nature COMMUNICATIONS

ARTICLE

Received 9 May 2014 | Accepted 23 Jul 2014 | Published 4 Sep 2014

DOI: 10.1038/ncomms5786

Widespread transient Hoogsteen base pairs in canonical duplex DNA with variable energetics

Heidi S. Alvey[1], Federico L. Gottardo[1], Evgenia N. Nikolova[2] & Hashim M. Al-Hashimi[3]

Hoogsteen (HG) base pairing involves a 180° rotation of the purine base relative to Watson–Crick (WC) base pairing within DNA duplexes, creating alternative DNA conformations that can play roles in recognition, damage induction and replication. Here, using nuclear magnetic resonance $R_{1\rho}$ relaxation dispersion, we show that transient HG base pairs occur across more diverse sequence and positional contexts than previously anticipated. We observe sequence-specific variations in HG base pair energetic stabilities that are comparable with variations in WC base pair stability, with HG base pairs being more abundant for energetically less favourable WC base pairs. Our results suggest that the variations in HG stabilities and rates of formation are dominated by variations in WC base pair stability, suggesting a late transition state for the WC-to-HG conformational switch. The occurrence of sequence and position-dependent HG base pairs provide a new potential mechanism for achieving sequence-dependent DNA transactions.

[1] Department of Chemistry and Biophysics, University of Michigan, 930 N. University Avenue, Ann Arbor, Michigan 48109, USA. [2] Department of Molecular Biology, The Scripps Research Institute, 10550 North Torrey Pines Road, La Jolla, California 92037, USA. [3] Department of Biochemistry and Chemistry, Duke University School of Medicine, 307 Research Dr, Nanaline H. Duke Building, Durham, North Carolina 27710, USA. Correspondence and requests for materials should be addressed to H.M.A.-H. (email: hashim.al.hashimi@duke.edu).

We recently showed[1–3] using nuclear magnetic resonance (NMR) $R_{1\rho}$ relaxation dispersion (RD)[4–6] that A•T and G•C Watson–Crick (WC) base pairs (bps) in CA/TG and TA/TA steps of duplex DNA transiently form Hoogsteen (HG) bps[7] with populations ranging between 0.14 and 0.49% and lifetimes between 0.3 and 2.5 ms at pH ~6 (Fig. 1a). HG bps form through 180° rotation of the purine base around the glycosidic bond from an *anti* to *syn* conformation (Fig. 1a). HG bps modify the structural and chemical presentation of DNA and thereby can play unique roles (reviewed in ref. 8) in DNA–protein recognition[9–12], damage induction[13] and repair[14,15] as well as replication[16,17]. For example, by narrowing the minor groove, HG bps have been shown to alter the DNA electrostatic potential 'seen' by DNA-binding proteins at the bp edges in the DNA grooves in the context of p53-DNA complexes[12,18,19]. HG bps can also transiently expose DNA sites that are otherwise inaccessible in WC bps, and thereby provide new mechanisms for damage induction. For example, using computational mapping, Bohnuud *et al.*[13] recently showed that transient G•C$^+$ HG bps could explain the susceptibility and accessibility of cytosines to hydroxymethylation by formaldehyde, thus explaining a long-standing mystery that has persisted for over 25 years. There is also structural and biochemical evidence that some members of the low-fidelity Y-family polymerases replicate DNA using HG pairing as the dominant mechanism, thus providing a mode for bypassing a variety of lesions on the WC face during replication[16,17,20]. Naked duplexes consisting entirely of HG bps have been reported in A–T-rich sequences that form parallel[21] and anti-parallel stranded[22,23] DNA. Computational studies on duplexes also suggest that the HG bp is a reasonable conformation, only slightly less stable than the canonical WC bp[24,25].

Considering that HG bps are an energetically favourable alternative to WC bps that can provide new mechanisms in a wide variety of DNA biochemical processes, it is of great interest to explore whether the transient HG bps observed by NMR are confined to flexible CA and TA steps, or occur more broadly across distinct sequence and position contexts in duplex DNA. Likewise, it is of interest to examine the sequence specificity of transient HG bp formation as this could provide new mechanisms for achieving sequence-specific DNA transactions that are based on shape[19].

By carrying out ^{13}C and ^{15}N $R_{1\rho}$ NMR RD measurements on 33 bps in eight distinct canonical DNA duplexes, we show that transient HG bps are not limited to CA and TA steps, but rather occur broadly across diverse sequence and positional contexts. We find that both the energetic stability and rates of HG bp formation exhibit a dependence on sequence and position, with HG bps forming faster and being more abundant in energetically less favourable WC bps.

Results

Widespread transient HG bps in duplex DNA. To more broadly examine the occurrence and sequence specificity of transient HG bps in canonical duplex DNA, we carried out ^{13}C and ^{15}N $R_{1\rho}$ NMR RD measurements targeting sugar C1′ and base C6/8 or N1/3 resonances in 20 A•T and 13 G•C bps in eight DNA duplexes that encompass a variety of sequence motifs, including $(A•T)_n$ repeats of varying length ($n=2$, 4, 5 and 6), a $(CA)_3$ repeat, a duplex sequence that forms HG bps on binding to the antibiotic echinomycin,[1,26,27] a $(CG)_3$ repeat capable of forming Z-DNA[28], and a B/Z junction forming sequence[29,30] (Fig. 1b). The targeted bps (Fig. 1b, highlighted in stars) encompass 6/10 dinucleotide steps that are positioned 1–6 bps away from the closest terminal end. $R_{1\rho}$ RD experiments were performed at pH 5.2–7.5 and 5.2–5.4 for A•T and G•C bps, respectively. We use low-pH conditions for transient (and protonated) G•C bps in order to increase the WC-to-HG chemical exchange signature detected by NMR $R_{1\rho}$ RD[3]. We recently reported a detailed analysis of the pH dependence of chemical exchange corresponding to transient HG bp formation and how measurements at such lower pH conditions can be qualitatively interpolated to assess exchange at higher pH[3].

The $R_{1\rho}$ NMR RD experiment measures the line broadening contribution to resonances of interest due to chemical exchange with a transient, lowly populated species[4,5]. In all cases, we measured significant ^{13}C and/or ^{15}N $R_{1\rho}$ RD (Supplementary Table 1) consistent with chemical exchange (Fig. 1c and Supplementary Fig. 1). A two-state analysis $\left(A \underset{k_B}{\overset{k_A}{\rightleftarrows}} B\right)$ of the $R_{1\rho}$ data[4–6] (see Methods) yielded populations ($p_B = \sim 0.08$–2.73%) and lifetimes ($\tau_B = \sim 0.12$–2.57 ms) for the transient state (Supplementary Fig. 2 and Supplementary Table 2) that are similar to those reported previously for transient HG bps ($p_B = \sim 0.14$–0.49% and $\tau_B = \sim 0.3$–2.5 ms)[1,2]. The chemical shifts (ω_B) of the transient state obtained using this analysis (Supplementary Fig. 3 and Supplementary Table 2) are also consistent with HG bps, including significantly downfield shifted purine C8 ($\Delta\omega \approx 2.72$ p.p.m.), purine C1′ ($\Delta\omega \approx 3.41$ p.p.m.), cytosine C6 ($\Delta\omega \approx 2.40$ p.p.m.) and upfield shifted imino N1/3 ($\Delta\omega \approx -1.84$ p.p.m.)[1,2]. Consistent with HG bps, we did not observe significant chemical exchange at adenine C2 and thymine C6, which do not experience large chemical shift changes upon HG bp formation[1,2] (Supplementary Fig. 1 and Supplementary Table 2). These results show that transient HG bps are not confined to CA/TG and TA/TA steps but rather occur broadly across a wide variety of sequence and positional contexts, including GA/TC, AA/TT, TA/TA, GG/CC, CG/CG and TG/CA dinucleotide steps (where the HG bp is underlined).

Position- and sequence-dependent energetic variability. We observe ~30-fold variations in the transient HG population (0.08–2.73% and 0.13–2.11% for A•T and G•C$^+$ bps, respectively) and ~20-fold variations in lifetimes (0.12–2.57 ms and 0.40–2.08 ms for A•T and G•C$^+$ bps, respectively) (Supplementary Fig. 2 and Supplementary Table 2). This corresponds to ~2.1 kcal mol^{-1} and ~2.8 kcal mol^{-1} variations in the relative thermodynamic stability ($\Delta\Delta G_{WC-HG}$ with $\Delta G_{WC-HG} = G_{HG} - G_{WC}$) (Fig. 2a) and forward free-energy barriers ($\Delta\Delta G^{\ddagger}_{WC-HG}$ with $\Delta G^{\ddagger}_{WC-HG} = G_{TS} - G_{WC}$, where TS is the transition state), respectively (Fig. 2b). These variations could reflect real sequence or position dependencies for transient HG bp formation. Alternatively, they could arise due to small differences in buffer conditions used for some of the duplexes (see Methods), particularly pH, which can affect the energetics of transient G•C$^+$ HG bp formation[1–3]. However, systematic deviations in ΔG_{WC-HG} and $\Delta G^{\ddagger}_{WC-HG}$ are not observed across DNA duplexes (Supplementary Fig. 4a). Furthermore, no correlations are observed between duplex melting temperatures measured by circular dichroism (CD) (Supplementary Fig. 4b,c and Supplementary Table 3) and either ΔG_{WC-HG} or $\Delta G^{\ddagger}_{WC-HG}$ (Supplementary Fig. 4a). Consistent with sequence- and/or position-dependent contributions, the variations in ΔG_{WC-HG} and $\Delta G^{\ddagger}_{WC-HG}$ are smaller (~1 kcal mol^{-1}) when comparing bps across different duplexes that share identical 5′ and 3′ neighbours and positions relative to duplex ends (indicated using horizontal lines in Fig. 2a,b; Supplementary Fig. 2 and Supplementary Table 2).

We observe significant variations in ΔG_{WC-HG} and $\Delta G^{\ddagger}_{WC-HG}$ for the same dinucleotide step, which may arise due to differences in position and/or differences in the broader sequence context.

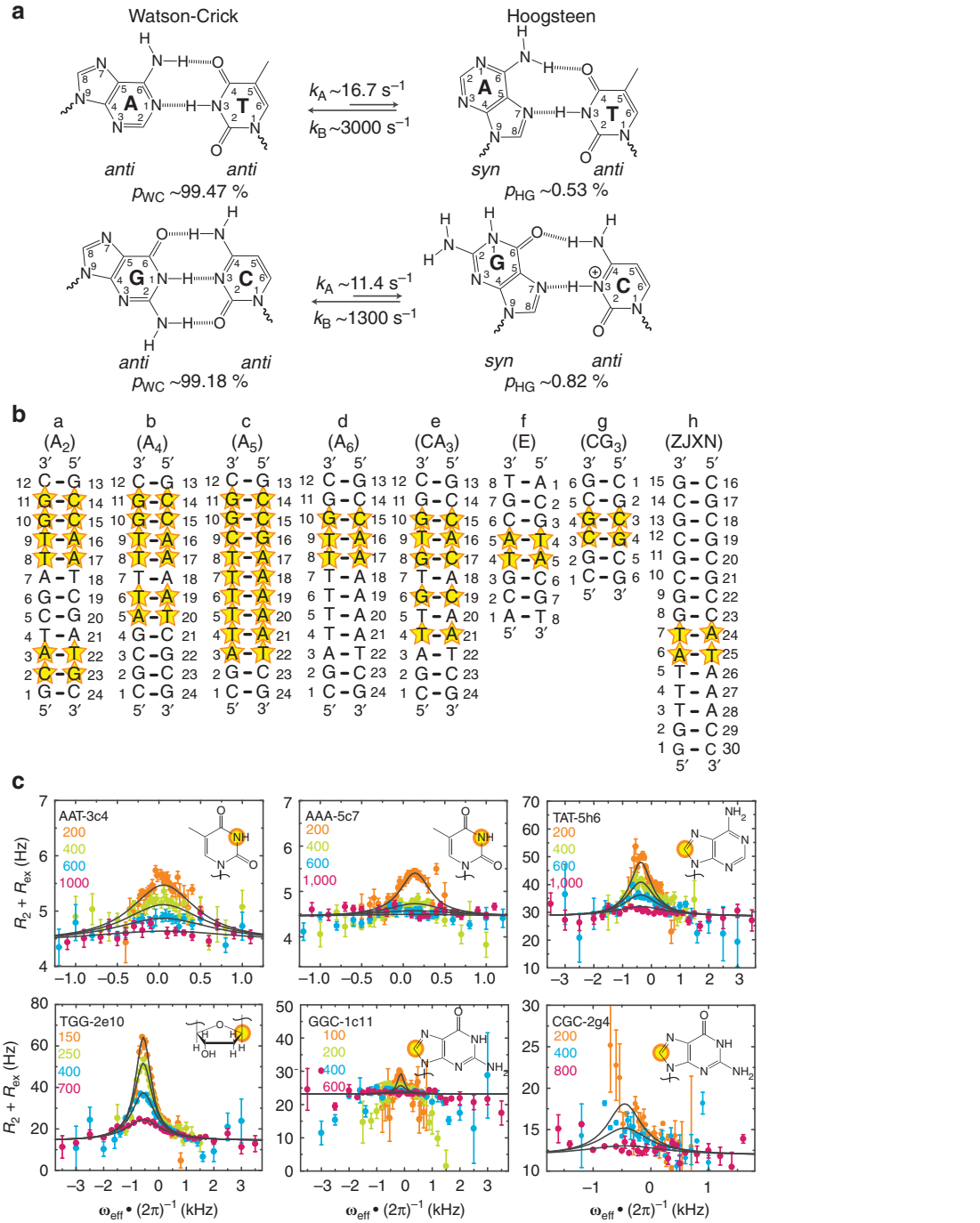

Figure 1 | Widespread occurrence of transient A•T and G•C$^+$ Hoogsteen base pairs in canonical duplex DNA. (**a**) The equilibrium between Watson–Crick and Hoogsteen base pairs. Shown are the average forward (k_A) and reverse (k_B) rate constants and populations of Watson–Crick (p_A) and Hoogsteen (p_B) base pairs obtained in this study. (**b**) DNA duplexes used in this study. Base pairs targeted for relaxation dispersion measurements are highlighted with a star. (**c**) Representative off-resonance ^{13}C and ^{15}N $R_{1\rho}$ relaxation dispersion profiles showing chemical exchange outside TA and CA steps of canonical duplex DNA. Spin lock powers (Hz) are shown in the inset. Data are fit to equation 2. Error bars represent experimental uncertainty (one standard deviation) estimated from monoexponential fitting of duplicate $R_{1\rho}$ data and analysis of signal-to-noise. See Methods for buffer conditions.

Notwithstanding these variations, the average stabilities of HG bps relative to WC bps (ΔG_{WC-HG}) calculated for individual dinucleotide steps (Fig. 2a) follow an order (TA/TA > AA/TT > CA/TG > GA/TC for A•T and TG/CA ≥ CG/CG ≥ GG/CC for G•C$^+$) that is nearly inverted relative to the well-documented WC dinucleotide stabilities (GA/TC ≥ CA/TG > AA/TT > TA/TA for A•T and CG/CG > GG/CC > TG/CA for G•C), which measure the stability of WC bps relative to the melted state[31]. Thus, transient HG bps seem to be more abundant in less stable WC dinucleotide steps such as CA/TG and TA/TA steps. The energetic preference for HG bps at TA/TA steps observed here is consistent with a large body of data showing that HG bps are favoured in A–T-rich sequences, particularly TA/TA steps (reviewed in ref. 8).

Origin of variable transient HG bp energetics. Strikingly, we observe a clear correlation ($R = 0.76$) between ΔG_{WC-HG} and

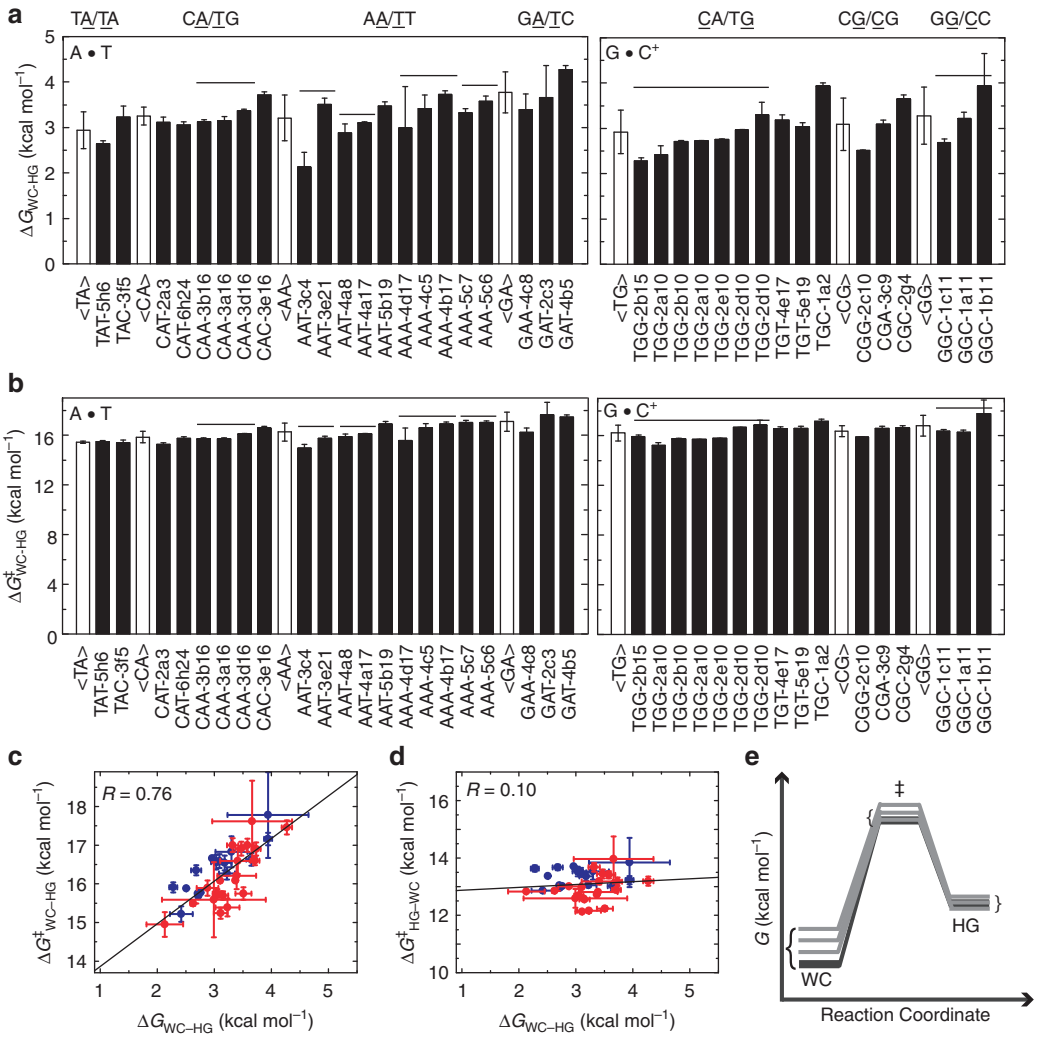

Figure 2 | Sequence- and position-dependent thermodynamic and kinetic parameters describing the Watson–Crick to Hoogsteen transition. (**a**) Free-energy difference (ΔG_{WC-HG}) and (**b**) forward free-energy barrier ($\Delta G^{\ddagger}_{WC-HG}$) for the Watson–Crick to Hoogsteen transition as derived from a two-state analysis of the $R_{1\rho}$ data. Error bars represent experimental uncertainty (one s.d.) estimated from propagation of errors from monoexponential fitting of duplicate $R_{1\rho}$ data and analysis of signal-to-noise. Average and standard deviations for dinucleotide steps are shown using white bars. Horizontal lines denote bps with same triplet sequence and positions relative to nearest terminal end. *X* axis labels denote the triple sequence context ('XYZ') with Hoogsteen base pair in the middle; the position relative to the closest terminal end ('−n'), and name of duplex as denoted in Fig. 1b. Correlation between (**c**) $\Delta G^{\ddagger}_{WC-HG}$ and ΔG_{WC-HG} as well as (**d**) $\Delta G^{\ddagger}_{HG-WC}$ and ΔG_{WC-HG}. A•T and G•C$^+$ base pairs are shown in red and blue, respectively. The best line of fit and corresponding Pearson coefficient (*R*) are shown. (**e**) Free-energy diagram of the Watson–Crick to Hoogsteen transition depicting relative variations in free energy of Watson–Crick, transition and Hoogsteen states.

$\Delta G^{\ddagger}_{WC-HG}$ (Fig. 2c) and a relatively uniform backward free-energy barrier ($\Delta G^{\ddagger}_{HG-WC} = G_{TS} - G_{HG}$) of $\sim$13 kcal mol^{-1} (Fig. 2d). Thus, changing the sequence context has little effect on the relative energetic stability of the TS and HG bp. This could either be because the stabilities of the TS and HG bp are not significantly affected by changes in sequence, or because their stabilities vary in a correlated manner. In contrast, a change in sequence context does change the relative stability of both the TS and HG bp relative to the WC bp (Fig. 2e).

One possibility is that sequence-specific variations are dominated by changes in variations in the WC bp without significantly affecting the stabilities of the TS and HG bp (Fig. 2e). Indeed, the observed variations in HG bp stability ($\sim$2.1 kcal mol^{-1}) are comparable in size with variations in WC bp stability ($\sim$2 kcal mol^{-1}) measured across dinucleotide steps using melting experiments[31]. This would also explain why transient HG bps are specifically more abundant at dinucleotide steps that have weakened WC stabilities (Fig. 2a,b). If the observed variations are indeed dominated by variations in WC stability, one might expect a similar correlation between the ΔG_{cl-op} and $\Delta G^{\ddagger}_{cl-op}$ values describing transitions between WC bps and the bp open state, especially since stability of the open state is not expected to vary significantly with sequence. Indeed, a previous analysis of ΔG_{cl-op} and $\Delta G^{\ddagger}_{cl-op}$ correlation reported by Coman and Russu[32] based on imino proton exchange measurements[33] reveals a comparably strong correlation ($R = 0.8$)[34]. Our results suggest that the free energy of the TS varies less relative to the HG bp with sequence/position as compared with the WC bp. If one were to assume that a similar sequence/position-dependent free energy implies a similar structure, even if the sequence/position dependence is very small, then these results would suggest that the TS is structurally more similar to the HG bp—consistent with a 'late' TS.

Φ-value analysis suggests a 'late' transition state. To quantify the extent to which the sequence-specific TS energetics are more similar to HG bps versus WC bps, we subjected the measured ΔG_{WC-HG} and $\Delta G^{\ddagger}_{WC-HG}$ values to Φ-value analysis[34,35] (Methods and Fig. 3). In this approach, one computes a Φ-value, which quantifies the relative magnitude of the sequence/position-dependent free-energy differences between the TS and WC bps and those between WC bps and HG bps,

$$\Phi = \Delta\Delta G_{TS-WC}/\Delta\Delta G_{WC-HG} \quad (1)$$

where $\Delta\Delta G_{TS-WC} = (G_{TS} - G_{WC})_{mut} - (G_{TS} - G_{WC})_{\Psi-WT}$ and $\Delta\Delta G_{WC-HG} = (G_{HG} - G_{WC})_{mut} - (G_{HG} - G_{WC})_{\Psi-WT}$ are the changes in the forward free-energy barrier and free-energy difference between WC and HG bps, respectively, on mutating (mut) the sequence/position of a reference (Ψ − WT) bp. It is instructive to consider two limiting cases to help understand how this analysis can be used to quantify the extent to which the sequence-specific TS energetics are more similar to HG bps versus WC bps and whether a TS is 'early' or 'late'. In the case that the TS and HG share identical sequence-specific energetics as might be expected for a late TS, a given sequence-specific perturbation equally affects G_{TS} and G_{HG} (that is, $G_{HG} - G_{TS}$ = constant for all mutants) and $\Delta\Delta G_{TS-WC} = \Delta\Delta G_{WC-HG}$ and $\Phi = 1$. On the other hand, if TS and WC share identical sequence-specific energetics as might be expected for an early TS, then $G_{TS} - G_{WC}$ = constant for all mutants and $\Delta\Delta G_{TS-WC} = 0$ and $\Phi = 0$. In practice, the Φ-value can range between 0 and 1, with intermediate Φ-values being more difficult to interpret in the context of a structural mechanism[14].

We arbitrarily assign reference A•T and G•C bps to be those having the smallest ΔG_{WC-HG} values, and which therefore form the most stable HG bps among those studied herein (Ψ − WT, Fig. 3a,b). Next, we computed Φ for each A•T and G•C$^+$ bp. This analysis was preformed separately for A•T and G•C$^+$ bps and repeated multiple times assuming a different bp as the designated wild-type reference (data not shown). In the vast majority of the cases, we measure Φ-values near 1, consistent with a 'late' TS (Fig. 3c and Supplementary Table 4). It should be noted that similar sequence/position-dependent energetics for the TS and HG bp does not have to imply similarity in structure, and that we cannot rule out the possibility of an early TS that has structural features similar to WC but sequence/position-dependent energetics that are more similar to HG.

Discussion

Our results suggest that HG bps can occur ubiquitously in canonical duplex DNA across different sequence and positional contexts. This can help explain how polymerases such as the human DNA polymerase iota can use HG pairing as a general mechanism to replicate DNA and thereby bypass lesions that diminish the ability to form WC bps[8,16]. Our findings also raise the possibility that HG bps may be widespread in genomic DNA, especially given that the energetic differences between WC and

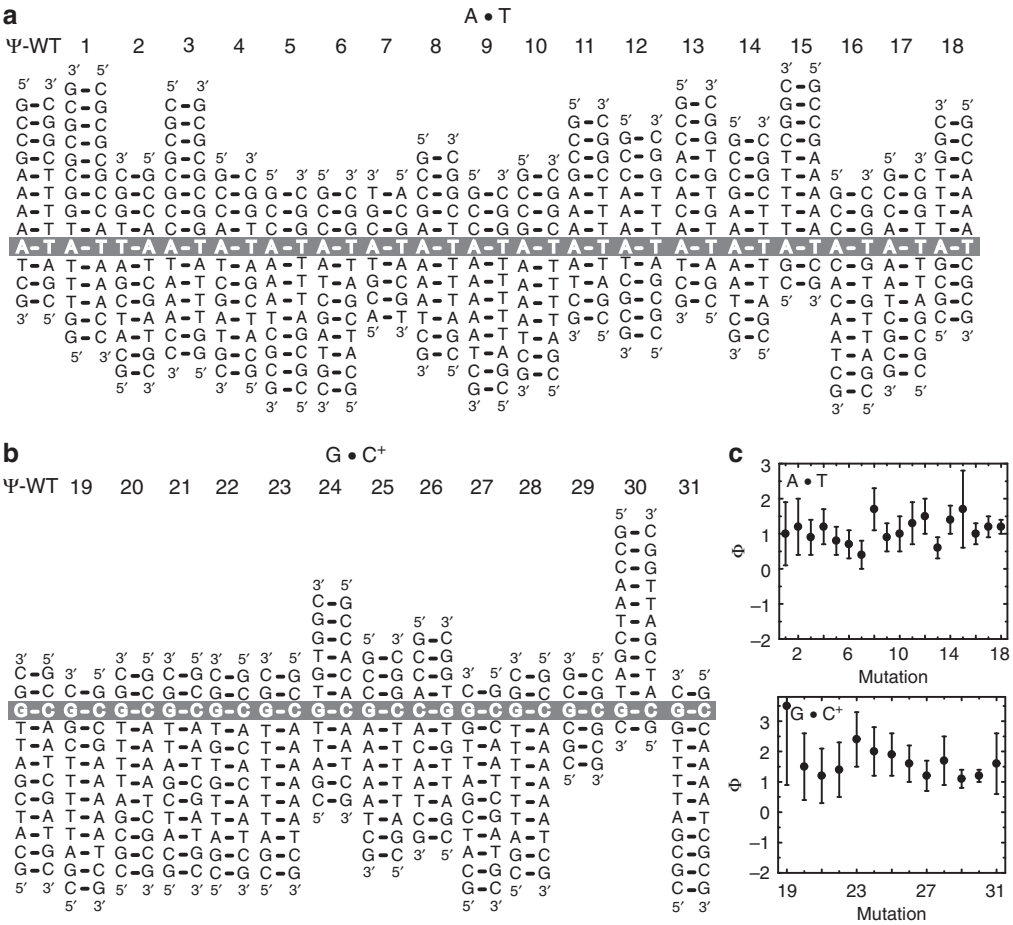

Figure 3 | Φ-value analysis. Perturbing the Watson–Crick to Hoogsteen equilibrium in (**a**) A•T and (**b**) G•C$^+$ base pairs by varying the sequence and/or positional context of individual base pairs. All duplexes are re-drawn after aligning each A•T or G•C$^+$ base pair with a reference sequence variant (Ψ − WT). (**c**) Φ-values for A•T and G•C$^+$ base pairs. The sequence variant number is shown above each secondary structure.

HG bps are small compared with forces in living cells, including those arising due to supercoiling, torsional stress, and protein binding. It is worth noting that difficulties in distinguishing between WC and HG bps based on X-ray crystallography data have been reported[8,36]. Our finding that transient HG bps can occur ubiquitously in duplex DNA calls for the re-examination of current X-ray structures of DNA to more critically assess for the occurrence of HG bps.

Our studies suggest that the occurrence of HG bps depends in a complex manner on both sequence context and position. This is consistent with the hypothesis by Honig and Rohs[37] that the observation of HG base pairing makes it is unlikely that protein–DNA binding is driven by a simple linear code. Nevertheless, our results together with prior studies[8] suggest that HG bps are likely to exist in greater abundance within unstable and structurally stressed environments, such as kinks and turns, which can destabilize WC bps, including stacking interactions. Indeed, while only a few X-ray structures have documented the existence of HG bps in DNA, in many cases HG bps occur near structurally stressed environments. For example, HG bps are observed near kinks or nicks in X-ray structures of DNA bound to TATA box-binding protein[11] and integration host factor[9], and near a hairpin loop in structures of DNA bound to TnpA transposase[38]. HG bps have also been observed for DNA in complex with antibiotics that contribute unique stacking interactions[27,39]. Previous studies[3] have shown that changes in counterion concentration have a measurable effect on the population and lifetime of HG bps. Future studies should therefore also examine how the sequence- and position-dependent HG energetics vary with increasing counterion concentration (Na^+ and Mg^{2+}). Studies so far suggest that Na^+ and Mg^{2+} stabilize A•T HG but destabilize G•C$^+$ HG bps in the case of CA/TG steps[3]. Although further studies are needed to more quantitatively understand the sequence and position dependence of transient HG bp formation, these energetic preferences may provide a new mechanism for shape-based DNA recognition[12,19] via indirect read-out mechanisms[40]. Our study has focused on the occurrence of single transient HG bps surrounded by WC bps. Additional studies are needed to examine sequence-specific propensities for forming longer stretches of HG bps, and HG tracts interspersed by WC bps. Such mixtures of WC and HG bps can endow genomic DNA with a new level of structural complexity similar to Z-DNA.

Our results suggest that the sequence- and position-specific variations in HG bp stabilities and lifetimes are dominated by variations in the WC bp stability and to a lesser extent by variations in the stabilities of the TS and HG bp. Interestingly, a similar trend has been reported for base opening[32]. Future studies should further explore the WC-to-HG transition pathways and examine whether they share a similar TS with base opening and whether there can be pathways toward HG that proceed via the base-opened state. Conjugate peak refinement simulations suggest a pathway in which the purine base rotates toward the major groove inside the double helix[1]; however, further experimental characterization is required. The observation that the sequence and/or position variations in the TS free energies are more similar to the HG bp versus the WC bp suggests the TS is structurally more similar to HG versus WC, consistent with a 'late' TS for the WC-to-HG transition. However, we cannot rule out an early TS that is structurally more similar to WC but has sequence/position-dependent energetics that are more similar to the HG bp. Although the structure of the TS remains unclear, an equally important question is why the energetic stabilities of HG bps appear to be only weakly dependent on sequence. Further studies are required to understand the structure and specific interactions that may help stabilize the TS and HG bp.

Methods

NMR samples and resonance assignments. Unlabelled DNA samples were purchased as single-stranded oligos from Integrated DNA Technologies (IDT) with standard desalting purification. The DNA oligos were resuspended to ~200 μM in 15 mM phosphate buffer with corresponding pH (see below), 25 mM NaCl, 0.1 mM EDTA. Duplexes were annealed by mixing an equimolar ratio of the complementary DNA strands, heating at 95 °C for 2 min followed by gradual cooling at room temperature for ~30 min. Unlabelled DNA duplexes were washed 3 × in resuspension buffer by microcentrifugation using an Amicon Ultra-4 centrifugal filter with a 3-kDa cutoff, concentrated to ~2–3 mM and ~250 μl, then supplied with 10% D_2O. Natural abundance CG_3 was resuspended in ~4 ml of milliQ H_2O and dialyzed against 2 l of milliQ H_2O with two exchanges for a total of 6 l of milliQ H_2O, using a dialysis tube from G-Biosciences with a 1-kDa cutoff. Dialyzed CG_3 was lyophilized and resuspended in NMR buffer to ~4 mM and supplied with 10% D_2O. Hemi-$^{13}C/^{15}N$-labelled A_5 duplex was prepared by annealing a uniformly $^{13}C/^{15}N$-labelled thymine-rich strand to a natural abundance adenosine-rich strand. Fully $^{13}C/^{15}N$-labelled DNA duplexes were prepared by annealing two labelled strands together. All labelled single strands were synthesized *in vitro* by the method of Zimmer and Crothers[41] using a DNA hairpin template with a 5′ overhang corresponding to the complement of the target labelled strand and a 3′ ribose (IDT). In this study, we used the same hairpin sequence as Zimmer and Crothers[41], Klenow fragment DNA polymerase (NEB) NEB2 buffer (NEB) and uniformly $^{13}C/^{15}N$-labelled dNTPs (Isotec, Sigma-Aldrich and Silantes). Base- and heat-catalyzed cleavage separated the hairpin template from the $^{13}C/^{15}N$-labelled synthesized product. The single-stranded DNA product was purified by 20% denaturing polyacrylamide gel electrophoresis, isolated by passive elution from crushed gel pieces and desalted on a C18 reverse-phase column (Sep-Pak, Waters). The oligo was lyophilized and suspended in NMR buffer. The semi-labelled DNA samples were prepared by titrating the unlabelled strand directly into an NMR tube containing the $^{13}C/^{15}N$-labelled strand and monitoring the disappearance of single-stranded DNA peaks using heteronuclear single quantum coherence (HSQC) experiments. The fully labelled samples were annealed in a similar fashion. 2D ^1H–^1H nuclear overhauser spectroscopy (NOESY) experiments at 26 °C (A_2, A_4, A_5, A_6, CA_3, E) or 25 °C (CG_3 and Z junction (ZJXN)) and pH 5.2 (A_5), 5.4 (A_2, A_4, A_6, CA_3, CG_3), 6.8 (E) or 7.5 (ZJXN) were used to assign resonances as described previously[1]. See NMR $R_{1\rho}$ relaxation dispersion for $R_{1\rho}$ RD pH values.

NMR $R_{1\rho}$ RD. All NMR experiments were performed on a Bruker Avance 600 MHz NMR spectrometer equipped with a 5-mm triple-resonance cryogenic probe. $R_{1\rho}$ RD experiments were performed at pH 5.2–7.5 and 5.2–5.4 for A•T and G•C$^+$ bps, respectively. Buffer conditions for $R_{1\rho}$ RD measurements are 15 mM sodium phosphate, 25 mM NaCl, 0.1 mM EDTA, 10% D_2O pH = 5.2 (A_5, A_4: C15 C6, A16 C1′, G10 C1′), pH = 5.4 (A_2, A_4, A_6, CA_3, CG_3), pH = 6.8 (E) and pH = 7.5 (ZJXN) at 26 °C (A_2, A_4, A_5, A_6, CA_3, E) or 25 °C (A_2 A3 C1′, CG_3, ZJXN). CG_3 and E are unlabelled DNA samples, the T-rich strand of A_5 is $^{13}C/^{15}N$ labelled while A_6, A_4, A_2, CA_3 and ZJXN are fully $^{13}C/^{15}N$-labelled. We use low-pH conditions for transient (and protonated) G•C$^+$ bps in order to increase the WC-to-HG chemical exchange signature detected by $R_{1\rho}$ RD[3]. We recently reported a detailed analysis of the pH dependence of chemical exchange corresponding to transient HG bp formation and how measurements at such lower pH conditions can be qualitatively interpolated to assess exchange at higher pH[3]. Carbon and nitrogen $R_{1\rho}$ RD profiles for guanine/adenine C8, guanine/adenine C1′, cytosine C6, guanine N1 and thymine N3 were measured using a one-dimensional acquisition scheme which uses selective Hartmann–Hahn polarization transfer[42] to selectively excite one C–H or N–H spin system at a time[6]. The spin lock power and offset frequencies are summarized in Supplementary Table 4. The following delays were used: A_2: A3 C1′, G10 C1′, A17 C1′; A_4: A5 C1′, G10 C1′, A16 C1′, A19 C1′; CA_3: G10 C1′; A_5: A3 C1′, C5 C1′ (0, 4, 8, 12, 18, 26, 34, 42, 12, 42); A_5 G11 C8 (0, 12, 32, 26, 32); A_4 A17 C8, A_6 A17 C8 (0, 4, 12, 32, 26, 32); A_2: C2 C6, T9 C6, G11 C8, A16 C8, A17 C2; CA_3: A16 C8, C17 C6, C19 C6, A21 C1′; A_4: C15 C6, G11 C8; A_5: C9 C6, G10 C8; A_6: G10 C8, A16 C8 (0, 4, 8, 12, 16, 20, 26, 32, 12, 32); A_2: T8 N3, G10 N1, G23 N1; A_5: T4 N3, T5 N3, T6 N3, T7 N3, T8 N3; A_6 G10 N1 (0, 8, 16, 24, 36, 48, 60, 80, 100, 16, 70, 100); CG_3 G4 C8 (0, 60, 60); ZJXN: A6 C8, A24 C8 (0, 4, 8, 12, 16, 20, 24, 30, 12, 30); E A5 C8: (0, 48, 48). Data points meeting C-C Hartmann–Hahn matching conditions were omitted as described previously[6]. Data were processed using NMRPipe[43] and $R_{1\rho}$ values were determined from monoexponential decay fits of the resonance intensities using a script[44] in Mathematica 9 (Wolfram Research). On- and off-resonance $R_{1\rho}$ data were fit to the Laguerre equation[45] (equation 2) using Origin 8.6 (OriginLab),

$$R_{1\rho} = R_1\cos^2\theta + R_2\sin^2\theta + \frac{\sin^2\theta p_A p_B \Delta\omega^2 k_{ex}}{\omega_A^2\omega_B^2/\omega_{eff}^2 + k_{ex}^2 - \sin^2\theta p_A p_B \Delta\omega^2\left(1 + \frac{2k_{ex}^2(p_A\omega_A^2 + p_B\omega_B^2)}{\omega_A^2\omega_B^2 + \omega_{eff}^2 k_{ex}^2}\right)} \quad (2)$$

Where $\omega_{eff}^2 = \Omega^2 + \omega_1^2$, $\omega_A^2 = (\Omega_A - \omega_{rf})^2 + \omega_1^2$ and $\omega_B^2 = (\Omega_B - \omega_{rf})^2 + \omega_1^2$. $\Delta\omega_{AB} = \Omega_B - \Omega_A$, where Ω_A and Ω_B are the chemical shifts of the ground (A) and transient (B) states, respectively, in Hz. R_1 and R_2 are the intrinsic longitudinal and transverse relaxation rate constants, respectively, and are assumed to be identical for the ground (A) and transient (B) states. $\theta = \arctan(\omega_1/\Omega)$ where ω_1 is the spin lock power strength. $\Omega = \Omega_{obs} - \omega_{rf}$ where Ω is the offset of the spin lock carrier

frequency (ω_{rf}) from the observed resonance frequency (Ω_{obs}). $\Omega_{obs} = p_A\Omega_A + p_B\Omega_B$ where p_A and p_B are the ground and transient state populations, respectively, and $p_A + p_B = 1$. k_{ex} is the chemical exchange rate constant for a two-state exchange process where $k_{ex} = k_A + k_B$, $k_A = k_{ex}p_B$ and $k_B = k_{ex}p_A$. k_A and k_B are the forward and reverse rate constants, respectively.

Plots of $R_{1\rho}$ data are presented as R_{2eff} ($R_{2eff} = R_2 + R_{ex}$) in Fig. 1c and Supplementary Fig. 1 where,

$$R_{2,\mathrm{eff}} = \frac{R_{1\rho}}{\sin^2\theta} - \frac{R_1}{\tan^2\theta} \quad (3)$$

In cases where large errors were accompanied by small R_{ex}, model selection comparing presence and absence of exchange was carried out using the F- and Akaike's Information Criterion (AIC)[46] tests to discriminate between exchange and no detectable exchange (data not shown). The F-test compares two nested models fit to the same data under the null hypothesis that the residual sum of squares of the less complex, restricted model is not significantly larger than that of the more complex, full model. The AIC test assesses the likelihood of a model given the data and seeks to minimize loss of information embedded within the data. For A_2 G23 N1, no exchange is the selected model under the conditions used. Because of lower sensitivity to exchange, we did not interpret absence of ^{15}N dispersion as evidence for absence of transient HG bps.

The free-energy difference between the WC GS and HG transient state (ΔG_{WC-HG}) was computed using,

$$\Delta G_{WC-HG} = -RT\left(\ln\left(\frac{k_1 h}{k_B T}\right) - \ln\left(\frac{k_2 h}{k_B T}\right)\right) \quad (4)$$

where k_1 and k_2 are the forward and reverse rate constants, respectively, h is Planck's constant, k_B is Boltzmann's constant, R is the gas constant and T is temperature. The forward barrier of the transition ($\Delta G^{\ddagger}_{WC-HG}$) was computed using,

$$\Delta G^{\pm}_{WC-HG} = -RT\ln\left(\frac{k_1 h}{\kappa k_B T}\right) \quad (5)$$

where κ is the transmission coefficient which is assumed to be unity.

CD melting. DNA duplexes (IDT) were prepared using ultracentrifugation as described in NMR Samples and Resonance Assignments in 15 mM phosphate buffer, 0.1 mM EDTA, 25 mM NaCl pH 5.4 and supplied with 10% D_2O. Duplexes were prepared by diluting complementary single-stranded stocks to 50 μM in the same tube with a final volume of 200 μl. Samples were denatured at 95 °C for 5 min followed by annealing of at least 10 min on the bench top. Samples were transferred to a 1-mm cuvette (Starna Cells), mineral oil was added to the top of the solution and the cuvette was capped. Melting experiments were performed on a Jasco Spectropolarimeter equipped with a recirculating water bath and Peltier temperature control unit. Temperature ramps were performed from 5 to 80 °C with a bandwidth of 5 nm (1 nm for ZJXN), ramp rate of 1 °C min^{-1}, equilibration time of 20 s and sensitivity of 100 millidegrees. Wavelength scans used the same sensitivity and bandwidth as melting runs. Spectral measurements were performed between 220 to 330 nm with a scan rate of 100 nm min^{-1}. Temperature ramp profiles were fit to the Boltzmann model[47], which has previously been used to determine nucleic acid melting temperatures[48],

$$\theta = \mathrm{LL} + \frac{\mathrm{UL} - \mathrm{LL}}{1 + e^{\frac{T_m - T}{a}}} \quad (6)$$

where θ is the elipticity at 254 nm normalized to the signal change magnitude, LL and UL are the lower and upper limits of the transition, respectively, a is the Hill slope, T_m is the melting temperature defined as the point of inflection of the melting curve and T is the independent variable temperature.

Phi (Φ)-value analysis. Φ-value analysis[34] was carried out by computing Φ using equation 1 ($\Phi = \Delta\Delta G_{TS-WC}/\Delta\Delta G_{WC-HG}$) where $\Delta\Delta G_{TS-WC} = \Delta G^{\ddagger}_{WC-HG,mut} - \Delta G^{\ddagger}_{WC-HG,\Psi-WT}$ and $\Delta\Delta G_{WC-HG} = \Delta G_{WC-HG,mut} - \Delta G_{WC-HG,\Psi-WT}$ are the change in the forward free-energy barrier ($\Delta G^{\ddagger}_{WC-HG}$) and free-energy difference between WC and HG bps (ΔG_{WC-HG}), respectively, on introduction of one or multiple mutations (mut) as compared with 'wild-type' (Ψ − WT). We defined Ψ-WT to be A_5 T4 N3 and A_2 G10 N1 for A•T and G•C$^+$ bps, respectively, given that they have the lowest ΔG_{WC-HG} values. To control for systematic errors in Φ arising due to use of a ^{15}N Ψ − WT reference resonance, we also performed Φ-value analysis assigning a ^{13}C resonance as Ψ − WT for A•T and G•C$^+$ bps, respectively. In all cases we observe Φ-values concentrated around ∼1 consistent with a 'late' TS. Note that $\Phi \approx 0$ when $\Delta\Delta G_{TS-WC} \approx 0$ relative to $\Delta\Delta G_{WC-HG}$ implying an early WC-like TS, whereas $\Phi \approx 1$ when $\Delta\Delta G_{TS-WC} \approx \Delta\Delta G_{WC-HG}$ and implies a late HG-like TS. Errors for calculated Φ-values for A_6 A17 C8, A_2 A3 C1′, A_4 C15 C6 and A_5 G10 C8 were larger than the corresponding values and thus could not be determined accurately.

References

1. Nikolova, E. N. *et al.* Transient Hoogsteen base pairs in canonical duplex DNA. *Nature* **470,** 498–502 (2011).
2. Nikolova, E. N., Gottardo, F. L. & Al-Hashimi, H. M. Probing transient Hoogsteen hydrogen bonds in canonical duplex DNA using NMR relaxation dispersion and single-atom substitution. *J. Am. Chem. Soc.* **134,** 3667–3670 (2012).
3. Nikolova, E. N., Goh, G. B., Brooks, 3rd C. L. & Al-Hashimi, H. M. Characterizing the protonation state of cytosine in transient G.C Hoogsteen base pairs in duplex DNA. *J. Am. Chem. Soc.* **135,** 6766–6769 (2013).
4. Palmer, 3rd A. G. Chemical exchange in biomacromolecules: past, present, and future. *J. Magn. Reson.* **241,** 3–17 (2014).
5. Sekhar, A. & Kay, L. E. NMR paves the way for atomic level descriptions of sparsely populated, transiently formed biomolecular conformers. *Proc. Natl Acad. Sci. USA* **110,** 12867–12874 (2013).
6. Hansen, A. L., Nikolova, E. N., Casiano-Negroni, A. & Al-Hashimi, H. M. Extending the range of microsecond-to-millisecond chemical exchange detected in labeled and unlabeled nucleic acids by selective carbon R(1rho) NMR spectroscopy. *J. Am. Chem. Soc.* **131,** 3818–3819 (2009).
7. Hoogsteen, K. The structure of crystals containing a hydrogen-bonded complex of 1-methylthymine and 9-methyladenine. *Acta Crystallogr.* **12,** 822–823 (1959).
8. Nikolova, E. N. *et al.* A historical account of hoogsteen base-pairs in duplex DNA. *Biopolymers* **99,** 955–968 (2013).
9. Rice, P. A., Yang, S., Mizuuchi, K. & Nash, H. A. Crystal structure of an IHF-DNA complex: a protein-induced DNA U-turn. *Cell* **87,** 1295–1306 (1996).
10. Aishima, J. *et al.* A Hoogsteen base pair embedded in undistorted B-DNA. *Nucleic Acids Res.* **30,** 5244–5252 (2002).
11. Patikoglou, G. A. *et al.* TATA element recognition by the TATA box-binding protein has been conserved throughout evolution. *Genes Dev.* **13,** 3217–3230 (1999).
12. Kitayner, M. *et al.* Diversity in DNA recognition by p53 revealed by crystal structures with Hoogsteen base pairs. *Nat. Struct. Mol. Biol.* **17,** 423–429 (2010).
13. Bohnuud, T. *et al.* Computational mapping reveals dramatic effect of Hoogsteen breathing on duplex DNA reactivity with formaldehyde. *Nucleic Acids Res.* **40,** 7644–7652 (2012).
14. Yang, W. Structure and mechanism for DNA lesion recognition. *Cell Res.* **18,** 184–197 (2008).
15. Yang, H., Zhan, Y., Fenn, D., Chi, L. M. & Lam, S. L. Effect of 1-methyladenine on double-helical DNA structures. *FEBS Lett.* **582,** 1629–1633 (2008).
16. Nair, D. T., Johnson, R. E., Prakash, S., Prakash, L. & Aggarwal, A. K. Replication by human DNA polymerase-iota occurs by Hoogsteen base-pairing. *Nature* **430,** 377–380 (2004).
17. Makarova, A. V. & Kulbachinskiy, A. V. Structure of human DNA polymerase iota and the mechanism of DNA synthesis. *Biochemistry (Moscow)* **77,** 547–561 (2012).
18. Harris, R. C. *et al.* Opposites attract: shape and electrostatic complementarity in protein-DNA complexes. in *Innovations in Biomolecular Modeling and Simulations* Vol 2 (ed. Schlick, T.) 53–80 (Royal Soc Chemistry, 2012).
19. Rohs, R. *et al.* The role of DNA shape in protein-DNA recognition. *Nature* **461,** 1248–1253 (2009).
20. Johnson, R. E., Prakash, L. & Prakash, S. Biochemical evidence for the requirement of Hoogsteen base pairing for replication by human DNA polymerase iota. *Proc. Natl Acad. Sci. USA* **102,** 10466–10471 (2005).
21. Liu, K., Miles, H. T., Frazier, J. & Sasisekharan, V. A novel DNA duplex. A parallel-stranded DNA helix with Hoogsteen base pairing. *Biochemistry* **32,** 11802–11809 (1993).
22. Abrescia, N. G., Gonzalez, C., Gouyette, C. & Subirana, J. A. X-ray and NMR studies of the DNA oligomer d(ATATAT): Hoogsteen base pairing in duplex DNA. *Biochemistry* **43,** 4092–4100 (2004).
23. Abrescia, N. G., Thompson, A., Huynh-Dinh, T. & Subirana, J. A. Crystal structure of an antiparallel DNA fragment with Hoogsteen base pairing. *Proc. Natl Acad. Sci. USA* **99,** 2806–2811 (2002).
24. Cubero, E., Luque, F. J. & Orozco, M. Theoretical study of the Hoogsteen-Watson-Crick junctions in DNA. *Biophys. J.* **90,** 1000–1008 (2006).
25. Cubero, E., Abrescia, N. G., Subirana, J. A., Luque, F. J. & Orozco, M. Theoretical study of a new DNA structure: the antiparallel Hoogsteen duplex. *J. Am. Chem. Soc.* **125,** 14603–14612 (2003).
26. Ughetto, G. *et al.* A comparison of the structure of echinomycin and triostin A complexed to a DNA fragment. *Nucleic Acids Res.* **13,** 2305–2323 (1985).
27. Gilbert, D. E., van der Marel, G. A., van Boom, J. H. & Feigon, J. Unstable Hoogsteen base pairs adjacent to echinomycin binding sites within a DNA duplex. *Proc. Natl Acad. Sci. USA* **86,** 3006–3010 (1989).
28. Schwartz, T., Rould, M. A., Lowenhaupt, K., Herbert, A. & Rich, A. Crystal structure of the Zalpha domain of the human editing enzyme ADAR1 bound to left-handed Z-DNA. *Science* **284,** 1841–1845 (1999).
29. Bothe, J. R., Lowenhaupt, K. & Al-Hashimi, H. M. Sequence-specific B-DNA flexibility modulates Z-DNA formation. *J. Am. Chem. Soc.* **133,** 2016–2018 (2011).

30. Ha, S. C., Lowenhaupt, K., Rich, A., Kim, Y. G. & Kim, K. K. Crystal structure of a junction between B-DNA and Z-DNA reveals two extruded bases. *Nature* **437,** 1183–1186 (2005).
31. SantaLucia, Jr J., Allawi, H. T. & Seneviratne, P. A. Improved nearest-neighbor parameters for predicting DNA duplex stability. *Biochemistry* **35,** 3555–3562 (1996).
32. Coman, D. & Russu, I. M. A nuclear magnetic resonance investigation of the energetics of basepair opening pathways in DNA. *Biophys. J.* **89,** 3285–3292 (2005).
33. Gueron, M., Kochoyan, M. & Leroy, J. L. A single mode of DNA base-pair opening drives imino proton exchange. *Nature* **328,** 89–92 (1987).
34. Fersht, A. R., Matouschek, A. & Serrano, L. The folding of an enzyme. I. Theory of protein engineering analysis of stability and pathway of protein folding. *J. Mol. Biol.* **224,** 771–782 (1992).
35. Neudecker, P., Zarrine-Afsar, A., Davidson, A. R. & Kay, L. E. Phi-value analysis of a three-state protein folding pathway by NMR relaxation dispersion spectroscopy. *Proc. Natl Acad. Sci. USA* **104,** 15717–15722 (2007).
36. Wang, J. DNA polymerases: Hoogsteen base-pairing in DNA replication? *Nature* **437,** E6–E7 discussion E7 (2005).
37. Honig, B. & Rohs, R. Biophysics: flipping Watson and Crick. *Nature* **470,** 472–473 (2011).
38. Ronning, D. R. *et al.* Active site sharing and subterminal hairpin recognition in a new class of DNA transposases. *Mol. Cell* **20,** 143–154 (2005).
39. Wang, A. H. *et al.* The molecular structure of a DNA-triostin A complex. *Science* **225,** 1115–1121 (1984).
40. Zhang, Y., Xi, Z., Hegde, R. S., Shakked, Z. & Crothers, D. M. Predicting indirect readout effects in protein-DNA interactions. *Proc. Natl Acad. Sci. USA* **101,** 8337–8341 (2004).
41. Zimmer, D. P. & Crothers, D. M. NMR of enzymatically synthesized uniformly 13C15N-labeled DNA oligonucleotides. *Proc. Natl Acad. Sci. USA* **92,** 3091–3095 (1995).
42. Pelupessy, P., Chiarparin, E. & Bodenhausen, G. Excitation of selected proton signals in NMR of isotopically labeled macromolecules. *J. Magn. Reson.* **138,** 178–181 (1999).
43. Delaglio, F. *et al.* Nmrpipe–a multidimensional spectral processing system based on unix pipes. *J. Biomol. NMR* **6,** 277–293 (1995).
44. Spyracopoulos, L. A suite of mathematica notebooks for the analysis of protein main chain 15N NMR relaxation data. *J. Biomol. NMR* **36,** 215–224 (2006).
45. Miloushev, V. Z. & Palmer, 3rd A. G. R(1rho) relaxation for two-site chemical exchange: general approximations and some exact solutions. *J. Magn. Reson.* **177,** 221–227 (2005).
46. Akaike, H. A new look at the statistical model identification. *IEEE Trans. Aut. Contr.* **19,** 716–723 (1974).
47. Schulz, M. N., Landstrom, J. & Hubbard, R. E. MTSA-a Matlab program to fit thermal shift data. *Anal. Biochem.* **433,** 43–47 (2013).
48. Doktycz, M. J., Morris, M. D., Dormady, S. J., Beattie, K. L. & Jacobson, K. B. Optical melting of 128 octamer DNA duplexes: effects of base pair location and nearest neighbors on thermal stability. *J. Biol. Chem.* **270,** 8439–8445 (1995).

Acknowledgements

We thank Dr Vivekanandan Subramanian for maintenance of the NMR instrument. We gratefully acknowledge Professor Ari Gafni for access to the CD instrument and Dr Joseph Schauerte for maintenance of the CD instrument. This work was supported by NIH grant GM089846 awarded to H.M.A.-H.

Author contributions

H.S.A., F.L.G. and H.M.A.-H. conceived the idea; H.S.A., F.L.G. and E.N.N. prepared samples and measured NMR data; H.S.A. carried out the data analysis with help from F.L.G., E.N.N. and H.M.A.-H.. H.S.A. and H.M.A.-H. wrote the manuscript with help from F.L.G. and E.N.N.

Additional information

Supplementary Information accompanies this paper at http://www.nature.com/naturecommunications

Competing financial interests: The authors declare no competing financial interests.

Reprints and permission information is available online at http://npg.nature.com/reprintsandpermissions/

How to cite this article: Alvey, H. S. *et al.* Widespread transient Hoogsteen base pairs in canonical duplex DNA with variable energetics. *Nat. Commun.* 5:4786 doi: 10.1038/ncomms5786 (2014).