

## SURVEY AND SUMMARY

# New insights into Hoogsteen base pairs in DNA duplexes from a structure-based survey

Huiqing Zhou<sup>1</sup>, Bradley J. Hintze<sup>1</sup>, Isaac J. Kimsey<sup>1</sup>, Bharathwaj Sathyamoorthy<sup>1</sup>, Shan Yang<sup>2</sup>, Jane S. Richardson<sup>1</sup> and Hashim M. Al-Hashimi<sup>1,3,\*</sup>

<sup>1</sup>Department of Biochemistry, Duke University, Durham, NC 27710, USA, <sup>2</sup>Department of Biochemistry, Stanford University, Stanford, CA 94305, USA and <sup>3</sup>Department of Chemistry, Duke University, Durham, NC 27708, USA

Received December 18, 2014; Revised February 24, 2015; Accepted March 01, 2015

### ABSTRACT

Hoogsteen (HG) base pairs (bps) provide an alternative pairing geometry to Watson–Crick (WC) bps and can play unique functional roles in duplex DNA. Here, we use structural features unique to HG bps (*syn* purine base, HG hydrogen bonds and constricted C1'–C1' distance across the bp) to search for HG bps in X-ray structures of DNA duplexes in the Protein Data Bank. The survey identifies 106 A•T and 34 G•C HG bps in DNA duplexes, many of which are undocumented in the literature. It also uncovers HG-like bps with *syn* purines lacking HG hydrogen bonds or constricted C1'–C1' distances that are analogous to conformations that have been proposed to populate the WC-to-HG transition pathway. The survey reveals HG preferences similar to those observed for transient HG bps in solution by nuclear magnetic resonance, including stronger preferences for A•T versus G•C bps, TA versus GG steps, and also suggests enrichment at terminal ends with a preference for 5'-purine. HG bps induce small local perturbations in neighboring bps and, surprisingly, a small but significant degree of DNA bending (~14°) directed toward the major groove. The survey provides insights into the preferences and structural consequences of HG bps in duplex DNA.

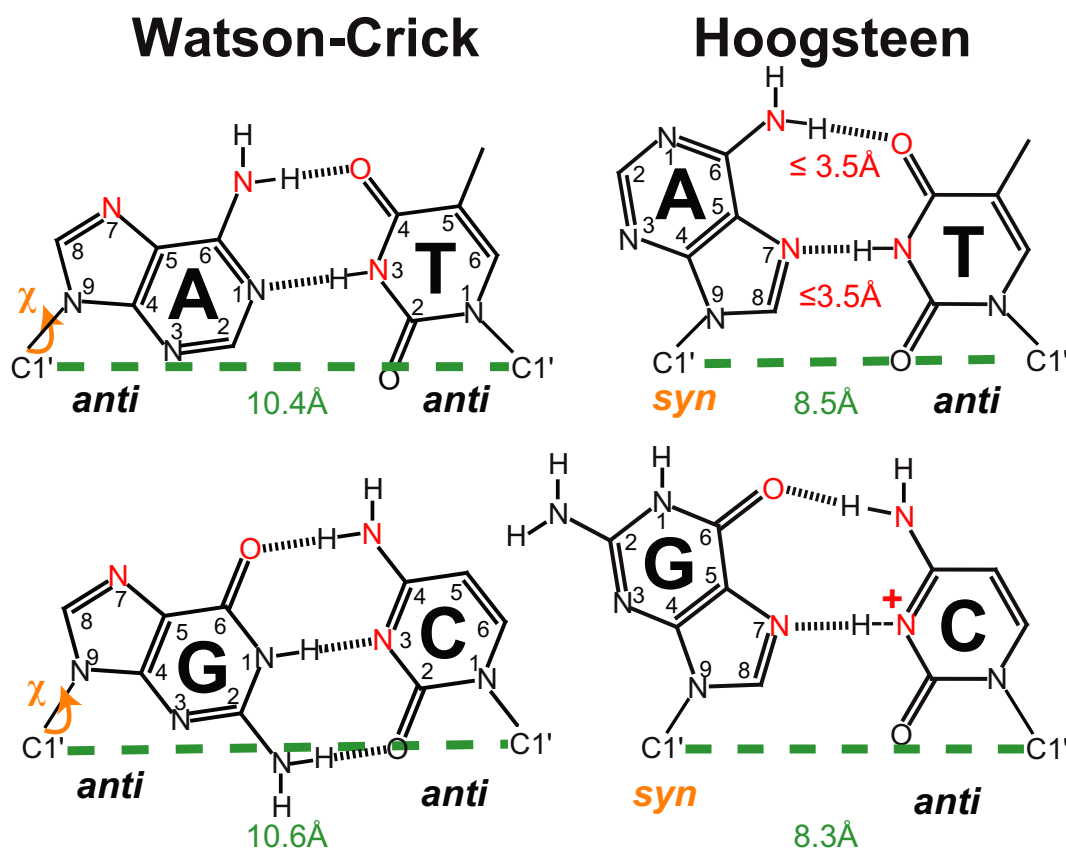
### INTRODUCTION

Hoogsteen (HG) base pairs (bps) were discovered in 1959 when Karst Hoogsteen used single crystal X-ray crystallography to visualize the pairing between 1-methylthymine and 9-methyladenine (1). Rather than observing a Watson–Crick (WC) bp, a unique pairing geometry was observed in

which the adenine base was flipped 180° to form a unique set of hydrogen bonds (H-bonds) (Figure 1). Two years earlier, Rich *et al.* had speculated that such a pairing scheme could explain how poly(rU) associates with poly(rA)-poly(rU) duplexes to form RNA triplexes (2). These bps are now referred to as 'HG' bps. Subsequent crystallographic studies on isolated pairs of A•U derivatives yielded HG bps (3–5) while, in contrast, G•C derivatives co-crystallized into the expected WC bps (6). The lack of experimental evidence for A•T/U WC bps raised considerable scepticism about the WC base-pairing scheme during 1960s and early 1970s as reviewed in (7). Computational studies on A•T base derivatives later showed that HG bps form H-bonds that are ~0.2–1 kcal/mol more stable than that corresponding WC pairing (8–10). The controversy over HG versus WC for A•T/U bps was partly resolved in 1973 when Rich *et al.* reported single crystal structures of dinucleoside phosphates showing WC A•U and G•C bps (11,12). Although WC bps became the dominant model for A•T and G•C bps in DNA, HG bps continued to surface in high-resolution structures of naked antiparallel DNA duplexes (13), DNA in complex with quinoxaline bis-intercalators (14) and proteins (15–17) as reviewed in (7). In parallel, computational studies (18,19) indicated that the stability of antiparallel HG duplex is comparable to that of the B-form WC helix and that the chimera WC–HG helix is an energetically accessible conformation. Studies based on circular dichroism (CD), nuclear magnetic resonance (NMR) and molecular dynamics simulations also showed A•T and G•C<sup>+</sup> HG bps in parallel double-stranded DNA (20–22), DNA triplexes (23) and dumbbell-like DNA structures (24). Here, we present a survey that is focused on HG bps in the context of antiparallel DNA duplexes including stem-loop DNA hairpins.

In duplex DNA, HG bps form by rotating the purine base 180° around the glycosidic bond, to adopt a *syn* rather than *anti* conformation (Figure 1). While A•T HG bps retain the

\*To whom correspondence should be addressed. Tel: +1 919 681 8804; Fax: +1 919 684 8885; Email: hashim.al.hashimi@duke.edu



**Figure 1.** HG-unique criteria. Shown are WC and HG A•T and G•C bps with highlighted key geometrical differences. Heavy atoms involved in HG hydrogen bonds (in red), *syn*  $\chi$  angle (in orange) and constricted C1'–C1' distances (in green). Average C1'–C1' distances from the survey are shown for each base-pair type.

WC A–N6–H—O4–T H-bond, they replace the WC N1—H–N3 H-bond with an N7—H–N3 H-bond (Figure 1). On the other hand, G•C HG bps retain the O6—H–N4 H-bond and replace the other two WC H-bonds (N1–H—N3 and N2–H—O2) with a single N7—H–N3<sup>+</sup> H-bond, which requires protonation of cytosine N3 (Figure 1). In addition, formation of HG-type H-bonds requires that the two bases come into closer proximity, thus constricting the C1'–C1' distance by  $\sim 2$  Å relative to WC bps (1,25) (Figure 1).

By modifying the structural presentation of sequence information in duplex DNA, HG bps can carry out unique functions (7,17,26–28). HG bps have been observed in several protein–DNA complexes (15–17,29) and DNA in complex with quinoxaline bis-intercalators (30–33) where they are thought to contribute to recognition. For example, in X-ray structures of bent TATA elements in complex with the TATA box-binding protein (TBP), two consecutive C•G HG bps help avoid a steric clash between the guanine exocyclic NH<sub>2</sub> group and a nearby leucine side chain (16). Two consecutive A•T HG bps in X-ray structures of DNA in complex with the tumor suppressor protein p53 are thought to contribute to a narrowed minor groove and a more negative electrostatic potential surface that may favor insertion of positively charged Arg248 (17). HG bps have also been observed in chemically modified DNA, including N2-propanoguanine (34), 1,N2-ethylguanine (35), and

N1-methyladenine (36), and 8-amino-purine (37–39) where they may contribute to damage accommodation, recognition and repair. There is also strong evidence that some members of the Y-family ‘low fidelity polymerases’ replicate DNA using HG pairing as the dominant mechanism, providing a means for bypassing lesions on the WC face during replication (27–28,40). In the complex structures of DNA-quinoxaline bis-intercalators (i.e. triostin A and echinomycin), HG bps are observed flanking the intercalation sites at both the interior and termini of the duplex. Solution NMR studies on this complex with various DNA sequences provided evidence for dynamic HG bps in solution (41,42). Theoretical studies suggested that the stabilization of HG bps flanking intercalation site and its dependence on DNA sequence could result from favorable van der Waals (43) and stacking (44) interactions.

Recent studies employing NMR spin relaxation in the rotating frame ( $R_{1\rho}$ ) (45–47) have shown that G•C<sup>+</sup> and A•T HG bps exist transiently in duplex DNA (7,48) across a variety of sequence and positional contexts (49). These transient HG bps form with strong sequence-specific energetic preferences that are comparable to the sequence-specific variations in WC stability (50) potentially providing a new basis for sequence-specific DNA transactions (50). These transient HG bps have populations of  $\sim 0.5\%$  and lifetimes of  $\sim 1$  ms (48,49) but can increase considerably in modified

bases such as inosine (51). Although small in population, these suggest that there can be as many as ~15 million HG bps in the human genome at a given time. In addition, it is well documented that WC and HG bps can be difficult to distinguish due to ambiguous electron densities (17,29,52–53) and yet they are often modeled by default as WC bps. This leaves open the possibility that there are actual HG bps mismodeled as WC. It is also possible that bps are often modeled as single states (HG or WC) when in fact partial occupancies of WC and HG might fit the density even better. These findings emphasize the need to rigorously study the occurrence and preferences of HG bps in duplex DNA.

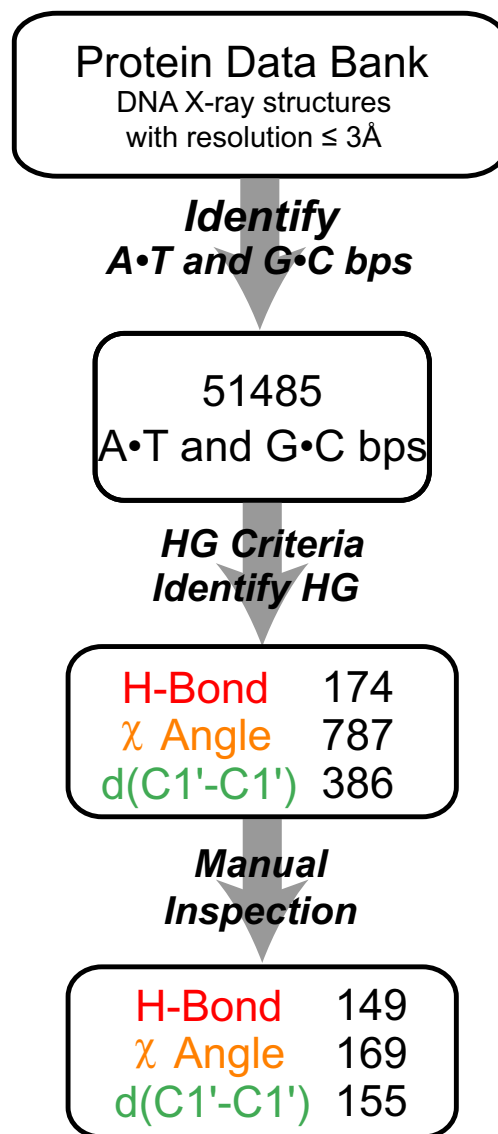
Here, we comprehensively examine the Protein Data Bank (PDB) (54) to survey the occurrence and structural features of HG bps in duplex DNA. We identify sequences and context preferences for HG bps, expose new types of HG-like bps that satisfy only one or two of the above three structural features (Figure 1) and show that HG bps induce local structural perturbations in adjacent WC bps as well as global DNA bending toward the major groove. This study provides a valuable framework for guiding future studies exploring the occurrence and functional importance of HG bps in duplex DNA.

## METHODS

### Survey protocol

DNA X-ray crystal structures (excluding DNA–RNA hybrids) with resolution  $\leq 3.0$  Å were downloaded as the biological assemblies from the PDB (54) on 4 September 2013 (Figure 2). For structures with palindromic double-stranded DNA that were deposited as single chains in the asymmetric unit, the chain ID of the symmetry mate was edited to be unique using a PyMOL (The PyMOL Molecular Graphics System, Schrödinger LLC) script for subsequent analyses. An in-house program was used to parse all structures using 3DNA (55) into a searchable database. The database contains all DNA bps in the biological unit and an accompanying list of structural descriptors defining those bps, including local base-pair parameters (shear, stretch, stagger, buckle, propeller, opening), the C1'–C1' distance across the bp, heavy atom distances in H-bonds and sugar-phosphodiester backbone torsion angles ( $\nu_0$ ,  $\nu_1$ ,  $\nu_2$ ,  $\nu_3$ ,  $\nu_4$ ,  $\alpha$ ,  $\beta$ ,  $\gamma$ ,  $\delta$ ,  $\epsilon$ ,  $\zeta$ ) (55,56). Next, the in-house program was used to identify bps that satisfy the following HG criteria (Figure 1):

- (i) HG hydrogen bonding: Both AN7—TN3 and AN6—TO4 distances  $\leq 3.5$  Å for A•T; both GN7—CN3 and GO6—CN4 distances  $\leq 3.5$  Å for G•C.
- (ii) Constricted C1'–C1' distance: The distance between the C1' atoms of the purine and pyrimidine pair is restricted to  $\leq 9.5$  Å, which is midway between the average distances observed for WC bps ( $\sim 10.5$  Å) (12) and HG bps ( $\sim 8.5$  Å) (14,25). (Note that the constricted C1'–C1' distance does not necessarily entail a shortened P–P distance across the helix.)
- (iii) *syn* purine: The *syn* glycosidic torsion angle ( $\chi$  angle) of the purine base is in the range  $0^\circ \leq \chi \leq 90^\circ$ .



**Figure 2.** The protocol employed to identify HG bps in DNA duplexes from the X-ray crystal structures. The number of identified HG bps satisfying each criterion is shown for the last two steps of the protocol. Manual inspection was used to exclude bps involved in base triples, tertiary interactions, bps between consecutive nucleotides on the same DNA strand, WC bps in Z-DNA and WC-like bps as described in Methods.

A list of bps was created that satisfy one, two or all three HG criteria listed above (Figure 2). Manual inspection was then carried out to exclude bps corresponding to base triples, tertiary interactions, bps between consecutive nucleotides on the same DNA strand, WC bps in Z-DNA and WC-like bps only satisfying the constricted C1'–C1' distance criterion which are better described as distorted WC bps (Figure 2). Average B-factors, if not specified, were calculated for each HG bp by averaging B-factors over all sugar-phosphate backbone and base heavy atoms (i.e. C, N, O and P). Finally, a total of 178 HG and HG-like bps (Figure 3A, Supplementary Table S1) were identified in antiparallel DNA duplexes including stem-loop hairpins, 12% of which are from naked DNA, 21% and 67% from DNA in

complex with small molecules and proteins, respectively. To assess potential modeling errors in the bps, we manually examined the electron density maps and average B-factors (averaged over all sugar-phosphate and base heavy atoms, i.e. C, N, O and P) of a given bp. For HG-like and terminal HG bps which are more susceptible to dynamics, we examined the B-factors, real-space correlation coefficient (RSCC) and  $\sigma$ -weighted  $2F_o - F_c$  map values for the base (Supplementary Table S3). The RSCC values range between 0 and 1 for zero and perfect correlation between the  $F_c$  and  $2F_o - F_c$  maps, respectively.

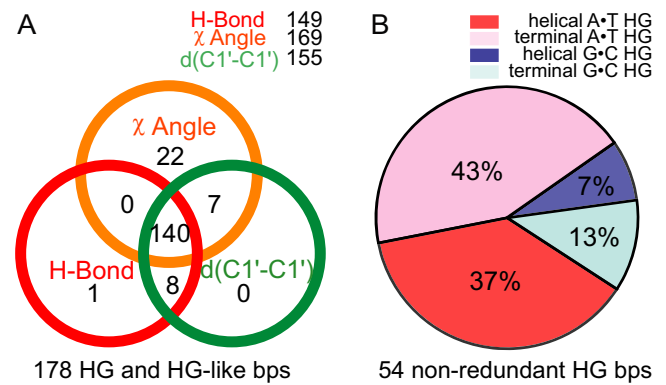
To statistically analyze the sequence- and position-specific preferences of HG bps, and their structural features, a set of 54 ‘non-redundant HG bps’ which satisfy all three criteria was generated (Supplementary Table S2, Table 1) by excluding redundant bps that are surrounded by identical adjacent WC bps in both 5' and 3' directions, involved in DNA duplexes with identical lengths, and when applicable, are bound to the same protein or ligand. A set of 43 ‘adjacent WC bps’ consisting of WC bps immediately adjacent to helical HG and HG-like bps was also generated along with a set of 186 ‘control WC bps’ consisting of helical WC bps that are in the same DNA duplexes containing HG bps but more than one bp away from the HG bps. For each of the three sets of bps (‘non-redundant HG’, ‘adjacent WC’ and ‘control WC’), population-weighted distributions were constructed for local base-pair parameters (i.e. shear, stretch, stagger, buckle, propeller, opening), C1'–C1' distances, heavy atom distances in H-bonds, as well as sugar ( $\nu_0 - \nu_4$ ) and phosphodiester backbone torsion angles ( $\alpha, \beta, \gamma, \delta, \epsilon, \zeta$ ).

### Analysis of local structure

To examine the impact of HG bps on the local DNA structure, we constructed 1D histogram distributions for local structural parameters (including sugar-phosphodiester torsion angles, local base-pair parameters, C1'–C1' distances and heavy atom distances in H-bonds) for the ‘non-redundant HG’, ‘adjacent WC’ and ‘control WC’ sets. We then used a recently introduced REsemble approach (57) to measure the similarity between ‘control WC’ distributions and those of ‘non-redundant HG’ and ‘adjacent WC’. Here, the overlap between two distributions (T and P) is computed using the square root of the Jensen–Shannon divergence ( $\Omega^2$ ) (58) given by Equation (1) (57):

$$\Omega^2(w_i^T(m), w_i^P(m)) = S\left(\frac{w_i^T(m) + w_i^P(m)}{2}\right) - \frac{1}{2}[S(w_i^T(m)) + S(w_i^P(m))] \quad (1)$$

where  $w_i^T(m)$  and  $w_i^P(m)$  are the corresponding population weights for the  $i$ th bin for a given bin size  $m$  of distributions T and P, i.e. HG bps or adjacent WC bps and control WC bps. The term  $S(w_i) = -\sum w_i(m) \log_2 w_i(m)$  is the information entropy (57). The value of  $\Omega$  is then computed as a function of bin size ( $m$ ) that is used to build the histogram distribution. The resulting values are summed over  $K$  different bin sizes and normalized against a zero-overlap condition ( $\Omega = 1$  for all bin sizes) according to Equation



**Figure 3.** (A) Venn diagram illustrating the number of identified HG and HG-like bps in duplex DNA, with the total numbers of HG bps that satisfy different combinations of the three criteria shown in the union of the circles. Examples of HG-like bps that satisfy one or two HG criteria are shown in Supplementary Figure S1. (B) Percentage distribution of the 54 non-redundant HG bps with respect to the base-pair type (A•T or G•C) and their position in the secondary structure (at termini or interior of the DNA helix).

(2),

$$\sum_K \Omega(w^T, w^P) = \frac{\sum_m \Omega(w_i^T(m), w_i^P(m))}{K} \quad (2)$$

The value of  $\sum_K \Omega(w^T, w^P)$  provides a measure of similarity between distributions T and P and ranges between 0 and 1 for perfect and zero similarity, respectively (57). REsemble was used to compare 1D distributions of sugar-phosphodiester backbone torsion angles ( $\alpha, \beta, \gamma, \delta, \epsilon, \zeta, \nu_0, \nu_1, \nu_2, \nu_3, \nu_4$ ) (see Figure 4A) and local base-pair parameters (shear, stretch, stagger, buckle, propeller twist, opening), C1'–C1' distances and heavy atom distances in H-bonds) (Supplementary Figures S2 and S3).

We previously used REsemble to compare 1D torsion angle distributions (57). Here, we extended the analysis to compare multi-dimensional probability distributions consisting of six phosphodiester backbone torsion angles, five sugar torsion angles and six local base-pair parameters. This was necessary because two pairs of 1D distributions (A and B versus a and b) could exhibit perfect overlap in 1D (A = a and B = b) yet exhibit zero overlap in 2D (AB ≠ ab). To maintain computational efficiency, 6D, 5D and 6D REsemble were used to compare similarities between adjacent WC bps (or HG bps if applicable) and the control WC bps of six phosphodiester backbone torsion angles ( $\alpha, \beta, \gamma, \delta, \epsilon, \zeta$ ), five sugar torsion angles ( $\nu_0, \nu_1, \nu_2, \nu_3, \nu_4$ ) and six local base-pair parameters (shear, stretch, stagger, buckle, propeller twist, opening) distributions, respectively (see Supplemental Information). Because HG bps have a distinct reference frame of local base-pair geometry relative to WC bps and by definition have a constricted C1'–C1' distance, the local base-pair parameters and C1'–C1' distances for HG bps were not computed and compared with control WC bps. As a control, we measured the similarity between the control WC bp distributions and distributions obtained by randomly picking entries from the same set such that the total number of entries equals that in the HG distribution.

To carry out multi-dimensional REsemble analyses, the translational local base-pair parameters (i.e. shear, stretch and stagger) between  $-2.5 \text{ \AA}$  and  $2.5 \text{ \AA}$  were linearly converted to the range of  $0^\circ$  to  $360^\circ$  to be consistent with the torsion angle distribution range. The multi-dimensional distribution was constructed by using the same bin size to bin each parameter. The bin size ( $m$ ) used in the REsemble analysis was varied between  $15^\circ$  and  $360^\circ$  with an increment of  $15^\circ$  ( $K = 24$ ). Note that data points near the edges of the angle distribution (i.e.  $-180^\circ$  and  $180^\circ$ ) can lead to overestimation of  $\sum_K \Omega$  (e.g.  $-179^\circ$  and  $179^\circ$  differ by  $358^\circ$  in binning but only differ by  $2^\circ$  in reality). All of the local structural parameters including the converted translational base-pair parameters are distributed in the middle of the range ( $-180^\circ$  to  $180^\circ$ ) except the backbone torsion angles  $\beta$  and  $\epsilon$ , which have major distributions near the edges (see Figure 4C). To minimize adverse effects from data near the edges, all backbone torsion angles ( $\alpha$ ,  $\beta$ ,  $\gamma$ ,  $\delta$ ,  $\epsilon$ ,  $\zeta$ ) between  $-180^\circ$  and  $0^\circ$  were reflected onto the  $180^\circ$  to  $360^\circ$  region by an addition of  $360^\circ$ , while those in the  $0^\circ$  to  $180^\circ$  region remain unchanged (Supplementary Figure S3a). This yielded a final distribution between  $0^\circ$  and  $360^\circ$  with all angles distributed away from the edges.

### Analysis of global structure

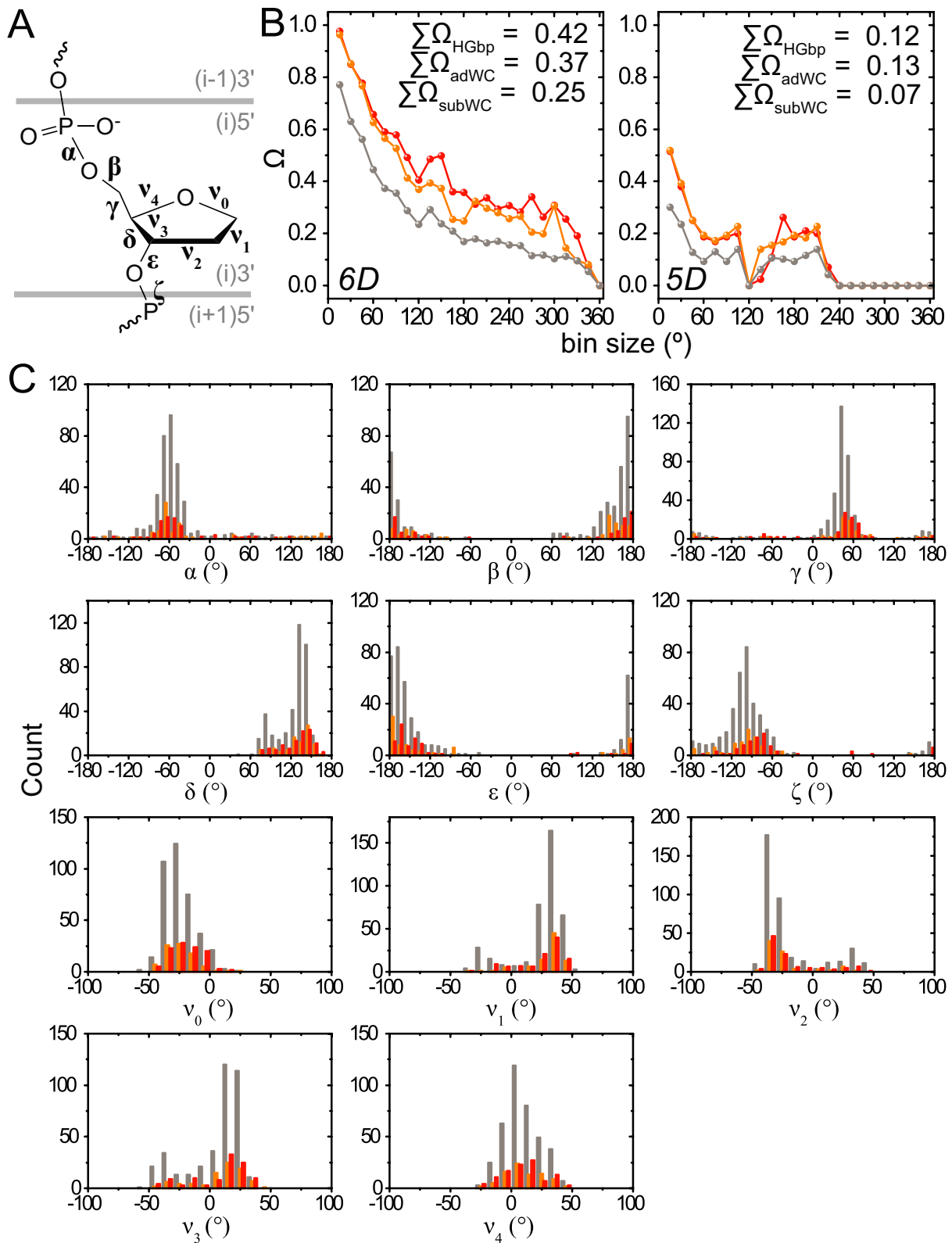
To assess the impact of HG bps on the DNA structure, we adopted the inter-helical Euler angle protocol developed for describing relative orientations of RNA A-form helices across junctions (59,60). Here, three inter-helical Euler angles ( $\alpha_h$ ,  $\beta_h$ ,  $\gamma_h$ ) are computed which describe the relative orientation of two helices across a given junction, in this case, a single or tandem HG/HG-like bps. For a given target DNA structure containing HG bps, we define a corresponding lower helix H1 and upper helix H2 to be the helices at the 5' and 3' sides, respectively, of the *syn* purine base in an HG bp (see Figure 5A). The inter-helical Euler angles describe the orientation of H2 relative to H1 across the junction of HG/HG-like bps and are determined by computing the rotation matrix that is required in order to rotate H2 so that it is in perfect coaxial alignment with H1. The approach has been described elsewhere in A-form RNA (59–61). Here we provide a brief description emphasizing those differences that relate to bending in B-form DNA.  $\beta_h$  is the inter-helical bend angle between H2 and H1, and ranges between  $0^\circ$  and  $180^\circ$ .  $\alpha_h$  and  $\gamma_h$  are defined as ‘twist’ and ‘arc’ angles of H2 around the H2 and H1 helical axes, respectively, and range between  $-180^\circ$  and  $180^\circ$  (see Figure 5B). The inter-helical Euler angles ( $\alpha_h$ ,  $\beta_h$ ,  $\gamma_h$ ) are computed relative to a reference idealized B-form linear helix with 10 bps per turn consisting of two consecutive and perfectly coaxial helices (iH1 and iH2). This reference B-form helix was constructed using the 3DNA fiber model (62) and the helix axis was oriented along the z-axis (see Figure 5B). The C1'–C1' vector across the WC bp in iH1 immediately neighboring the junction was oriented along the y-axis with the major groove facing the +x direction (see Figure 5C). H1 in the target DNA structure was superimposed onto iH1 using heavy atoms (i.e. C, N, O, P) in the sugar-phosphodiester backbone. Next, reference helix iH2 was superimposed onto the resulting target helix H2 to yield iH2'. A rotation matrix

$R(\alpha_h, \beta_h, \gamma_h)$  was then computed to transform iH2' back to iH2 using the EULER-RNA program (<https://sites.google.com/site/hashimigroup/resources>) (60,63). The direct output ( $\alpha_h^0$ ,  $\beta_h^0$ ,  $\gamma_h^0$ ) from EULER-RNA was then converted based on the current definition of inter-helical Euler angles by

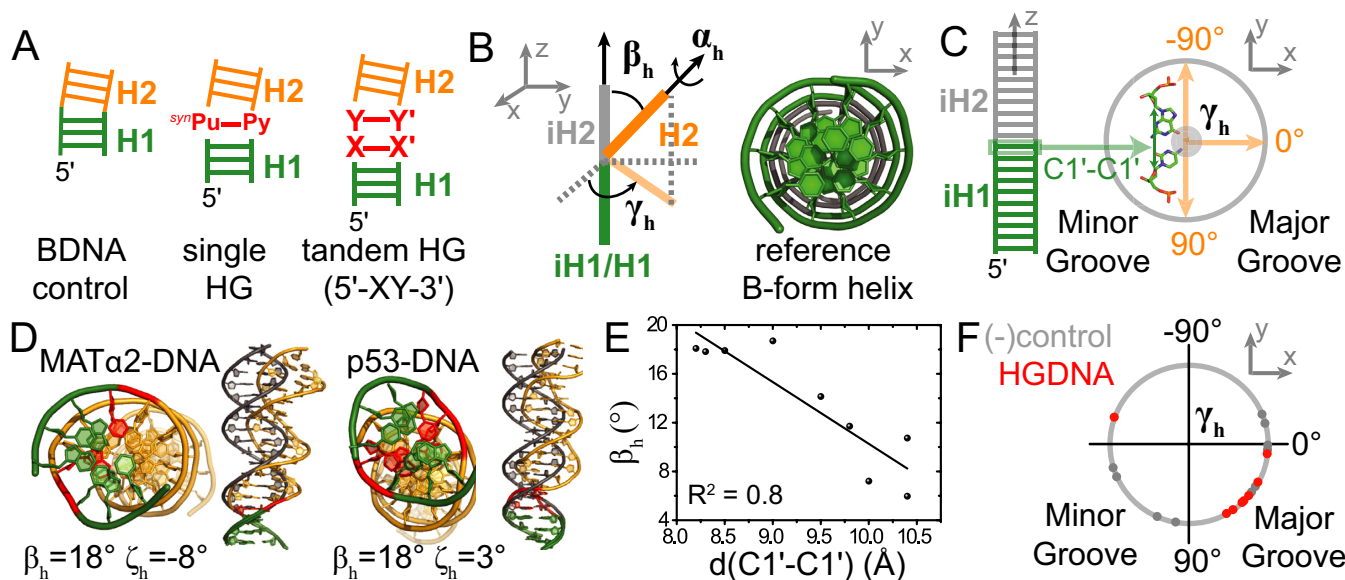
$$\begin{aligned} \text{if } \beta_h^0 \geq 0, (\alpha_h, \beta_h, \gamma_h) &= (\alpha_h^0, \beta_h^0, \gamma_h^0); \\ \text{if } \beta_h^0 < 0, (\alpha_h, \beta_h, \gamma_h) &= (\alpha_h^0 \pm 180^\circ, -\beta_h^0, \gamma_h^0 \pm 180^\circ) \end{aligned}$$

In this reference frame,  $\gamma_h$  corresponds to the angle between the x-axis of the reference frame and the projection of the H2 helix axis onto the x–y plane (see Figure 5B) and represents the bending direction of H2 relative to the aligned WC bp (see Figure 5C);  $-90^\circ \leq \gamma_h \leq 90^\circ$  indicates bending toward the major groove whereas  $-180^\circ < \gamma_h < -90^\circ$  and  $90^\circ < \gamma_h < 180^\circ$  reflect bending toward the minor groove (see Figure 5C). Note that the bending direction (major or minor groove) may vary depending on the choice of the reference bp. For example, the direction may be different relative to a reference bp in H2, where  $\alpha_h$  and not  $\gamma_h$  specifies the direction of bending of H1 relative to H2. A complete description of the bending direction requires all three Euler angles. The inter-helical twist angle  $\zeta_h = \alpha_h + \gamma_h$  describes the relative twist between H1 and H2, and is equal to zero for a perfectly coaxial helix in B-form DNA.  $\zeta_h > 0^\circ$  and  $\zeta_h < 0^\circ$  represents under- and over-twisting, respectively (59). The three angles  $\beta_h$ ,  $\gamma_h$  and  $\zeta_h$  provide a complete angular description of the two helices.

The above approach for computing bend and twist angles assumes an idealized B-form geometry for the two helices. In RNA, the A-form geometry has been shown to be highly robust across different sequence contexts, and to a very good approximation, WC bps surrounding WC bps can be modeled assuming an idealized A-form geometry (59–61). There can be greater variability in local structural parameters in B-form DNA based on analyses of X-ray structures (64) and molecular dynamics simulation (MD) trajectories (65). We previously showed that the computed inter-helical Euler angles will not be reliable if the target helices superimpose with idealized helices with RMSD  $> 2 \text{ \AA}$  (59). In the current study, six out of 15 structures yielded superposition RMSD  $> 2 \text{ \AA}$  and were excluded from analysis (Table 2 and Supplementary Figure S5b). Among the remaining nine structures, five contained helices with terminal or non-canonical bps, which were used in the superposition (Supplementary Figure S5c). To evaluate the robustness of this approach, we compared the inter-helical angles computed when varying the number of bps (2 versus 3 bps) and types of heavy atoms (with or without sugar atoms C1'/C2'/O4') used in the superposition and found very small variations ( $\leq 2^\circ$  for  $\beta_h$  and  $\leq 6^\circ$  for  $\alpha_h$  and  $\gamma_h$ ) (data not shown). In addition, as a negative control, we computed inter-helical angles for 11 X-ray structures of B-form naked DNA duplexes that do not contain HG bps and show no significant localized bending (Supplementary Table S4). These structures were used in a prior survey of duplex DNA structures (66). Each duplex was sub-divided into two coaxially stacked 3 bp helices denoted H1 and H2 (see Figure 5A) as indicated in Supplementary Table S4. Inter-helical Euler angles ( $\alpha_h$ ,  $\beta_h$ ,  $\gamma_h$ ) were then computed for H1 and H2 in all 11 struc-



**Figure 4.** (A) Phosphodiester backbone ( $\alpha$ ,  $\beta$ ,  $\gamma$ ,  $\delta$ ,  $\epsilon$ ,  $\zeta$ ) and sugar ( $\nu_0$ – $\nu_4$ ) torsion angles within a nucleotide ‘i’. (B) 6D and 5D REsemble plots of  $\Omega$  versus bin size (in the unit of degree) for phosphodiester backbone (left) and sugar torsion angles (right) with the normalized  $\Sigma\Omega$  reported for each dataset (HG bps in red, adjacent WC bps in orange and randomly selected subset of control WC bps in gray) compared against the control WC bps (see Methods). (C) Overlay of histograms obtained from 1D distributions of sugar-phosphodiester backbone torsion angles for non-redundant HG bps (in red), adjacent WC bps (in orange) and control WC bps (in gray).



**Figure 5.** Global DNA bending about HG bps (A) Definition of H1 and H2 in HG bp containing DNA duplexes used to compute inter-helical Euler angles. (B) Reference frame used to define the inter-helical Euler angles ( $\alpha_h$ ,  $\beta_h$ ,  $\gamma_h$ ) of the target helix along with a reference idealized B-form DNA helix consisting of lower (iH1, in green) and upper (iH2, in gray) helices. The lower helix (H1, in green) of the target helix is superimposed onto iH1. (C) Definition of the rotation axis around the Z direction for the calculation of inter-helical Euler angles. The C1'–C1' vector of the WC bp of iH1 nearest to iH2 is oriented along the y-axis. This orientation makes it possible to distinguish bending toward major or minor groove by the value of  $\gamma_h$ . (D) Examples of DNA bending at single (PDBID: 1K61) and tandem HG bps (PDBID: 3KZ8) in DNA–protein complexes. Structures are overlaid on reference B-form DNA helix (in gray). (E) Correlation plot between the inter-helical bend angle ( $\beta_h$ ) and corresponding C1'–C1' distance across the HG or HG-like bp. (F) The bending directions ( $\gamma_h$ ) of nine structures containing helical HG and HG-like bps (in red) compared to those of 11 linear control B-form DNA structures (in gray).

tures. As a positive control, we computed inter-helical angles for two bent DNA structures that do not contain HG bps, including an A6-T6 A-tract sequence (PDBID: 1FZX) (67) and the nucleosome particle (PDBID: 3UT9) (68). The inter-helical Euler angles were then computed at the bent site and compared to the bending angle reported in literature as determined by other methods (e.g. Curves (69)) (Supplemental Information).

## RESULTS AND DISCUSSION

### Structural polymorphism in HG bps

Our survey identified a total of 106 A•T and 34 G•C HG bps that satisfy all three HG criteria in DNA duplexes. In addition, the survey identified 22 A•T and 16 G•C HG-like bps that satisfy one or two of the HG criteria (Figure 3A). Note that 91% of HG and 74% of HG-like bps have averaged B-factors over all heavy atoms (i.e. C, N, O and P) on the base and sugar-phosphodiester backbone  $\leq 60$  (Supplementary Figure S1b), indicating that they are reasonably well defined by the crystallography data.

The vast majority of the HG and HG-like bps (88%) are found in structures of duplex DNA in complex with proteins and/or ligands. Among these bps, most cases (96%) are not in direct contact with the protein or ligand and 66% are located at the duplex terminal ends. In contrast, all 16 HG bps observed in naked DNA duplexes correspond to pure HG helices of AT-repeats. No HG bps neighbored by WC bps on either end in naked DNA duplexes are observed. Many of these HG and HG-like bps, especially those lo-

cated at duplex termini, do not appear to be documented in the primary literature.

The survey identifies a total of 178 HG and HG-like bps (128 A•T and 50 G•C bps) which correspond to  $\sim 0.3\%$  of all 51485 A•T and G•C DNA bps in the PDB (as of 4/9/2013). Interestingly, this overall abundance of HG bps compares favorably to the population  $\sim 0.5\%$  (at pH  $\sim 6.8$ ) measured by NMR relaxation dispersion for transient HG bps in duplex DNA in solution (48). This suggests that the HG bps captured by NMR and X-ray crystallography are subject to similar energetic forces and that the differences in the experimental conditions do not lead to substantial changes in the overall abundance of HG versus WC bps. However, we cannot rule out that the environmental factors influence the distribution of HG bps and their specific location.

The 38 HG-like bps exclude the entries that only satisfy the constricted C1'–C1' distance criterion, which we consider 'distorted WC' bps. Most of the HG-like bps satisfy only the *syn* purine criterion (22 HG<sup>*syn*</sup>) or both the *syn* purine and constricted C1'–C1' distance (7 HG<sup>*syn+C1'*</sup>). The HG<sup>*syn*</sup> represents a partially open HG bp in which the purine and pyrimidine bases are not brought into proximity following the purine flip, explaining the absence of HG-type H-bonding (Supplementary Figure S1c, top). Interestingly, the HG<sup>*syn*</sup> conformation falls along a WC-to-HG transition pathway previously proposed based on peak conjugate refinement simulations (48) and by  $\Phi$ -value analysis (49). The HG<sup>*syn+C1'*</sup> bps feature deviations in shear that do not support HG-type H-bonds (Supplementary Figure S1c). Ap-

Table 1. Summary of DNA structures containing HG and HG-like bps

	PDBID	Sequence Context	Resolution (Å)	Biological Context	Context Function	Crystal Contact
<i>HG base pairs</i>	1RSB	5'-ATATAT-3'	2.17	naked DNA	N.A.	blunt-end stacking (HH)
	1XVK	5'-GCGTACGC-3'	1.26	echinomycin	antitumor antibiotic	blunt-end stacking (HH)
	1XVN	5'-ACGTACGT-3'	1.50	echinomycin	antitumor antibiotic	blunt-end stacking (HH)
	3EY0	5'-ATA	2.52	pentamidine	antiprotozoal drug	blunt-end stacking (HH)
	1VS2	5'-GCGTACGC-3'	2.00	trioestin A	antitumor antibiotic	blunt-end stacking (HH)
	3H80	T(MA7)G	2.00	alpha-ketoglutarate-dependent dioxygenase alkB homolog 2 (ABH2)	oxidoreductase; methylation lesion repair	
	3IGM	GCA-3'	2.20	apicomplexan apetala2 (AP2) domain	transcription regulator protein	
	2XCS	5'-AGC	2.10	S. aureus gyrase and antibacterial agent	type IIA topoisomerase; introduction of negative supercoils in DNA	protein contact
	4BUL	5'-AGC	2.60	S. aureus gyrase and antibacterial agent	type IIA topoisomerase; introduction of negative supercoils in DNA	protein contact
	1F2I	5'-ATG	2.35	fusion Cys2His2 zinc-finger protein	DNA binding	blunt end stacking (HH)
	3VOK	5'-ATG	2.00	heme-regulated transporter regulator (HrTR)	transcription regulator	
	1T3N	5'-AGG	2.30	human polymerase $\epsilon$	error-prone Y family polymerase; replicate DNA	
	2ALZ	5'-TGG	2.50	human polymerase $\epsilon$	error-prone Y family polymerase; replicate DNA	
	3GV5	TTC	2.00	human polymerase $\epsilon$	error-prone Y family polymerase; replicate DNA	
	4EBD	5'-CTG	2.57	human polymerase $\epsilon$	error-prone Y family polymerase; replicate DNA	
	4EYI	5'- <u>(DG)GG</u>	2.90	human polymerase $\epsilon$	error-prone Y family polymerase; replicate DNA	
	1K6I	TAA	2.10	Mating-type protein $\alpha$ -2 (MAT $\alpha$ 2) homeodomain	DNA binding	
	4ATI	AAC-3'	2.60	Microphthalmia-associated transcription factor (MITF)	DNA binding	
	2ATA	5'-AAG	2.20	human p53 core domain	tumor suppressor protein	blunt-end stacking (HH)
	3IGK	CATG	1.70	human p53 core domain	tumor suppressor protein	
	3KZ8	CATG	1.91	human p53 core domain	tumor suppressor protein	
	2ODI	5'-AAC	1.45	BcnI	typeII restriction endonuclease; recognize and excise DNA	sticky-end stacking (WH)
	3N7B	5'-AGT	2.65	SgrAI	excise DNA	blunt-end stacking (WH)
	2IBK	5'-TCATGA	2.25	sulfolobus solfataricus P2 DNA polymerase IV (Dpo4)	error-prone Y family polymerase; replicate DNA	
	3V6J	<u>(EFG)GA</u>	2.30	sulfolobus solfataricus P2 DNA polymerase IV (Dpo4)	error-prone Y family polymerase; replicate DNA	
	1QN3	ACGG	1.95	TATA-box binding protein	transcription initiation factor TFIID-1; specifically binds TATA-box DNA	
	1QN6	AGG	2.10	TATA-box binding protein	transcription initiation factor TFIID-1; specifically binds TATA-box DNA	
	1QNB	TGG	2.23	TATA-box binding protein	transcription initiation factor TFIID-1; specifically binds TATA-box DNA	
2VIH	CTTTAG	2.10	IS608 transposase	DNA binding and excision		
2VJU	CTTAG	2.40	IS608 transposase	DNA binding and excision		
2XM3	CTTAG	2.30	ISDra2 transposase	DNA binding and excision		
1ODG	5'-TAGGC(5CM)TG	2.80	very-short-patch repair (Vsr) enzyme	nucleotide excision repair of G•T mismatches		
<i>HG-like base pairs</i>	1QP5	CGG	2.60	naked DNA	N.A.	
	239D	GGG-3'	2.05	naked DNA	N.A.	
	2PIS	GAA(FFD)TT	2.80	naked DNA	N.A.	
	329D	CGG	2.70	naked DNA	N.A.	
	4E8X	CGG-3'	2.18	ruthenium complex	N.A.	
	1LWW	TAC	2.10	human 8-oxoguanine DNA glycosylase	base excision DNA repair	
	3GYH	TAG	2.80	alkyltransferase-like (ATL) protein	DNA alkylation damage repair	
	2WT7	5'-AAT	2.30	heterodimeric MafB: cFos	leucine zipper transcription factor	blunt-end stacking (WH)
	1K7A	5'-ACA	2.80	Ets domain of Ets-1	Ets family transcription activator	
	4I2O	TAT	1.77	FixK2 protein	transcription regulator	
	1IHF	CAA	2.50	integration host factor (IHF)	DNA binding; architectural factor	protein contact
	4AUW	5'-TAA	2.90	bZIP homodimeric MafB	transcription factor	sticky-end stacking (HT)
	1DE9	5'- <u>(3DR)GA</u>	3.00	human major apurinic/apyrimidinic endonuclease (APE1)	base excision DNA repair of apurinic/apyrimidinic DNA	
	3G2D	5'-CAG	2.30	Mth212 exodeoxyribonuclease	uridine/abasic endonuclease, 3'->5' exonuclease	
	3ODH	ATA-3'	2.30	OkrAI endonuclease	restriction endonuclease	stacking (WH)
	1S0M	5'- <u>AT(BPA)A</u>	2.70	sulfolobus solfataricus P2 DNA polymerase IV (Dpo4)	error-prone Y family polymerase; replicate DNA	
3V6H	(EFG)GA	2.30	polymerase IV (Dpo4)	error-prone Y family polymerase; replicate DNA		
1OZJ	5'-GTA	2.40	Smad3 MH1 DNA binding domain	transcription factor; DNA binding		

Shown are non-redundant HG bps that satisfy all three HG criteria (top) and also HG-like bps (bottom) as defined in Methods. Only the sequence of one strand surrounding the HG bp (colored in red) is shown and strand termini are indicated using '5-' and '3-'. Modifications with abbreviations from the PDB are shown in parentheses in the sequence. Residues that do not form bps in the helix are underlined. Interactions involving HG bps in crystal contacts are listed in the last column with types of stacking shown in the parentheses including HG-to-HG (HH), WC-to-HG (WH) and HG-to-TT mismatch (HT).



**Table 2.** Inter-helical Euler angles ( $\alpha_h$ ,  $\beta_h$ ,  $\gamma_h$ ) and the inter-helical twist angle ( $\zeta_h$ ) calculated across the junction of HG and HG-like bps

PDB	Residue	Sequence	H1 RMSD (Å)	H2 RMSD (Å)	$\alpha_h$ (°)	$\beta_h$ (°)	$\gamma_h$ (°)	$\zeta_h$ (°)
1K61	E. 4:10 F. 40:34	<b>TGT</b> AATT AC <b>ATT</b> AA	0.8	0.9	-18	18	46	-8
3KZ8	C. 2:9 D.19:12	<b>GGC</b> ATGCC CC <b>GTA</b> CGG	0.8	0.8	20	18	56	3
3IGL	A. 3:10 B. 10:3	<b>GGC</b> ATGCC CC <b>GTA</b> CGG	0.8	0.8	20	18	56	4
1QNB	E. 207:213 F. 222:216	<b>AAT</b> GGGC TT <b>AC</b> CCG	<u>2.4</u>	1.3	17	60	-1	-21
1QN3	E. 206:213 F. 223:216	<b>AAAC</b> GGGC TT <b>TGC</b> CCG	<u>2.9</u>	1.2	33	77	10	-29
1QN6	E. 207:213 F. 222:216	<b>ATAG</b> GGC T <b>ATC</b> CCG	<u>2.4</u>	1.1	17	61	-3	-22
1IHF	C. -32:-26 D. 32:26	<b>AGC</b> AATG TC <b>G(n)</b> TTAC	<u>2.8</u>	0.7	20	57	-9	-25
3H8O	B. 266:270 C. 278:274	<b>AT</b> ( <b>MA7</b> )GC T <b>AT</b> CG	1.2	0.8	-7	19	41	-3
1VS2	A. 2:7 B. 7:2	<b>CGT</b> ACG <b>GC</b> ATGC	<u>2.9</u>	<u>2.9</u>	-16	23	20	-69
1XVK	A. 2:7 B. 7:2	<b>CGT</b> ACG <b>GC</b> ATGC	<u>3.1</u>	<u>3.1</u>	-14	24	22	-65
2PIS	C. 3:7 D. 24:20	<b>CGA</b> (FFD) <b>GCT</b> T(FFD)	0.8	0.9	-29	13	61	-4
2PIS	D. 17:21 C. 10:6	<b>GA</b> (FFD)T <b>CT</b> T(FFD)A	0.6	0.9	10	6	29	2
1LWW	E. 26:30 D. 5:1	<b>CT</b> ACC <b>GAT</b> GG	1.0	1.0	-10	7	48	2
4I2O	X. 3:7 W. 26:22	<b>CT</b> ATC <b>GAT</b> AG	0.7	0.8	-162	11	-160	1
329D	B. 20:24 A. 5:1	<b>GCG</b> GT <b>CGC</b> CA	0.7	0.8	29	12	7	0

Sequences are shown corresponding to the given residues where the lower helices (H1) are marked in bold and HG or HG-like bps shown in red. Helices with superposition RMSD > 2 Å to the idealized B-form helix upon aligning heavy atoms in the sugar-phosphodiester backbone are underlined. The inter-helical twist angle is computed by  $\zeta_h = \alpha_h + \gamma_h - n \times 36^\circ$ , where  $n$  corresponds to the number of HG bps in the junction, assuming that each HG bp in the junction of H1 and H2 contributes to +36° helical twist change as in the B-form DNA.

proximately 62% of the HG<sup>syn</sup> and HG<sup>syn+Cl'</sup> bps are located at or near the duplex terminal ends (Supplementary Table S1). Another 8 bps satisfy the HG H-bonds and constricted Cl'-Cl' distance but feature *anti* rather than *syn* purine base (HG<sup>Hbond+Cl'</sup>). Seven of these HG<sup>Hbond+Cl'</sup> bps are located at a nicked site in the integration host factor (IHF)-DNA complex and another one occurs at the terminal end of a left-handed DNA duplex (Supplementary Figure S1c). We also find one bp that only satisfies the HG H-bonding criterion (Supplementary Figure S1c, bottom). This corresponds to an *anti* purine base with a C3'-endo sugar pucker that forms HG H-bonds through rearrangement of the sugar-phosphodiester backbone in a manner analogous to Z-DNA. Note that we cannot rule out that some of these bps arise due to modeling errors and ambiguous density. Indeed, it was previously shown that an open G·U pair without hydrogen bonding in the ribozyme structure (PDBID: 1CX0) could be more confidently assigned to be a reverse wobble G·U bp after single-residue remodeling with ERRASER (70).

### Structure and sequence preferences of HG bps

We examined a set of 54 non-redundant HG bps (see Methods, Table 1) to assess the statistical significance of their position and sequence preferences. In this non-redundant set, HG bps are ~4-fold more enriched in A•T versus G•C bps at both helical (by ~5-fold) and terminal (by ~3-fold) sites (Figure 3B). Similar preferences are observed when considering all 140 HG bps. The increased preference for A•T versus G•C<sup>+</sup> HG bps is consistent with the ~8-fold greater abundance of transient A•T versus G•C<sup>+</sup> HG bps measured in duplex DNA in solution by NMR relaxation dispersion (48,71). The lower abundance of G•C<sup>+</sup> HG bps can be attributed to the loss of one H-bond as well as by the requirement to protonate cytosine N3 (pKa ~7.2) (71). The average pH and standard deviations in the crystallization conditions for structures containing A•T or G•C HG both have pH~6±1 as compared to the overall average pH~7±1 of all DNA structures in the PDB with resolution ≤3.5 Å (Supplementary Figure S6). Interestingly, three structures of DNA-quinoxaline bis-intercalator complexes with G•C

HG bps are collected under rather acidic conditions (pH = 4.5) (31,72). In these cases, it is possible that the lower pH contributes toward the stabilization of the HG bps. Other structures with only A•T HG bps from similar complexes are observed at higher pH~6.

The overall ratio between bps in the interior of DNA helices and bps at helix termini is ~6:1. The same ratio for non-redundant HG bps identified in the survey is ~1:1 (Figure 3B), implying that HG bps are enriched at terminal ends. The terminal bps should be treated with caution given that they have increased susceptibility to dynamics and intrinsic structural noise. However, close examination of the electron density maps (see Methods) at these terminal HG bps reveals that most have good electron density and low to moderate average B-factors (Supplementary Table S3). The examination of non-redundant HG and HG-like bps reveals that 66% of all terminal HG and HG-like bps are involved in crystal contacts with nearby DNA or proteins in the crystal lattice while the remaining 34% of terminal HG bps are observed within active sites of polymerases, endonucleases or as the apical loop closing bps showing no contact with nearby molecules. It is possible that these terminal HG bps are stabilized, at least in part, by crystal packing forces. All of the HG bps in crystal contact with nearby DNA molecules show end-to-end stacking, 86% of which involves very similar blunt-ended HG-to-HG stacking with 2-fold symmetry while 14% are attributed to HG-to-WC stacking with translational symmetry in the crystal contact (Table 1). Enrichment of one specific type of stacking in crystal contacts indicates that favorable stacking could be a reason for the enrichment of HG bps at terminal ends. Nevertheless, one should not solely consider these HG bps as crystallographic artifacts. It may well be that packing in the crystal lattice could have parallels *in vivo* including in nucleosomes, chromatin, and possibly other stressed cellular DNA. This suggests that HG bps may play unique roles at terminus-involved DNA biochemical transactions such as active sites of DNA polymerases, nucleases, transposases and ligases involved in DNA replication, recombination and various damage repair pathways. This also suggests that transient HG bps may also be more abundant at terminal ends, possibly as intermediates that have been observed accompanying end-fraying events (50,73). This preference for A•T HG bps at terminal ends is consistent with early studies showing that isolated 9-methyladenine and 1-methylthymine bases prefer to associate as HG rather than WC bps (1,7).

Interestingly the *syn* purine bases in terminal HG bps show a 12-fold preference for being at the 5' versus the 3'-end. A recent study showed that sequence-specific variations in the stabilities of HG bps in duplexes can be attributed to variations in WC stability (49,50). These A•T HG preferences may also mirror variations in WC stabilities and indeed prior NMR studies suggest weaker stability for 3' versus 5' terminal guanine WC bps (74). End-to-end intermolecular stacking could also favor 5'-purine A•T HG bps as this results in a TA dinucleotide step formed in the crystal contact which favors HG (19,49).

Although the limited number of HG bps does not allow a statistically rigorous analysis of sequence-specific preferences, we note a few observations. First, we only observe either single HG bps surrounded by WC bps or two tan-

dem (TA, AT or CG) HG bps that are palindromic. There are also three entire helices that are formed exclusively of HG bps in sequences consisting of three (13,75–76) or five AT repeats (77). We do not observe three or more consecutive HG bps that are surrounded by WC bps. In addition, the TA step (HG bp is underlined) is the most frequently observed HG step representing 47% of all single HG bp-involved steps while the GG step is completely absent. These sequence-specific preferences are in good agreement with studies showing that HG bps favor AT-rich sequences (7,48–49) and NMR relaxation dispersion studies showing the greatest abundance of transient HG bp is in TA steps and lowest abundance is in GG steps (49).

### Impact of HG bps on local B-form DNA structure

Prior structural (29,75–76) and computational studies (18,19) have shown that HG bps can be accommodated within B-form helices of WC bps without significantly distorting the base-pairing geometry and sugar-phosphodiester backbone of neighboring WC bps. Nevertheless, studies have reported small perturbations induced by HG pairing including  $\alpha/\gamma$  torsion angles with *gauche*<sup>+</sup>/*gauche*<sup>-</sup> rather than the common *gauche*<sup>-</sup>/*gauche*<sup>+</sup> at the HG bp (29,75). Our survey provides an opportunity to examine any local perturbations that may be induced by formation of HG bps.

To this end, we compared 1D histogram distributions of sugar ( $\nu_0$ – $\nu_4$ ) and phosphodiester backbone ( $\alpha$ ,  $\beta$ ,  $\gamma$ ,  $\delta$ ,  $\epsilon$ ,  $\zeta$ ) torsion angles (Figure 4A) for three data sets (i) non-redundant HG bps ('HG'), (ii) WC bps adjacent to HG ('adjacent WC') and (iii) WC bps surrounded by WC bps ('control WC'). Visual inspection reveals relatively broad distributions with no discernable differences between the three sets of distributions (Figure 4C). To more quantitatively compare the similarities between the distributions, we used the REsemble approach (57), which was recently developed to measure the extent of similarity between histogram distributions (see Methods). In this approach, the similarity between two distributions is measured by computing  $\Sigma\Omega$  which ranges between 0 and 1 for maximum and minimum similarity, respectively (see Methods) (57). In general, we observe high similarity between the three datasets with  $\Sigma\Omega \leq 0.2$  for HG versus control WC and adjacent WC versus control WC with the values being generally lower for sugar torsion angles ( $\Sigma\Omega < 0.1$ ) (Supplementary Figure S3b). However, among these small differences, the relatively larger deviations ( $\Sigma\Omega > 0.1$ ) are in  $\gamma$  and  $\zeta$  for HG versus control WC and in  $\alpha$  for adjacent WC versus control WC, suggesting that HG bps more likely induce changes in these torsion angles (Supplementary Figure S3b).

Although only small deviations are apparent when comparing 1D (Supplementary Figure S3b) and 2D (Supplementary Figure S4) distributions of local torsion angles, there could be more significant deviations induced by HG bps that are not captured because they involve small correlated variations in different torsion angles. To examine this, we used REsemble to compare 5D sugar ( $\nu_0$ ,  $\nu_1$ ,  $\nu_2$ ,  $\nu_3$ ,  $\nu_4$ ) and 6D phosphodiester ( $\alpha$ ,  $\beta$ ,  $\gamma$ ,  $\delta$ ,  $\epsilon$ ,  $\zeta$ ) torsion angle distributions across the various sets of bps. Note that carrying out higher dimensionality comparisons (e.g.

11D encompassing both sugar and phosphodiester torsion angles) is very computationally costly (see Methods). Although the probability for overlap decreases rapidly for multi-dimensional distributions, this inherent decrease in overlap is taken into account by evaluating the overlap between the control WC distribution and sub-distributions of its own (see Methods). Based on this analysis, we find that 6D distributions (see Methods) of six phosphodiester backbone torsion angles for both HG bps ( $\Sigma\Omega_{\text{HGbp}} = 0.42$ ) and adjacent WC bps ( $\Sigma\Omega_{\text{adWC}} = 0.37$ ) deviate significantly from control WC bps as compared to the subset WC bps taken from the control WC bps ( $\Sigma\Omega_{\text{subWC}} = 0.25$ ) (Figure 4B). This indicates that torsion angles in HG bps and adjacent WC bps deviate from the control WC bps even when taking into account intrinsic statistical deviations in the multi-dimensional distribution of torsion angles for the control WC bps (Figure 4B). Similar but smaller deviations are observed for the 5D distributions of sugar torsion angles ( $\nu_0, \nu_1, \nu_2, \nu_3, \nu_4$ ) (Figure 4B).

Examination of histogram distributions and REsemble analyses for local base-pair parameters (shear, stretch, stagger, buckle, propeller, opening), C1'–C1' distances, heavy atom distances in H-bonds reveals slight differences in base-pair geometries between adjacent WC and control WC datasets. In particular, although the differences are small based on 1D REsemble analysis ( $\Sigma\Omega < 0.2$ ), 6D REsemble analysis shows local base-pair parameters in the adjacent WC deviate from control WC bps ( $\Sigma\Omega_{\text{adWC}} = 0.31$ ), compared to that of a subset taken from control WC bps ( $\Sigma\Omega_{\text{subWC}} = 0.20$ ) (Supplementary Figure S2c). The largest deviations are observed for shear, opening, C1'–C1' distances as well as heavy atom distances in H-bonds for adjacent WC bps compared to control WC bps (Supplementary Figure S2b). Based on 1D histogram distributions, the adjacent WC bps tend to have larger shear and opening, together with shorter heavy atom distances in H-bonds as compared to control WC bps (Supplementary Figure S2a). These perturbations on the WC bps adjacent to HG are consistent with the observation of exchange broadening of aromatic resonances in WC bps adjacent to HG bps that are trapped by N1-methylation which suggests enhanced dynamics (48,71).

### Impact of HG bps on global B-form DNA structure

In crystal structures of DNA–IHF (e.g. PDBID: 1IHF) and DNA–TBP (e.g. PDBID: 1QN3) complexes, HG bps are observed near sharp kinks in the DNA (15,16). Interestingly, kinking across HG helices with sticky ends has also been observed in the absence of proteins or ligands in a coiled-coil DNA structure with sequence d(CGATATATATAT) (PDBID: 2AF1) where the (AT)<sub>5</sub> HG bps form a linear HG helix but there is a kink between two HG helices at the junction of two intermolecular G–C bps formed by the sticky ends (77). To examine whether HG bps are more generally associated with DNA bending, we manually examined all DNA structures (total of 15) containing HG bp(s) flanked by at least two WC bps.

Among these 15 DNA structures, 10 contain a single HG bp while five contain tandem HG bps. Interestingly, we find evidence for bending across all single and tandem HG bps.

In some cases, the ability to characterize global bending is obscured by local structural distortions arising due to the presence of nicks in DNA strands and protein/ligand interactions. However, clear signs of bending are observed for the complexes of DNA–MAT $\alpha$ 2 (PDBID: 1K61) (Figure 5D), DNA–p53 (PDBID: 3KZ8) (Figure 5D) and DNA–AlkBH2 (PDBID: 3H8O) that show very little local deviation from the reference B-form structure with superposition RMSD of H1 and H2 < 2 Å (see below and Supplementary Figure S5a and c).

To put these observations on a quantitative footing, we adapted our approach for computing inter-helical Euler angles in RNA junctions to compute inter-helical Euler angles in DNA (59,60) (see Methods). We first benchmarked this approach on 13 DNA structures that do not contain HG bps and that show either the absence or presence of bending (Supplementary Table S4). The computed inter-helical Euler angles were consistent with previous analyses of these DNA duplexes. The linear DNA duplexes yielded an average computed inter-helical bend angle of  $\beta_h = 6^\circ$  with a narrow standard deviation ( $\sigma = 2^\circ$ ) and average inter-helical twist angle of  $\zeta_h = -3^\circ$  with a narrow  $\sigma = 2^\circ$  (Supplementary Table S4). The computed averages and standard deviations of  $\beta_h$  and  $\zeta_h$  are insensitive to having one or two WC bps as the junction between H1 and H2 (data not shown). As expected these angles are near zero when considering the uncertainty in the computed inter-helical bend angle of  $\sim 5^\circ$  arising due to superposition inaccuracy for short helical segments (59). For the bent duplexes, the computed inter-helical Euler angles capture the degree and direction of bending previously reported in the solution structure of A6–T6 A-tract DNA (PDBID: 1FZX) (67) and the nucleosome DNA (PDBID: 3UT9) (68) (Supplemental Information).

By comparison with the negative control results ( $\beta_h = 6^\circ \pm 2^\circ$ ), we observed consistently higher degrees of bending in the HG-containing duplexes overall (average  $\beta_h = 37^\circ$  with a range of  $\sigma = 23^\circ$ ) (Table 2). Substantial bending (average  $\beta_h = 14^\circ$  with  $\sigma = 5^\circ$ ) is observed for structures that show little local distortion (superposition RMSD < 2 Å) from the idealized B-form helix (Table 2). A smaller degree of bending is observed for two A–T bps in two similar DNA–p53 complexes when they adopt WC ( $\beta_h \sim 8^\circ$  in structure with PDBID: 3KMD) rather than HG ( $\beta_h \sim 18^\circ$  in structure with PDBID: 3KZ8) geometry. Significant bending ( $\beta_h \sim 14^\circ$ ) is also observed at a HG-like bp, which does not satisfy HG H-bonding found in a naked DNA duplex (PDBID: 2PIS) containing the modification of 3-fluorobenzene (78).

These results imply that HG bps are associated with a modest degree of DNA bending though contributions from environmental factors such as protein/ligand interaction and crystal packing cannot be ruled out. This is consistent with spin relaxation dispersion showing that the population of transient HG bps in CA steps increases in longer A-tracts, which are known to induce DNA bending (48). This increase in global bending may arise in part from the correlated local variations in sugar and phosphodiester torsion angles identified by REsemble (Figure 4B). It should be noted that the computed inter-helical bend angles are in principle subject to uncertainties arising due to a small number of bps flanking the HG bp available for superposition as

well as local distortions in the helical structure (see Methods). However, control calculations examining the effect of varying the number of bps and the types of atoms used in the superposition suggest that the observed bending can be robustly defined (see Methods).

Interestingly, the HG bending is consistently directed toward the major groove ( $-90^\circ \leq \gamma_h \leq 90^\circ$ ) as compared to the more random bending directions observed for the negative control linear B-form DNA structures (Figure 5F). This directional bending is observed consistently across naked DNA and diverse DNA-protein complexes containing HG and HG-like bps (Table 2) under different crystallization conditions. The only exception is an HG-like bp (PDBID: 4I2O) showing bending toward the minor groove that also features the least constricted C1'-C1' distance (see below). In addition, we observe an inverse correlation ( $R^2 \sim 0.8$ ) between the degree of bending ( $\beta_h$ ) and the C1'-C1' distance of HG and HG-like bps in intact DNA helices (Figure 5E). This trend holds even for duplexes that have only minor local distortions (superposition RMSD to idealized B-form  $< 2$  Å). This suggests a mechanism for correlating HG and bending. In particular, it becomes increasingly difficult to accommodate a WC bp geometry under more constricted C1'-C1' distances due to steric clashes that arise between the bases. This steric clash can be released by forming HG bps, which in turn makes possible a range of conformations with variable constricted C1'-C1' distances, sugar-backbone distortions and DNA bend angles. It is interesting to note that constricted C1'-C1' distances naturally lead to a narrowed minor groove of the bp which can result in favorable electrostatic interactions through minor groove recognition as observed for example with Arg248 in the structure of the p53-DNA complex (17,79). It is also important to note that DNA bending does not necessarily require constriction of the C1'-C1' distance; bending may arise also due to local translation or rotation of WC base-pair steps (e.g. in roll, tilt and/or propeller twist) as proposed for bending in the nucleosome DNA (68) and A-tract DNA (67,80). For example, the average C1'-C1' distance in the curved nucleosome structure (PDBID: 1KX5) is  $\sim 10.6$  Å  $\pm 0.3$  Å.

In contrast to the bend angle, HG bps do not apparently lead to significant changes in the inter-helical twist angle  $\zeta_h$  (see Methods). For example, a single HG bp leads to over-twisting by  $\sim 8^\circ$  in the complex structure of DNA-MAT $\alpha$ 2 (PDBID: 1K61) and  $\sim 3^\circ$  in the complex structure of DNA-AlkBH2 (PDBID: 3H8O) (Table 2). Additional data is needed to examine the effects of single HG bps, though based on these two observations, one would predict that single HG bps are favored by positive DNA supercoiling as arises for example, in front of RNA polymerases during transcription and in front of the replication fork during DNA replication. On the other hand, the observed inter-helical twist angle ( $\zeta_h = -3^\circ$ ) across tandem HG bps in the p53-DNA complex (Figure 5D) is within error of control B-form DNA structures ( $\zeta_h = -3^\circ \pm 2^\circ$ ); thus no evident tendency to over- or under-twisting of tandem HG bps can be concluded from the current survey. Similarly in five other structures containing single HG-like bps (Table 2), the computed inter-helical twist angles range from  $-4^\circ$  to  $2^\circ$ , which shows no preference for over- or under-twisting.

HG-mediated DNA bending may provide a new mechanism for indirect DNA sequence-specific recognition. Many DNA binding proteins bind DNA as oligomers and interact with multiple sites along the DNA. HG-induced bending could play topological roles defining the geometrical presentation of distant binding sites along the DNA duplex.

## CONCLUSION

The survey reveals that HG bps exist in B-form DNA at a similar level of abundance relative to WC bps ( $\sim 0.3\%$ ), as is observed transiently in canonical duplex DNA under solution conditions by NMR. The HG bps are  $\sim 4$ -fold more abundant in A•T versus G•C<sup>+</sup> bps, in good agreement with the  $\sim 8$ -fold larger populations observed for transient A•T versus G•C<sup>+</sup> HG bps in solution by NMR. Likewise they appear to be more abundant in AT-rich sequences, consistent with sequence-specific preferences observed by NMR. The survey also reveals that HG bps are  $\sim 6$ -fold more abundant at helix termini versus the helix interior with a preference for the *syn* purine base being located at the 5'-terminal end. Such terminal HG bps may play unique roles in DNA biochemical transactions. Future NMR studies should examine the abundance of transient HG bps at terminal ends. In addition, the survey uncovers HG-like bps that only satisfy one or two of the HG criteria. These conformations provide insights into the potential mechanisms for WC-to-HG transitions. The survey also suggests that HG bps induce small local perturbations, particularly on phosphodiester backbone torsion angles  $\alpha$ ,  $\gamma$  and  $\zeta$ . Perhaps the most significant finding in the survey is the observation that HG bps induce significant degrees of DNA bending ( $\sim 14^\circ$ ) and in a manner that is inversely proportional to the C1'-C1' distance across the bp. This HG-mediated DNA bending may provide a new mechanism for indirect DNA sequence-specific recognition.

## SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

## ACKNOWLEDGEMENTS

We thank Dr Yi Xue for helping with PyMOL scripts for preparing input structure files of double-stranded DNA deposited as single chains in one asymmetric unit in the PDB for the 3DNA program, Dr Evgenia N. Nikolova, Dr Heidi S. Alvey and all Al-Hashimi lab members for valuable discussions and suggestions. We also thank Shouri Gottiparthi, our high school research fellow, who assisted in gathering modeled HG bps from the PDB. Figure 5 and Supplementary Figure S5 were generated in part with PyMOL (The Pymol Molecular Graphics System, Schrödinger LLC).

## FUNDING

National Institutes of Health [R01-GM089846 to H.M.A., R01-GM073919 to D.C.R., P01-GM063210 to J.S.R.]. Funding for open access charge: Agilent Thought Leader Award to H.M.A.

*Conflict of interest statement.* None declared.

## REFERENCES

- Hoogsteen, K. (1959) The structure of crystals containing a hydrogen-bonded complex of 1-methylthymine and 9-methyladenine. *Acta Crystallogr.*, **12**, 822–823.
- Felsenfeld, G. and Rich, A. (1957) Studies on the formation of two- and three-stranded polyribonucleotides. *Biochim. Biophys. Acta*, **26**, 457–468.
- Haschemeyer, A.E. and Sobell, H.M. (1963) The crystal structure of an intermolecular nucleoside complex: adenosine and 5-bromouridine. *Proc. Natl Acad. Sci. U.S.A.*, **50**, 872–877.
- Mathews, F.S. and Rich, A. (1964) The molecular structure of a hydrogen bonded complex of N-ethyl adenine and N-methyl uracil. *J. Mol. Biol.*, **8**, 89–95.
- Haschemeyer, A.E. and Sobell, H.M. (1965) The crystal structure of a hydrogen bonded complex of adenosine and 5-bromouridine. *Acta Crystallogr.*, **18**, 525–532.
- Haschemeyer, A.E. and Sobell, H.M. (1965) The crystal structure of a hydrogen bonded complex of deoxyguanosine and 5-bromodeoxycytidine. *Acta Crystallogr.*, **19**, 125–130.
- Nikolova, E.N., Zhou, H., Gottardo, F.L., Alvey, H.S., Kimsey, I.J. and Al-Hashimi, H.M. (2013) A historical account of Hoogsteen base-pairs in duplex DNA. *Biopolymers*, **99**, 955–968.
- Pranata, J., Wierschke, S.G. and Jorgensen, W.L. (1991) OPLS potential functions for nucleotide bases. Relative association constants of hydrogen-bonded base pairs in chloroform. *J. Am. Chem. Soc.*, **113**, 2810–2819.
- Trollope, K.I., Gould, I.R. and Hillier, I.H. (1993) Modelling of electrostatic interactions between nucleotide bases using distributed multipoles. *Chem. Phys. Lett.*, **209**, 113–116.
- Gould, I.R. and Kollman, P.A. (1994) Theoretical investigation of the hydrogen bond strengths in guanine-cytosine and adenine-thymine base pairs. *J. Am. Chem. Soc.*, **116**, 2493–2499.
- Day, R.O., Seeman, N.C., Rosenberg, J.M. and Rich, A. (1973) A crystalline fragment of the double helix: the structure of the dinucleoside phosphate guanylyl-3',5'-cytidine. *Proc. Natl Acad. Sci. U.S.A.*, **70**, 849–853.
- Rosenberg, J.M., Seeman, N.C., Kim, J.J., Suddath, F.L., Nicholas, H.B. and Rich, A. (1973) Double helix at atomic resolution. *Nature*, **243**, 150–154.
- Abrescia, N.G., Thompson, A., Huynh-Dinh, T. and Subirana, J.A. (2002) Crystal structure of an antiparallel DNA fragment with Hoogsteen base pairing. *Proc. Natl Acad. Sci. U.S.A.*, **99**, 2806–2811.
- Wang, A.H., Ughetto, G., Quigley, G.J., Hakoshima, T., van der Marel, G.A., van Boom, J.H. and Rich, A. (1984) The molecular structure of a DNA-triostin A complex. *Science*, **225**, 1115–1121.
- Rice, P.A., Yang, S., Mizuuchi, K. and Nash, H.A. (1996) Crystal structure of an IHF-DNA complex: a protein-induced DNA U-turn. *Cell*, **87**, 1295–1306.
- Patikoglou, G.A., Kim, J.L., Sun, L., Yang, S.H., Kodadek, T. and Burley, S.K. (1999) TATA element recognition by the TATA box-binding protein has been conserved throughout evolution. *Genes Dev.*, **13**, 3217–3230.
- Kitayner, M., Rozenberg, H., Rohs, R., Suad, O., Rabinovich, D., Honig, B. and Shakked, Z. (2010) Diversity in DNA recognition by p53 revealed by crystal structures with Hoogsteen base pairs. *Nat. Struct. Mol. Biol.*, **17**, 423–429.
- Cubero, E., Abrescia, N.G.A., Subirana, J.A., Luque, F.J. and Orozco, M. (2003) Theoretical study of a new DNA structure: the antiparallel Hoogsteen duplex. *J. Am. Chem. Soc.*, **125**, 14603–14612.
- Cubero, E., Luque, F.J. and Orozco, M. (2006) Theoretical study of the Hoogsteen-Watson-Crick junctions in DNA. *Biophys. J.*, **90**, 1000–1008.
- Cubero, E., Luque, F.J. and Orozco, M. (2001) Theoretical studies of d(A:T)-based parallel-stranded DNA duplexes. *J. Am. Chem. Soc.*, **123**, 12018–12025.
- Hashem, G.M., Wen, J.D., Do, Q. and Gray, D.M. (1999) Evidence from CD spectra and melting temperatures for stable Hoogsteen-paired oligomer duplexes derived from DNA and hybrid triplexes. *Nucleic Acids Res.*, **27**, 3371–3379.
- Escudé, C., Mohammadi, S., Sun, J.S., Nguyen, C.H., Bisagni, E., Liquier, J., Taillandier, E., Garestier, T. and Hélène, C. (1996) Ligand-induced formation of Hoogsteen-paired parallel DNA. *Chem. Biol.*, **3**, 57–65.
- Sklenár, V. and Felgon, J. (1990) Formation of a stable triplex from a single DNA strand. *Nature*, **345**, 836–838.
- Escaja, N., Gómez-Pinto, I., Rico, M., Pedroso, E. and González, C. (2003) Structures and stabilities of small DNA dumbbells with Watson-Crick and Hoogsteen base pairs. *ChemBioChem*, **4**, 623–632.
- Hoogsteen, K. (1963) The crystal and molecular structure of a hydrogen-bonded complex between 1-methylthymine and 9-methyladenine. *Acta Crystallogr.*, **16**, 907–916.
- Lu, L., Yi, C., Jian, X., Zheng, G. and He, C. (2010) Structure determination of DNA methylation lesions N1-meA and N3-meC in duplex DNA using a cross-linked protein-DNA system. *Nucleic Acids Res.*, **38**, 4415–4425.
- Nair, D.T., Johnson, R.E., Prakash, S., Prakash, L. and Aggarwal, A.K. (2004) Replication by human DNA polymerase- $\iota$  occurs by Hoogsteen base-pairing. *Nature*, **430**, 377–380.
- Johnson, R.E., Prakash, L. and Prakash, S. (2005) Biochemical evidence for the requirement of Hoogsteen base pairing for replication by human DNA polymerase  $\iota$ . *Proc. Natl Acad. Sci. U.S.A.*, **102**, 10466–10471.
- Aishima, J., Gitti, R.K., Noah, J.E., Gan, H.H., Schlick, T. and Wolberger, C. (2002) A Hoogsteen base pair embedded in undistorted B-DNA. *Nucleic Acids Res.*, **30**, 5244–5252.
- Quigley, G.J., Ughetto, G., van der Marel, G.A., van Boom, J.H., Wang, A.H. and Rich, A. (1986) Non-Watson-Crick G-C and A-T base pairs in a DNA-antibiotic complex. *Science*, **232**, 1255–1258.
- Cuesta-Seijo, J.A. and Sheldrick, G.M. (2005) Structures of complexes between echinomycin and duplex DNA. *Acta Crystallogr. D Biol. Crystallogr.*, **61**, 442–448.
- Cuesta-Seijo, J.A., Weiss, M.S. and Sheldrick, G.M. (2006) Serendipitous SAD phasing of an echinomycin-(ACGTACGT)<sub>2</sub> bisintercalation complex. *Acta Crystallogr. D Biol. Crystallogr.*, **62**, 417–424.
- Pföh, R., Cuesta-Seijo, J.A. and Sheldrick, G.M. (2009) Interaction of an echinomycin-DNA complex with manganese ions. *Acta Crystallogr. F*, **65**, 660–664.
- Singh, U.S., Moe, J.G., Reddy, G.R., Weisenseel, J.P., Marnett, L.J. and Stone, M.P. (1993) <sup>1</sup>H NMR of an oligodeoxynucleotide containing a propanodeoxyguanosine adduct positioned in a (CG)<sub>3</sub> frameshift hotspot of *Salmonella typhimurium* hisD3052: Hoogsteen base-pairing at pH 5.8. *Chem. Res. Toxicol.*, **6**, 825–836.
- Shanmugam, G., Kozekov, I.D., Guengerich, F.P., Rizzo, C.J. and Stone, M.P. (2008) Structure of the 1,N<sub>2</sub>-ethenedeoxyguanosine adduct opposite cytosine in duplex DNA: Hoogsteen base pairing at pH 5.2. *Chem. Res. Toxicol.*, **21**, 1795–1805.
- Yang, H., Zhan, Y., Fenn, D., Chi, L.M. and Lam, S.L. (2008) Effect of 1-methyladenine on double-helical DNA structures. *FEBS Lett.*, **582**, 1629–1633.
- García, R.G., Ferrer, E., Macías, M.J., Eritja, R. and Orozco, M. (1999) Theoretical calculations, synthesis and base pairing properties of oligonucleotides containing 8-amino-2'-deoxyadenosine. *Nucleic Acids Res.*, **27**, 1991–1998.
- Soliva, R., García, R.G., Blas, J.R., Eritja, R., Asensio, J.L., González, C., Luque, F.J. and Orozco, M. (2000) DNA-triplex stabilizing properties of 8-aminoguanine. *Nucleic Acids Res.*, **28**, 4531–4539.
- Cubero, E., Aviñó, A., de la Torre, B.G., Frieden, M., Eritja, R., Luque, F.J., González, C. and Orozco, M. (2002) Hoogsteen-based parallel-stranded duplexes of DNA. Effect of 8-Amino-purine Derivatives. *J. Am. Chem. Soc.*, **124**, 3133–3142.
- Makarova, A.V. and Kulbachinskiy, A.V. (2012) Structure of human DNA polymerase  $\iota$  and the mechanism of DNA synthesis. *Biochemistry (Moscow)*, **77**, 547–561.
- Gao, X. and Patel, D.J. (1988) NMR studies of echinomycin bisintercalation complexes with d(A1-C2-G3-T4) and d(T1-C2-G3-A4) duplexes in aqueous solution: sequence-dependent formation of Hoogsteen A1.cntdot.T4 and Watson-Crick T1.cntdot.A4 base pairs flanking the bisintercalation site. *Biochemistry*, **27**, 1744–1751.
- Gilbert, D.E., van der Marel, G.A., van Boom, J.H. and Feigon, J. (1989) Unstable Hoogsteen base pairs adjacent to echinomycin binding sites within a DNA duplex. *Proc. Natl Acad. Sci. U.S.A.*, **86**, 3006–3010.

43. Singh, U.C., Pattabiraman, N., Langridge, R. and Kollman, P.A. (1986) Molecular mechanical studies of d(CGTAACG)<sub>2</sub>: complex of triostin A with the middle A - T base pairs in either Hoogsteen or Watson-Crick pairing. *Proc. Natl Acad. Sci. U.S.A.*, **83**, 6402–6406.
44. Gallego, J., Ortiz, A.R. and Gago, F. (1993) A molecular dynamics study of the bis-intercalation complexes of echinomycin with d(ACGT)<sub>2</sub> and d(TCGA)<sub>2</sub>: rationale for sequence-specific Hoogsteen base pairing. *J. Med. Chem.*, **36**, 1548–1561.
45. Palmer, A.G. III (2014) Chemical exchange in biomacromolecules: past, present, and future. *J. Magn. Reson.*, **241**, 3–17.
46. Mittermaier, A.K. and Kay, L.E. (2009) Observing biological dynamics at atomic resolution using NMR. *Trends Biochem. Sci.*, **34**, 601–611.
47. Sekhar, A. and Kay, L.E. (2013) NMR paves the way for atomic level descriptions of sparsely populated, transiently formed biomolecular conformers. *Proc. Natl Acad. Sci. U.S.A.*, **110**, 12867–12874.
48. Nikolova, E.N., Kim, E., Wise, A.A., O'Brien, P.J., Andricioaei, I. and Al-Hashimi, H.M. (2011) Transient Hoogsteen base pairs in canonical duplex DNA. *Nature*, **470**, 498–502.
49. Alvey, H.S., Gottardo, F.L., Nikolova, E.N. and Al-Hashimi, H.M. (2014) Widespread transient Hoogsteen base pairs in canonical duplex DNA with variable energetics. *Nat. Commun.*, **5**, 4786.
50. Santa Lucia, J. Jr, Allawi, H.T. and Seneviratne, P.A. (1996) Improved nearest-neighbor parameters for predicting DNA duplex stability. *Biochemistry*, **35**, 3555–3562.
51. Nikolova, E.N., Stull, F. and Al-Hashimi, H.M. (2014) Guanine to inosine substitution leads to large increases in the population of a transient G•C Hoogsteen base pair. *Biochemistry*, **53**, 7145–7147.
52. Wang, J. (2005) DNA polymerases: Hoogsteen base-pairing in DNA replication? *Nature*, **437**, E6–E7.
53. Ethayathulla, A.S., Tse, P.W., Monti, P., Nguyen, S., Inga, A., Fronza, G. and Viadiu, H. (2012) Structure of p73 DNA-binding domain tetramer modulates p73 transactivation. *Proc. Natl Acad. Sci. U.S.A.*, **109**, 6066–6071.
54. Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N. and Bourne, P.E. (2000) The Protein Data Bank. *Nucleic Acids Res.*, **28**, 235–242.
55. Lu, X.J. and Olson, W.K. (2003) 3DNA: a software package for the analysis, rebuilding and visualization of three-dimensional nucleic acid structures. *Nucleic Acids Res.*, **31**, 5108–5121.
56. Olson, W.K., Bansal, M., Burley, S.K., Dickerson, R.E., Gerstein, M., Harvey, S.C., Heinemann, U., Lu, X.J., Neidle, S., Shakked, Z. et al. (2001) A standard reference frame for the description of nucleic acid base-pair geometry. *J. Mol. Biol.*, **313**, 229–237.
57. Yang, S., Salmon, L. and Al-Hashimi, H.M. (2014) Measuring similarity between dynamic ensembles of biomolecules. *Nat. Methods*, **11**, 552–554.
58. Fisher, C.K., Huang, A. and Stultz, C.M. (2010) Modeling intrinsically disordered proteins with bayesian statistics. *J. Am. Chem. Soc.*, **132**, 14919–14927.
59. Bailor, M.H., Mustoe, A.M., Brooks, C.L. III and Al-Hashimi, H.M. (2011) 3D maps of RNA interhelical junctions. *Nat. Protoc.*, **6**, 1536–1545.
60. Bailor, M.H., Sun, X. and Al-Hashimi, H.M. (2010) Topology links RNA secondary structure with global conformation, dynamics, and adaptation. *Science*, **327**, 202–206.
61. Musselman, C., Pitt, S., Gulati, K., Foster, L.L., Andricioaei, I. and Al-Hashimi, H. (2006) Impact of static and dynamic A-form heterogeneity on the determination of RNA global structural dynamics using NMR residual dipolar couplings. *J. Biomol. NMR*, **36**, 235–249.
62. Lu, X.J. and Olson, W.K. (2008) 3DNA: a versatile, integrated software system for the analysis, rebuilding and visualization of three-dimensional nucleic-acid structures. *Nat. Protoc.*, **3**, 1213–1227.
63. Bailor, M.H., Musselman, C., Hansen, A.L., Gulati, K., Patel, D.J. and Al-Hashimi, H.M. (2007) Characterizing the relative orientation and dynamics of RNA A-form helices using NMR residual dipolar couplings. *Nat. Protoc.*, **2**, 1536–1546.
64. Olson, W.K., Gorin, A.A., Lu, X.J., Hock, L.M. and Zhurkin, V.B. (1998) DNA sequence-dependent deformability deduced from protein–DNA crystal complexes. *Proc. Natl Acad. Sci. U.S.A.*, **95**, 11163–11168.
65. Dans, P.D., Pérez, A., Faustino, I., Lavery, R. and Orozco, M. (2012) Exploring polymorphisms in B-DNA helical conformations. *Nucleic Acids Res.*, **40**, 10668–10678.
66. Svozil, D., Kalina, J., Omelka, M. and Schneider, B. (2008) DNA conformations and their sequence preferences. *Nucleic Acids Res.*, **36**, 3690–3706.
67. MacDonald, D., Herbert, K., Zhang, X., Pologruto, T. and Lu, P. (2001) Solution structure of an A-tract DNA bend. *J. Mol. Biol.*, **306**, 1081–1098.
68. Chua, E.Y., Vasudevan, D., Davey, G.E., Wu, B. and Davey, C.A. (2012) The mechanics behind DNA sequence-dependent properties of the nucleosome. *Nucleic Acids Res.*, **40**, 6338–6352.
69. Lavery, R. and Sklenar, H. (1988) The definition of generalized helicoidal parameters and of axis curvature for irregular nucleic acids. *J. Biomol. Struct. Dyn.*, **6**, 63–91.
70. Kapral, G.J., Jain, S., Noeske, J., Doudna, J.A., Richardson, D.C. and Richardson, J.S. (2014) New tools provide a second look at HDV ribozyme structure, dynamics and cleavage. *Nucleic Acids Res.*, **42**, 12833–12846.
71. Nikolova, E.N., Goh, G.B., Brooks, C.L. III and Al-Hashimi, H.M. (2013) Characterizing the protonation state of cytosine in transient G•C Hoogsteen base pairs in duplex DNA. *J. Am. Chem. Soc.*, **135**, 6766–6769.
72. Wang, A.H.J., Ughetto, G., Quigley, G.J. and Rich, A. (1986) Interactions of quinoxaline antibiotic and DNA.: the molecular structure of a Triostin A—d(GCGTACGC) complex. *J. Biomol. Struct. Dyn.*, **4**, 319–342.
73. Santa Lucia, J. Jr (1998) A unified view of polymer, dumbbell, and oligonucleotide DNA nearest-neighbor thermodynamics. *Proc. Natl Acad. Sci. U.S.A.*, **95**, 1460–1465.
74. Kojima, C., Ono, A., Kainosho, M. and James, T.L. (1998) DNA duplex dynamics: NMR relaxation studies of a decamer with uniformly <sup>13</sup>C-labeled purine nucleotides. *J. Magn. Reson.*, **135**, 310–333.
75. Abrescia, N.G., González, C., Gouyette, C. and Subirana, J.A. (2004) X-ray and NMR studies of the DNA oligomer d(ATATAT): Hoogsteen base pairing in duplex DNA. *Biochemistry*, **43**, 4092–4100.
76. Pous, J., Urpi, L., Subirana, J.A., Gouyette, C., Navaza, J. and Campos, J.L. (2008) Stabilization by extra-helical thymines of a DNA duplex with Hoogsteen base pairs. *J. Am. Chem. Soc.*, **130**, 6755–6760.
77. De Luchi, D., Tereshko, V., Gouyette, C. and Subirana, J.A. (2006) Structure of the DNA coiled coil formed by d(CGATATATATAT). *ChemBioChem*, **7**, 585–587.
78. Matsuda, S., Fillo, J.D., Henry, A.A., Rai, P., Wilkens, S.J., Dwyer, T.J., Geierstanger, B.H., Wemmer, D.E., Schultz, P.G., Spraggon, G. et al. (2007) Efforts toward expansion of the genetic alphabet: structure and replication of unnatural base pairs. *J. Am. Chem. Soc.*, **129**, 10466–10473.
79. Cho, Y., Gorina, S., Jeffrey, P.D. and Pavletich, N.P. (1994) Crystal structure of a p53 tumor suppressor–DNA complex: understanding tumorigenic mutations. *Science*, **265**, 346–355.
80. Nelson, H.C., Finch, J.T., Luisi, B.F. and Klug, A. (1987) The structure of an oligo(dA)•oligo(dT) tract and its biological implications. *Nature*, **330**, 221–226.