GBshape: a genome browser database for DNA shape annotations

Tsu-Pei Chiu[1,†], Lin Yang[1,†], Tianyin Zhou[1], Bradley J. Main[1], Stephen C.J. Parker[2], Sergey V. Nuzhdin[1], Thomas D. Tullius[3] and Remo Rohs[1,4,*]

[1]Molecular and Computational Biology Program, Department of Biological Sciences, University of Southern California, Los Angeles, CA 90089, USA, [2]Departments of Computational Medicine and Bioinformatics and Human Genetics, University of Michigan, Ann Arbor, MI 48109, USA, [3]Department of Chemistry and Program in Bioinformatics, Boston University, Boston, MA 02215, USA and [4]Departments of Chemistry, Physics, and Computer Science, University of Southern California, Los Angeles, CA 90089, USA

ABSTRACT

Many regulatory mechanisms require a high degree of specificity in protein-DNA binding. Nucleotide sequence does not provide an answer to the question of why a protein binds only to a small subset of the many putative binding sites in the genome that share the same core motif. Whereas higher-order effects, such as chromatin accessibility, cooperativity and cofactors, have been described, DNA shape recently gained attention as another feature that fine-tunes the DNA binding specificities of some transcription factor families. Our Genome Browser for DNA shape annotations (GBshape; freely available at http://rohslab.cmb.usc.edu/GBshape/) provides minor groove width, propeller twist, roll, helix twist and hydroxyl radical cleavage predictions for the entire genomes of 94 organisms. Additional genomes can easily be added using the GBshape framework. GBshape can be used to visualize DNA shape annotations qualitatively in a genome browser track format, and to download quantitative values of DNA shape features as a function of genomic position at nucleotide resolution. As biological applications, we illustrate the periodicity of DNA shape features that are present in nucleosome-occupied sequences from human, fly and worm, and we demonstrate structural similarities between transcription start sites in the genomes of four *Drosophila* species.

INTRODUCTION

DNA shape analysis has been established in recent years as an approach that reveals determinants of protein-DNA binding specificity beyond the primary nucleotide sequence (1–4). Interactions between nucleotides within a binding site or its flanks are implicitly contained in the 3D structure of a DNA binding site. DNA shape is influenced by the core motif (5) and its flanking sequences (6) and therefore potentially characterizes binding sites with higher precision. In addition to taking into account interrelationships between nucleotide positions, DNA shape integrates over diverse nucleotide sequences that can give rise to similar DNA shapes, a phenomenon known as degeneracy of DNA sequence and structure. As a consequence, DNA shape was found to be evolutionarily conserved to a higher degree than is DNA sequence (7).

Based on these findings it seems advantageous to incorporate DNA shape features in motif scanning and *de novo* motif discovery methods (8–11). Another application for DNA shape analysis would be in the functional evaluation of genetic variation, which is commonly described in terms of nucleotide sequence (12,13). These and other applications will require the mapping of DNA shape features for entire genomes. To make the necessary data available we developed GBshape. Prediction of DNA shape features from nucleotide sequence is based on high-throughput methods for deriving DNA shape features, by using pentamers to mine results from all-atom Monte Carlo simulations of DNA fragments (14–16), and by predicting hydroxyl radical cleavage patterns based on an experimental dataset (17).

GBshape is a multi-species database currently containing whole-genome data for 94 organisms from groups of diverse species (Table 1). For each organism the database provides four genome browser tracks with annotations for Minor Groove Width (MGW), Propeller Twist (ProT), Roll and Helix Twist (HelT) (14). In a fifth track, GBshape shows hydroxyl radical cleavage annotations from the ·OH Radical Cleavage Intensity Database for double-stranded DNA (ORChID2) (18). These five DNA shape annotations were

---

*To whom correspondence should be addressed. Tel: +1 213 7400552; Fax: +1 213 8214257; Email: rohs@usc.edu
†The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.

Table 1. Current number of genomes from diverse species in GB-shape listed by UCSC Genome Browser organism group with additional groups added.

Organism group	Genome count
Mammals	47
Vertebrates	19
Deuterostomes	3
Insects	14
Nematodes	6
Fungi	1
Plants	1
Protists	1
Bacteria	1
Others	1
Total	94

generated with the high-throughput prediction platform. DNA shape data can be visualized either qualitatively or downloaded as quantitative values via the GBshape user interface. GBshape contains DNA shape annotations for 91 genomes taken from the UCSC Genome Browser (19) and three additional genomes from plants, parasitic protists and bacteria (Table 1).

We demonstrate the value of analyzing DNA shape annotations using GBshape by comparing the structural features of *in vivo* nucleosome binding sites from worm, fly and human (20) and the evolutionary conservation of DNA shape at transcription start sites (TSSs) across multiple *Drosophila* species (21). The GBshape database completes the family of DNA shape tools that includes DNAshape, a web server for high-throughput prediction of DNA shape features for up to 1 million base pairs (14), TFBSshape, a database of DNA shape features of transcription factor binding site motifs (22), and ORChID, a database and prediction tool for hydroxyl radical cleavage patterns (17,18).

DATABASE

Database architecture and methodology

The core of our database is a high-throughput prediction platform (Figure 1) that we developed to generate DNA shape data for storage in GBshape. Whole genome sequence files (in FASTA format) for multiple species are subjected to the high-throughput prediction programs DNAshape (14) and ORChID2 (18) that are embedded in the GBshape platform. The GBshape prediction platform was designed to be extendable by plug-ins of other whole-genome annotation programs (Figure 1). The results of high-throughput prediction programs are converted to the bigWig data format, which can be displayed in a genome browser. The platform was developed in C++ and runs on a high-performance computing cluster (HPCC).

The GBshape database combines DNA shape data with standard UCSC Genome Browser annotations (19) and whole-genome sequence data. The sequences of 91 genomes (Supplementary Table S1) and corresponding standard annotations were downloaded from the UCSC Genome Browser (19). Although several genome assembly versions are available for many of these species, at this stage of development the most recent genome assembly for each species was chosen. The reference genome from *Saccharomyces*
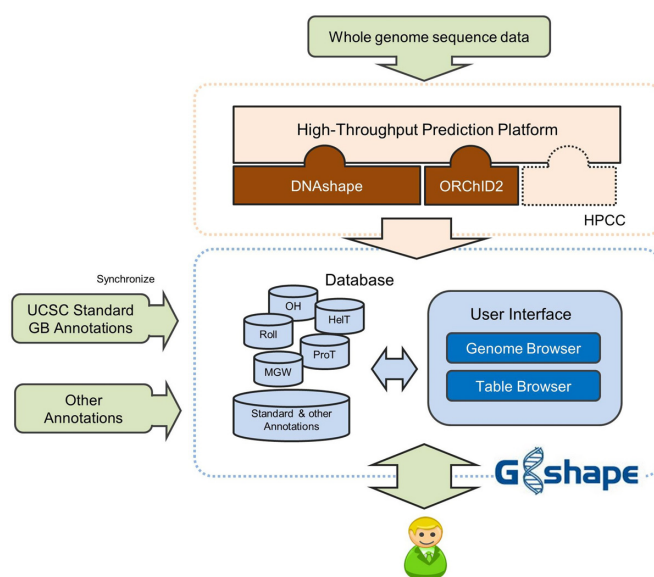

Figure 1. Architecture of the GBshape database. GBshape consists of a high-throughput prediction platform, data depositories and a user interface. DNA shape annotations of entire genomes can be generated by the high-throughput prediction platform, which runs on a high-performance computing cluster (HPCC), and, together with genome sequences and UCSC Genome Browser standard annotations, stored in the data depositories. The user interface provides multiple functionalities for users to either visualize or download structural annotations.

cerevisiae was identical with the one provided by the Saccharomyces Genome Database (23). Three additional reference genomes from *Arabidopsis thaliana* (24), *Plasmodium falciparum* (25) and *Escherichia coli* (26) that were not present in the UCSC Genome Browser were added to GBshape (Supplementary Table S1). The GBshape framework enables an easy expansion to additional genome assemblies, and users can submit a web form requesting the addition of specific genomes to our database. The GBshape database runs on MySQL (Figure 1).

The GBshape tracks for MGW, ProT, Roll and HelT were generated using our high-throughput method DNAshape (14). These DNA shape features were selected based on prior experimental studies demonstrating their important role in protein-DNA recognition, and include MGW (27–29), ProT (6), Roll (30) and HelT (28). Using pentamers as sliding windows, DNAshape mines all-atom Monte Carlo simulations (15,31) of 2121 DNA fragments of 10–27 bp in length. Each of the 512 unique pentamers is assigned the average value of all of its occurrences in the dataset at the central nucleotide for MGW and ProT and at the two central base pair (bp) steps of the pentamer for Roll and HelT. Each pentamer occurs on average 44 times in our Monte Carlo-generated dataset. The DNAshape method was validated against experimental data from X-ray crystallography, nuclear magnetic resonance spectroscopy and hydroxyl radical cleavage measurements (14).

We have shown that the hydroxyl radical, a small, uncharged, highly reactive molecule, reacts with the backbone of naked DNA in a manner that reflects the solvent accessible surface areas of the hydrogen atoms of the deoxyribose sugar, thus providing an experimental image of DNA
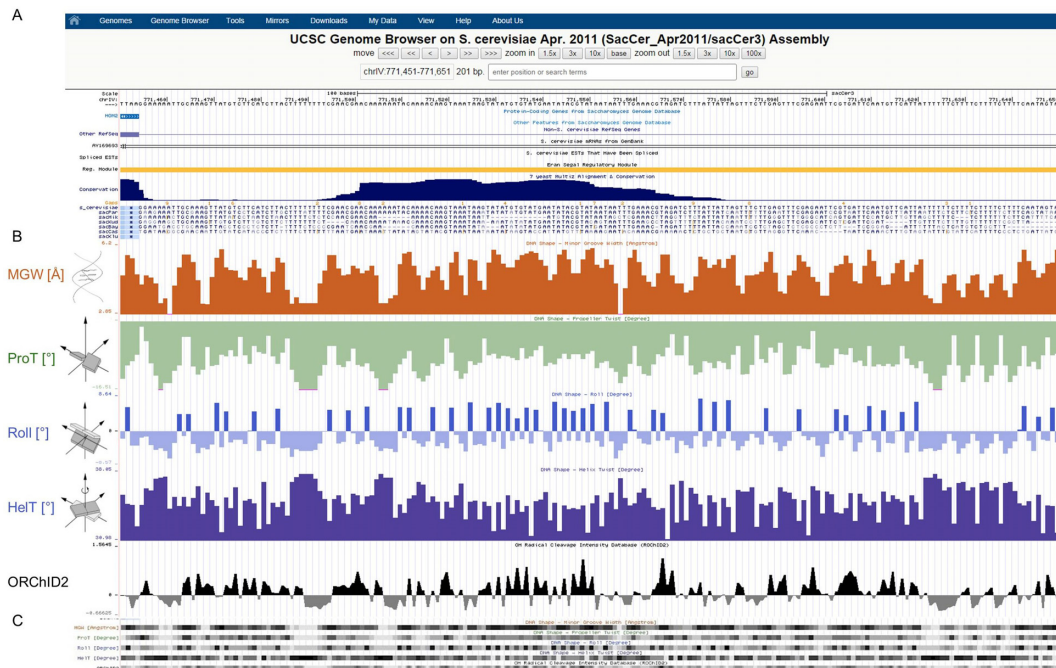
Figure 2. Visual display of GBshape annotations in the genome browser for a specific position in the *S. cerevisiae* genome. (A) Genome positions and UCSC Genome Browser standard annotation tracks. (B) DNA shape annotation tracks for MGW, ProT, Roll, HelT and hydroxyl radical cleavage intensity (ORChID2). (C) Heat map views for DNA shape annotations.

backbone shape (18,32). To develop this chemical approach into a high-throughput method we performed hydroxyl radical cleavage experiments on 150 diverse DNA fragments of 40 bp in length. We devised a prediction algorithm, based on this database of experimental cleavage patterns, that uses a sliding tetramer window to predict the cleavage pattern for DNA sequences of any length (17). We subsequently extended this method by averaging the predicted cleavage patterns of both DNA strands to develop OR-ChID2, which we previously showed to be correlated with MGW and electrostatic potential (18). Thus, the ORChID2 pattern provides an experiment-based prediction of minor groove shape, which complements the Monte Carlo-based DNA shape features as an additional annotation track in GBshape.

User interface

The GBshape user interface is a customized version of the UCSC Genome Browser that is hosted on our local server. The user interface contains some important functionalities of the UCSC Genome Browser, including the genome browser, table browser, the Basic Local Alignment Search Tool-like alignment tool (BLAT) and the 'add custom tracks' tool. The GBshape interface runs on a Linux-operated dual-core IBM server with Apache.

GBshape consists of two major tools—a genome browser and a table browser. The genome browser provides a graphical representation of DNA shape annotations along with standard genome browser annotations. The genome browser also supports text and sequence search functions to provide easy access to genomic regions of interest. The table browser enables data manipulation, downloads of multiple

records and basic statistical analyses, which cannot be performed with the genome browser function.

To visualize DNA shape annotations the user clicks on 'Genome Browser' in the navigation bar on the left of the GBshape homepage. On the Genome Browser Gateway page the user chooses an organism group, species genome, genome assembly, genome position and search terms of interest. After the 'submit' button is pressed, consolidated results for DNA shape annotations, together with standard genome annotations (Figure 2A), are shown on the display page. The shape annotations MGW, ProT, Roll, HelT and ORChID2 can be shown as quantitative plots (Figure 2B) or condensed into heat maps (Figure 2C).

The sequence-alignment tool, BLAT, can be used to search specific regions of the genome based on sequence similarity. To use BLAT, click on 'Tools' in the navigation bar at the top of the Genome Browser Gateway page, select 'Blat' in the pull-down menu, select a genome, assembly, query type, sort output and output type, and then press the 'submit' button. Genomic annotations can be viewed by clicking on the 'browser' link at the left of the search results. Supplementary Table S2 provides information on genomes for which BLAT supports a sequence search.

The view of the genome browser can be adjusted by using the buttons located near the top of the display page to move along the genome sequence, zoom in or zoom out, or by dragging and zooming the genomic position. The display type of an annotation track can be changed by selecting the pull-down menu from the track control panel at the bottom of the page. A heat map view can be shown for a track by setting the display type as 'dense' on the corresponding control panel. Users can upload their own tracks to compare with
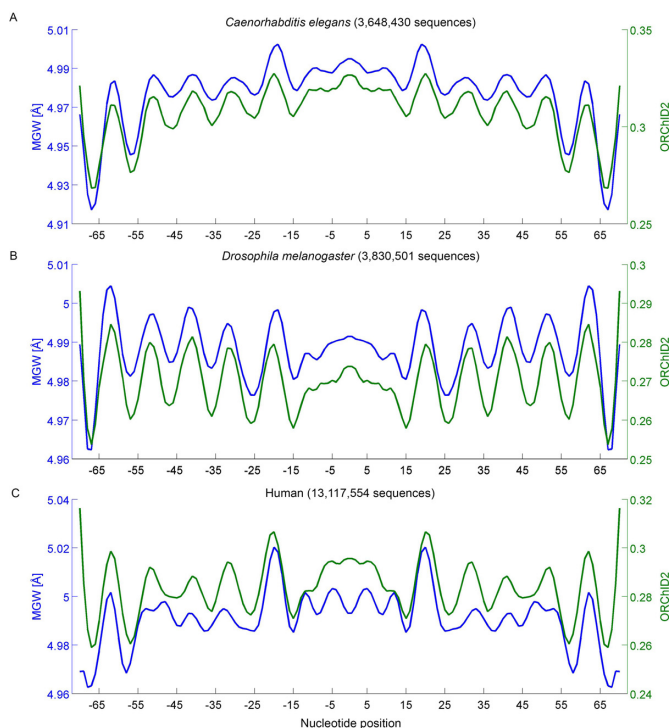
Figure 3. Variation in MGW (blue) and ORChID2 (green) signals on average in nucleosome sequences from the (A) *Caenorhabditis elegans*, (B) *Drosophila melanogaster* and (C) human genomes. Numbering of the nucleotide position starts with −1 and 1 for the central two base pairs, respectively.
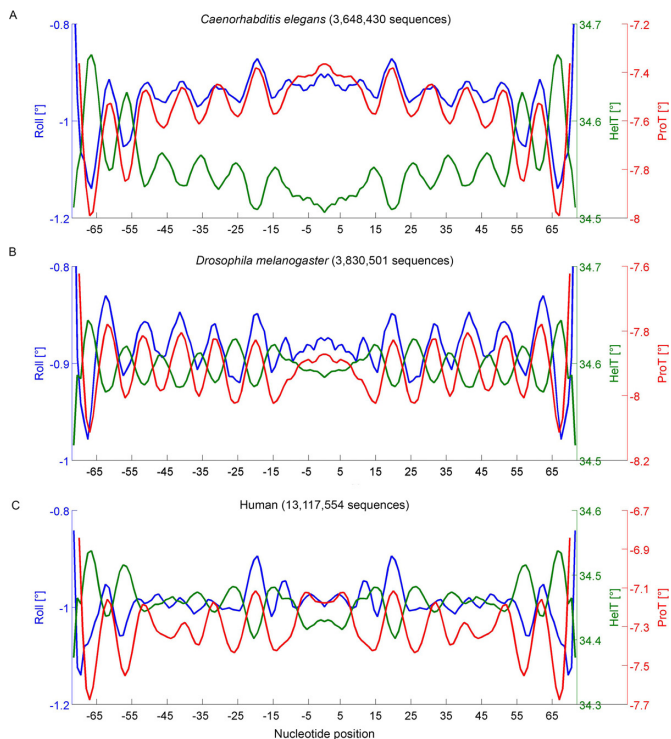

Figure 4. Variation in Roll (blue), HelT (green) and ProT (red) on average in nucleosome sequences from the (A) *Caenorhabditis elegans*, (B) *Drosophila melanogaster* and (C) human genomes. Numbering of the nucleotide position starts with −1 and 1 for the central two base pairs, respectively.

the existing annotations by using the function 'add custom tracks'.

The table browser supports downloading and analysis of quantitative DNA shape annotations. To access these functions, click on 'Table Browser' in the navigation bar on the GBshape homepage. The Table Browser can also be found under the 'Tools' link in the navigation bar of the Genome Browser Gateway page. One can download DNA shape annotations for an entire genome or for a specified genomic region by setting parameters on the Table Browser page. Download of data for multiple regions is specified by setting 'define regions'. Users can download data that match certain criteria by setting the 'filter' function, or manipulate data from different datasets by using the 'intersection' function. Output data can be exported in a variety of formats for further analysis or for use in other applications. Statistical correlations can be calculated over selected datasets, such as the correlation between data in different shape annotation tracks.

BIOLOGICAL APPLICATIONS

Nucleosome binding sites

Periodicity in nucleotide sequence has been detected in DNA sequences that wrap around histone octamers to form nucleosome core particles (33). The 10-bp periodicity of dinucleotide occurrence (34) and A-tract composition (1) mirrors the variation in width of the DNA minor groove as it is directed toward the histone core once every helical turn. We reported that the minor groove in nucleosome-bound DNA exhibits a 10-bp periodicity in MGW and electrostatic potential, and concluded that contacts of histone arginines with narrow minor groove regions are stabilized by the 10-bp shape-dependent periodicity in electrostatic potential (1).

A question that arises from these observations is whether periodic patterns in dinucleotide occurrence result in DNA shape features that guide nucleosome formation. Genome-wide nucleosome occupancy maps with thousands of nucleosome binding sites have been experimentally constructed by digesting intact chromatin with micrococcal nuclease followed by sequencing the underlying protected DNA fragments (MNase-seq) (34,35). We have used GBshape to infer structural features of these nucleosome-bound sequences.

We previously analyzed DNA shape features of 23 076 nucleosome-bound sequences from *Saccharomyces cerevisae* (34) and 25 654 from *Drosophila melanogaster* (35). We showed that analysis of shape profiles generated by the DNAshape and ORChID2 algorithms reveals a pronounced 10-bp periodicity in structural properties of nucleosomal DNA (14,18). The modENCODE consortium recently generated more extensive lists of nucleosome-bound sequences of much higher quality for human, *Drosophila melanogaster* and *Caenorhabditis elegans* (20).

We have now used GBshape to predict MGW and compare this structural property to the ORChID2 pattern for these massive lists of 3.6 million from *Caenorhabditis elegans*, 3.8 million nucleosome-bound sequences from *Drosophila melanogaster* and 13.1 million from the human genome (Figure 3). The strong correlation between MGW
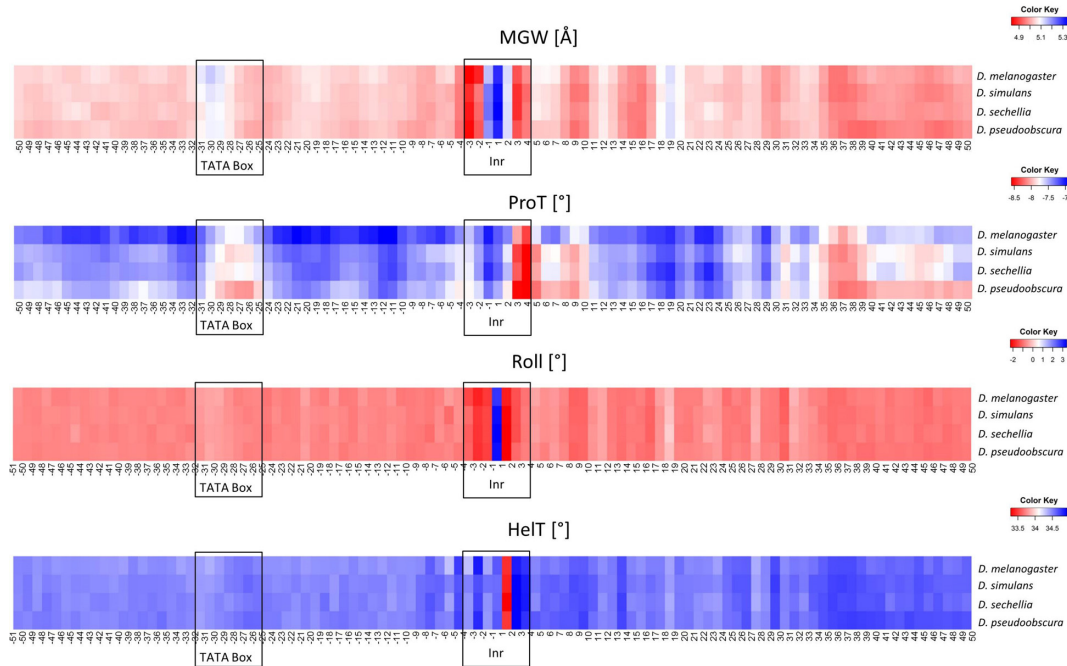
Figure 5. Average heat maps for four DNA shape features of TSSs and 50 bp up- and downstream in four fly species. The analysis is based on 3823 TSSs from the *D. melanogaster*, 6909 TSSs from the *D. simulans*, 7234 TSSs from the *D. sechellia* and 7397 TSSs from the *D. pseudoobscura* genomes. Column numbers in each heat map indicate the nucleotide position relative to the TSS. Black frames mark the locations of the Initiatior (Inr) element and TATA box.

and ORChID2 for all three organisms served as a validation of GBshape due to the independent approaches used to generate these predictions. Whereas the 10-bp periodicity was shared between human, fly and worm, details of the DNA shape profiles of nucleosomal DNA varied across species due to the different nucleotide compositions of these genomes. Analysis of the other DNA shape features Roll, HelT and ProT further confirmed the shared 10-bp periodicity as well as distinctions in DNA shape between nucleosome-bound sequences in these genomes (Figure 4). The maxima and minima of the MGW, Roll and ProT patterns overlapped, whereas the troughs in the HelT patterns matched the peaks in the other parameters, indicating a local helix unwinding at positions where a more positive Roll locally widens the minor groove.

TSSs

TSSs are located at the 5′ end of genes where contacts with RNA polymerase II initiate transcription. A long-standing question in the field is how these positions can be identified in a genome using computational methods (36). Whereas the presence of conserved sequence elements, such as the TATA box and the Initiator (Inr) element, represent one possibility for identifying TSSs, nucleotide composition varies in Inr elements and in regions surrounding TSSs. Previous reports suggested that structural features, including DNA bending and melting, enhance protein binding at TSSs (36). We used GBshape as a high-throughput approach to annotate DNA shape features at TSSs of four different *Drosophila* species.

We derived data from paired-end cap analysis for gene expression experiments (21) to identify TSSs in the *D. melanogaster*, *D. simulans*, *D. sechellia* and *D. pseudoobscura* genomes. Transcription initiates from a range of positions at a given promoter, resulting in a frequency distribution that varies from 'broad' to 'sharp' between promoters (37). Depending on the analysis, a single representative TSS position for each promoter can be chosen based on the median position or the position with the maximum number of initiation events (peaks) within a given TSS distribution. For this analysis, we chose the peak position within each local frequency distribution in order to maximize the 5′ sequence alignments.

Our DNA shape analysis of *Drosophila* TSSs revealed a clear structural signature for the Inr element despite the nucleotide sequence variation of this element. Moreover, specific DNA shape annotations of TSS regions were apparent for MGW, ProT, Roll and HelT (Figure 5). For each DNA shape feature the patterns were similar among the four *Drosophila* species, suggesting an evolutionary role of DNA structure. Whereas this effect merits further investigation, GBshape provides a platform that enables studies in which one can easily navigate between DNA sequence and shape information for a very large genomic datasets.

CONCLUSIONS

We have developed a database of DNA shape annotations for whole genomes of 94 organisms. Given the emerging literature on the importance of DNA structural features in refining transcription factor binding specificities (2), this tool provides a framework for integrating DNA shape in

whole-genome analyses. GBshape currently includes tracks for five structural features: MGW, ProT, Roll and HelT using DNAshape predictions (14), and hydroxyl radical cleavage intensity derived from ORChID2 (18). To demonstrate the utility of GBshape we analyzed structural features of nucleosome binding sites and TSSs. The availability of DNA shape annotations for entire genomes will enable the integration of DNA structure into genome analyses that currently are based only on nucleotide sequence.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online. See Supplementary Data for detailed author contributions.

ACKNOWLEDGEMENTS

FUNDING

REFERENCES

1. 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.
2. Slattery,M., Zhou,T., Yang,L., Dantas Machado,A.C., Gordân,R. and Rohs,R. (2014) Absence of a simple code: how transcription factors read the genome. *Trends Biochem. Sci.*, **39**, 381–399.
3. Rohs,R., Jin,X., West,S.M., Joshi,R., Honig,B. and Mann,R.S. (2010) Origins of specificity in protein-DNA recognition. *Annu. Rev. Biochem.*, **79**, 233–269.
4. Parker,S.C. and Tullius,T.D. (2011) DNA shape, genetic codes, and evolution. *Curr. Opin. Struct. Biol.*, **21**, 342–347.
5. Slattery,M., Riley,T., Liu,P., Abe,N., Gomez-Alcala,P., Dror,I., Zhou,T., Rohs,R., Honig,B., Bussemaker,H.J. *et al.* (2011) Cofactor binding evokes latent differences in DNA binding specificity between Hox proteins. *Cell*, **147**, 1270–1282.
6. Gordân,R., Shen,N., Dror,I., Zhou,T., Horton,J., Rohs,R. and Bulyk,M.L. (2013) Genomic regions flanking E-box binding sites influence DNA binding specificity of bHLH transcription factors through DNA shape. *Cell Rep.*, **3**, 1093–1104.
7. Parker,S.C., Hansen,L., Abaan,H.O., Tullius,T.D. and Margulies,E.H. (2009) Local DNA topography correlates with functional noncoding regions of the human genome. *Science*, **324**, 389–392.
8. Meysman,P., Dang,T.H., Laukens,K., De Smet,R., Wu,Y., Marchal,K. and Engelen,K. (2011) Use of structural DNA properties for the prediction of transcription-factor binding sites in Escherichia coli. *Nucleic Acids Res.*, **39**, e6.
9. Maienschein-Cline,M., Dinner,A.R., Hlavacek,W.S. and Mu,F. (2012) Improved predictions of transcription factor binding sites using physicochemical features of DNA. *Nucleic Acids Res.*, **40**, e175.
10. Hooghe,B., Broos,S., van Roy,F. and De Bleser,P. (2012) A flexible integrative approach based on random forest improves prediction of transcription factor binding sites. *Nucleic Acids Res.*, **40**, e106.
11. Greenbaum,J.A., Parker,S.C. and Tullius,T.D. (2007) Detection of DNA structural motifs in functional genomic elements. *Genome Res.*, **17**, 940–946.
12. Maurano,M.T., Wang,H., Kutyavin,T. and Stamatoyannopoulos,J.A. (2012) Widespread site-dependent buffering of human regulatory polymorphism. *PLoS Genet.*, **8**, e1002599.
13. Maurano,M.T., Humbert,R., Rynes,E., Thurman,R.E., Haugen,E., Wang,H., Reynolds,A.P., Sandstrom,R., Qu,H., Brody,J. *et al.* (2012) Systematic localization of common disease-associated variation in regulatory DNA. *Science*, **337**, 1190–1195.
14. Zhou,T., Yang,L., Lu,Y., Dror,I., Dantas Machado,A.C., Ghane,T., Di Felice,R. and Rohs,R. (2013) DNAshape: a method for the high-throughput prediction of DNA structural features on a genomic scale. *Nucleic Acids Res.*, **41**, W56–W62.
15. Rohs,R., Sklenar,H. and Shakked,Z. (2005) Structural and energetic origins of sequence-specific DNA bending: Monte Carlo simulations of papillomavirus E2-DNA binding sites. *Structure*, **13**, 1499–1509.
16. Zhang,X., Dantas Machado,A.C., Ding,Y., Chen,Y., Lu,Y., Duan,Y., Tham,K.W., Chen,L., Rohs,R. and Qin,P.Z. (2014) Conformations of p53 response elements in solution deduced using site-directed spin labeling and Monte Carlo sampling. *Nucleic Acids Res.*, **42**, 2789–2797.
17. Greenbaum,J.A., Pang,B. and Tullius,T.D. (2007) Construction of a genome-scale structural map at single-nucleotide resolution. *Genome Res.*, **17**, 947–953.
18. Bishop,E.P., Rohs,R., Parker,S.C., West,S.M., Liu,P., Mann,R.S., Honig,B. and Tullius,T.D. (2011) A map of minor groove shape and electrostatic potential from hydroxyl radical cleavage patterns of DNA. *ACS Chem. Biol.*, **6**, 1314–1320.
19. Karolchik,D., Barber,G.P., Casper,J., Clawson,H., Cline,M.S., Diekhans,M., Dreszer,T.R., Fujita,P.A., Guruvadoo,L., Haeussler,M. *et al.* (2014) The UCSC Genome Browser database: 2014 update. *Nucleic Acids Res.*, **42**, D764–D770.
20. Ho,J.W., Jung,Y.L., Liu,T., Alver,B.H., Lee,S., Ikegami,K., Sohn,K.A., Minoda,A., Tolstorukov,M.Y., Appert,A. *et al.* (2014) Comparative analysis of metazoan chromatin organization. *Nature*, **512**, 449–452.
21. Main,B.J., Smith,A.D., Jang,H. and Nuzhdin,S.V. (2013) Transcription start site evolution in Drosophila. *Mol. Biol. Evol.*, **30**, 1966–1974.
22. Yang,L., Zhou,T., Dror,I., Mathelier,A., Wasserman,W.W., Gordân,R. and Rohs,R. (2014) TFBSshape: a motif database for DNA shape features of transcription factor binding sites. *Nucleic Acids Res.*, **42**, D148–D155.
23. Cherry,J.M., Hong,E.L., Amundsen,C., Balakrishnan,R., Binkley,G., Chan,E.T., Christie,K.R., Costanzo,M.C., Dwight,S.S., Engel,S.R. *et al.* (2012) Saccharomyces Genome Database: the genomics resource of budding yeast. *Nucleic Acids Res.*, **40**, D700–D705.
24. Schneeberger,K., Ossowski,S., Ott,F., Klein,J.D., Wang,X., Lanz,C., Smith,L.M., Cao,J., Fitz,J., Warthmann,N. *et al.* (2011) Reference-guided assembly of four diverse Arabidopsis thaliana genomes. *Proc. Natl. Acad. Sci. U.S.A.*, **108**, 10249–10254.
25. Aurrecoechea,C., Brestelli,J., Brunk,B.P., Dommer,J., Fischer,S., Gajria,B., Gao,X., Gingle,A., Grant,G., Harb,O.S. *et al.* (2009) PlasmoDB: a functional genomic database for malaria parasites. *Nucleic Acids Res.*, **37**, D539–D543.
26. Riley,M., Abe,T., Arnaud,M.B., Berlyn,M.K., Blattner,F.R., Chaudhuri,R.R., Glasner,J.D., Horiuchi,T., Keseler,I.M., Kosuge,T. *et al.* (2006) Escherichia coli K-12: a cooperatively developed annotation snapshot–2005. *Nucleic Acids Res.*, **34**, 1–9.
27. Joshi,R., Passner,J.M., Rohs,R., Jain,R., Sosinsky,A., Crickmore,M.A., Jacob,V., Aggarwal,A.K., Honig,B. and Mann,R.S. (2007) Functional specificity of a Hox protein mediated by the recognition of minor groove structure. *Cell*, **131**, 530–543.
28. Chang,Y.P., Xu,M., Dantas Machado,A.C., Yu,X.J., Rohs,R. and Chen,X.S. (2013) Mechanism of origin DNA recognition and assembly of an initiator-helicase complex by SV40 large tumor antigen. *Cell Rep.*, **3**, 1117–1127.
29. Eldar,A., Rozenberg,H., Diskin-Posner,Y., Rohs,R. and Shakked,Z. (2013) Structural studies of p53 inactivation by DNA-contact mutations and its rescue by suppressor mutations via alternative protein-DNA interactions. *Nucleic Acids Res.*, **41**, 8748–8759.
30. Lazarovici,A., Zhou,T., Shafer,A., Dantas Machado,A.C., Riley,T.R., Sandstrom,R., Sabo,P.J., Lu,Y., Rohs,R.,

Stamatoyannopoulos,J.A. *et al.* (2013) Probing DNA shape and methylation state on a genomic scale with DNase I. *Proc. Natl. Acad. Sci. U.S.A.*, **110**, 6376–6381.

31. Rohs,R., Bloch,I., Sklenar,H. and Shakked,Z. (2005) Molecular flexibility in ab initio drug docking to DNA: binding-site and binding-mode transitions in all-atom Monte Carlo simulations. *Nucleic Acids Res.*, **33**, 7048–7057.

32. Balasubramanian,B., Pogozelski,W.K. and Tullius,T.D. (1998) DNA strand breaking by the hydroxyl radical is governed by the accessible surface areas of the hydrogen atoms of the DNA backbone. *Proc. Natl. Acad. Sci. U.S.A.*, **95**, 9738–9743.

33. Segal,E., Fondufe-Mittendorf,Y., Chen,L., Thastrom,A., Field,Y., Moore,I.K., Wang,J.P. and Widom,J. (2006) A genomic code for nucleosome positioning. *Nature*, **442**, 772–778.

34. Field,Y., Kaplan,N., Fondufe-Mittendorf,Y., Moore,I.K., Sharon,E., Lubling,Y., Widom,J. and Segal,E. (2008) Distinct modes of regulation by chromatin encoded through nucleosome positioning signals. *PLoS Comput. Biol.*, **4**, e1000216.

35. Mavrich,T.N., Jiang,C., Ioshikhes,I.P., Li,X., Venters,B.J., Zanton,S.J., Tomsho,L.P., Qi,J., Glaser,R.L., Schuster,S.C. *et al.* (2008) Nucleosome organization in the Drosophila genome. *Nature*, **453**, 358–362.

36. Bansal,M., Kumar,A. and Yella,V.R. (2014) Role of DNA sequence based structural features of promoters in transcription initiation and gene expression. *Curr. Opin. Struct. Biol.*, **25**, 77–85.

37. Rach,E.A., Yuan,H.Y., Majoros,W.H., Tomancak,P. and Ohler,U. (2009) Motif composition, conservation and condition-specificity of single and alternative transcription start sites in the Drosophila genome. *Genome Biol.*, **10**, R73.1–R73.24.