Supporting Information Appendix

Tautomerism provides a molecular explanation for the mutagenic properties of the anti-HIV nucleoside 5-aza-5,6-dihydro-2'-deoxycytidine

Deyu Li[a,b,c,1], Bogdan I. Fedeles [a,b,c,1], Vipender Singh [a,b,c,1], Chunte Sam Peng[a,1,2], Katherine J. Silvestre[a], Allison K. Simi[b], Jeffrey H. Simpson[a], Andrei Tokmakoff[a,2,3] & John M. Essigmann[a,b,c,3]

Departments of [a]Chemistry and [b]Biological Engineering, [c]Center for Environmental Health Sciences, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA. [1]These authors contributed equally to this work.

[2]Current address: Department of Chemistry, the University of Chicago, Chicago, IL 60637.

[3]To whom correspondence should be addressed. E-mail: jessig@mit.edu and tokmakoff@uchicago.edu.

Table of Contents

Supporting Results

Supporting Materials & Methods

Supporting Results

Table S1. Polymerase bypass efficiencies (reported as a percentage relative to unmodified G) of C, KP1212 and m3C as determined by the CRAB assay. The data tabulated below are shown in **Fig. 2*D***.

Lesion	HK82		HK82SOS		HK81	
	Avg.	Std. Dev.	Avg.	Std. Dev.	Avg.	Std. Dev.
C	100%	4%	100%	16%	100%	12%
KP1212	91%	9%	124%	14%	128%	11%
m3C	14%	1%	45%	4%	113%	12%

Table S2. Mutagenicity of C, KP1212, GATC, and m3C as determined by the REAP assay. The data tabulated below in part (a) are shown in **Fig. 2*E***.

(a) HK82

Lesion/base	**Average**				**Standard deviation**			
	% G	**% A**	**% T**	**% C**	**% G**	**% A**	**% T**	**% C**
C	0.0	0.0	0.0	100.0	0.0	0.0	0.0	0.0
KP1212	0.4	0.0	10.5	89.1	0.6	0.0	1.3	1.0
m3C	1.3	30.6	54.9	13.2	0.6	3.0	3.7	1.8
GATC	17.2	25.8	34.4	22.6	4.8	1.0	3.5	1.6

(b) HK82-SOS

Lesion/base	**Average**				**Standard deviation**			
	% G	**% A**	**% T**	**% C**	**% G**	**% A**	**% T**	**% C**
C	0.0	0.1	0.0	99.9	0.0	0.0	0.0	0.1
KP1212	0.2	0.1	10.3	89.4	0.2	0.1	3.1	3.2
m3C	3.2	33.6	57.5	5.7	1.3	1.4	1.9	2.6
GATC	20.5	29.5	33.1	16.9	0.6	0.4	0.7	0.3

(c) HK81

Lesion/base	**Average**				**Standard deviation**			
	% G	**% A**	**% T**	**% C**	**% G**	**% A**	**% T**	**% C**
C	0.0	0.1	0.0	99.9	0.0	0.1	0.0	0.1
KP1212	0.1	0.2	9.5	90.2	0.0	0.1	0.3	0.2
m3C	0.0	0.0	0.0	100.0	0.0	0.0	0.0	0.0
GATC	20.8	32.8	31.6	14.8	1.9	0.8	0.9	2.5

Table S3. ^{1}H chemical shift assignments of KP1212 in DMF-d_7 at 20 °C.

^{1}H Position	^{1}H Chemical Shift (δ) in DMF	^{1}H Peak Multiplicity and Coupling Constant (J in Hz)
6 a	4.57	Quartet, J = 11.5, 2.0
6 b	4.57	Quartet, J = 11.5, 2.0
1′	6.27	Triplet, J = 7.3
2′ a	2.15	Multiplet
2′ b	1.87	Multiplet
3′	4.26	Singlet (br)
3′-OH	5.20	Singlet (br)
4′	3.71	Multiplet
5′ a	3.57	Singlet (br)
5′ b	3.57	Singlet (br)
5′-OH	4.91	Singlet (br)
HOD in DMF	3.53	Singlet (br)
DMF (aldehyde)	8.03	Singlet
DMF (methyl a)	2.92	Quintet
DMF (methyl b)	2.75	Quintet

Table S4. ^{13}C chemical shift assignments of KP1212 in DMF-d_7 at 20 °C.

^{13}C Position	^{13}C Chemical Shift (δ) in DMF	^{13}C Chemical Shift (δ) in D_2O
2	N/A	161.84 (from HMBC)
4	N/A	160.91 (from HMBC)
6	N/A	52.30 (from HSQC)
1′	83.20 (from HSQC)	85.11
2′	37.06	36.48
3′	72.41	72.41
4′	87.39	86.31
5′	63.63	63.11
DMF (aldehyde)	163.15	
DMF (methyl a)	35.65	
DMF (methyl b)	30.52	

Table S5. Chemical shifts, areas, and assignments of the active protons on the nucleobase part of KP1212 in DMF-d_7 at -50 °C.

Peak	Chemical Shift (ppm)	Simulated area	Assignment	Calculated area	Difference
i	11.44 (imino+amido)	0.41	2e+5n+5m+3g	0.41	0.00
ii	8.43 (enol)	0.38	2d	0.38	0.00
iii	8.06 (enol)	0.51	1c	0.51	0.00
iv	7.43 (amino)	0.40	2f+3h	0.40	0.00
v	7.08 (amino)	0.63	1a+4l+5o+3i	0.62	-0.01
vi	6.48 (amino)	0.67	1b+4j+4k	0.68	+0.01

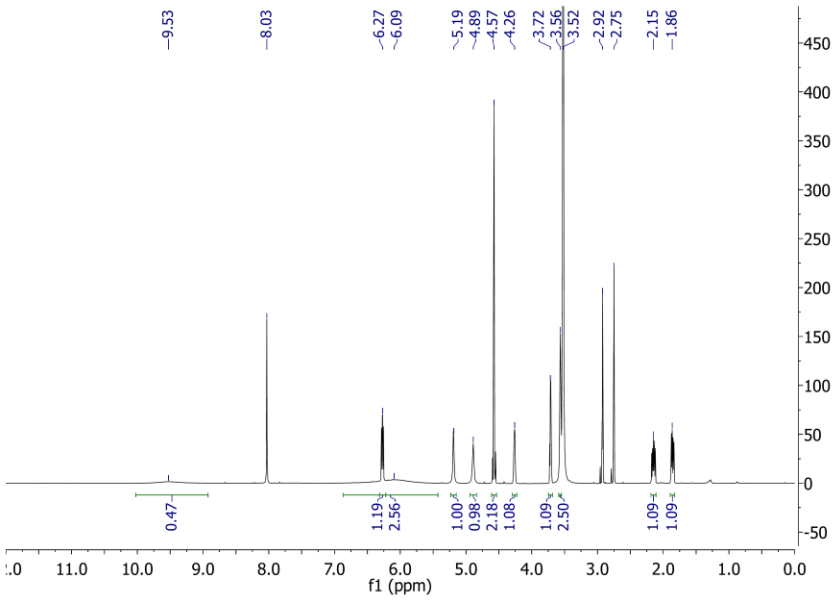

Fig. S1. ^{1}H NMR spectrum of KP1212 in DMF-d_7 at 20 °C.

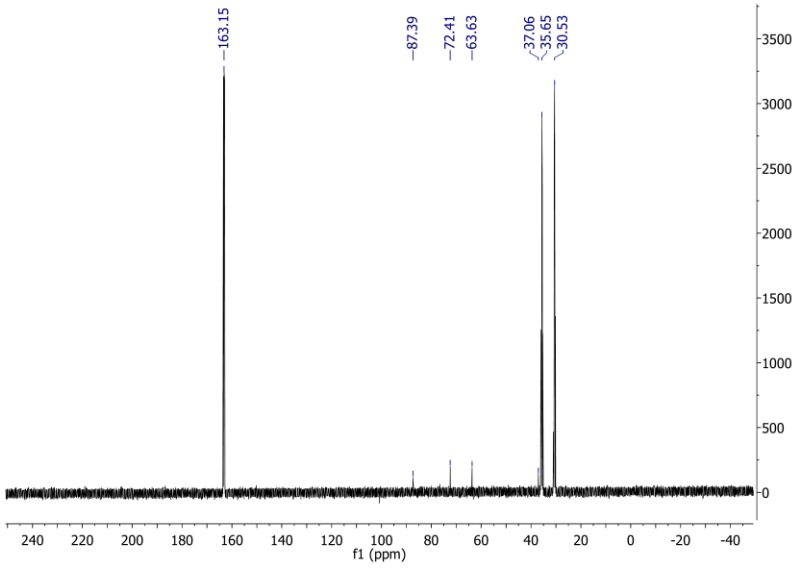

Fig. S2. ^{13}C NMR spectrum of KP1212 in DMF-d_7 at 20 °C.

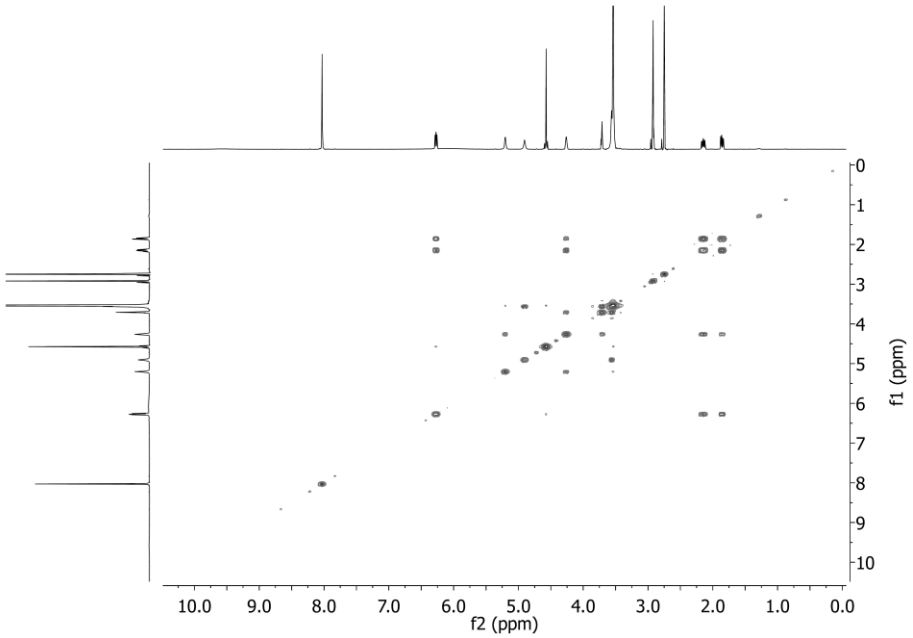

Fig. S3. ^{1}H COSY NMR spectrum of KP1212 in DMF-d_7 at 20 °C.

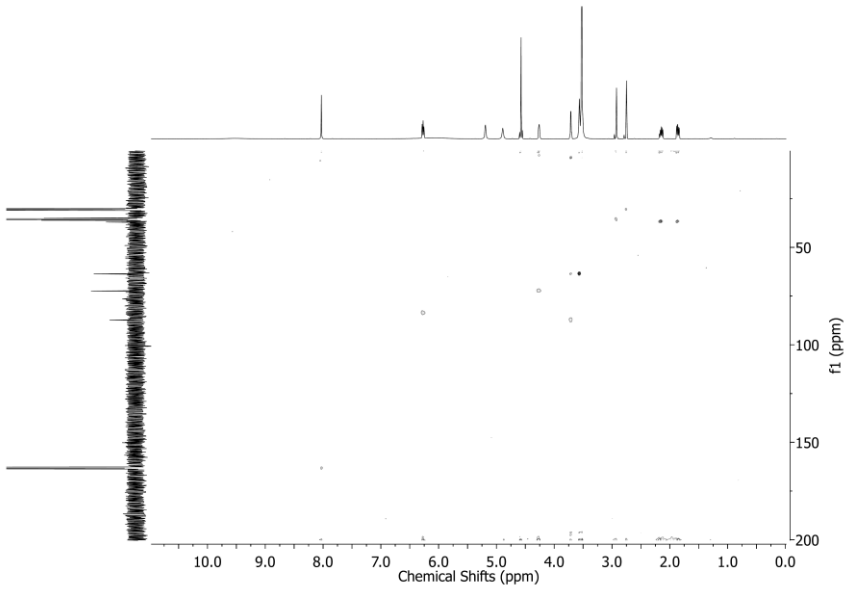

Fig. S4. ^{1}H-^{13}C HSQC NMR spectrum of KP1212 in DMF-d_7 at 20 °C.

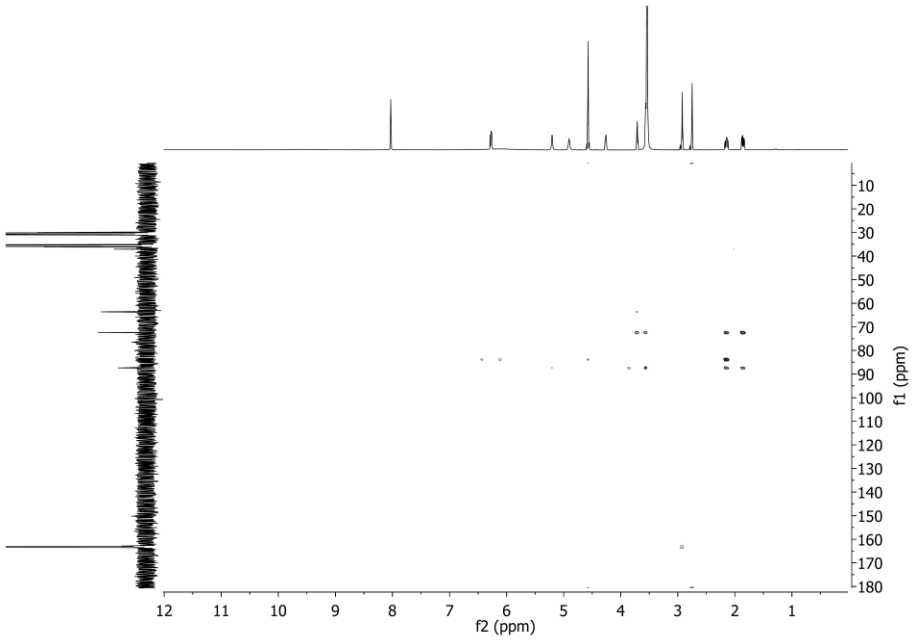

Fig. S5. ^{1}H-^{13}C HMBC NMR spectrum of KP1212 in DMF-d_7 at 20 °C.

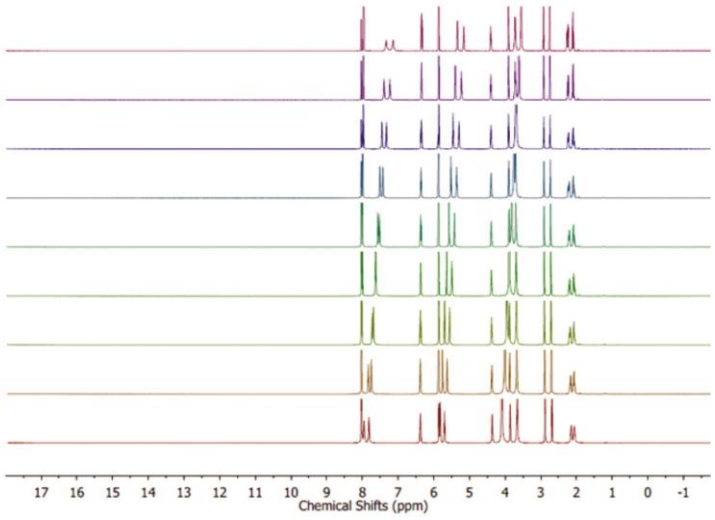

Fig. S6. Variable temperature ^{1}H NMR of deoxycytidine in DMF-d_7 from 20 °C (top) to -60 °C (bottom) in 10 °C decrements. There is no signal above 8.5ppm, which indicates no imino or enol tautomer exists under these conditions.

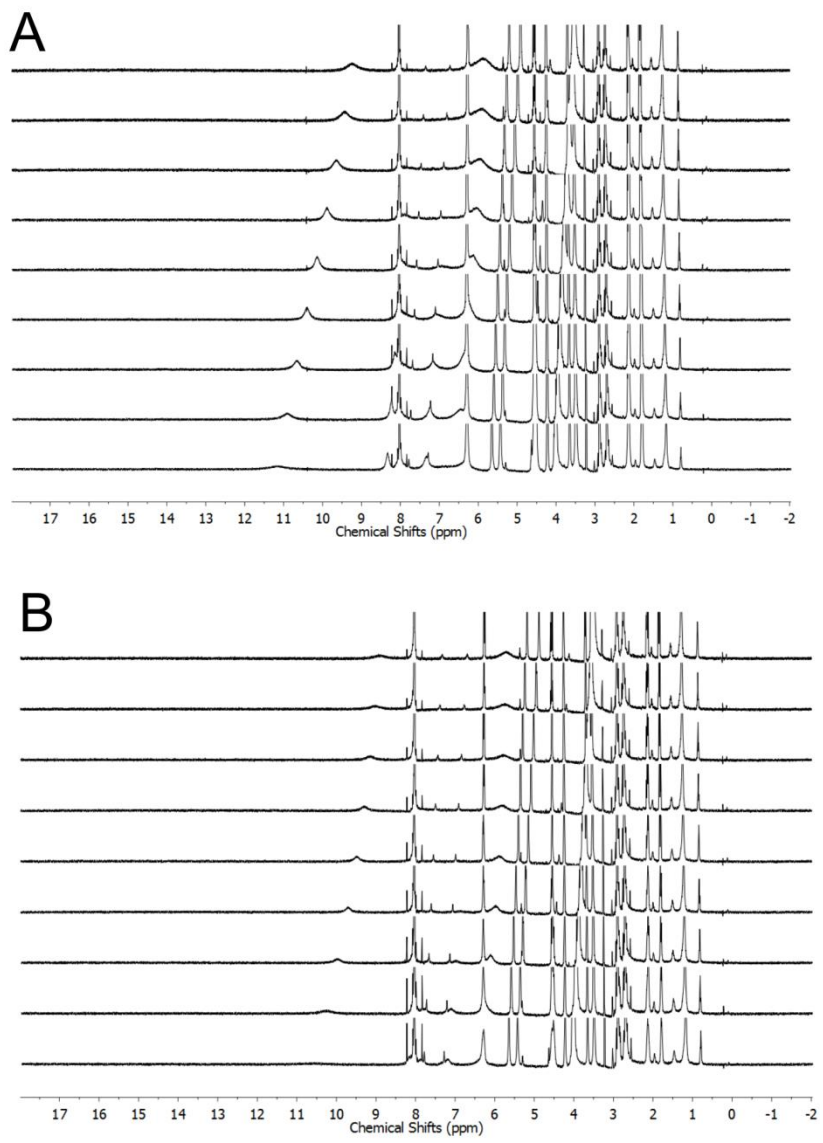

Fig. S7. Variable temperature ^{1}H NMR spectra of KP1212 in DMF-d_7 at different concentrations. (**A**) 3.2 mg/ml. (**B**) 1.0 mg/ml. In each panel, temperature varies from 20 °C (top) to -60 °C (bottom) in 10 °C decrements.

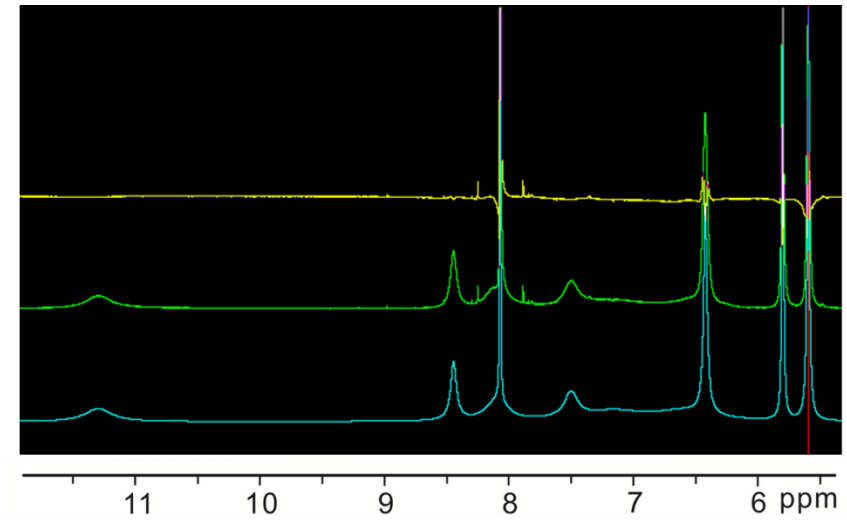

Fig. S8. Simulated vs. experimental NMR spectra. NMR spectrum (green, 5.5 to 12.0 ppm) of KP1212 at -50 °C in DMF-d_7, simulated NMR spectrum (blue), and the difference spectrum (yellow) between them.

Supporting Note

KP1212 NMR Data analysis from experiments and simulations.

In the variable temperature NMR experiments of KP212, the signals from the imino protons could easily be separated from the enol and amino signals. However, the spectral information was not sufficient to distinguish between the imino protons of regioisomers **2Z** and **2E** (**Fig. 4*A***) because the only difference between the two isomers is the configuration of the imino double

bond. For the assignment of NMR signals, the isomers **2Z** and **2E** were treated as one species (**2**); similarly, the isomers **5E** and **5Z** were treated as one species (**5**) (**Fig. 4*A***). According to the rules of keto-enol and amino-imino tautomerism of nucleic acid bases, five possible tautomers of KP1212 (**Fig. 4*A***) were possible: enol-amino tautomer (form **1**), enol-imino tautomer (form **2**), keto-amino tautomer (form **3**), keto-amino tautomer (form **4**), and keto-imino tautomer (form **5**). The active protons on the nucleobase part of the five tautomers are labeled as lower case **a** to **o** (**Fig. 4*A***).

At -50 °C six distinct proton resonances (**Fig. 3**, **i** to **vi**) were observed, attributable to protons bound either to nitrogen or to oxygen atom (for only the base portion of the KP1212 molecule). Integration of the six broad peaks permitted estimation of the relative ratios of the five tautomeric species that exist in solution (**Fig. 4*A***). Because each tautomer can generate only three discreet amino, imino, amido, and/or enol resonances, the observation of six proton resonances indicated the presence of multiple tautomeric forms of the KP1212 molecule in the solution state in DMF.

The total area of the six peaks is about 3.00, which corresponds exactly to the three active protons in the nucleobase of KP1212. By simulating the six peaks with Gaussian and Lorentzian functions and fitting the total areas to 3.00 (**Fig. S8**), the individual area of each peak was obtained (**Table S5)**. The accuracy of those simulations could be achieved to about ±1%. From the above analyses, the total area of the possible 15 active protons is 3.00. Each tautomer has three equivalent protons (which have equal integrated areas); adding together the proportions of the five tautomers should yield 1.00 or 100% in percentage.

Active proton peak assignments.

There are four types of protons in the structures of the five tautomeric species (**Fig. 4*A***): amido (**5m** and **3g**), imino (**5n** and **2e**), enol (**2d** and **1c**), and amino (**4j**, **4k**, **4l**, **5o**, **2f**, **1a**, **1b**, **3h**, **and 3i**). By comparing the chemical shifts with the NMR data of model molecules from the NMR database established by Prof. Hans J. Reich (http://www.chem.wisc.edu/areas/reich/chem605/), peak **i** at 11.44 ppm was assigned to protons from amido and imino groups; peak **ii** (8.43 ppm) and **iii** (8.06 ppm) to enol groups; peak **iv** (7.43 ppm), **v** (7.08 ppm); and **vi** (6.48 ppm) to amino protons (**Fig. 3**). (For deoxycytidine in DMF-d_7, the two aromatic amino protons on the nucleobase maintained a 1:1 area ratio at all temperatures (20 to -60 °C, **Fig. S6**) and their chemical shifts at 20 °C are 7.34 and 7.14 ppm.)

Peaks at 8.43 (the area equals 0.38, future usage of areas will be in brackets) and 8.06 (0.51) ppm are from enols. Because there are only two enol protons (**1c** and **2d**) in the five tautomers, one peak must be **1c** and the other must be **2d**.

2e is an imino peak, which should be in peak **i** at 11.44 ppm (0.41). Then **2e** should ≤0.41. Because **2e** and **2d** are in the same tautomer, their area should be the same (the three active protons on the base in any one tautomer should have the same area). Thus the enol peak **2d** could only be peak **ii** (0.38) at 8.43 ppm; the other enol peak **1c** must be peak **iii** (0.51) at 8.06 ppm.

From above, **1c**=0.51=**1a**=**1b**; and **1a** and **1b** are amino protons, which correspond to peaks **iv**-**vi**. Peak **iv** has an area of 0.40, which is ≤0.51; so **1a** and **1b** must be in peak **v** (0.63) and **vi** (0.67) separately. For chemical shift of amino protons in the same tautomer of KP1212, we assign a ring proton with bigger chemical shift than an exocyclic proton; an exocyclic proton on

the N3 side is higher than the one on the N5 side (they are slowly exchangeable). So **1a** (0.51) is in peak **v** (0.63) and **1b** (0.51) is in peak **vi** (0.67).

Because **2d**=0.38=**2e**=**2f** and **2f** is an amino peak, **2f** can only exist in peak **iv**, **v**, or **vi**. However, **2f** cannot coexist with either **1a** (in peak **v**) or **1b** (in peak **vi)** due to the fact that the total of **2** (0.38) and **1** (0.51) would be 0.89, much bigger than the area of either peak **v** (0.63) or **vi** (0.67). Therefore, **2f** could only be in peak **iv** (0.40).

For the amido and imino peak **i**, **5m**+**3g**+**5n**+**2e**=0.41. Because **2e**=0.38, then **5** should be ≤(0.41-0.38)/2=0.015; **3** should be ≤0.41-0.38=0.03. Because **2**=0.38, **1**=0.51, **5**≤0.015, and **3**≤0.03 and **4**=1.00-**1-2-3-5**, then **4**≥0.065.

The amino protons **4j**, **4k**, and **4l** could only be in peak **iv**, **v**, and **vi**. Peak **iv** (0.40) contains **2f** (0.38) and an additional 0.02 area, which could not include a peak from **4** (≥0.065). Peak **vi** (0.67) contains **1b** (0.51) and an additional 0.16 area, which cannot contain all three amino protons from **4** (≥0.065*3=0.195). Thus, there should be at least one proton from **4** in peak **v**. Peak **v** (0.63) contains **1a** (0.51) and an additional 0.12 area, which can contain no more than one proton from **4** (≥0.065). By comparing the chemical environments of **4j**, **4k**, and **4l**, we assign **4l**>**4j**>**4k** in chemical shift. So **4l** is in peak **v** and **4j** and **4k** are in peak **vi**.

Peak **iv** (0.40) contains **2f** (0.38) and 0.02 of other peak(s). There are only three amino protons left unassigned that could be in peak **iv**: **5o**, **3h**, and **3i**. The area 0.02 can thus contain one, two, or three of these. The several possibilities for the assignments of peak **5o**, **3h**, and **3i** are discussed below one by one.

Possibility 1. The residual 0.02 in peak **iv** is from **three** peaks, namely **5o**, **3h**, and **3i**. From peak **v**, **4l**=0.63-**1a**=0.63-0.51=0.12. However, from peak **vi**, **4j**=**4k**=(0.67-**1b**)/2=(0.67-

0.51)/2=0.08. Therefore, the two areas generated from peak **v** and **vi** for **4** are not consistent with each other, thus ruling out this possibility.

Possibility 2. The residual 0.02 in peak **iv** is from **two** peaks among **5o**, **3h**, and **3i**.

2.1. Assume the residual 0.02 of peak **iv** is from **5o** and **3h** (Because **3h** should have a higher chemical shift than **3i**, it is not possible to have a combination of **5o** and **3i** for peak **iv** and leave **3h** for peak **v** or **vi**.). Then **3i** should be in either peak **v** or **vi**.

Let us assume **3i** (≤0.03) is in peak **v**. From peak **vi**, **4j**=**4k**=(0.67-**1b**)/2=(0.67-0.51)/2=0.08. Then from peak **v** (0.63) and **4j**=**4k**=**4l**=0.08, we get **3i**=0.63-**1a**-**4l**=0.63-0.51-0.08=0.04, which is not consistent with the assumption that **3i** ≤0.03. Also, when **3i**=0.04=**3g** as part of peak **i**, we get **5m**=**5n**=(0.41-**2e**-**3g**)/2=(0.41-0.38-0.04)/2= -0.005, which is not possible for an NMR area.

If **3i** is not in peak **v**, it must be in peak **vi**; then **4l**=0.63-**1a**=0.63-0.51=0.12 from peak **v**. But from peak **vi**, **4j**=**4k**=(0.67-**1b**-**3i**)/2, which is ≤(0.67-**1b**)/2=(0.67-0.51)/2=0.08. These two calculations of the area of **4** contradict each other, ruling out possibility *2.1*.

2.2. Assume the residual 0.02 in peak **iv** is from **3h** and **3i**. Then **3h**=**3i**=0.02/2=0.01=**3g**. **5o** should be either in peak **v** or **vi**. From peak **i**, **5m**=**5n**=(0.41-**2e**-**3g**)/2=(0.41-0.38-0.01)/2=0.01. Now, if **5o** is in peak **v**, **4l**=0.63-**1a**-**5o**=0.63-0.51-0.01=0.11 from peak **v**. However, **4j**=**4k**=(0.67-**1b**)/2=(0.67-0.51)/2=0.08 in peak **vi**. Again, those two values of **4** from peak **v** and **vi** do not support each other.

If **5o** is in peak **vi**, **4l**=0.63-**1a**=0.63-0.51=0.12 from peak **v**. However, **4j**=**4k**=(0.67-**1b**-**5o**)/2=(0.67-0.51-0.01)/2=0.075 in peak **vi**. Those two numbers of tautomer **4** from peak **v** and **vi** do not support each other, indicating possibility *2.2* is also not correct.

Possibility 3. The residual 0.02 in peak **iv** is from **one** peak, either **5o** or **3h** (Because **3h** should have a higher chemical shift than **3i**, it is not possible to have **3i** in peak **iv**; **3i** could only be in peak **v** or **vi**).

3.1. Assume the residual 0.02 in peak **iv** is from **5o**, then **5o**=0.02=**5m**=**5n**. We can calculate from peak **i** that **3g**=0.41-**2e**-**5m**-**5n**=0.41-0.38-0.02-0.02= -0.01, which is not a possible number for an NMR area.

3.2. Assume the residual 0.02 in peak **iv** is from **3h**, then **3h**=0.02=**3i**=**3g**. From peak **i**, **5m**=**5n**=(0.41-**2e**-**3g**)/2=(0.41-0.38-0.02)/2=0.005. Then **4**=1.00-**1**-**2**-**3**-**5**=1.00-0.51-0.38-0.02-0.005=0.085. Peak **vi** contains at least three peaks: **1b**, **4j**, and **4k**. Therefore, the total area of peak **vi** is **1b**+**4j**+**4k**=0.51+0.085+0.085=0.68, which is similar to the simulated area of peak **vi** (0.67). Because 0.68 is already bigger than 0.67, the last two amino peaks, **5o** and **3i**, should both be in peak **v**. Then the total area of peak **v** from NMR data is **1a**+**4l**+**5o**+**3i**=0.51+0.085+0.005+0.02=0.62, very close to the simulated area of peak **v** (0.63).

From the above assignments, the ratios of the five tautomers were determined: 51.0% of **1**, 38.0% of **2**, 2.0% of **3**, 8.5% of **4**, and 0.5% of **5**. The relative abundance of the five tautomers is (from high to low): **1**>**2**>**4**>**3**>**5** (**Table S5**).

Supporting Materials & Methods

In vivo and *in vitro* mutagenicity, and *in vivo* lesion bypass assays.

The REAP and CRAB assays described here have been adapted from the work of Delaney and Essigmann(1, 2). Please refer to the original method for more details.

Cell strains

All *E. coli* strains used in this work contain the F′ episome, which enables infection by M13 phage. GW5100 strain was used for large scale preparation of M13 phage DNA, SCS110 (JM110, *end A1*) was used for amplification of progeny phage post-electroporation, and NR9050 was the strain of choice for double agar plating with X-gal for blue-clear detection of plaques. The *E. coli* strains used to replicate lesion-containing phage were HK81 (as AB1157, but *nalA*) and HK82 (as AB1157, but *nalA alkB22;* AlkB-deficient).

Oligonucleotides

All oligonucleotides and primers were obtained from Integrated DNA Technologies (IDT) unless otherwise specified. Oligonucleotides of the sequence 5′-GAAGACCTXGGCGTCC-3′, where X is either KP1212 or m3C were synthesized and purified as previously described. DNA concentration was measured by UV absorbance using the extinction coefficients (ε) at 260 nm. For lesion-containing oligonucleotides, the extinction coefficient was approximated with that of an oligonucleotide containing normal C instead of KP1212 or m3C. Sixteen-mer oligonucleotides with the same sequence but with X = G, A, T, or C, were used as controls. The 19-mer ‘competitor’ oligonucleotide of the sequence 5′-GAAGACCTGGTAGCGCAGG-3′ was

used in the CRAB assay. Scaffold oligonucleotides (5′-GGTCTTCCACTGAATCATGGTCA TAC-3′ and 5′-AAAACGACGGCCAGTGAATTGGACGC-3′) were used to align and ligate the 16-mers containing lesions into the *Eco*RI-cleaved single-stranded M13 vector during genome construction. The following primers were used to PCR-amplify the DNA of the progeny phage for the REAP and CRAB assays: 5′-YCAGCTATGACCATGATTCAGTGGAAGAC-3′ (REAP and CRAB forward primer); 5′-YCAGGGTTTTCCCAGTCACGACGTTGTAA-3′ (CRAB reverse primer); 5′-YTGTAAAACGACGGCCAGTGAATTGGACG-3′ (REAP reverse primer). Y denotes an aminoethoxyethyl ether group present at the 5′ ends to prevent their labeling with [γ-^{32}P]-ATP in subsequent reactions.

Enzymes and chemicals

*Eco*RI, *Hae*III, *Bbs*I, *Hin*FI, T4 DNA Ligase, T4 DNA polymerase, BSA, and their enzyme reaction buffers were from New England Biolabs. Shrimp alkaline phosphatase (SAP) was from Roche. P1 nuclease, 5-bromo-4-chloro-3-indolyl-beta-D-galactopyranoside (X-gal) and isopropyl-β-D-1-thiogalactopyranoside (IPTG) were from Sigma Aldrich. Optikinase was from Affymetrix. Sephadex G-50 Fine resin was from Amersham Biosciences. Hydroxylapatite resin, 19:1 acrylamide:bisacrylamide solution, and *N*,*N*,*N*′,*N*′-tetra-methyl-ethylenediamine (TEMED) were from Bio-Rad. Phenol:chloroform:isoamyl alcohol (25:24:1; pH 8) was from Invitrogen. [γ-^{32}P]-ATP was from Perkin Elmer. ATP (cold) was from GE Healthcare Lifesciences.

Double agar overlay plaque method for phage analysis

This method, adapted from Delaney *et al.* (2004)(2), was used for determining phage titers to ensure statistical robustness throughout the experiment, and in particular to insure that the electroporation of the constructed genomes produced a sufficient number (10^5-10^6) of initial events. This method was not used for mutational analyses. Briefly, a 1:6 dilution of a saturated overnight culture of NR9050 was grown in 2×YT medium for 1 h at 37 °C with aeration. A mixure of 300 µl of the dilute overnight culture, 10 µl IPTG (24 mg/ml), 25 µl of 1 % thiamine, and 40 µl X-Gal (40 mg/ml in DMF) was added to 2.5 ml top agar maintained in a molten state at 52 °C. The resulting mixture was used to plate both electroporated cells and appropriate dilutions of supernatants containing phage particles onto B-broth plates. These plates were allowed to solidify for 10 min at room temperature and then incubated overnight at 37 °C to obtain dark blue, light blue, or clear plaques.

M13 phage DNA

M13mp7(L2) phage single-stranded DNA starting material was isolated as follows. A well-isolated plaque from an M13 stock plated using the double agar overlay method was plugged using a sterile Pasteur pipette and vortexed in 1 ml LB, 200 µl of which was used to make a starter culture (grown overnight) by mixing with 10 µl of an overnight saturated culture of GW5100 cells in 10 ml LB. One milliliter of this phage starter culture was then used to inoculate GW5100 cells, which had been grown using 500 µl of an overnight saturated culture in 250 ml of fresh 2×YT medium for 2 h at 37 °C and shaken at 275 rpm. The inoculated culture was grown further for 8 h at 37 °C with aeration, after which the cells were pelleted and discarded. The supernatant was supplemented with 4% PEG 8000 MW and 0.5 M NaCl, which

allowed the precipitation of the phage, over 24 h at 4 °C. The phage were pelleted, resuspended in 5 ml TE pH 8, and extracted four times with 3 ml 25:24:1 phenol:chloroform:isoamyl alcohol. The aqueous phase was passed through a 0.5 g hydroxylapatite column (BioRad), washed with 5 ml TE, and eluted in 1 ml fractions with 12 ml of 0.16 M phosphate buffer. The DNA-containing fractions were identified by spotting on an agarose gel plate containing ethidium bromide. The phosphate buffer in those fractions was then exchanged for TE by three washes in Microsep 100K spin dialysis columns (Pall Lifesciences). The DNA obtained was at a yield of ≥ 1 pmol/ml of 2×YT large culture, and was stored at -20 °C until further use.

Construction of genomes

The M13mp7(L2) phage features a hairpin structure that contains an *Eco*RI site, for easy linearization. Twenty pmol of M13 single-stranded DNA were linearized by incubation with 40 U of *Eco*RI for 8 h at 23 °C. Scaffolds (25 pmol in 1 µl each) were annealed to the ends of the linearized genome by incubation at 50 °C for 5 min followed by cooling to 0 °C over 50 min. In addition, 30 pmol of each 16-mer oligonucleotide insert was 5′ phosphorylated using 15 U of Optikinase, in Optikinase buffer supplemented with 1 mM ATP, at 37 °C for 1 h. The phosphorylated oligonucleotide was subsequently ligated into the linearized genome by incubation with 1 mM ATP, 10 mM DTT, 25 ug/ml BSA, and 800 U T4 DNA ligase for 8 h at 16 °C. The powerful exonuclease activity of T4 DNA polymerase (0.25 U/µl, 4 h at 37 °C) was then used to degrade the scaffolds. Finally, the constructed M13 genome was purified by extraction with 100 µl 25:24:1 phenol:chloroform:isoamyl alcohol, followed by removal of residual phenol and salts using a Qiaquick column (Qiagen). Recovery yields of 30-45% were obtained.

Genome validation and normalization

The incorporation of lesion-containing oligonucleotides in M13 genomes was confirmed and the relative concentration of the constructed genomes was determined and normalized using the following procedure. A 10-fold molar excess of scaffolds (previously used in constructing the genomes) were annealed to ~0.35 pmol of each genome in 5 μl. The genomes were cleaved with 10 U of *Hin*FI in the presence of 1 U SAP (which dephosphrylates the newly formed 5′ ends) by incubating at 37 °C for 1 h followed by phosphatase inactivation at 80 °C for 5 min and a slow (0.2 °C/sec) cooling down to 20 °C. The 5′ ends were then labeled at 37 °C for 1 h with 1.66 pmol [γ-^{32}P]-ATP (6000 Ci/mmol) in a 12 μl reaction containing NEBuffer 2, 5 mM DTT, 150 pmol cold ATP, 5 U Optikinase and 10 U *Hae*III. After quenching the reaction with 12 μl Maxam-Gilbert loading dye (98% formamide, 10 mM EDTA, 0.025% bromophenol blue and 0.025% xylene cyanol FF), the products were resolved using 20% denaturing PAGE until the xylene cyanol dye migrated a distance of 12 cm. The bands corresponding to fully-ligated genomes were then quantified using phosphorimagery and normalized with respect to one another. The genomes were then diluted to the same final concentration. The band generated from the competitor genome was used as a marker.

Post-normalization, a test electroporation was performed in HK81 electrocompetent cells using the control genome mixed in different ratios with the competitor genome. The results of the test electroporation were determined by plating the cells immediately after electroporation using the phage-overlay method to yield a dark blue (control): light blue (competitor) plaque count ratio. The ratio that yielded a 80:20 dark blue:light blue phage count was selected as the formulation ratio for the bypass assay of the lesion-containing genomes.

Preparation of electrocompetent cells

A saturated overnight culture of the desired strain was diluted 1:100 in LB medium and grown at 37 °C with shaking at 275 rpm until it reached early log phase (OD_{600} ~0.4-0.5), typically 2-3 h. The cells were then pelleted by centrifugation at 9500 rpm (Sorvall GSA rotor), resuspended and washed three times with 175 ml cold sterile water, and finally resuspended in a minimal volume of 10% glycerol. Four hundred ml culture typically yielded 5 ml of electrocompetent cells, which were aliquoted and stored at -80 °C prior to use.

CRAB assay

Lesion-containing genomes were mixed with the competitor genome (80:20 molar ratio) and electroporated in triplicate into 100 µl competent cells using a 2 mm-gap cuvette and a BTX electroporator set to 2.5 kV and 125 Ω. Typical pulse lengths were 4.9 msec. The cells were immediately transferred to 10 ml LB and an aliquot was immediately plated using the agar overlay method to verify that a minimum of 10^5 independent initial electroporation events occurred in 10 ml of culture. Incubation for 6 h at 37 °C with aeration generated the progeny phage, which were then reamplified in SCS110 cells (10 µl overnight culture and 100 µl of the 6 h supernatant in 10 ml LB) to dilute out the residual genomic DNA used for electroporation. Single-stranded M13 progeny phage DNA was isolated from 1.4 ml of supernatant using a QIAprep Spin M13 Kit with final DNA suspension in 100 µl elution buffer. CRAB forward and reverse primers were used to amplify the region of interest from 15 µl per QIAprep elution sample in a total volume of 25 µl containing 50 pmol of each primer, 1× *Pfu Turbo* buffer, 1.25 U *Pfu Turbo* DNA polymerase and 25 mM of each dNTP. The PCR program used 30 cycles of

67 °C (1 min) annealing and 72 °C (1 min) extensions. After purification using Qiaquick PCR purification columns, 4 μl of the PCR product were treated with *Bbs*I (1.5 U) and SAP (0.3 U) in a total volume of 6 μl by incubating at 37 °C for 4 h. After phosphatase inactivation, (80 °C for 5 min), the 5′ ends were then radiolabeled using [γ -^{32}P]-ATP (1.66 pmol, 10 uCi/μl, 6000 Ci/mmol), cold ATP (16 pmol) and 5 U of Optikinase at 37 °C for 30 min. Following the kinase inactivation at 65 °C for 20 min, the labeled product was then trimmed by *Hae*III (10 U in a final volume of 10 μl) at 37 °C for 2 h, and diluted 1:1 in formamide loading dye. PAGE separation on a 20% denaturing gel, run for ~3.5 h at 550 V, until the xylene cyanol dye migrated 10.5 cm, yielded two bands for each sample: an 18-mer originating from the lesion-containing genome and a 21-mer originating from the competitor genome. The gels were visualized on a phosphorimager (Typhoon 7000, GE Healthcare Life Sciences); the band intensities were quantified using ImageJ software (National Institutes of Health). Lesion bypass was calculated as the percentage ratio of the intensity of the 18-mer band (lesion signal) to the intensity of the 21-mer band (competitor signal).

REAP assay

The REAP assay methodology is identical to the CRAB assay except for the PCR primers, which amplified only the progeny DNA originating from the lesion-carrying genomes. Following electrophoresis, the 18-mer bands were excised from the gel, crushed and extracted with 200 μl water. After desalting with Sephadex G-50 Fine resin spin columns, the samples were lyophilized to dryness, resuspended in 5 μl containing 1 μg P1 Nuclease in 30 mM sodium acetate (pH 5.3) and 10 mM zinc chloride, and incubated at 50 °C for 1 h. One μl of each sample was then spotted onto PEI-TLC plates and separated using 200 ml of a saturated solution of

$(NH_4)_2HPO_4$ adjusted to pH 5.8. After 24 h of development, the TLC plates were air-dried and quantified using phosphorimagery.

In vitro replication assay

The single-stranded, lesion-containing genomes, constructed above were also used as templates for *in vitro* polymerase extension assays. Fifty fmol of each genome were mixed with 50 pmol of REAP forward and reverse primers, 1.25 U *Pfu Turbo* DNA polymerase and 5 nmol of each dNTP in 1× *Pfu Turbo* buffer. The first cycle of a PCR program (annealing at 67 °C, extension at 72 °C) allowed the *Pfu Turbo* polymerase to synthesize the complementary strand and place a base opposite the lesion. As replication occurs at 72 °C and *Pfu Turbo* is one of the most accurate polymerases known, possessing a robust 3′-5′ exonuclease activity for proofreading, this assay captures the most stringent intrinsic base-pairing preference of the DNA lesion. Subsequently, 29 PCR cycles amplified the newly formed strand and generated a product that was then analyzed using the exact downstream steps of the REAP assay described above.

Supporting References

1. Delaney JC, Essigmann JM (2006) Assays for determining lesion bypass efficiency and mutagenicity of site-specific DNA lesions in vivo. *Meth Enzymol* 408:1–15.

2. Delaney JC, Essigmann JM (2004) Mutagenesis, genotoxicity, and repair of 1-methyladenine, 3-alkylcytosines, 1-methylguanine, and 3-methylthymine in alkB Escherichia coli. *Proc Natl Acad Sci USA* 101:14051–14056.