Supplementary Tables

Supplementary Table 1. Models of increasing complexity when applied to reference position 54 in the 8-oxoG template described in the main text capture more information resulting in statistically significant increases in the likelihoods (from least to most complex models). The different models tested are given along the columns. In going from a multisite likelihood model (Multi-site Mixture) to a CRF with nearest neighbor interactions (CRF2), the –loglikelihood drops from 4010.8 to 4003.1, corresponding to a p-value of 1.12e-4 as determined by the loglikelihood ratio test.

Model Type	Single Site Homog.	Single Site Mixture	Multi-site Homogeneous	Multi-site Mixture	CRF1	CRF2
Window Size	1	1	8	8	8	8
# Kinetic Rate Parameters	1	2	8	16	16	16
# Mixing Proportion Parameters	0	1	0	8	8	8
# Interaction Parameters	-	-	0	0	1	2
# Total Parameters	1	3	8	24	25	26
- log likelihood	591.4	592.8	4038.2	4010.8	4010.6	4003.1
χ^2 p-value	-	1.00	-	3.89e-6	0.52	1.12e-4

Supplementary Table 2. 95% confidence intervals for the GATC sites tested for 6-mA modifications in the pRRS plasmid in M.EcoK*dam* (as described in the main text).

Position	Strand	Cas Mean IPD	Control Mean IPD	Supervised Loglikelihood Ratio	Model Based P value (-log10)	Mixing Proportion Estimate	Lower Bound of 95% CI	Upper Bound of 95% CI
3122	reverse	13.69	1.74	2205.81	480.75	0.22	0.16	0.28
307	forward	12.63	2.26	1874.98	408.88	0.23	0.16	0.30
1219	reverse	13.91	1.93	3099.18	674.82	0.24	0.17	0.30
2940	reverse	13.45	1.76	2439.40	531.50	0.24	0.17	0.30
308	reverse	14.47	2.00	2830.52	616.46	0.22	0.17	0.26
344	reverse	13.37	2.13	2694.46	586.91	0.24	0.18	0.30
3315	reverse	13.15	2.07	1856.69	404.91	0.24	0.18	0.30
1140	forward	11.61	1.64	3408.69	742.05	0.24	0.18	0.31
3121	forward	13.28	2.04	1954.50	426.16	0.24	0.18	0.31
3128	reverse	12.65	2.04	2054.60	447.91	0.24	0.18	0.30
664	forward	13.68	1.79	2236.49	487.42	0.24	0.19	0.29
1128	forward	11.66	1.60	2984.46	649.90	0.30	0.19	0.41
2574	forward	9.70	1.40	2584.03	562.92	0.26	0.19	0.34
2444	forward	12.05	1.65	2456.17	535.14	0.25	0.19	0.30
3314	forward	16.00	2.29	1998.36	435.69	0.26	0.20	0.33
360	forward	10.86	1.38	2655.24	578.39	0.26	0.20	0.32
3127	forward	14.73	2.50	1765.80	385.16	0.27	0.20	0.33
2900	forward	12.74	1.61	2430.56	529.58	0.27	0.20	0.33
3109	reverse	15.24	2.51	1955.27	426.33	0.27	0.20	0.34
2808	forward	10.89	1.33	2590.47	564.32	0.26	0.21	0.32
682	forward	12.32	1.81	2259.95	492.52	0.26	0.21	0.32
1237	forward	11.20	1.35	3572.79	777.70	0.27	0.21	0.33
1227	reverse	12.51	1.73	3143.23	684.39	0.26	0.21	0.32
1226	forward	15.92	2.25	2942.07	640.70	0.25	0.21	0.30
343	forward	14.26	1.76	2717.85	591.99	0.30	0.21	0.40
1313	reverse	11.61	1.46	2985.20	650.06	0.27	0.21	0.34
361	reverse	15.37	1.70	2547.12	554.90	0.26	0.21	0.31
1238	reverse	13.40	1.81	2972.19	647.24	0.27	0.21	0.33
1141	reverse	14.07	1.93	3113.37	677.90	0.26	0.21	0.32
2901	reverse	13.82	2.02	2207.48	481.12	0.27	0.22	0.33
2939	forward	16.63	2.89	2060.82	449.26	0.29	0.22	0.36
683	reverse	11.84	1.49	2554.09	556.42	0.27	0.22	0.33
618	forward	13.69	1.89	2493.81	543.32	0.28	0.22	0.34
3186	reverse	14.38	2.13	2250.24	490.41	0.29	0.23	0.36
619	reverse	13.68	1.96	2422.79	527.89	0.28	0.23	0.33
1312	forward	14.88	2.34	2762.38	601.66	0.31	0.23	0.38
1024	reverse	14.15	2.46	2246.02	489.49	0.28	0.23	0.34
1129	reverse	10.80	1.45	3486.92	759.04	0.29	0.23	0.34
1023	forward	14.29	1.88	2630.61	573.04	0.30	0.24	0.37
3108	forward	11.24	1.84	2052.24	447.39	0.29	0.24	0.35
1218	forward	16.39	2.33	3013.97	656.31	0.30	0.24	0.36
3185	forward	14.47	2.16	2116.75	461.41	0.30	0.24	0.36
2445	reverse	12.84	1.73	2457.05	535.33	0.30	0.24	0.37

2575	reverse	13.61	1.41	3114.26	678.10	0.32	0.25	0.38
665	reverse	15.38	3.08	2017.33	439.81	0.37	0.26	0.48
2809	reverse	19.80	2.65	2048.16	446.51	0.35	0.27	0.42

Supplementary Table 3. Number of significant detections (FDR < 0.05) made for each nucleotide type on the light and heavy strands of the mitochondrial genome. The numbers in bold parentheses in the light or heavy rows of the G column indicate the number of sites that were enriched for sequencing errors. The numbers in parentheses in the last row indicate the fold enrichment of observed divided the number expected from the background distribution of the four nucleotides.

Strand	**G**	**A**	**T**	**C**	**Total**
heavy	99 **(9)**	12	36	25	172
light	42 **(3)**	22	18	48	130
Both	141	34	54	73	302
Expected Count Both Strands	67 (2.1x)‡	85 (0.4x)†	90 (0.6x)†	72 (1.01x)	

‡ Enrichment p value << 0.01

† Under-enrichment p value << 0.01

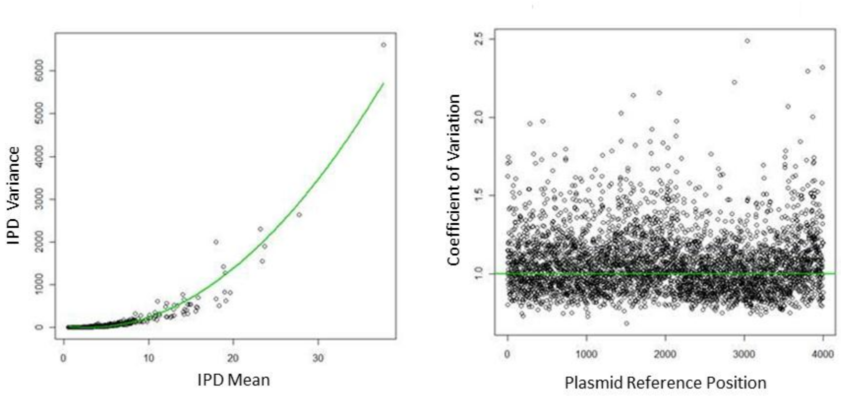
Supplementary Table 4. Adenosine residues in the mtDNA genome detected as significantly kinetically varying and greater than 20 bases away from the nearest neighboring kinetic variation site.

Strand	mtDNA Position	Closest C Kinetic Variation Event	Closest G Kinetic Variation Event	Closest T Kinetic Variation Event	Minimum Distance
heavy	2579	58	69	135	58
heavy	2607	75	41	163	41
heavy	1644	33	60	63	33
heavy	3862	35	32	47	32
heavy	2252	29	102	76	29
heavy	1637	28	53	56	28
heavy	1124	28	29	90	28
light	4150	77	82	94	77
light	1898	401	70	97	70
light	4108	101	87	52	52
light	1008	45	53	41	41
light	5237	30	28	35	28
light	3309	40	26	66	26
light	5159	24	106	30	24

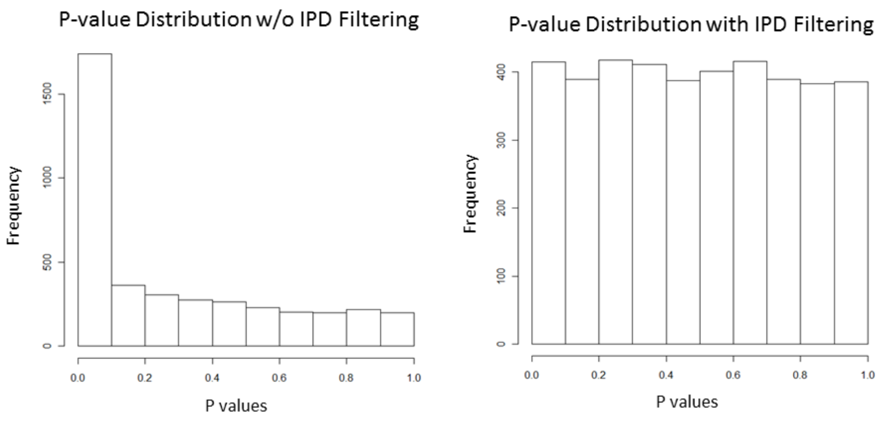
Supplementary Table 5. The twelve mtDNA kinetic variation events at A residues validated using synthetic oligonucleotides.

Position in the mtDNA Genome	**Strand**	**Log-Likelihood Ratio – log10 (P value)**	**21 Base Context (x is the KVE A residue)**	**Full Oligonucleotide Sequence**
8003	+	12.4123	TGACGTTGACxATCGAGTAGT	/5Phos/cccgCCTCTACCTAxAACTCACAGCTGACGTTGACxATCGAGTAGTcaac
782	+	11.807	ATGCAGCTCAxAACGCTTAGC	/5Phos/cccgCTTTAGCAATxAACGAAAGTTATGCAGCTCAxAACGCTTAGCcaac
8548	-	9.9688	GCAATGAATGxAGCGAACAGA	/5Phos/cccgGGTGGCACGGxGAATTTTGGAGCAATGAATGxAGCGAACAGAcaac
15751	+	9.1899	TCCTCATTCTxACCTGAATCG	/5Phos/cccgTCCTCATTCTxACCTGAATCGTACACAATCAxAGACGCCCTCcaac
839	+	8.8488	CTTTAGCAATxAACGAAAGTT	/5Phos/cccgCTTTAGCAATxAACGAAAGTTATGCAGCTCAxAACGCTTAGCcaac
13795	+	8.2646	CCTCTACCTAxAACTCACAGC	/5Phos/cccgCCTCTACCTAxAACTCACAGCTGACGTTGACxATCGAGTAGTcaac
15426	+	8.0786	TACACAATCAxAGACGCCCTC	/5Phos/cccgTCCTCATTCTxACCTGAATCGTACACAATCAxAGACGCCCTCcaac
1583	+	7.4617	AGTGTACTGGxAAGTGCACTT	/5Phos/cccgCTTCGCTTCGxAGCGAAAAGTAGTGTACTGGxAAGTGCACTTcaac
15963	+	7.2953	ATTTCTGAAAAAGAGACTAAA	/5Phos/cccgCTAAGATTCTxATTTAAACTAAAATCAGAGAxAAAGTCTTTAcaac
7338	+	7.1054	CTTCGCTTCGxAGCGAAAAGT	/5Phos/cccgCTTCGCTTCGxAGCGAAAAGTAGTGTACTGGxAAGTGCACTTcaac
16006	+	6.9358	CTAAGATTCTxATTTAAACTA	/5Phos/cccgCTAAGATTCTxATTTAAACTAAAATCAGAGAxAAAGTCTTTAcaac
4373	-	6.8701	GGTGGCACGGxGAATTTTGGA	/5Phos/cccgGGTGGCACGGxGAATTTTGGAGCAATGAATGxAGCGAACAGAcaac

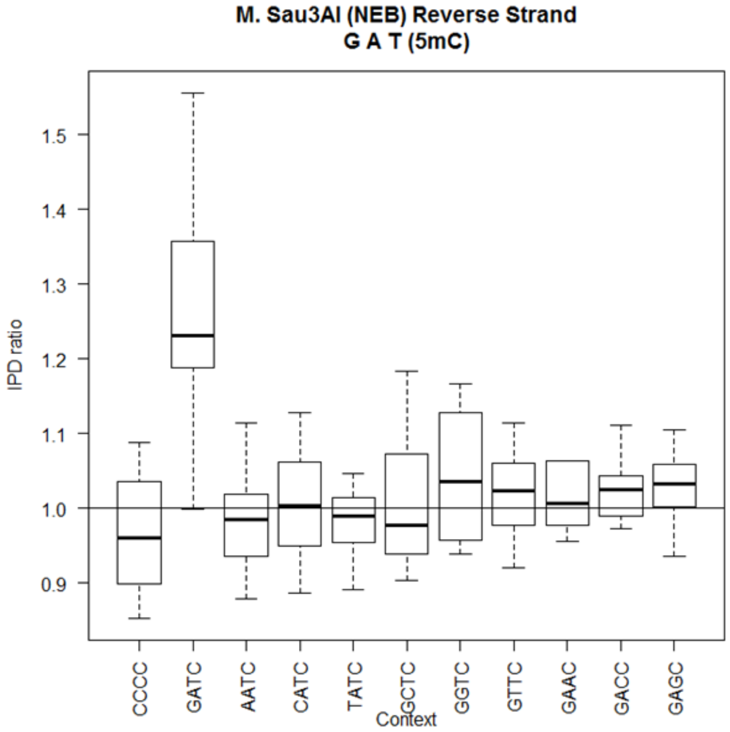
Supplementary Figures



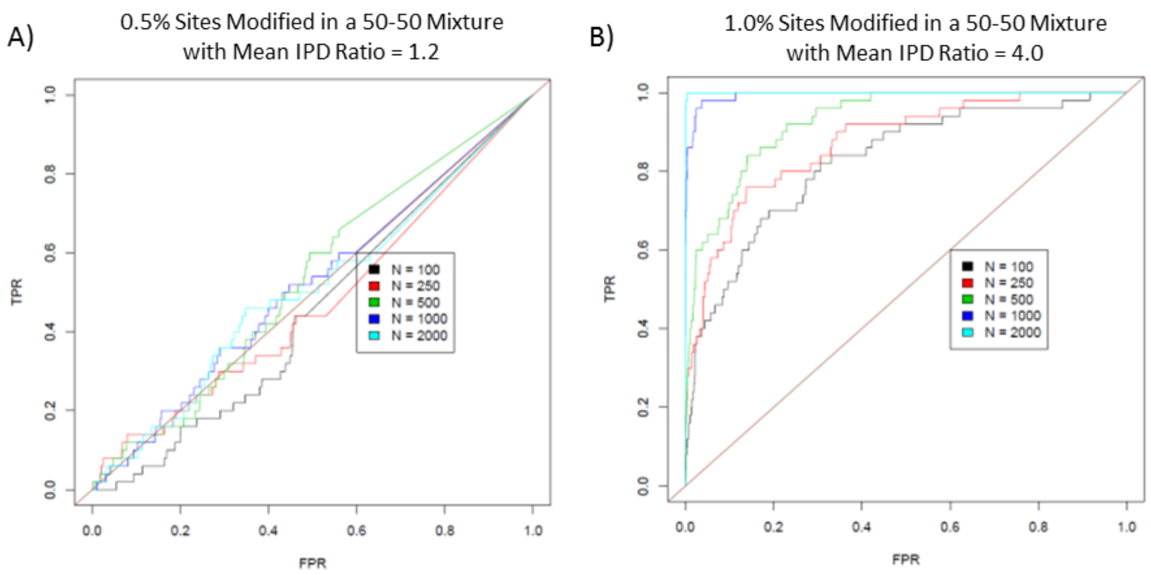
Supplementary Figure 1. Assessing the relationship between mean and variance in interpulse durations (IPD). The left plot compares IPD variance to IPD mean using the M.Sau3AI plasmid data. The quadratic relationship between the mean and variance measures is the relationship expected for an exponential distribution. The right plot examines the coefficient of variation for IPD values as a function of position in the plasmid for the M.Sau3AI data. The average value across the genome is approximately 1.06, close to the expected value of 1 if the data were exponentially distributed.



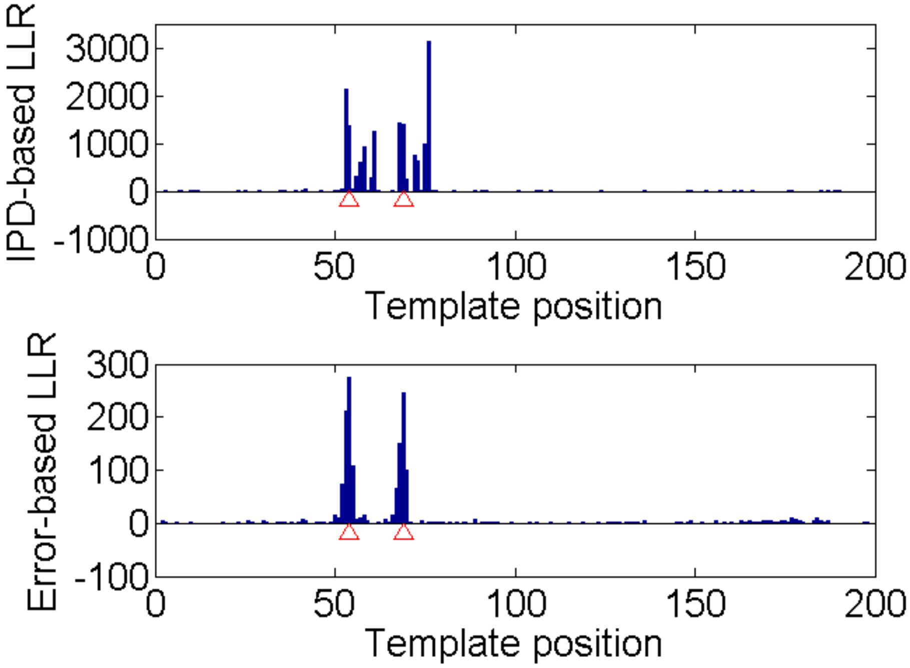
Supplementary Figure 2. P value distribution for the single site exponential model. **A)** P value distribution without filtering out very long IPD values. **B)** P value distribution after filtering out very long IPD values.



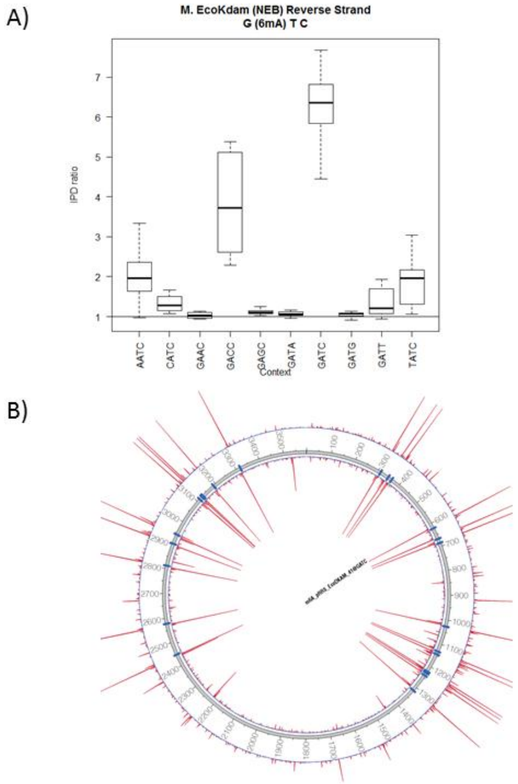
Supplementary Figure 3. Box plot of the IPD ratios for all 4mer contexts in the pRRS plasmid ending with a C residue. Only the GATC context is observed to give rise to IPD ratios that are significantly greater than one, confirming the specificity of the methyltransferase Sau3AI for the GATC context.



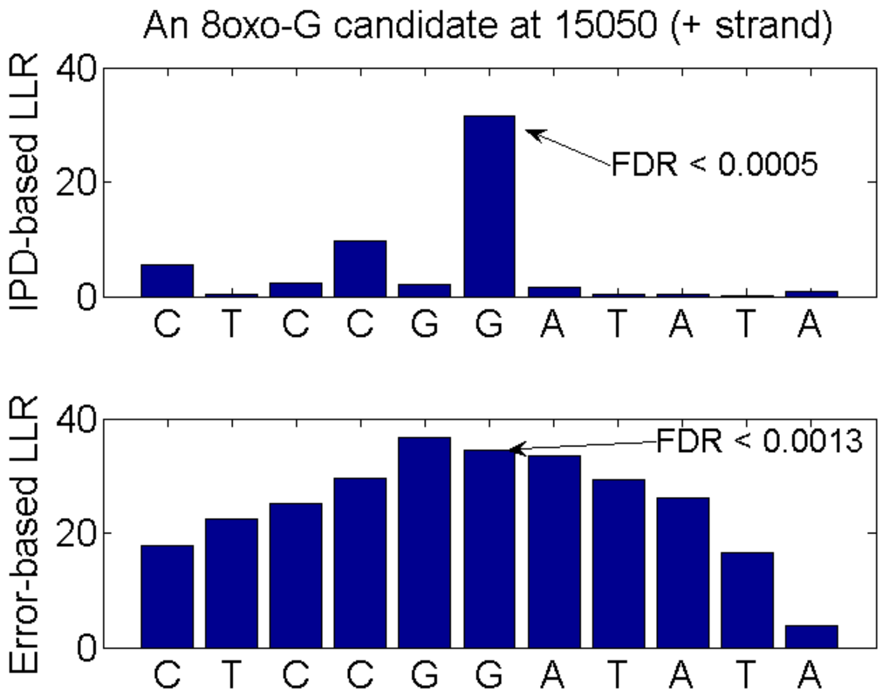
Supplementary Figure 4. ROC curves for simulated 5-methyl-C data using the full CRF model allowing for nearest neighbor interactions. **A)** Using the plasmid sequence described in the main text, 0.5% of the sites were simulated as having the 5-methyl-C modification in a 50-50 mixture (50% of the sequences with respect to a given site having the modification and 50% having no modification), with the mean IPD ratio between modified and unmodified sites simulated to be 1.2, the mean ratio we observed in the M.Sau3AI data set between amplified and observed sequence data. Even as the sample size is increased from 100 to 2000, the area under the ROC curve does not increase significantly beyond 0.5 (what we would expect by chance). **B)** Similar to panel A), but now we have increased the mean IPD ratio to 4.0 and increased the number of modified sites to 1%. In this case, with very significantly increased signal to noise, the unsupervised model readily identifies the mixtures at the different modified sites. By the time the sample size hits 2000, we are able to perfectly identify the affected sites using the unsupervised model without any false positives.



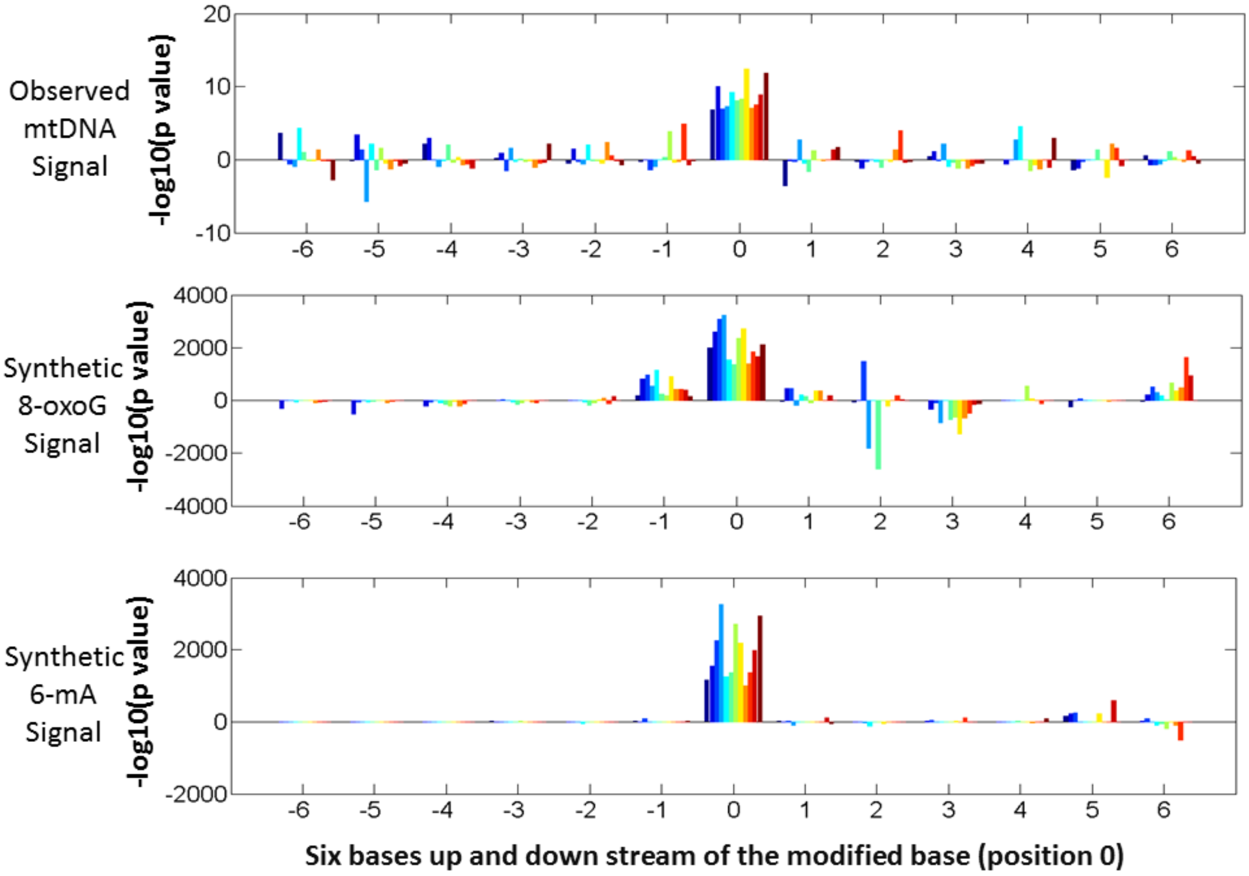
Supplementary Figure 5. IPD and error profiles predict 8-oxo-G events. In the upper panel the IPD-based loglikelihood ratio (LLR) statistics are plotted for each position in the 200bp artificial template in which 8-oxo-G lesions were induced at two positions (red triangles). The only significant signal occurs at the locations of the 8-oxo-G events, with IPDs at many of the sites in the neighborhood of the lesions affected. The lower panel depicts the LLR test statistics for the error profile at each position. Single molecule allelic differences at a given site compared to the reference sequence adds to the significance of this statistic. The high LLR values in this panel indicates very significantly rates of error at and around the location of 8-oxo-G events.



Supplementary Figure 6. Detection of M.EcoKdam induced modification events. **A)** Box plots for the IPD ratios of the A residues in the indicated 4mer contexts for the M.EcoKdam data. The IPD ratios for the GATC context as expected are significantly greater than 1, but also unexpectedly for the GACC and AATC contexts. **B)** Plasmid pRRS depicted as a circos plot, with the inside of the annulus representing the coordinates of the plasmid, the blue hash marks indicating A residues in a GATC context, and the two red curves representing –log10(p value) for the single site likelihood model for the two DNA strands. The p values are based on the full 500 coverage of the plasmid genome.



Supplementary Figure 7. Example of a putative 8-oxo-G event at position 15,050 on the positive strand of the mtDNA genome (displayed in 3' to 5' orientation). The upper panel highlights a very significant loglikelihood ratio test statistic at this position (p ~ 5e-8), indicating that the IPDs at this position in the native sample were significantly longer than the IPDs in the control sample. The bottom panel indicates a significantly elevated error rate at this position and in the neighborhood of this position.



Supplementary Figure 8. Classifying kinetic variation events by comparing observed KVE signals to KVE signals induced by known modification types in the same context. Twelve putative 6-mA events were identified from the mtDNA KVE signature for validation. The KVE signal for these twelve different events are depicted in the top graph, with the 12 color bars for each of the 13 positions depicted representing the signal for each KVE. The x-axis represents the 6 bases upstream and 6 downstream of the KVE of interest (at position 0). The y-axis represents the log likelihood ratio for the kinetic variation at the site detected (position 0) and for the positions flanking the site detected. The second and third graphs represent the kinetic variation signature for the original 12 KVEs identified in the mtDNA, but with oligos for the corresponding flanking sequences synthesized so that position 0 harbored 8-oxo-A and 6-mA modification events corresponding to the second and third graphs, respectively. A comparison of the first and second graphs indicates secondary peaks at position -1, +1, and +6 for the 8-oxo-A graph but not for the observed mtDNA graph, whereas the signal for the third graph with strong, consistent signal in both cases only apparent at position 0, consistent with a 6-mA event.