SUPPLEMENTARY DISCUSSION

Visualizing Transient Watson-Crick Like Mispairs in DNA and RNA Duplexes

Isaac J. Kimsey, Katja Petzold, Bharathwaj Sathyamoorthy, Zachary W. Stein and Hashim M. Al-Hashimi*

Department of Biochemistry and Chemistry

Duke University Medical Center

Durham, NC 27710, USA

Email: hashim.al.hashimi@duke.edu

Tel: 919-684-4325, Fax: 919-684-8885

Supplementary Discussion 1 | Observed excited states are inconsistent with Hoogsteen, base opened, and other states.

Excited state 1 (ES1) and excited state 2 (ES2) chemical shifts (CSs) are inconsistent with Hoogsteen (HG) bps. Whereas for HG dG•dC and dA•dT bps the dG-N1 and dT-N3 experience small upfield CSs ($\Delta\omega_{N1/N3} = \omega_{ES} - \omega_{GS}$ = -1 to -3 ppm)[57], the excited states reported here for dG•dT and rG•rU show very large downfield CSs ($\Delta\omega_{N1/N3}$ +15-56 ppm, Supplementary Table 1). Furthermore, in HG bps the purine base (dA or dG) rotates by ~180° about X, giving rise to large downfield sugar C1′ CSs for the purine ($\Delta\omega_{C1'}$ ~3-4 ppm)[24]. For dG•dT ES1 we observe a near absence of chemical exchange for both dG-C1′ and dT-C1′, and for dG•dT ES2 the population is too small to fit the ^{13}C $R_{1\rho}$ relaxation dispersion (RD) data. Finally, dG•dC and dA•dT HG bps are strongly favored at low pH[24,78], whereas dG•dT and rG•rU ES1 are largely pH-independent, while the dG•dT and rG•rU ES2s are strongly favored at high pH.

In addition, we are able to rule out a Hoogsteen dG^{enol}•dT pair (Extended Data Fig. 4) as ES1 based on the absence of chemical exchange on the purine C1′. It has been previously shown that the *anti*-to-*syn* transition of dG in a canonical dG•dC pair to form a $dG•dC^+$ HG bp[24] gives rise to significant chemical exchange on dG-C1′ which is not observed in the dG•dT ES1.

ES1 and ES2 are also inconsistent with a base opened state. Though the base opened state model and lifetimes for dG•dT and rG•rU mispairs are contested (<3.8 μs by NMR[79] and ~10 ms by single-molecule studies[80]), both the sign and magnitude of the observed dG-N1, dT-N3, rG-N1 and rU-N3 CSs for all dG•dT and rG•rU transient states reported herein are inconsistent with a base opened state. In the base opened state, solvent-exposed base opened H-bond donors would lead to small upfield ^{15}N CSs on either

d/rG-N1, dT-N3 or rU-N3 imino nitrogens (Extended Data Fig. 4) owing to a loss of H-bonding[30,57]. Moreover, the absence of any detectable chemical exchange at C1′ in the dG•dT pair (Extended Data Fig. 2) is inconsistent with base opening which would be expected to change the glycosidic bond angle and/or sugar pucker[81] and therefore give rise to large changes in C1′ CSs[80]. The observed exchange rates for all states of dG•dT and rG•rU pairs also differ by >1-2 orders of magnitude relative to values reported for base opening[79,80]. Furthermore, a base opened ES1 in which dG-N1 or dT-N3 is deprotonated via tautomerization or ionization is expected to be exceedingly energetically unfavorable given the cumulative energetic cost of base opening coupled with deprotonation without a gain of stable H-bonding. Finally, the thermodynamic parameters for dG•dT and rG•rU ES1 (Extended Data Fig. 6), particularly the enthalpy differences for the transition to ES1, are inconsistent with an opened state, as argued previously for transient HG bps[24]. Finally, chemical modifications (^{8Br}dG•dT and dG•^{5Br}dU) which destabilize the WB GS (apparent by lowered melting temperature and higher imino proton exchange) do not increase the population of ES1, as would be expected if ES1 was a base opened state.

For dG•dT ES2 we have also considered a state in which dT is anionic but does not adopt WC-like geometry. Because the ionization of dT does not permit the formation of a third H-bond as in ES1, we have to consider that the WC-like dG•dT$^-$ can transiently slip in to an inverted wobble (iWB) geometry that has been previously observed by X-ray crystallography[3,34] and computational studies[35] (Fig. 3a and Extended Data Fig. 4). Here, dG-N1 is H-bonded to dT-O4 and dT-N3 is H-bonded to dG-N2[3,34,35]. However, the predicted CSs for the anionic iWB ($\Delta\omega_{dT\text{-}N3}$ +46 ppm) are in poor agreement with the experimentally measured values ($\Delta\omega_{dT\text{-}N3}$ +56 ppm), which are in better agreement with values predicted for the WC-like dG•dT$^-$ mispair ($\omega_{dT\text{-}N3}$ +54 ppm). Nevertheless, we

cannot rule out the possibility that the WC-like dG•dT$^-$ exists in dynamic equilibrium with such a minor iWB or other species that fall outside the detection limits of the RD experiments.

We are also able to rule out that the observed ESs represent a stable dG^{enol}•dT^{enol} WB "double tautomer"[82]. Computational studies suggest this species would be greatly destabilized relative to the canonical GS WB by 14.5 kcal/mol[82], which is an order of magnitude less stable than the ESs observed here (~3-5 kcal/mol, Extended Data Fig. 6). Moreover, such a species would entail 100% tautomeric species of both dG and dT. Full deprotonation of both bases would be expected to give rise to significantly downfield shifted imino nitrogens for both dG-N1 and dT-N3 on the order of ~45-60 ppm (Extended Data Fig. 4-5). Our observed dG•dT ES1 and ES2 chemical shifts are strongly inconsistent with simultaneous deprotonation of both dG-N1 and dT-N3 (Extended Data Fig. 4).

Finally, we have also considered an alternative tautomeric form in which dG-N7 is protonated and dG-N1 is deprotonated, resulting in an overall charge neutral WC-like structure analogous to that observed for the ^{m7}dG•dT mispair[83]. Based on protonation of dGTP at pH 2.1 (data not shown), protonation of dG-N7 should result in an upfield shift in dG-C8 ($\Delta\omega_{C8}$ -1.9 ppm). However, dG•dT ES1 features a small downfield CS on dG-C8 ($\Delta\omega_{C8}$ +1.6 ppm). In contrast, this downfield shift is consistent with a dG^{enol}•dT form (Fig. 3a).

Supplementary Discussion 2 | Additional exchange contributions at low pH.

For hp-GT DNA at pH 6.0, we observe both the expected dG•dT ES1 tautomeric form (p_{ES1} ~0.17%), but also another unique exchange process characterized by a much

larger population (p_{ES2*} ~0.97%) that increases with decreasing pH, and a slower exchange rate (k_{ex} 766 s^{-1}). This alternative second excited state (referred to as ES2*) is distinct from the anionic dG•dT^- ES2 apparent at high pH. ES2* is apparent only at low pH and does not feature CSs characteristic with deprotonation of dT-N3 or dG-N1. Rather, ES2* exchange parameters and its pH dependence are in very strong agreement with exchange parameters reported for transient HG bps in canonical dG•dC WC bps[24]. The ES2* therefore represents the dG•dT mispair remotely sensing transitions toward HG bps in the adjacent dG•dC bps. The ES CSs feature very small changes in CSs for dG-N1 ($\Delta\omega_{N1}$ +1.31 ppm), dG15-C8 ($\Delta\omega_{C8}$ +0.75 ppm), dT5-N3 ($\Delta\omega_{N3}$ -0.86 ppm) and dT5-C6 ($\Delta\omega_{C6}$ +0.86 ppm) (Supplementary Table 1). While these are inconsistent with formation of a HG bp, they are in strong agreement with changes in CS expected for a bp *adjacent* to a dG•dC bp undergoing exchange to a HG dG•dC^+ bp[24,78]. Indeed, two dG•dC pairs flank the central site-labeled dG•dT mispair, and studies show that the population of ES dG•dC^+ HG bps increase with lowering pH[24]. Thus, this exchange process most likely reflects the dG•dT mispair *passively* sensing the WC-to-HG transition in neighboring bps. Note that exchange contributions from WC-to-HG transitions in dG•dC^+ bps are negligible at neutral pH[24,78], and are therefore not expected to contribute to the measured RD at pH ≥6.9.

Supplementary Discussion 3 | Hydrogen bonding in the WC-like excited states. Imino nitrogen CSs are highly sensitive to H-bonding[30,68]. It is well documented that a loss of N-H- -N or N-H- -O hydrogen bonding in WC pairs leads to a small upfield shift in the ^{15}N CSs of protonated imino N1/3 H-bond donors and a larger downfield shift for the non-protonated N1/N3 H-bond acceptors (Extended Data Fig. 4)[30,57,68]. The ES1 dT-N3 and ES2 dG-N1 CSs are downfield shifted relative to the GS WB, consistent with a gain in WC-like H-bonding during the WB-to-WC transition. For the H-bond acceptors N1 and

N3, the dominant factor is deprotonation, which leads to a large downfield shift for ES1 dG-N1 and ES2 dT-N3. Here, a gain in H-bonding can be expected to attenuate the magnitude of the downfield shift. Indeed, this is what is observed when comparing ES1 dG-N1 and ES2 dT-N3 (Extended Data Fig. 4) with the isolated dGTP-N1 and dTTP-N3 CSs (Extended Data Fig. 5). Rapid averaging between multiple WC-like states (dG^{enol}•dT⇌dG•dT^{enol} or dG^{-}•dT⇌dG•dT^{-}) is another important factor that can contribute to the observed ^{15}N CSs. Both H-bonding effects and rapid averaging among tautomeric/anionic states are taken into account in the presented DFT analysis of CSs (see Methods).

Supplementary Discussion 4 | Inter-base distance dependent CSs.

It has been previously shown by *ab initio* studies that constriction of the N1- -N3 distance in a WC G•C base pair can give rise to large changes in the N1 and N3 ^{15}N CSs[68]. We performed similar DFT studies on a WC-like dG^{enol}•dT mispair (see Methods). Another potential explanation for the observed downfield shifted dG-N1 and dT-N3 ES1 CSs could be that they reflect a dG^{enol}•dT with constricted N1- -N3 distance (Extended Data Fig. 4) rather than a rapid equilibrium with dG•dT^{enol} (Fig. 3a). Although DFT calculations (see Methods) indicate that constricting the dG-N1 to dT-N3 distance from the optimized distance of ~2.86Å to 2.54Å could explain the observed N1/N3 CSs for dG•dT ES1 (Extended Data Fig. 4), the resulting distance is significantly smaller than that normally observed in nucleic acid structures (~2.89±0.17Å, based on a survey of 7749 canonical WC dG•dC pairs in X-ray structures of naked DNA and DNA-protein complexes). Furthermore, in the case of RNA, no single distance between rG-N1 and rU-N3 in rG^{enol}•rU (data not shown) can explain the near equivalent rG•rU ES1 ^{15}N CSs (Fig. 4b). In contrast, all of the dG•dT and rG•rU ES1 CSs can be explained by a rapid exchange

with dG•dT^{enol} or rG•rU^{enol}, respectively. Independent evidence for dG^{enol}•dT⇌dG•dT^{enol} exchange is also available based on computational studies[31,69]. Finally, the rapid exchange between dG^{enol}•dT and dG•dT^{enol} is consistent with the absence of CS perturbations with the modification ^{5Br}dU, which could impact other aspects of base pairing.

Supplementary Discussion 5 | CS discrepancy for dT-C6 in ^{m6}dG•dT mispair.

The observed CSs for ^{m6}dG-C8, ^{m6}dG-C1′, and dT-C1′ of our ^{m6}dG•dT mispairs (Fig. 3b and Extended Data Fig. 7) are consistent with a dominant dG^{enol}•dT bp. The discrepancy with the dT-C6 CS can be attributed to deviations from an ideal WC-like geometry. Here, the ^{m6}dG•dT mispair geometry shown in Fig. 3 is based on previous NMR and X-ray crystallography[12,30,84] studies on ^{m6}dG•dT which show that the dT is opened away from the paired ^{m6}dG. The H-bonding pattern was established in solution via a ^{15}N NMR study[30] that showed direct evidence for a stable H-bond between ^{m6}dG-N2 and dT-O2, and the lack of a stable H-bond between ^{m6}dG-N1 and dT-N3. However, a crystallographic study has provided indirect evidence for a CH- -O H-bond between the m6-methyl protons and dT-O4[12].

DFT calculations on ^{m6}dG•dT show that relative to a geometry-optimized dG•dT wobble, this distortion is predicted to minimally affect the ^{m6}dG-C8 CSs, but will induce a more significant upfield shift in dT-C6 ($\Delta\omega_{dT\text{-}C6(DFT\ m6dG\bullet dT\ –\ DFT\ dGenol\bullet dT)}$ -2.7 ppm), consistent with the CSs observed for the ^{m6}dG•dT mispair ($\Delta\omega_{dT\text{-}C6(m6dG\bullet dT\ –\ ES1\ dGenol\bullet dT)}$ -2.1 ppm) (Fig. 3b).

Supplementary Discussion 6 | Potential for a more complex exchange process in rG•rU mispairs.

The rG•rU ES1 population and CSs are largely independent of temperature at pH 6.9 (Fig. 4c and Supplementary Table 1) as observed for DNA at the same pH (Fig. 2a-b). Indeed, we find excellent agreement between the thermodynamic parameters describing dG•dT and rG•rU ES1 at pH 6.9 (Extended Data Fig. 6). However, upon increasing the pH to ≥7.9 in the hp-GU-20 construct we observe deviations in CS for both ES1 rG-N1 ($\Delta\omega_{ES1(pH6.9)\rightarrow ES1(pH7.9)}$ +6.1 ppm) and rU-N3 ($\Delta\omega_{ES1(pH6.9)\rightarrow ES1(pH7.9)}$ +3.6 ppm) that are not observed in DNA (Extended Data Fig. 9). These results suggest a more complex exchange process at higher pH.

Supplementary Discussion 7 | Evidence for uridine deprotonation in a guanine riboswitch.

^{15}N $R_{1\rho}$ RD data was collected for two uridines (rU17 and rU69) involved in rG•rU pairing in a 69 nucleotide *Bacillus subtilis* guanine riboswitch (termed xpt-G)[71]. RD data collected at pH 7.9 reveals strongly downfield shifted ES CSs for rU17-N3 and rU69-N3 that are consistent with deprotonation (Extended Data Fig. 9). The large downfield shift and small population are consistent with a WC-like rG•rU$^-$ mispair as observed in RNA duplexes. On-resonance ^{15}N RD was collected for rU17-N3 and rU69-N3 at pH 6.7 and suggests there is limited chemical exchange, but it cannot be ruled out that the same exchange process occurs further off-resonance from the carrier position as is seen at pH 7.9 (Extended Data Fig. 8).

Supplementary Discussion 8 | Probabilities of misincorporation and base substitutions.

The misincorporation and base substitution probabilities shown in Fig. 5a and Extended Data Fig. 10 span a variety of reaction conditions[47,48], template/primer sequence contexts[73] and polymerases[46] including high-fidelity prokaryotic and eukaryotic polymerases and viral reverse-transcriptases with limited or no proof-reading capabilities. The DNA misincorporation and base substitution data points for these figures were taken from 9 different publications[46-48,72-77]. Care was taken to ensure the correct reported pH value was assigned to each respective misincorporation probability, though in some cases the authors omitted directly stating the reaction pH and instead reference conditions in another paper. Note that we can expect that buffers will have slightly different pH values at differing temperatures, and we were unable to account for this if the original authors have not already. Moreover, the selection of kinetic misincorporation and base substitution studies used is not meant to be comprehensive. It can be expected that some studies which report on different polymerases at different conditions may not fall within a range governed by dG•dT ES1 or ES2. This is most likely for polymerase families (ex. Family Y, etc) that do not rely as stringently on correct WC geometry for incorporation (ex. Pol ι). Moreover, we expect that differing sequence contexts, temperatures, pH, and ionic strength will affect the relative stabilities of these ESs[3] such that it may help to explain the broad range of misincorporation probabilities(Fig. 5a and Extended Data Fig. 10)[46-48,73]. Initial evidence for this can be seen in the differing probabilities of forming dG•dT ES1 for hp-GT (p_{ES1} 0.17%) and Dickerson-GT ES1 (p_{ES1} 0.25%)(Supplementary Table 1).

While DNA kinetic misincorporation and base substitution studies using exonuclease-deficient polymerases can report on the role of initial dNTP selection on the overall

fidelity of replication, deconvoluting the contribution of different fidelity checks[45,85] during translation presents a challenge. It is likely that the range of amino acid misincorporation shown[44] (Fig. 5b) represents a combination of initial tRNA selection in addition to proofreading[43,45,86]. While we observe transient WC-like rG•rU mispairs wth probabilities on the order of ~10^{-3}-10^{-4}, a number of the reported amino acid misincorporation probabilities are on the order of ~10^{-5}. It is likely that ribosomal proofreading capabilities explain the 10^{-1} difference observed for the lower limit of amino acid misincorporation (10^{-5}) versus ES1/ES2 probability (10^{-4}). Though it should also be noted that the probability of forming ES2 is strongly pH-dependent and thus a lower pH could account for the probability of ES2 formation in the range of 10^{-5}.

The rG•rU-dependent amino acid misincorporation data (Fig. 5b) was taken from Zhang *et al.*[44]. The data shown for amino acid misincorporation probabilities reflect that of proteins taken from recombinant *E. coli*, Chinese Hamster ovary cells (antibodies), and from humans (human serum albumin). Because there is no reported specific temperature or pH value associated with the amino acid misincorporation rates, we could not specifically correlate our rG•rU ES1 or ES2 probabilities at a given pH or temperature. Instead, we can observe that the amino acid misincorporation probabilities observed in cells is largely spanned by the probability with which rG•rU mispairs transiently adopt a WC-like tautomeric or anionic state. Further studies are needed to explore more direct correlations between amino acid misincorporation and the WC-like rG•rU ES1 and ES2 probabilities.

Supplementary Discussion 9 | WC-like dG•T mispairs and their role in misincorporation.

These results indicate that there are potentially multiple misincorporation pathways funneling through either tautomeric (dG^{enol}•dT and dG•dT^{enol}) or anionic (dG•dT^{-} and dG^{-}•dT) mispairs, with WC-like enol tautomers being generally more strongly discriminated against during initial selection or becoming energetically less favorable within the polymerase active site.

In contrast to the probabilities of misincorporation and base substitutions during replication (10^{-3}-10^{-6}), the probability of ionization[47] or tautomerization[36,87] of NTPs in free solution is expected to deviate by several orders of magnitude from the observed misincorporation probabilities. Therefore, it is unlikely that misincorporation is driven by tautomerization or ionization of isolated NTPs.

While the data presented in Fig. 5 shows clear correlations between the probability of dGTP•dT/dG•dTTP misincorporation and the probability of forming a WC-like dG•dT ES in a duplex DNA, additional studies are needed to better understand how the complex environment within cells might modify the intrinsic duplex energetics. Moreover, while *in vitro* studies have established that base ionization is possible within a DNA polymerase active site[1,47,50,88], and crystallographic studies have established that WC dG•dT and rG•rU mispairs can form within polymerase and ribosome active sites through either tautomerization or ionization[1-3,6], to the best of our knowledge, it is not known if base ionization and/or tautomerization is possible within the active sites of polymerase and ribosomes within living cells.

Supplementary Discussion 10 | The effect of high-pH on misincorporation and base substitution probabilities.

The strong correlation between experimentally measured kinetic misincorporation probabilities of dGTP•dT/dG•dTTP and the predicted probability of forming dG•dT^- in a duplex (Fig. 5c) begins to diverge at less physiologically relevant pH values (pH 9-9.5)(Extended Data Fig. 10). This suggests a more complicated model, possibly due to ionization of other protein and nucleic acid functional groups. We also note that divergent misincorporation probabilities have been reported at pH ≥9, with kinetic misincorporation studies using AMV RT reporting an upper limit of dGTP•dT/dG•dTTP misincorporation[47] while base substitution studies performed using an exonuclease-deficient Klenow polymerase shows a continuing exponential increase in base substitution frequencies[48].