with other PATTR-like sequences, such as chromosomes 17 (t(11;17)(q11;q11.21) (Kehrer-Sawatzki et al., 1997; Kurahashi et al., 2006) and 8 (t(8;22)(q24.13;q11.21) (Sheridan et al., 2010). Taken together, these observations lend support to

the hypothesis that large inverted repeats may fold into cruciform structures in the human genome, which are then cleaved at the single-stranded loops and joined with similar structures. Thus, cruciform-mediated recombination represents a novel mechanism of genetic instability associated with recurrent constitutional translocations (Kurahashi et al., 2006).

In addition to constitutional translocations, non-B DNA-forming motifs have been noted at or near breakpoint cluster regions (BCR) of somatic rearrangements associated with cancer. In pediatric B-cell precursor acute lymphoblastic leukemia (BCP-ALL), alternating purine-pyrimidine tracts with the potential to form left-handed (Z-DNA) helices have been found in the BCR of the ets variant 6 (*ETV6*, *TEL*) gene, whose fusions with the runt-related transcription factor 1 (*RUNX1*, *AML1*) gene are recognized as the most frequent (~25%) type of translocation (t(12;21)(p13;q22)) in PCB-ALL (Thandla et al., 1999; Wiemels and Greaves, 1999). The intrachromosomal amplification of chromosome 21 (iAMP21) characterizes a subgroup of pediatric PCB-ALL with poor prognosis on standard therapy, and is generally restricted to older children (Sinclair et al., 2011). Sequence analyses of a cohort of iAMP21 patients identified a recurrent BCR in intron 1 of the phosphodiesterase 9A (*PDE9A*) gene proximal to a long stretch (~1 kb) of CA·TG (Z-DNA) repeats (Sinclair et al., 2011). Juxtaposition of the 5' flanking region of the B-cell CLL/lymphoma 2 (*BCL2*) gene to various sequences on the immunoglobulin light chain loci, as well as translocations involving the 3' end of the v-myc myelocytomatosis viral oncogene homolog (avian) (*MYC*) protooncogene, represent common alterations associated with chronic lymphocytic leukemia (CLL) and lymphoma (Adachi and Tsujimoto, 1990; Seite et al., 1993). Both the *BCL2* and *MYC* BCRs, as well as BCRs at other loci (Boehm et al., 1989), have been seen to encompass multiple alternating purine-pyrimidine elements (Adachi and Tsujimoto, 1990; Rimokh et al., 1991; Seite et al., 1993). These observations suggest that Z-DNA-forming motifs serve to promote genomic rearrangements associated with lymphoid tumors.

The introduction of high-resolution comparative genomic hybridization arrays (aCGH) has been playing a critical role in elucidating the genomic architecture of complex genomic rearrangements associated with human inherited disorders (Stankiewicz and Lupski, 2010). The recent analysis of families with genomic rearrangements of chromosome Xq28 revealed the unexpected configurations of triplicated ge-nomic segments nested within duplicated genomic regions. The triplicated segments were invariably inverted with respect to their flanking duplicated segments (DUP-TRP/INV-DUP) and one of the duplication/triplication junctions was characterized by the presence of inverted repeats (Carvalho et al., 2011). The current molecular mechanism for DUP-TRP/INV-DUP consists of a model based on replication fork stalling at the inverted repeats, followed by template switching and resumption of DNA synthesis on the original DNA template. Thus, repetitive DNA motifs are likely to be involved in mediating ge-nomic instability through several mechanisms, in addition to their transient formation of alternative DNA structures.

In recent years, several web servers and databases have been developed to search for individual non-B DNA-forming motifs. Greglist (Zhang et al., 2008), GRSDb (Kostadinov et al., 2006) and Quadbase (Yadav et al., 2008) are pre-computed G-quadruplex databases. QGRS Mapper (Kikin et al., 2006) and Quadfinder (Scaria et al., 2006) are web servers for G-quadruplex prediction. TFO (Gaddis et al., 2006), TRACTS (Gal et al., 2003), and TTS (Jenjaroenpun and Kuznetsov, 2009) are web servers searching for triplex forming sequences, whereas Z-Hunt-II (Schroth et al., 1992) and Z-Catcher (Li et al., 2009) search for motifs predicted to form left-handed Z-DNA. Although there are several databases and web servers dedicated to the search of non-B DNA motif, a web server that can process multiple sequence motifs was lacking.

Advantages

The nBMST offers to the user the advantage of customizing the search for multiple motifs with a single click, without the need to visit multiple web servers (Table 3). It also enables users to upload larger sequences (up to 20 MB) than the existing tools (1–3 MB). The 20 MB upload size has been implemented with the goal of allowing whole genome searches for a variety of organisms, including most bacteria and viruses. Other significant advantages of the nBMST include dynamic visualization, batch capability, results storage of up to six months for registered users, user-friendly graphical user interfaces, various downloadable file formats that can be used in further analyses, and extensive Help and Frequently Asked Questions (FAQs) contents. Finally, by allowing relatively loose definitions combined with more restrictive subsets (cruciform and triplex DNA, for example) where appropriate, we provide a way for the users to adjust the false negative vs. false positive tradeoff to best suit their needs. The nBMST could be used for any type of DNA sequences and therefore its applications are expected to assist in various aspects of genome-wide analyses.

ANTICIPATED RESULTS

The nBMST allows users to find repetitive motifs that have the potential to form non-B DNA structures in any nucleotide sequences. It aims at assisting researchers to elucidate any potential association between non-B DNA-forming motifs and cancer or inherited human diseases or pathogenicity. Here we demonstrate the application of nBMST to two published case studies.

Bacterial genome - *Neisseria gonorrhoeae*

Cahoon and Seifert (Cahoon and Seifert 2009) reported a 16 base pair intergenic guanine (G)-rich sequence (5'-GGGTGGGTTGGGTGGG-3') upstream of the *PilE* locus that was required for pilin antigenic variation in the human pathogen *Neisseria gonorrhoeae*, and which formed a guanine quadruplex(G4) structure *in vitro*. Targeted mutations at any of the 12 G-C bp within the 16 G-rich sequence suppressed antigenic variation by inhibiting recombination, thereby mimicking the effects of naturally-occurring transposon insertions upstream of the *PilE* gene.

The nucleotide sequences of two *Neisseria gonorrhoeae* genomes, NCCP11945 (2.23 Mbp) and FA 1090 (2.15 Mbp), were downloaded from http://www.ncbi.nlm.nih.gov/nuccore/NC_011035 and http://www.ncbi.nlm.nih.gov/nuccore/NC_002946 respectively. The nBMST was used to find 103 and 94 G-quadruplex forming repeats in NCC11945 and FA1090 respectively. The G-quadruplex predictions made by the nBMST were found to be comparable to those obtained from other tools, including QGRS Mapper (Kikin et al., 2006) and QuadFinder (Scaria et al., 2006) (*result not shown*). The nBMST (Figure 12) confirmed the presence of the 16 bp G-quadruplex forming repeat, as well as a 17 bp version of the same repeat that was found to also form a quadruplex structure when the third G from the 3' end in the sequence above was mutated (Cahoon and Seifert 2009).

Pediatric brain tumor sequences

Forty-three sequences (accession numbers FR799511-FR799553) from *RAF* fusion genes associated with low-grade gliomas and astrocytomas from patients aged between 1–20 years (Lawson et al., 2011) were downloaded from the European Nucleotide Archive <http://www.ebi.ac.uk/ena/>. Of the 43 cases, the sequence from one patient (pilocytic astrocytoma, isolate PA28) was excluded because the fusion breakpoint(s) could not be unambiguously identified. Of the remaining 42 patients, 40 had simple *KIAA1549-BRAF*

rearrangements while 2 (PA27 and PA30) had more complex genomic rearrangements with large insertions.

Using nBMST, all of the six non-B DNA motifs were searched in the rearranged tumor sequences and a total of 21 non-B DNA-forming motifs were found. Twelve (57.14%) resulted from the two patients, PA27 (FR799551) and PA30 (FR799552-3), who displayed complex rearrangements. The *KIAA1549-BRAF* complex rearrangement in PA27 also contained a 67-bp insertion that could not be mapped to a single genomic locus. This 67 bp *de novo* insertion (Figure 13) was found to comprise several non-B DNA-forming motifs, including inverted repeats (potentially forming cruciforms) and slipped motifs, raising the possibility that these repetitive sequences might have provided microhomology and assisted in the repair process (Simsek et al., 2011). Thus, nBMST identified novel non-B DNA forming motifs in the rearranged sequences, supporting the use of nBMST as a tool for detecting potential DNA secondary structures that may be implicated in biological function.

CRITICAL PARAMETERS

Currently, the only selectable parameter for nBMST users is the non-B DNA motif types (Table 1), which are defined based on a variety of experimental and theoretical studies (Cer et al. 2011). By selecting subset motifs (such as Cruciform or Slipped), the user is able to separate those results that retain the highest probability of forming secondary structures *in vivo*, based on experimental studies. Likewise, omitting these more restrictive motif subsets minimizes the chance of potentially missing relevant results. We anticipate that future versions of nBMST will allow users to adjust most parameters, such as lengths of the repeats spacers and the presence of mismatches.

TIME CONSIDERATIONS

The size of the sequence(s) submitted, the number and types of the non-B DNA motifs selected, and the computing resources available at the ABCC at the time of job submission determine the turnaround time for nBMST results. Table 4 shows turnaround time for five different bacterial genomes with sizes ranging from 2 to 15 Mbp and the differences in turnaround time when mirror repeats and inverted repeats are excluded in the search. It should be noted that this turnaround time includes not only the time to annotate non-B DNA motifs in the sequences submitted but also the time to create static PNG images and dynamic PolyBrowse tracks. As shown in Table 4, the algorithms for mirror and inverted repeats are the most computationally intensive, and therefore take more time to complete than the rest of the motifs. When quick results are desired for large sequences, it is recommended that separate jobs be submitted for these two motif types while one can analyze the results for other motifs.

Acknowledgments

The authors thank Scientific Web Programming Group, System Administrators Toni Harbaugh and Michael Gurski for their help. This work was supported by the Center for Biomedical Informatics and Information Technology (CBIIT)/Cancer Biomedical Informatics Grid (caBIG) ISRCE yellow task #09-260 to NCI-Frederick and National Cancer Institute/National Institutes of Health contract HHSN261200800001E (to A.B.).

LITERATURE CITED

Adachi M, Tsujimoto Y. Potential Z-DNA elements surround the breakpoints of chromosome translocation within the 5' flanking region of bcl-2 gene. *Oncogene*. 1990; 5(11):1653–1657. [PubMed: 2176281]

- Akagi K, Li J, Stephens RM, Volfovsky N, Symer DE. Extensive variation between inbred mouse strains due to endogenous L1 retrotransposition. *Genome Res.* 2008; 18(6):869–80. [PubMed: 18381897]
- Akagi K, Stephens RM, Li J, Evdokimov E, Kuehn MR, Volfovsky N, Symer DE. MouseIndelDB: a database integrating genomic indel polymorphisms that distinguish mouse strains. *Nucleic Acids Res.* 2009; 38:D600–6. [PubMed: 19933259]
- Bacolla A, Wells RD. Non-B DNA conformations, genomic rearrangements, and human disease. *J Biol Chem.* 2004; 279(46):47411–47414. [PubMed: 15326170]
- Bacolla A, Wells RD. Non-B DNA conformations as determinants of mutagenesis and human disease. *Mol Carcinog.* 2009; 48(4):273–285. [PubMed: 19306308]
- Boehm T, Mengle-Gaw L, Kees UR, Spurr N, Lavenir I, Forster A, Rabbitts TH. Alternating purine-pyrimidine tracts may promote chromosomal translocations seen in a variety of human lymphoid tumours. *EMBO J.* 1989; 8(9):2621–2631. [PubMed: 2531086]
- Brouwer JR, Willemsen R, Oostra BA. Microsatellite repeat instability and neurological disease. *Bioessays.* 2009; 31(1):71–83. [PubMed: 19154005]
- Cahoon LA, Seifert HS. An alternative DNA structure is necessary for pilin antigenic variation in *Neisseria gonorrhoeae*. *Science.* 2009; 325(5941):764–767. [PubMed: 19661435]
- Carvalho CM, Ramocki MB, Pehlivan D, Franco LM, Gonzaga-Jauregui C, et al. Inverted genomic segments and complex triplication rearrangements are mediated by inverted repeats in the human genome. *Nat Genet.* 2011; 43(11):1074–1081. [PubMed: 21964572]
- Cer RZ, Bruce KH, Mudunuri US, Yi M, Volfovsky N, Luke BT, Bacolla A, Collins JR, Stephens RM. Non-B DB: a database of predicted non-B DNA-forming motifs in mammalian genomes. *Nucleic Acids Res.* 2011; 39(Database issue):D383–391. [PubMed: 21097885]
- Cooper DN, Bacolla A, Ferec C, Vasquez KM, Kehrer-Sawatzki H, Chen JM. On the sequence-directed nature of human gene mutation: the role of genomic architecture and the local DNA sequence environment in mediating gene mutations underlying human inherited disease. *Hum Mutat.* 2011; 32(10):1075–1099. [PubMed: 21853507]
- Dragileva E, Hendricks A, Teed A, Gillis T, Lopez ET, Friedberg EC, Kucherlapati R, Edlmann W, Lunetta KL, MacDonald ME, Wheeler VC. Intergenerational and striatal CAG repeat instability in Huntington's disease knock-in mice involve different DNA repair genes. *Neurobiol Dis.* 2009; 33(1):37–47. [PubMed: 18930147]
- Emanuel BS. Molecular mechanisms and diagnosis of chromosome 22q11.2 rearrangements. *Dev Disabil Res Rev.* 2008; 14(1):11–18. [PubMed: 18636632]
- Entezam A, Usdin K. ATR protects the genome against CGG. CCG-repeat expansion in Fragile X premutation mice. *Nucleic Acids Res.* 2008; 36(3):1050–1056. [PubMed: 18160412]
- Fleming K, Riser DK, Kumari D, Usdin K. Instability of the fragile X syndrome repeat in mice: the effect of age, diet and mutations in genes that affect DNA replication, recombination and repair proficiency. *Cytogenet Genome Res.* 2003; 100(1–4):140–146. [PubMed: 14526174]
- Foiry L, Dong L, Savouret C, Hubert L, te Riele H, Junien C, Gourdon G. Msh3 is a limiting factor in the formation of intergenerational CTG expansions in DM1 transgenic mice. *Hum Genet.* 2006; 119(5):520–526. [PubMed: 16552576]
- Gaddis SS, Wu Q, Thames HD, DiGiovanni J, Walborg EF, MacLeod MC, Vasquez KM. A web-based search engine for triplex-forming oligonucleotide target sequences. *Oligonucleotides.* 2006; 16(2):196–201. [PubMed: 16764543]
- Gal M, Katz T, Ovadia A, Yagil G. TRACTS: A program to map oligopurine, oligopyrimidine and other binary DNA tracts. *Nucleic Acids Res.* 2003; 31(13):3682–3685. [PubMed: 12824393]
- Gomes-Pereira M, Fortune MT, Ingram L, McAbney JP, Monckton DG. Pms2 is a genetic enhancer of trinucleotide CAG. CTG repeat somatic mosaicism: implications for the mechanism of triplet repeat expansion. *Hum Mol Genet.* 2004; 13(16):1815–1825. [PubMed: 15198993]
- Inagaki H, Ohye T, Kogo H, Kato T, Bolor H, Taniguchi M, Shaikh TH, Emanuel BS, Kurahashi H. Chromosomal instability mediated by non-B DNA: cruciform conformation and not DNA sequence is responsible for recurrent translocation in humans. *Genome Res.* 2009; 19(2):191–198. [PubMed: 18997000]

- Jenjaroenpun P, Kuznetsov VA. TTS mapping: integrative WEB tool for analysis of triplex formation target DNA sequences, G-quadruplets and non-protein coding regulatory DNA elements in the human genome. *BMC Genomics*. 2009; 10(Suppl 3):S9. [PubMed: 19958507]
- Kehrer-Sawatzki H, Haussler J, Krone W, Bode H, Jenne DE, Mehnert KU, Tummers U, Assum G. The second case of a t(17;22) in a family with neurofibromatosis type 1: sequence analysis of the breakpoint regions. *Hum Genet*. 1997; 99(2):237–247. [PubMed: 9048928]
- Kikin O, D'Antonio L, Bagga PS. QGRS Mapper: a web-based server for predicting G-quadruplexes in nucleotide sequences. *Nucleic Acids Res*. 2006; 34(Web Server issue):W676–682. [PubMed: 16845096]
- Kostadinov R, Malhotra N, Viotti M, Shine R, D'Antonio L, Bagga P. GRSDb: a database of quadruplex forming G-rich sequences in alternatively processed mammalian pre-mRNA sequences. *Nucleic Acids Res*. 2006; 34(Database issue):D119–124. [PubMed: 16381828]
- Kurahashi H, Inagaki H, Hosoba E, Kato T, Ohye T, Kogo H, Emanuel BS. Molecular cloning of a translocation breakpoint hotspot in 22q11. *Genome Res*. 2007; 17(4):461–469. [PubMed: 17267815]
- Kurahashi H, Inagaki H, Ohye T, Kogo H, Kato T, Emanuel BS. Chromosomal translocations mediated by palindromic DNA. *Cell Cycle*. 2006; 5(12):1297–1303. [PubMed: 16760666]
- Kurahashi H, Inagaki H, Ohye T, Kogo H, Tsutsumi M, Kato T, Tong M, Emanuel BS. The constitutional t(11;22): implications for a novel mechanism responsible for gross chromosomal rearrangements. *Clin Genet*. 2010; 78(4):299–309. [PubMed: 20507342]
- Lawson AR, Hindley GF, Forshew T, Tatevossian RG, Jamie GA, Kelly GP, Neale GA, Ma J, Jones TA, Ellison DW, Sheer D. RAF gene fusion breakpoints in pediatric brain tumors are characterized by significant enrichment of sequence microhomology. *Genome Res*. 2011; 21(4):505–514. [PubMed: 21393386]
- Li H, Xiao J, Li J, Lu L, Feng S, Droge P. Human genomic Z-DNA segments probed by the Z alpha domain of ADAR1. *Nucleic Acids Res*. 2009; 37(8):2737–2746. [PubMed: 19276205]
- Lin Y, Wilson JH. Transcription-induced DNA toxicity at trinucleotide repeats: double bubble is trouble. *Cell Cycle*. 2011; 10(4):611–618. [PubMed: 21293182]
- Lopez Castel A, Cleary JD, Pearson CE. Repeat instability as the basis for human diseases and as a potential target for therapy. *Nat Rev Mol Cell Biol*. 2010; 11(3):165–170. [PubMed: 20177394]
- Manley K, Shirley TL, Flaherty L, Messer A. Msh2 deficiency prevents in vivo somatic instability of the CAG repeat in Huntington disease transgenic mice. *Nat Genet*. 1999; 23(4):471–473. [PubMed: 10581038]
- McMurray CT. Mechanisms of trinucleotide repeat instability during human development. *Nat Rev Genet*. 2010; 11(11):786–799. [PubMed: 20953213]
- Messaed C, Rouleau GA. Molecular mechanisms underlying polyalanine diseases. *Neurobiol Dis*. 2009; 34(3):397–405. [PubMed: 19269323]
- Mirkin SM. Expandable DNA repeats and human disease. *Nature*. 2007; 447(7147):932–940. [PubMed: 17581576]
- Orr HT, Zoghbi HY. Trinucleotide repeat disorders. *Annu Rev Neurosci*. 2007; 30:575–621. [PubMed: 17417937]
- Pearson CE, Nichol Edamura K, Cleary JD. Repeat instability: mechanisms of dynamic mutations. *Nat Rev Genet*. 2005; 6(10):729–742. [PubMed: 16205713]
- Phan AT, Kuryavyi V, Patel DJ. DNA architecture: from G to Z. *Current Opinion in Structural Biology*. 2006; 16(3):288–298. [PubMed: 16714104]
- Punga T, Buhler M. Long intronic GAA repeats causing Friedreich ataxia impede transcription elongation. *EMBO Mol Med*. 2010; 2(4):120–129. [PubMed: 20373285]
- Rimokh R, Rouault JP, Wahbi K, Gadoux M, Lafage M, Archimbaud E, Charrin C, Gentilhomme O, Germain D, Samarut J, et al. A chromosome 12 coding region is juxtaposed to the MYC protooncogene locus in a t(8;12)(q24;q22) translocation in a case of B-cell chronic lymphocytic leukemia. *Genes Chromosomes Cancer*. 1991; 3(1):24–36. [PubMed: 2069907]
- Scaria V, Hariharan M, Arora A, Maiti S. Quadfinder: server for identification and analysis of quadruplex-forming motifs in nucleotide sequences. *Nucleic Acids Res*. 2006; 34(Web Server issue):W683–685. [PubMed: 16845097]

- Schroth GP, Chou PJ, Ho PS. Mapping Z-DNA in the human genome. Computer-aided mapping reveals a nonrandom distribution of potential Z-DNA-forming sequences in human genes. *J Biol Chem*. 1992; 267(17):11846–11855. [PubMed: 1601856]
- Seite P, Leroux D, Hillion J, Monteil M, Berger R, Mathieu-Mahul D, Larsen CJ. Molecular analysis of a variant 18;22 translocation in a case of lymphocytic lymphoma. *Genes Chromosomes Cancer*. 1993; 6(1):39–44. [PubMed: 7680220]
- Shelbourne PF, Keller-McGandy C, Bi WL, Yoon SR, Dubeau L, Veitch NJ, Vonsattel JP, Wexler NS, Arnheim N, Augood SJ. Triplet repeat mutation length gains correlate with cell-type specific vulnerability in Huntington disease brain. *Hum Mol Genet*. 2007; 16(10):1133–1142. [PubMed: 17409200]
- Sheridan MB, Kato T, Haldeman-Englert C, Jalali GR, Milunsky JM, Zou Y, Klaes R, Gimelli G, Gimelli S, Gemmill RM, Drabkin HA, Hacker AM, Brown J, Tomkins D, Shaikh TH, Kurahashi H, Zackai EH, Emanuel BS. A palindrome-mediated recurrent translocation with 3:1 meiotic nondisjunction: the t(8;22)(q24.13;q11.21). *Am J Hum Genet*. 2010; 87(2):209–218. [PubMed: 20673865]
- Sinclair PB, Parker H, An Q, Rand V, Ensor H, Harrison CJ, Strefford JC. Analysis of a breakpoint cluster reveals insight into the mechanism of intrachromosomal amplification in a lymphoid malignancy. *Hum Mol Genet*. 2011; 20(13):2591–2602. [PubMed: 21487021]
- Stankiewicz P, Lupski JR. Structural variation in the human genome and its role in disease. *Annu Rev Med*. 2010; 61:437–455. [PubMed: 20059347]
- Stein LD, Mungall C, Shu S, Caudy M, Mangone M, Day A, Nickerson E, Stajich JE, Harris TW, Arva A, Lewis S. The generic genome browser: a building block for a model organism system database. *Genome Res*. 2002; 12(10):1599–1610. [PubMed: 12368253]
- Thandla SP, Ploski JE, Raza-Egilmez SZ, Chhalliyil PP, Block AW, de Jong PJ, Aplan PD. ETV6-AML1 translocation breakpoints cluster near a purine/pyrimidine repeat region in the ETV6 gene. *Blood*. 1999; 93(1):293–299. [PubMed: 9864173]
- Tong M, Kato T, Yamada K, Inagaki H, Kogo H, Ohye T, Tsutsumi M, Wang J, Emanuel BS, Kurahashi H. Polymorphisms of the 22q11.2 breakpoint region influence the frequency of de novo constitutional t(11;22)s in sperm. *Hum Mol Genet*. 2010; 19(13):2630–2637. [PubMed: 20392709]
- van den Broek WJ, Nelen MR, Wansink DG, Coerwinkel MM, te Riele H, Groenen PJ, Wieringa B. Somatic expansion behaviour of the (CTG)_n repeat in myotonic dystrophy knock-in mice is differentially affected by Msh3 and Msh6 mismatch-repair proteins. *Hum Mol Genet*. 2002; 11(2):191–198. [PubMed: 11809728]
- Vandyke DL, Weiss L, Roberson JR, Babu VR. The frequency and mutation-rate of balanced autosomal rearrangements in man estimated from prenatal genetic-studies for advanced maternal age. *Am J Hum Genet*. 1983; 35(2):301–308. [PubMed: 6837576]
- Wells RD. Non-B DNA conformations, mutagenesis and disease. *Trends Biochem Sci*. 2007; 32(6):271–278. [PubMed: 17493823]
- Wells RD. DNA triplexes and Friedreich ataxia. *FASEB J*. 2008; 22(6):1625–1634. [PubMed: 18211957]
- Wells RD, Dere R, Hebert ML, Napierala M, Son LS. Advances in mechanisms of genetic instability related to hereditary neurological diseases. *Nucleic Acids Res*. 2005; 33(12):3785–3798. [PubMed: 16006624]
- Wheeler VC, Lebel LA, Vrbanc V, Teed A, te Riele H, MacDonald ME. Mismatch repair gene Msh2 modifies the timing of early disease in Hdh(Q111) striatum. *Hum Mol Genet*. 2003; 12(3):273–281. [PubMed: 12554681]
- Wiemels JL, Greaves M. Structure and possible mechanisms of TEL-AML1 gene fusions in childhood acute lymphoblastic leukemia. *Cancer Res*. 1999; 59(16):4075–4082. [PubMed: 10463610]
- Yadav VK, Abraham JK, Mani P, Kulshrestha R, Chowdhury S. QuadBase: genome-wide database of G4 DNA--occurrence and conservation in human, chimpanzee, mouse and rat promoters and 146 microbes. *Nucleic Acids Res*. 2008; 36(Database issue):D381–385. [PubMed: 17962308]
- Zhang R, Lin Y, Zhang CT. Greglist: a database listing potential G-quadruplex regulated genes. *Nucleic Acids Res*. 2008; 36(Database issue):D372–376. [PubMed: 17916572]

- Zhao J, Bacolla A, Wang G, Vasquez KM. Non-B DNA structure-induced genetic instability and evolution. *Cell Mol Life Sci.* 2010; 67(1):43–62. [PubMed: 19727556]
- Zu T, Gibbens B, Doty NS, Gomes-Pereira M, Huguët A, et al. Non-ATG-initiated translation directed by microsatellite expansions. *Proc Natl Acad Sci U S A.* 2011; 108(1):260–265. [PubMed: 21173221]

Internet Resources

- Non-B DB, a database resource for integrated annotations and analysis of non-B DNA-forming motifs.
<http://nonb.abcc.ncifcrf.gov>
- PolyBrowse, ABCC genome browser for variations and annotations.
http://pbrowse3.abcc.ncifcrf.gov/cgi-bin/gb2/gbrowse/Human_37/
- Tandem Repeats Finder. <http://tandem.bu.edu/trf/trf.submit.options.html>
- QuadFinder to find cruciform DNA. <http://miracle.igib.res.in/quadfinder/crux.html>
- QuadBase, a database of quadruplex motifs. <http://quadbase.igib.res.in/>
- Greglist, a database of G-quadruplex regulated genes. <http://tubic.tju.edu.cn/greglist/>
- GRSDB, a database of G-Rich sequences. <http://bioinformatics.ramapo.edu/GRSDB2/>
- Quadruplex forming G-Rich Sequences (QGRS) Mapper.
<http://bioinformatics.ramapo.edu/QGRS/index.php>
- Inverted Repeat Finder, a command line version of the IRF algorithm used to investigate inverted repeat structure of the human genome. <http://tandem.bu.edu/irf/irf.download.html>
- Triplex Target DNA Site (TTS) Mapping.
http://ggeda.bii.a-star.edu.sg/~piroonj/TTS_mapping/TTS_mapping.php
- Triplex-Forming Oligonucleotide Target Sequence Search program. <http://spi.mdanderson.org/tfo/>
- The Tracts program to detect and analyze binary tracts in a DNA sequence.
<http://biportal.weizmann.ac.il/tracts/tracts.html>
- Z-Hunt tool to find Z-DNA. <http://gac-web.cgrb.oregonstate.edu/zDNA/>

non-B DB
A Database for Integrated Annotations and Analysis of non-B DNA Forming Motifs

ABCC

Home | About Us | Collaborators | FAQs | Help | References | Resources

Genomic Database Search Tools

- Search by Feature
- Search by Location
- Search by Feature Attributes
- non-B DNA Motif Search Tool
- Database Statistics
- Contact Us

View in PolyBrowse

non-B DNA Motif Search Tool (nBMST)

Version 1.0

1. Enter an email address (optional).
2. Select one or more non-B DNA motifs.
3. Select one of the two methods of submitting input DNA sequence(s) in [FASTA format](#). Or use single or multiple sequence samples for testing purposes.
4. Enter Captcha information, if not logged in.
5. Click Submit.

For more help with nBMST, visit the [Help](#) and [FAQ](#) pages.

A field with an asterisk (*) before it is a required field.

Email address:

*Non-B DNA motif(s):

- ☒ A-Phased Repeats
- ☒ Direct Repeats and Slipped Motifs
- ☒ G-Quadruplex Forming Repeats
- ☒ Inverted Repeats and Cruciform Motifs
- ☒ Mirror Repeats and Triplex Motifs
- ☒ Z-DNA Motifs

☐ Use the single sequence sample
☐ Use the multi sequence sample
☐ Enter sequence(s)
☒ Upload sequence(s)

*File: (Maximum file size allowed is 20 MB)

*Captcha

Type the characters seen in the Image below to continue.

You are not logged in. If you have registered and wish this query to be recorded for future viewing you must be logged in.
[Register](#) | [Login](#)

Click [here](#) to learn the search criteria used for these motifs.

Figure 1. The nBMST submission page. The 5 steps involved in the submission process are shown. An email address is entered and all the non-B DNA motifs are selected (*grayed area*). A FASTA sequence, NC_007530.2 *Bacillus anthracis* str. 'Ames Ancestor', in this example is uploaded. The captcha characters (4AkA4 in this instance) are entered since the user did not log in.

Home About Us Collaborators FAQs Help References Resources

Genomic Database Search Tools

- Search by Feature
- Search by Location
- Search by Feature Attributes

non-B DNA Motif Search Tool

Database Statistics

Contact Us

View in PBrowse: ▾

nBMST results

Job ID: 8c32923d93
Total non-B motifs found: 6720
Results will be stored until: 10/01/2011 11:45am
Link: <http://nonb.abcc.ncifcrf.gov/apps/nBMST/pickup/8c32923d93>
Dynamic Visualization: [PolyBrowse](#)

Download all files for this job

[Expand all results](#)

▼	a-phased repeats (1189 Results)	Download all files for this motif
▼	direct repeats (280 Results)	Download all files for this motif
▼	g-quadruplex forming repeats (6 Results)	Download all files for this motif
▼	inverted repeats (3945 Results)	Download all files for this motif
▼	mirror repeats (1147 Results)	Download all files for this motif
▼	z-dna motifs (29 Results)	Download all files for this motif

Figure 2.

The nBMST results page. The upper section includes the major statistics of the nBMST run and the lower section displays expandable individual results for each motif type.

Figure 3. Static visualization of the direct repeats as a PNG image. The PNG image may be saved either by right-clicking on it or by clicking on “Download all files for this motif” on the upper right.

Figure 4.

Dynamic visualization of the direct repeats on PolyBrowse page. The PolyBrowse page is created uniquely for each nBMST job submitted and is visible only to the user who submitted the job. In this example, job ID 8c32923d93, a 5 Mbp region of the *Bacillus anthracis* results is displayed here.

Figure 5.

A smaller region, 6.001 kbp, of Figure 4 is zoomed in and all the non-B DNA motif tracks are turned on. The $-/+$ sign (*left of the motif tracks*) may be used to toggle between hiding and showing the results.

Figure 6. Dynamic visualization of multiple FASTA files. When multiple FASTA sequences are submitted, PolyBrowse displays multiple sub-links (*red box*), each representing a specific sequence. Clicking on each sub-link displays the non-B DNA motifs found for that specific sequence.

Figure 7.

PolyBrowse general page layout displaying *File* and *Help* options, the *Landmark or Region* where coordinates of around human c-MYC gene is entered and Human v37 selected as *Data Source*. Also shown are *Gene* tracks including *Refseq Genes*, *mRNAs* and *Promoters*; *Non-B DNA motif* track *G-Quadruplex Forming Repeat* and *Polymorphism* track *dbSNP*. Note that out of the three G-quadruplex repeat motifs (shown in blue glyph), two are predicted within the promoter, MYC_Prom (orange glyph) and one within the first exon of mRNA, NM_002467.3 (gray glyph) of the MYC gene. One dbSNP entry, rs13250910 (red glyph), falls within one G-quadruplex motif.

☒ Synteny ☐ All on ☐ All off		
☒ 1k LiftOver Blocks		
☒ Non-B DNA motifs ☐ All on ☐ All off		
☐ Cruciform Motif	☐ Inverted Repeat	☐ Triplex Motif
☐ Direct Repeat	☐ Mirror Repeat	☐ Z-DNA_Motif
☒ G-Quadruplex Forming Repeat	☐ Slipped Motif	

Figure 8.

PolyBrowse page showing *Non-B DNA motifs* section where annotation track for *G-Quadruplex Forming Repeat* is turned on. Another track turned on is *1k LiftOver Blocks* from *Synteny* section.

Figure 9. PolyBrowse page showing “*trace Gplexes clusters*” tracks where “*trace*” refers to Sanger trace reads which have been mapped against the human genome reference (version 37.1).

Figure 10.

PolyBrowse. 1k LiftOver blocks available for cross species comparison. Note that when *chr8_128748001_L01k* is hover upon, the syntenic regions for other species are displayed. Clicking on a new species takes the user to the selected species. In the example in Figure 13, we clicked on

Figure 11.

PolyBrowse page showing the MYC gene in Chimp v2 as *Data Source*. Note that three G-quadruplex forming repeat motifs are all conserved in the chimpanzee genome sequence. Similar to human, out of the three G-quadruplex repeat motifs (*shown in blue glyph*), two are predicted within the promoter, MYC_Prom (*orange glyph*) and one within the first exon of mRNA, XM_519958.2 (*gray glyph*) of the MYC gene.

Figure 12. The 16 and 17 bp G4 DNA-forming structures reported in Cahoon and Seifert (2009) are captured by nBMST. (A) Details of the G4 motifs in the *Neisseria gonorrhoeae* genome NCCP11945 (B) Details of the G4 motifs in the *Neisseria gonorrhoeae* genome FA 1090 (C) The genomic region of NCCP11945 containing both the 16 and 17 bp G4 DNA-forming repeats as seen in PolyBrowse (D) The details of the 17 bp G4 DNA-forming repeat is obtained by clicking on the green track in (C).

Table 1

The non-B DNA forming motif search criteria

Non-B DNA Motif type (subtypes in bold)	Criteria	Details on non-B DNA subtypes/Notes	Example
Inverted Repeats Cruciform DNA	10–100 nt with reverse complement separated by 0–100 nt spacer	Flagged as Cruciform_Motif if spacer = 0–3 nt	
G-Quadruplex G4 Runs	4 or more G-tracts (3–5 G's) separated by 1–7 nt spacers	Structural preference for short spacers with C's and/or T's	
Direct Repeats Slipped DNA	10–50 nt repeat separated by 0–5 nt spacer	Flagged as Slipped_Motif if spacer = 0 nt	
Mirror Repeats Triplex DNA	10–100 nt mirrored within 0–100 nt spacer	Flagged as Triplex_Motif if 90% Purine or Pyrimidine and spacer = 0–8 nt	
Z-DNA G-Y Runs	G followed by Y (C or T) for at least 10 nt	One strand must contain alternating Gs	GAGCGTGTGTGCGCGCCA
Bent DNA A-Phased Repeats	3 or more A-tracts (3–5 As) 10 nt on center each	Spacers between equal sized A-tracts must contain some non As	

Table 2

Acronyms used in PolyBrowse and nBMST

Acronym	Non-B DNA motifs
DR	Direct Repeat
SM	Slipped Motif
GQFR	G-Quadruplex Forming Repeat
IR	Inverted Repeat
CM	Cruciform Motif
MR	Mirror Repeat
TM	Triplex Motif
ZDM	Z-DNA Motif
APR	A-Phased Repeat

Table 3
The non-B DNA motifs and other DNA features covered in nBMST versus other existing tools

Non-B DNA motif/feature	nBMST	QuadFinder	QGRS Mapper	TTS Mapping	TFO	Z-Hunt II	IRF
Direct repeats	Yes	No	No	No	No	No	No
G-Quadruplex forming repeats	Yes	Yes	Yes	Yes	No	No	No
Inverted repeats	Yes	No	No	No	No	No	Yes
Cruciform motifs	Yes	Yes	No	No	No	No	No
Mirror repeats	Yes	No	No	No	No	No	No
Triplex forming repeats	Yes	No	No	Yes	Yes	No	No
Z-DNA motifs	Yes	No	No	No	No	Yes	No
A-Phased repeats (static bends)	Yes	No	No	No	No	No	No

Table 4

Time taken by nBMST for different sizes of DNA sequence data

Species strain	RefSeq No.	Data size (MB ³)	Seq length (Mbp ⁴)	Total Time ¹	Total motifs ¹	Total Time ²	Total motifs ²
<i>Neisseria gonorrhea</i> FA 1090	NC_002946.2	2.1	2.2	1 min	650	0 hr 53 min	3196
<i>Mycobacterium tuberculosis</i> H37Rv	NC_000962.2	4.3	4.4	1 min	1493	1 hr 50 min	3787
<i>Bacillus anthracis</i> Ames	NC_003997.3	5.1	5.2	2 min	1319	2 hrs 9 min	6704
<i>Pseudomonas fluorescens</i> SBW25	NC_012660.1	6.6	6.7	2 min	1327	2 hrs 45 min	5433
<i>Caenorhabditis elegans</i> chromosome I	NC_003279.6	14.5	15.1	4 min	5528	6 hrs 10 min	57883

¹APR, DR, G-QUADRUPLIX, AND Z-DNA MOTIFS

²APR, DR, G-QUADRUPLIX, Z-DNA MOTIFS, IR, AND MR MOTIFS

³MB, MEGABYTE

⁴MBP, MEGA BASE PAIR