Published online 20 November 2010 *Nucleic Acids Research, 2011, Vol. 39, Database issue* ***D383–D391***
doi:10.1093/nar/gkq1170

Non-B DB: a database of predicted non-B DNA-forming motifs in mammalian genomes

Regina Z. Cer[1], Kevin H. Bruce[1], Uma S. Mudunuri[1], Ming Yi[1], Natalia Volfovsky[1], Brian T. Luke[1], Albino Bacolla[1,2], Jack R. Collins[1] and Robert M. Stephens[1,*]

[1]Advanced Biomedical Computing Center, Information Systems Program, SAIC-Frederick, Inc., NCI-Frederick, Frederick, MD 21702 and [2]Department of Carcinogenesis, The University of Texas M.D. Anderson Cancer Center, Science Park-Research Division, 1808 Park Road 1C, Smithville, TX 78957, USA

Received August 15, 2010; Revised August 28, 2010; Accepted November 1, 2010

ABSTRACT

Although the capability of DNA to form a variety of non-canonical (non-B) structures has long been recognized, the overall significance of these alternate conformations in biology has only recently become accepted *en masse*. In order to provide access to genome-wide locations of these classes of predicted structures, we have developed non-B DB, a database integrating annotations and analysis of non-B DNA-forming sequence motifs. The database provides the most complete list of alternative DNA structure predictions available, including Z-DNA motifs, quadruplex-forming motifs, inverted repeats, mirror repeats and direct repeats and their associated subsets of cruciforms, triplex and slipped structures, respectively. The database also contains motifs predicted to form static DNA bends, short tandem repeats and homo(purine•pyrimidine) tracts that have been associated with disease. The database has been built using the latest releases of the human, chimp, dog, macaque and mouse genomes, so that the results can be compared directly with other data sources. In order to make the data interpretable in a genomic context, features such as genes, single-nucleotide polymorphisms and repetitive elements (SINE, LINE, etc.) have also been incorporated. The database is accessed through query pages that produce results with links to the UCSC browser and a GBrowse-based genomic viewer. It is freely accessible at http://nonb.abcc.ncifcrf.gov.

INTRODUCTION

The ability of certain DNA sequences to adopt alternative conformations, in addition to the canonical Watson–Crick right-handed double helix, has long been recognized (1). Indeed, a large number of studies have documented the formation of alternative (non-B) DNA structures by biophysical methods, including X-ray crystallography (2–4), nuclear magnetic resonance (NMR) spectroscopy (5) and circular dichroism (6). Other methods, such as the detection of single-stranded bases upon non-B DNA structure formation by chemical and enzymatic probes and the relaxation of negative supercoiling by two-dimensional gel electrophoresis have played a major role in revealing the formation of non-B DNA conformations in biological systems (7–9).

Repetitive DNA motifs may fold into non-B DNA structures. Specifically, inverted repeats can adopt cruciform structures, runs of alternating purine–pyrimidine bases are able to switch from the right-handed B- to the left-handed Z-DNA helix, homo(purine•pyrimidine) tracts with mirror repeat symmetry may fold into several types of intramolecular triplexes, four sets of three, four or five guanines, each interrupted by ~1–7 bases, can form highly stable, polymorphic, quadruplex structures and direct repeats can give rise to loops or hairpins through the misalignment of complementary strands, also known as slipped structures (10).

A number of bioinformatic searches have been conducted with the aim of identifying the biological relevance of putative non-B DNA structures in mammalian and other genomes (1). These studies support the notion that the secondary structure conformational domain, rather than the underlying sequence symmetry, often contributes to the control of diverse biological functions, including replication, transcription, immune response

*To whom correspondence should be addressed. Tel: +1 301 846 5787; Fax: +1 301 846 5762; Email: stephensr@mail.nih.gov

(11), recombination and antigenic variation in human pathogens (1,12). Concomitant to this notion, a number of studies have provided circumstantial evidence for the involvement of DNA secondary structures in inducing genetic instability, both in model systems (13–15) and in association with human genetic disease (16–20), including genomic regions that do not contain known genes, suggesting that deeper functional annotation across these regions is warranted. Therefore, the need has arisen to provide the scientific community with a tool that offers a systematic cataloguing of all predicted sequences currently known to potentially form alternative DNA conformations. The non-B DB database bridges this gap by providing a resource for searching, mapping and comparing non-B DNA-forming motifs among various mammalian species.

RESULTS

Non-B DB versus existing databases

To date, several reports have detailed methods aimed at enumerating and evaluating predicted non-B DNA-forming elements from genomic sequences, including QuadBase (21), TTS (22), TRF (23) and others (documented at http://nonb.abcc.ncifcrf.gov/Resources/). These reports use various consensus-based scanning methods for identifying one specific class of predicted non-B DNA structure. In some cases, the identified motifs are screened for the presence of other overlapping functional motifs, such as Sp1 binding sites and CpG islands (24). In other cases, the resulting motifs can be searched by genomic position and scanned for the presence of other nearby non-B DNA predicted features [e.g. triplex sequences near quadruplexes (22)]. More recently, analyses that incorporate thermodynamic values into the overall scoring method (25–27) have been reported. Together, these resources provide an important, yet partial, view into the complexities of locating and characterizing the many different sequence motifs that have the potential of forming non-B DNA structures. Our database expands on these functionalities by including all classes of predicted non-B DNA-forming sequences and by using the latest genome assemblies of human, mouse and other mammalian species. The non-B DNA data are available with current genomic annotation data and polymorphism information. Importantly, non-B DB provides the capacity to visualize the data in a genomic context that is fully integrated with other genomic features, such as genes and single-nucleotide polymorphisms (SNPs). The same interface allows for the users to upload their own annotation data, which are displayed alongside the in-house data through the PolyBrowse and UCSC interfaces.

One of the main difficulties in developing and evaluating algorithms that predict the likely candidates for each class of non-B structures is the lack of large collections of experimental data that have validated their formation *in vivo*. Although most non-B DNA structures can be formed under *in vitro* conditions, the identification of such conformations *in vivo* and the elucidation of parameters that govern their B to non-B equilibria have presented formidable challenges. In addition, these equilibria are influenced by local superhelical density, the presence of nearby DNA unwinding element complexes (DUEs) (28), the transcriptional status, nucleosome assembly and other tissue/temporally regulated biological processes. In light of these considerations, we have taken the approach of using rather broad and general identification methods based exclusively on sequence features; thus, although subsequent filtering of the sampled data is straightforward because of the flexibility provided by the database, our current criteria are expected to include a subset of both false positive and negative hits.

Non-B DB: key features

We have previously reported the construction of a database containing information on mouse indel polymorphisms (30). Herein, we have extended that system to include motifs with the potential to form non-B DNA structures. A number of studies *in vitro* (31–34) and *in vivo* (29,35–38) have indicated that the structural transition from B to non-B DNA is assisted by unrestrained negative supercoiling. In mammalian cells, the global steady-state levels of negative supercoiling vary depending on chromosomal location (39), but are expected to increase transiently by processes, such as transcription, replication and repair, that entail separation of the complementary strands and thus affect nucleosome occupancy (29,38,40–42). However, because the kinetics of these processes may vary among cell types and various developmental stages, an assessment of the probability that a defined chromosomal sequence might exist in the non-B form is currently not available. Indeed, only limited overlap has been reported between the predicted Z-DNA formation based on *in silico* thermodynamic predictions and genomic loci bound to the Zα domain of ADAR1, which displays high specificity for Z-DNA (43). Thus, a combination of factors, including nucleosome occupancy, negative supercoiling, matrix attachment sites, replication, transcription and repair may underlie B to non-B equilibria *in vivo*. In the absence of such information, our search algorithms were based solely on sequence relationships derived from *in vitro* data.

The general approach involves running a scanning application for each specific predicted non-B DNA class against each chromosome (Table 1), including G-quadruplex motifs, alternating purine–pyrimidine sequences, mirror repeats, inverted repeats and direct repeats. Although the 'Mirror Repeat' class as a whole has not been reported to form specific non-B DNA structures, it is included in the database as it is used as a first step in the identification of triplex-forming motifs, i.e. the subset of mirror repeats with purine/pyrimidine content.

The output file in GFF format (http://nonb.abcc.ncifcrf.gov/FAQs/) is then loaded into a MySQL database. The data from all such scans are merged and can be queried and displayed using our local instance of GBrowse called PolyBrowse (44) at http://pbrowse3.abcc.ncifcrf.gov/cgi-bin/gb2/gbrowse/human_37 and several

Table 1. Criteria for predicting non-B DNA-forming motifs in non-B DB

DNA feature	Search criteria	Subset of 'DNA feature' forming non-B DNA	Search criteria for 'Subset of DNA feature'
Inverted repeat	Repeat: 10–100 nt Spacer: 0–100 nt	Cruciform motif	Repeat: 10–100 nt Spacer: 0–3 nt
Mirror repeat	Repeat: 10–100 nt Spacer: 0–100 nt	Triplex motif	Repeat: 10–100 R or Y nt Spacer: 0–8 nt
Direct repeat	Repeat: 10–50 nt Spacer: 0–5 nt	Slipped motif	Repeat: 10–50 nt Spacer: 0 nt
Z-DNA repeat	≥5 units of CG/TG or CG/CA repeats	Whole set	As per the whole set
G-quadruplex forming repeat	Four identical blocks of (3–7) G nt, each block separated by 1–7 nt	Whole set	As per the whole set
A-phased repeat	≥3 runs of A-tracts with 10-bp phasing	Whole set	As per the whole set

Inverted repeat: a pair of DNA sequences, each 10–100 nt in length and separated by a spacer of 0–100 nt, whose sequence composition on the same strand of DNA is such that the bases of the first repeat, when read in the 5′→3′ orientation, are complementary to those of the second repeat read in the 3′→5′ orientation. The term 'complementary' refers to the Watson–Crick hydrogen bonding scheme, whereby A only pairs with T and C only pairs with G. Only perfect inverted repeats that conform to this Watson–Crick pairing scheme are considered.
Cruciform motif: the subset of inverted repeat sequences in which the 'Spacer' comprises 0–3 bases; due to their proximity, this subset of inverted repeat sequences may fold-back and form intramolecular, antiparallel, double helices stabilized by Watson–Crick hydrogen bonds, i.e. a cruciform structure (1,34).
Mirror repeat: a pair of DNA sequences, each 10–100 nt in length and separated by a spacer of 0–100 nt, whose sequence composition on the same strand of DNA is such that the bases of the first repeat, when read in the 5′→3′ orientation, are identical to those of the second repeat read in the 3′→5′ orientation (palindrome); only perfectly matching repeats are included.
Triplex motif: the subset of mirror repeat sequences comprising only purines (R = A and G) [or pyrimidines (Y = C and T)] on the same strand of DNA, and which are separated by few (0–8) nt ('Spacer'). These motifs are able to form various intramolecular three-stranded (triplex, H-DNA) isoforms stabilized by Hoogsteen hydrogen bonds (1,52,53). Only R•Y-containing mirror repeats that may yield A:A•T and G:G•C base triplets (colon indicates Hoogsteen hydrogen bonded bases; dot indicates Watson–Crick hydrogen bonded bases) for the R:R•Y type of intramolecular triplexes and T:A•T and C^+:G•C triplets for the Y:R•Y type of intramolecular triplexes are included since these are considered the most stable triplet combinations.
Direct repeat: two tracts of DNA, each comprising 10–50 nt and separated by 0–5 nt, having the same sequence composition.
Slipped motif: the subset of direct repeat sequences without a spacer (tandem repeats); when aligned in an out-of-register fashion, tandem repeats may give rise to single-stranded loops and/or hairpins (1).
Z-DNA motif: five or more tandem repeats, each comprising an alternating pyrimidine–purine dinucleotide motif, in which the pattern YG is maintained on at least one of the DNA strands; examples include $(CG•CG)_6$, $(CA•TG)_5$ and $[(TG)_3(CG)_4•(CG)_4(CA)_3]$; these motifs may adopt the left-handed Z-DNA conformation (3,54).
G-quadruplex-forming repeat: four blocks, each containing the same number (*n*) of G bases (*n* can vary from 3 to 7), on the plus or minus strand, separated by 1–7 nt; this type of DNA sequence may adopt quadruplex structures (2); overlapping tracts of four G-blocks are also considered.
A-phased repeat: three runs of A bases (A-tracts) in phase with the helical pitch of the DNA double-helix, i.e. 10 bp; an A-tract is defined as a set of A•T base-pairs without a TpA step (47,55–57); three or more tracts of $A_{3–7}$, $T_{3–7}$, AAATTT, AAATTTT and AAAATTT (in any combination) on the plus or minus strand, whose centers are separated by 10 bases, are considered; since A-tracts induce static bends in the DNA double helix, the overall DNA superhelix is expected to display either a left-handed or a right-handed writhe (47,55–57); as mentioned, all the search criteria used herein do not allow for interruptions in the repeats and no thermodynamic information was factored-in in the algorithms used.

GFF-based query tools at http://nonb.abcc.ncifcrf.gov (Figure 1). Importantly, the result pages produced from the queries contain links that allow the user to switch to the genome browser view of that feature, as well as a view that provides the sequence and other annotations for each feature.

These data represent the basis for the non-B DNA annotation information for each species. The scanning criteria do not allow for mismatches within the repeat segments; however, this feature may be added as information becomes available as to the acceptable structural tolerances for each mismatch case. Also, currently not included are very large palindromes (>100 kb), such as those that characterize the Y chromosome and whose recombination is known to lead to spermatogenic failure (45,46). Nevertheless, some aspects related to the presence of mismatches are presented in the polymorphism analysis described below.

After scanning across different mammalian genomes, the numbers of each of the predicted classes of non-B DNA structure-forming motifs appear to be quite variable (Table 2).

As the overall base composition between different mammalian genomes is rather similar (data not shown), the observed differences in the numbers of predicted non-B DNA motifs could simply result from the altered arrangement of bases from one species to another. Alternatively, variations in the population of classes of repetitive elements (SINE, LINE, etc.) among species, or other unknown features, might also contribute to the observed differences. This interspecies variability appears to be uniformly distributed along the entire chromosomes, rather than concentrated in large repetitive clusters (data not shown). Whether these differences play any role or contribute to conferring species-specific differences remains to be investigated.

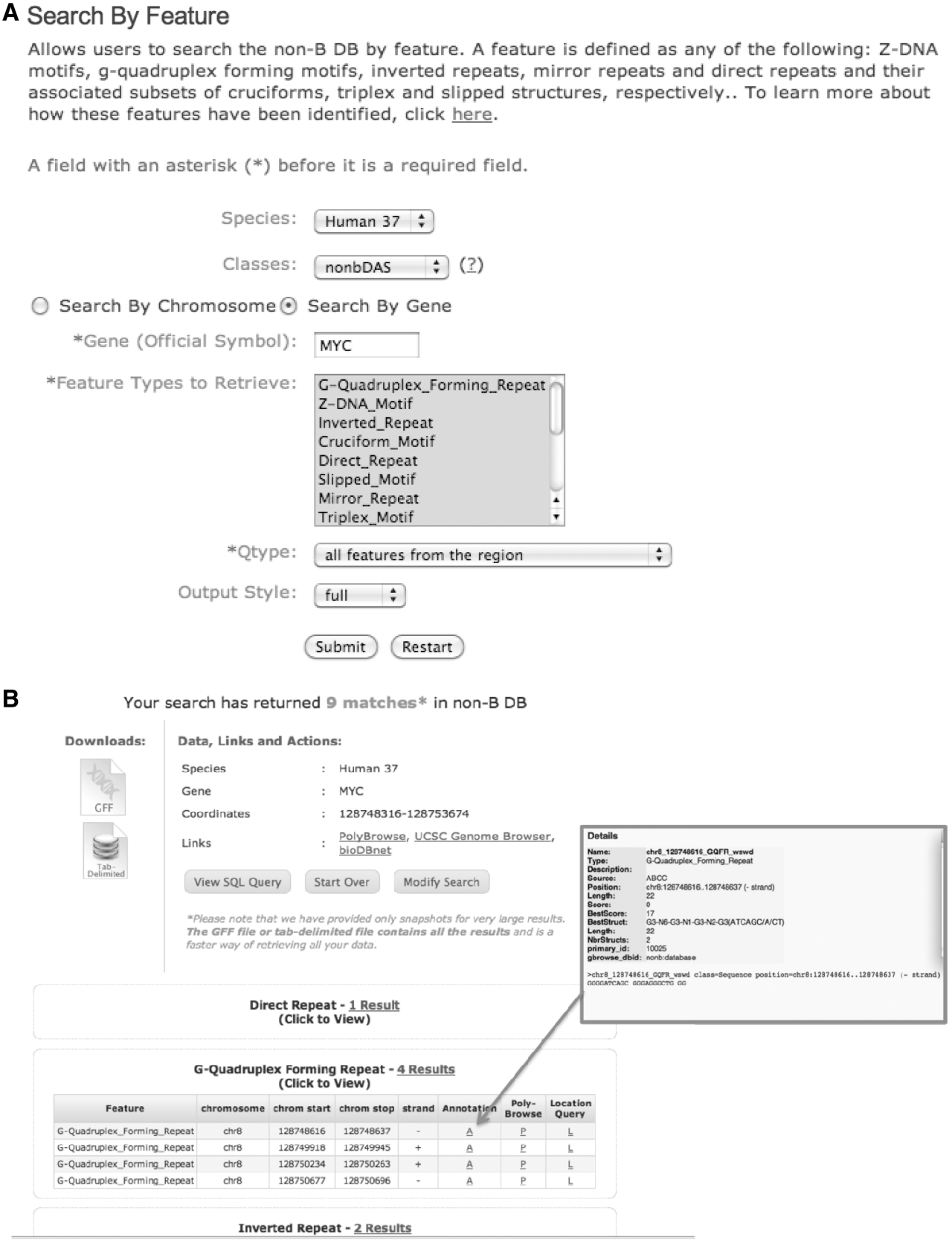

Figure 1. (**A**) Non-B DB user interface. (**B**) Query results page. Non-B DB user interface (A) three query modes are available at the Genomic Database Search Tools page of the non-B DB web interface: Search by Feature (shown), Search by Feature Attributes, and Search by Location (a feature browser). In this example, all non-B DNA motifs were queried by the gene symbol *MYC*. (B) Query results and links to PolyBrowse, UCSC Genome Browser, bioDBnet (51), etc. Users may download the results in GFF format and tab delimited format. Direct access to the sequence and other annotation information for each of the features is available by clicking on the 'A' link (red box).

Table 2. Statistics for DNA repeats and non-B DNA-forming motifs in non-B DB

DNA feature	Human 37	Mouse 37	Dog 2	Chimp 2	Macaque 1
G-quadruplex forming repeat	374 545	559 280	492 535	314 171	298 142
Inverted repeat	1 044 533	801 242	814 080	998 249	843 889
Cruciform motif	197 910	188 532	172 032	190 736	128 334
Direct repeat	871 045	1 593 107	968 955	787 335	765 798
Slipped motif	347 969	695 150	404 750	314 516	305 285
Mirror repeat	16 51 723	3 431 486	1 829 867	14 85 135	14 55 025
Triplex motif	1 79 623	618 928	336 642	1 05 640	1 40 580
Z-DNA repeat	294 320	690 276	261 012	278 928	280 982
A-phased repeat	1 130 731	9 09 653	1 241 082	1 085 591	1 098 030

For the current releases of the five mammalian genomes indicated, the motif searches were performed and the number of features for each class was counted. According to Table 1, the cruciform motifs represent a subset of the inverted repeat class, the slipped motifs represent a subset of the direct repeat class and the triplex motif represents a subset of the mirror repeat class.

A caveat concerning the simple assessment and comparison of the number of non-B DNA-forming repeats among species relates to the criteria used and the counting method. For example, in the G-quadruplex forming sequences, the pattern of a run of 3Gs followed by 1–7 bases repeated four times can be extended, as long as more runs of Gs are encountered, resulting in a single cluster that has the potential to form many substructures. This circumstance needs to be considered when comparing between different reports or methods. Although our approach identifies this finding as a single cluster in the database, separate database tables are provided, in which all possible permutations of the sequence that satisfies the consensus sequence are reported.

In addition to the non-B DNA predicted motifs, the database contains other features of the DNA, such as phased A-tracts that impart static bends to the double-helix and may be involved in nucleosome assembly (47), simple tandem repeats (STR) including triplet repeats whose expansions cause a number of neuromuscular disorders (20) and poly(purine•pyrimidine) tracts, which are characterized by high stacking interactions (48). In addition, NCBI-derived features, such as genes, SNPs and RepeatMasker (http://www.repeatmasker.org/) elements are also included. This integrated information is critical not only for guiding the user visually, but also for enabling queries that combine 'classes', such as 'exons' containing predicted 'Z-DNA' forming sequences, etc.

Cross-species comparisons

One of the main features of the non-B DB is the ability to compare different mammalian genomes for the presence of non-B DNA-forming motifs. This allows for conservation of the predicted elements to be evaluated visually. Figure 2 illustrates this capability by comparing the presence of G-quadruplex forming motifs in the region upstream of the *MYC* locus across the human, chimp, macaque, dog and mouse reference genomes. In order to view syntenic regions in other genomes, the liftOver application from the UCSC website was used to map 1-kb fragments along each chromosome to the corresponding other genomes. These mapped features are called liftOver1k. Areas where a syntenic match failed to be identified (i.e. that region was absent in the other genome, or mapped redundantly) do not show a link to that species. Other non-B DNA tracks available in PolyBrowse will be described in more detail elsewhere (Cer *et. al.*, manuscript in preparation).

Polymorphism analysis

The computed non-B DNA forming elements are likely to be under-represented in our reference genome as their underlying repeats may be polymorphic among individuals. Because this type of information may be critical in the context of gene regulation or predisposition to disease (48), we used a specific parser to scan both the reference human genome as well as additional sequence sources, such as trace reads from the trace archive (http://www.ncbi.nlm.nih.gov/Traces/trace.cgi) and contigs (http://www.ncbi.nlm.nih.gov/projects/WGS/WGSprojectlist.cgi) from personal genome projects (49), for matches to the non-B DNA motifs. Each match found in either the reference or alternate source is then scored for being polymorphic or not. Of the sites identified as polymorphic, a second evaluation is made to determine whether the polymorphism would affect the motif underlying the putative non-B DNA structure. The results of this scan are incorporated into the database as a series of separate tracks (Figure 3B, trace GPlex tracks). Additional information can be gathered by extending this type of analysis to sequence alignments using closely related species. Currently, only the G-quadruplex forming motif supports this type of query.

In order to provide access to the back-end database, we have leveraged two existing tools from the bioinformatics community. The first is a BioPerl (50) set of methods, which is used to query genome databases in various ways, such as by position, by class, or by attribute. This same set of utilities is used within the context of PolyBrowse (44), so that visualization of the genomic features is made available. In addition to linking the outputs from the query tools to the browser for visualization, we also provide links allowing the returned data to be displayed in the familiar UCSC (http://genome.ucsc.edu/) genome browser (Figure 3C), as well as links to

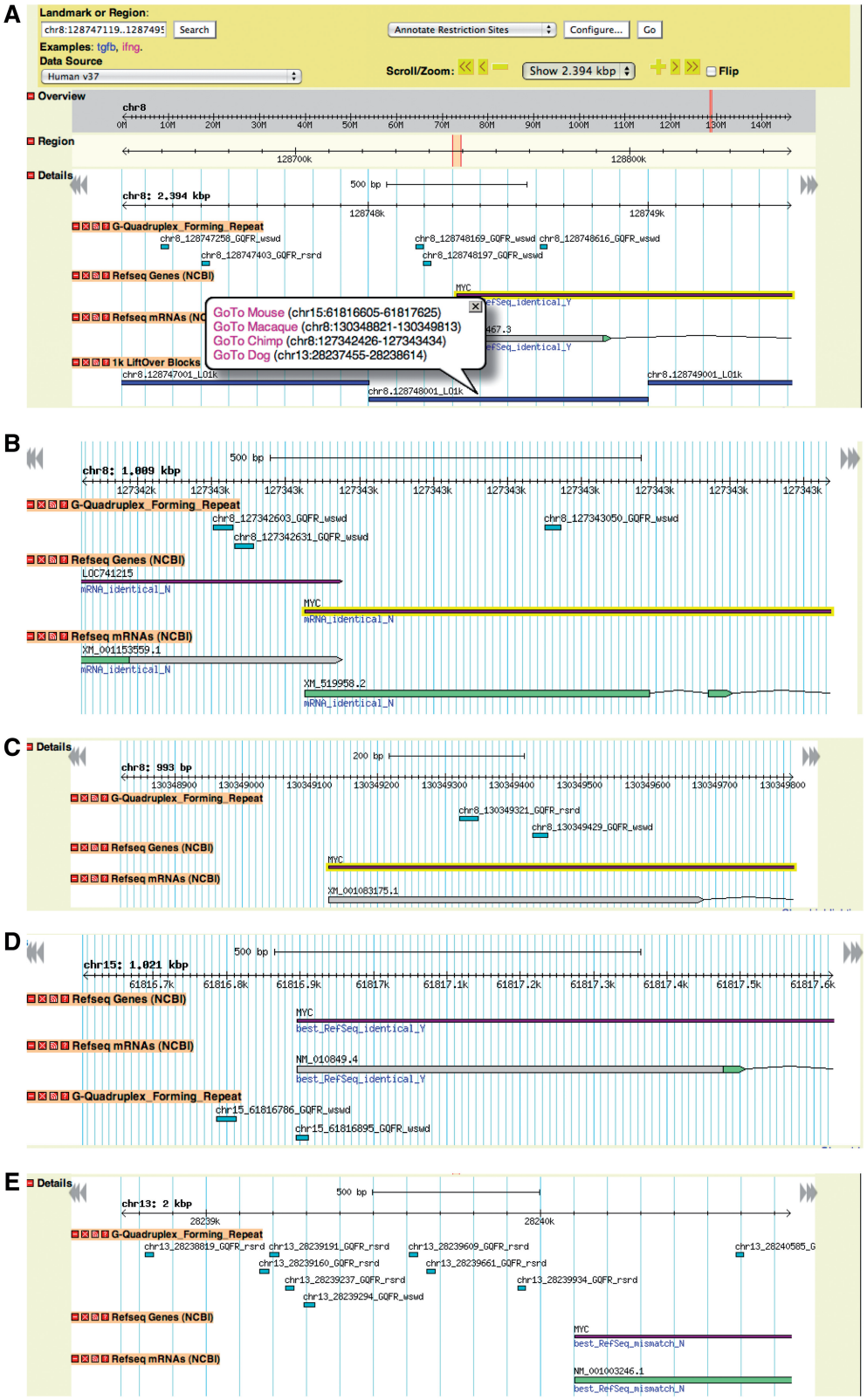

Figure 2. (**A**) PolyBrowse view of G-quadruplex forming repeats in the *MYC* gene with species syntenic information. (**B**) Syntenic region of the human *MYC* gene in chimp. (**C**) Syntenic region of the human *MYC* gene in macaque. (**D**) Syntenic region of the human *MYC* gene in mouse. (**E**) Syntenic region of the human *MYC* gene in dog. Visualization of cross-species information in PolyBrowse: the genomic region shown in Figure 1 is displayed herein to illustrate the syntenic capability. In this case, the G-quadruplex forming repeat track shows the locations of these features in the region near the beginning of the *MYC* gene. The additional track illustrating the liftOver1k blocks is also reported. In that track, when the user moves the cursor over the objects, a popup window appears that contains links (A) to the syntenic locations in the other available species. (B–E) Results upon clicking on the links for chimp, macaque, mouse and dog, respectively. Currently, the liftOver feature only allows moving from the human to the other species but not among the non-human species.

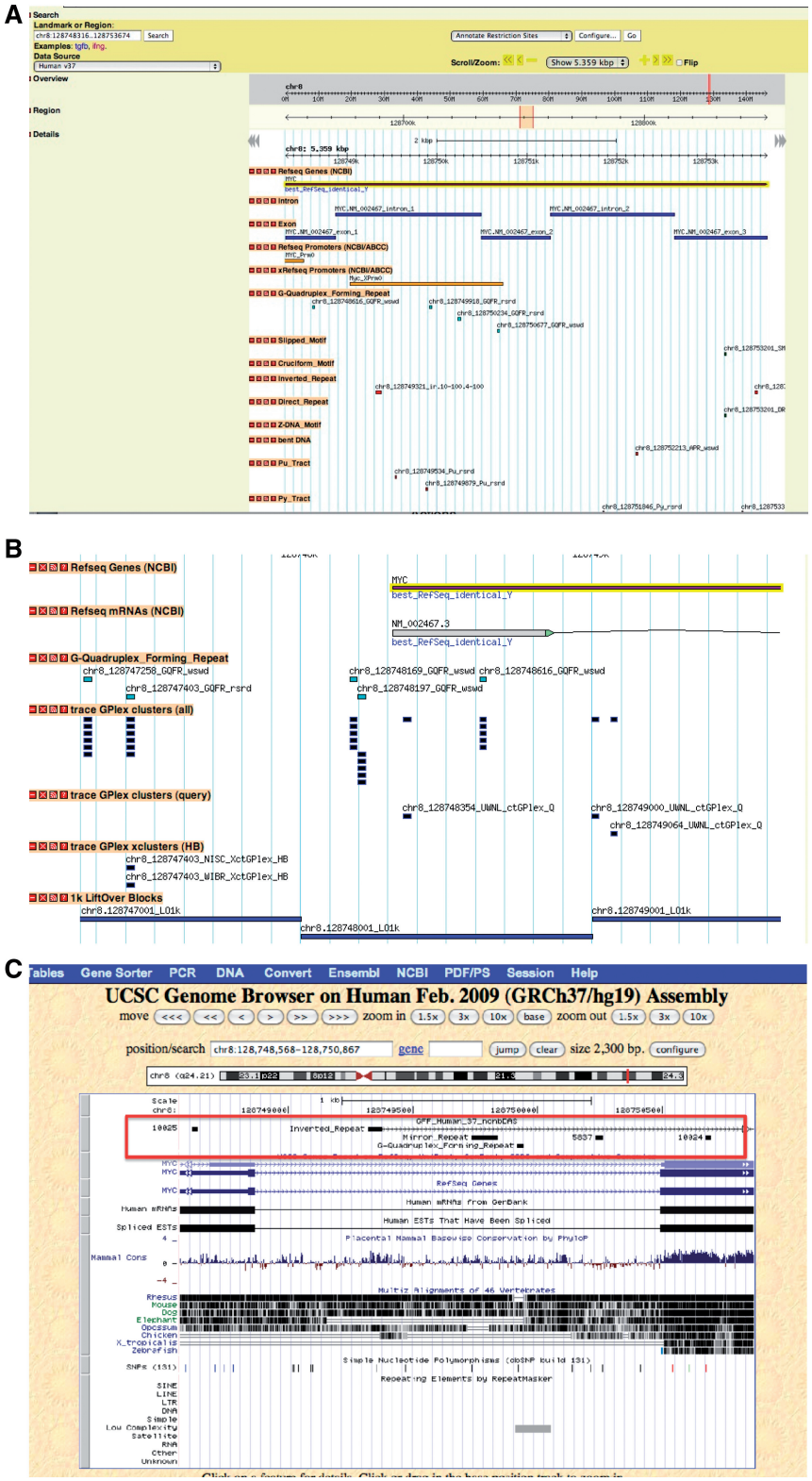

Figure 3. (**A**) PolyBrowse view of some of the non-B DNA features in the human *MYC* gene. (**B**) G-Quadruplex polymorphism tracks. (**C**) UCSC view of tracks. Links to PolyBrowse and UCSC browsers: (A) display of PolyBrowse rendering the *MYC* gene and its promoter with some of the non-B DNA motifs tracks with the PuPy [poly(purine•pyrimidine)] and polymorphism tracks turned on. Several other tracks include features computed at the Advanced Biomedical Computing Center (ABCC), such as STRs, base composition, physical DNA characteristics, mapping, as well as the NCBI derived features, including genes, SNPs, cytogenic markers assembly information, RepeatMasker elements, etc. (data not shown). (B) Display of the *MYC* gene in PolyBrowse showing some of the polymorphism information in non-B DB. Below the gene and mRNA tracks, the searched motifs in the reference sequence are displayed (teal). Below them, the alignments from the trace archive are shown (blue). Some of the predicted motifs are found only in the trace sequences. An additional track shows specific G-quadruplex motifs in which all of the observed trace files contained mutations disrupting the G-quadruplex motif. (C) Display of the *MYC* gene at the UCSC human genome browser (GRCh37/hg19 assembly) as linked to non-B DB from the search shown in Figure 2B. Some of the non-B DNA motifs from non-B DB can be seen in the red rectangular box.

our bioDBnet database warehouse (51), which contains gene-centric information derived from several sources, and additional links.

CONCLUSIONS

Herein, we present a database containing the locations of motifs predicted to adopt the most common non-B DNA structures. The database can be used to browse specific genomic regions for the possible contribution of non-B DNA-forming elements to inherent biological observations derived from the region. In addition to the locations of predicted motifs, the database also contains polymorphism information about each of the test sequences, as well as additional candidate sequences not present within the reference genomes. The database is accessible using both query pages and PolyBrowse. Additional genomes are in the process of being added to the system and will continue to be updated and added as they become available. Input from the community regarding the addition of other tracks, enhanced algorithms for the detection or scoring of the identified motifs or additional query tools are welcome and will be incorporated into the system as appropriate. Further additions, such as a community-based curation capability and the addition of other validation information through literature mining approaches are also under consideration.

We anticipate that significant improvements to our methods will be made in the future by incorporating energetic, and other secondary metrics, to the current predictive algorithms. Although significant biological knowledge would be required, such as localized superhelical density, nucleosome positioning, etc. (see above), the overall goal is to associate a likelihood index with each of the predicted locations for each of the non-B DNA-forming classes. Finally, as reliable methods are expected to be developed that identify genome-wide data on non-B DNA structures *in vivo* and some of the biological parameters involved, the resulting data sets can be used to train the prediction tools, resulting in improved predictive capabilities for each type of non-B-forming classes.

ACKNOWLEDGEMENTS

We thank Dr. Robert Wells for many useful suggestions and Dr. Karen Vasquez for assistance and sharing unpublished data. We also acknowledge the many valuable contributions from the participants at the FASEB Summer Conference on 'Biological Impact of Alternative DNA Structures' held at Steamboat Springs, CO in July 2010. 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.

FUNDING

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.). Funding for open access charge: National Cancer Institute/National Institutes of Health contract HHSN261200800001E.

Conflict of interest statement. None declared.

REFERENCES

1. Zhao,J., Bacolla,A., Wang,G. and Vasquez,K.M. (2010) Non-B DNA structure-induced genetic instability and evolution. *Cell. Mol. Life Sci.*, **67**, 43–62.
2. Neidle,S. and Parkinson,G.N. (2008) Quadruplex DNA crystal structures and drug design. *Biochimie*, **90**, 1184–1196.
3. Wang,A.J., Quigley,G.J., Kolpak,F.J., van der Marel,G., van Boom,J.H. and Rich,A. (1981) Left-handed double helical DNA: variations in the backbone conformation. *Science*, **211**, 171–176.
4. Chandrasekhar,S., Naik,T.R., Nayak,S.K. and Row,T.N. (2010) Crystal structure of an intermolecular 2:1 complex between adenine and thymine. Evidence for both Hoogsteen and 'quasi-Watson–Crick' interactions. *Bioorg. Med. Chem. Lett.*, **20**, 3530–3533.
5. Patel,D.J., Phan,A.T. and Kuryavyi,V. (2007) Human telomere, oncogenic promoter and 5′-UTR G-quadruplexes: diverse higher order DNA and RNA targets for cancer therapeutics. *Nucleic Acids Res.*, **35**, 7429–7455.
6. Kypr,J., Kejnovska,I., Renciuk,D. and Vorlickova,M. (2009) Circular dichroism and conformational polymorphism of DNA. *Nucleic Acids Res.*, **37**, 1713–1725.
7. Lilley,D.M.J. and Dahlberg,J.E. (eds), (1992), DNA Structures part B: chemical and electrophoretic analysis of DNA, *Methods Enzymol*, Vol. 212. Elsevier/Academic Press, San Diego, CA, pp. 139–155.
8. Mirkin,S.M. (2008) Discovery of alternative DNA structures: a heroic decade (1979–1989). *Front. Biosci.*, **13**, 1064–1071.
9. Rich,A. and Zhang,S. (2003) Timeline: Z-DNA: the long road to biological function. *Nat. Rev. Genet.*, **4**, 566–572.
10. Bacolla,A. and Wells,R.D. (2004) Non-B DNA conformations, genomic rearrangements, and human disease. *J. Biol. Chem.*, **279**, 47411–47414.
11. Ha,S.C., Kim,D., Hwang,H.Y., Rich,A., Kim,Y.G. and Kim,K.K. (2008) The crystal structure of the second Z-DNA binding domain of human DAI (ZBP1) in complex with Z-DNA reveals an unusual binding mode to Z-DNA. *Proc. Natl Acad. Sci. USA*, **105**, 20671–20676.
12. Hill,S.A. and Davies,J.K. (2009) Pilin gene variation in *Neisseria gonorrhoeae*: reassessing the old paradigms. *FEMS Microbiol. Rev.*, **33**, 521–530.
13. Glickman,B.W. and Ripley,L.S. (1984) Structural intermediates of deletion mutagenesis: a role for palindromic DNA. *Proc. Natl Acad. Sci. USA*, **81**, 512–516.
14. Akgun,E., Zahn,J., Baumes,S., Brown,G., Liang,F., Romanienko,P.J., Lewis,S. and Jasin,M. (1997) Palindrome resolution and recombination in the mammalian germ line. *Mol. Cell. Biol.*, **17**, 5559–5570.
15. Gordenin,D.A., Lobachev,K.S., Degtyareva,N.P., Malkova,A.L., Perkins,E. and Resnick,M.A. (1993) Inverted DNA repeats: a source of eukaryotic genomic instability. *Mol. Cell. Biol.*, **13**, 5315–5322.
16. Sheridan,M.B., Kato,T., Haldeman-Englert,C., Jalali,G.R., Milunsky,J.M., Zou,Y., Klaes,R., Gimelli,G., Gimelli,S., Gemmill,R.M. *et al.* (2010) A palindrome-mediated recurrent translocation with 3:1 meiotic nondisjunction: the t(8;22)(q24.13;q11.21). *Am. J. Hum. Genet.*, **87**, 209–218.
17. Kurahashi,H., Inagaki,H., Ohye,T., Kogo,H., Tsutsumi,M., Kato,T., Tong,M. and Emanuel,B.S. (2010) The constitutional t(11;22): implications for a novel mechanism responsible for gross chromosomal rearrangements. *Clin. Genet.*, **78**, 299–309.
18. Carvalho,C.M., Zhang,F., Liu,P., Patel,A., Sahoo,T., Bacino,C.A., Shaw,C., Peacock,S., Pursley,A., Tavyev,Y.J. *et al.*

(2009) Complex rearrangements in patients with duplications of MECP2 can occur by fork stalling and template switching. *Hum. Mol. Genet.*, **18**, 2188–2203.
19. D'Angelo,C.S., Gajecka,M., Kim,C.A., Gentles,A.J., Glotzbach,C.D., Shaffer,L.G. and Koiffmann,C.P. (2009) Further delineation of nonhomologous-based recombination and evidence for subtelomeric segmental duplications in 1p36 rearrangements. *Hum. Genet.*, **125**, 551–563.
20. Wells,R.D. and Ashizawa,T. (2006) *Genetic Instabilities and Neurological Diseases*, 2nd edn. Elsevier/Academic Press, San Diego, CA.
21. Yadav,V.K., Abraham,J.K., Mani,P., Kulshrestha,R. and Chowdhury,S. (2008) QuadBase: genome-wide database of G4 DNA–occurrence and conservation in human, chimpanzee, mouse and rat promoters and 146 microbes. *Nucleic Acids Res.*, **36**, D381–D385.
22. Jenjaroenpun,P. and Kuznetsov,V.A. (2009) 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*, **10 (Suppl. 3)**, S9.
23. Benson,G. (1999) Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res.*, **27**, 573–580.
24. Eddy,J. and Maizels,N. (2008) Conserved elements with potential to form polymorphic G-quadruplex structures in the first intron of human genes. *Nucleic Acids Res.*, **36**, 1321–1333.
25. Ho,P.S., Ellison,M.J., Quigley,G.J. and Rich,A. (1986) A computer aided thermodynamic approach for predicting the formation of Z-DNA in naturally occurring sequences. *EMBO J.*, **5**, 2737–2744.
26. Ho,P.S. (2009) Thermogenomics: thermodynamic-based approaches to genomic analyses of DNA structure. *Methods*, **47**, 159–167.
27. Schroth,G.P., Chou,P.J. and Ho,P.S. (1992) Mapping Z-DNA in the human genome. Computer-aided mapping reveals a nonrandom distribution of potential Z-DNA-forming sequences in human genes. *J. Biol. Chem.*, **267**, 11846–11855.
28. Chowdhury,A., Liu,G., Kemp,M., Chen,X., Katrangi,N., Myers,S., Ghosh,M., Yao,J., Gao,Y., Bubulya,P. *et al.* (2010) The DNA unwinding element binding protein DUE-B interacts with Cdc45 in preinitiation complex formation. *Mol. Cell. Biol.*, **30**, 1495–1507.
29. Wittig,B., Dorbic,T. and Rich,A. (1989) The level of Z-DNA in metabolically active, permeabilized mammalian cell nuclei is regulated by torsional strain. *J. Cell Biol.*, **108**, 755–764.
30. Akagi,K., Stephens,R.M., Li,J., Evdokimov,E., Kuehn,M.R., Volfovsky,N. and Symer,D.E. (2010) MouseIndelDB: a database integrating genomic indel polymorphisms that distinguish mouse strains. *Nucleic Acids Res.*, **38**, D600–D606.
31. Singleton,C.K., Klysik,J., Stirdivant,S.M. and Wells,R.D. (1982) Left-handed Z-DNA is induced by supercoiling in physiological ionic conditions. *Nature*, **299**, 312–316.
32. Courey,A.J. and Wang,J.C. (1988) Influence of DNA sequence and supercoiling on the process of cruciform formation. *J. Mol. Biol.*, **202**, 35–43.
33. Collier,D.A., Griffin,J.A. and Wells,R.D. (1988) Non-B right-handed DNA conformations of homopurine.homopyrimidine sequences in the murine immunoglobulin C alpha switch region. *J. Biol. Chem.*, **263**, 7397–7405.
34. Lilley,D.M., Gough,G.W., Hallam,L.R. and Sullivan,K.M. (1985) The physical chemistry of cruciform structures in supercoiled DNA molecules. *Biochimie*, **67**, 697–706.
35. Dayn,A., Malkhosyan,S. and Mirkin,S.M. (1992) Transcriptionally driven cruciform formation in vivo. *Nucleic Acids Res.*, **20**, 5991–5997.
36. Krasilnikov,A.S., Podtelezhnikov,A., Vologodskii,A. and Mirkin,S.M. (1999) Large-scale effects of transcriptional DNA supercoiling in vivo. *J. Mol. Biol.*, **292**, 1149–1160.
37. Bacolla,A., Jaworski,A., Connors,T.D. and Wells,R.D. (2001) *PKD1* unusual DNA conformations are recognized by nucleotide excision repair. *J. Biol. Chem.*, **276**, 18597–18604.
38. Kouzine,F., Sanford,S., Elisha-Feil,Z. and Levens,D. (2008) The functional response of upstream DNA to dynamic supercoiling *in vivo*. *Nat. Struct. Mol. Biol.*, **15**, 146–154.
39. Kramer,P.R. and Sinden,R.R. (1997) Measurement of unrestrained negative supercoiling and topological domain size in living human cells. *Biochemistry*, **36**, 3151–3158.
40. Jimenez-Ruiz,A., Zhang,Q. and Shen,C.K. (1995) In vivo binding of trimethylpsoralen detects DNA structural alterations associated with transcribing regions in the human beta-globin cluster. *J. Biol. Chem.*, **270**, 28978–28981.
41. Leonard,M.W. and Patient,R.K. (1991) Evidence for torsional stress in transcriptionally activated chromatin. *Mol. Cell. Biol.*, **11**, 6128–6138.
42. Ristic,D., Wyman,C., Paulusma,C. and Kanaar,R. (2001) The architecture of the human Rad54–DNA complex provides evidence for protein translocation along DNA. *Proc. Natl Acad. Sci. USA*, **98**, 8454–8460.
43. Li,H., Xiao,J., Li,J., Lu,L., Feng,S. and Droge,P. (2009) Human genomic Z-DNA segments probed by the Z alpha domain of ADAR1. *Nucleic Acids Res.*, **37**, 2737–2746.
44. Stein,L.D., Mungall,C., Shu,S., Caudy,M., Mangone,M., Day,A., Nickerson,E., Stajich,J.E., Harris,T.W., Arva,A. *et al.* (2002) The generic genome browser: a building block for a model organism system database. *Genome Res.*, **12**, 1599–1610.
45. Kuroda-Kawaguchi,T., Skaletsky,H., Brown,L.G., Minx,P.J., Cordum,H.S., Waterston,R.H., Wilson,R.K., Silber,S., Oates,R., Rozen,S. *et al.* (2001) The AZFc region of the Y chromosome features massive palindromes and uniform recurrent deletions in infertile men. *Nat. Genet.*, **29**, 279–286.
46. Lange,J., Skaletsky,H., van Daalen,S.K., Embry,S.L., Korver,C.M., Brown,L.G., Oates,R.D., Silber,S., Repping,S. and Page,D.C. (2009) Isodicentric Y chromosomes and sex disorders as byproducts of homologous recombination that maintains palindromes. *Cell*, **138**, 855–869.
47. Rohs,R., West,S.M., Sosinsky,A., Liu,P., Mann,R.S. and Honig,B. (2009) The role of DNA shape in protein–DNA recognition. *Nature*, **461**, 1248–1253.
48. Bacolla,A., Larson,J.E., Collins,J.R., Li,J., Milosavljevic,A., Stenson,P.D., Cooper,D.N. and Wells,R.D. (2008) Abundance and length of simple repeats in vertebrate genomes are determined by their structural properties. *Genome Res.*, **18**, 1545–1553.
49. Molla,M., Delcher,A., Sunyaev,S., Cantor,C. and Kasif,S. (2009) Triplet repeat length bias and variation in the human transcriptome. *Proc. Natl Acad. Sci. USA*, **106**, 17095–17100.
50. Stajich,J.E. (2007) An introduction to BioPerl. *Methods Mol. Biol.*, **406**, 535–548.
51. Mudunuri,U., Che,A., Yi,M. and Stephens,R.M. (2009) bioDBnet: the biological database network. *Bioinformatics*, **25**, 555–556.
52. Wells,R.D., Collier,D.A., Hanvey,J.C., Shimizu,M. and Wohlrab,F. (1988) The chemistry and biology of unusual DNA structures adopted by oligopurine.oligopyrimidine sequences. *FASEB J.*, **2**, 2939–2949.
53. Frank-Kamenetskii,M.D. and Mirkin,S.M. (1995) Triplex DNA structures. *Annu. Rev. Biochem.*, **64**, 65–95.
54. Ho,P.S. (1994) The non-B-DNA structure of d(CA/TG)n does not differ from that of Z-DNA. *Proc. Natl Acad. Sci. USA*, **91**, 9549–9553.
55. Barbic,A., Zimmer,D.P. and Crothers,D.M. (2003) Structural origins of adenine-tract bending. *Proc. Natl Acad. Sci. USA*, **100**, 2369–2373.
56. Stefl,R., Wu,H., Ravindranathan,S., Sklenar,V. and Feigon,J. (2004) DNA A-tract bending in three dimensions: solving the dA4T4 vs. dT4A4 conundrum. *Proc. Natl Acad. Sci. USA*, **101**, 1177–1182.
57. Lankas,F., Spackova,N., Moakher,M., Enkhbayar,P. and Sponer,J. (2010) A measure of bending in nucleic acids structures applied to A-tract DNA. *Nucleic Acids Res.*, **38**, 3414–3422.