Electronic Supplementary Information

Advantages and challenges associated with bisulfite-assisted nanopore direct RNA sequencing for modifications

Aaron M. Fleming,* Judy Zhu, Vilhelmina K. Done, and Cynthia J. Burrows*

Department of Chemistry, University of Utah, 315 S. 1400 East, Salt Lake City, UT, 84112-0850, United States

Item **Page**

Figure S1. RNA sequences studied and their characterization.

The sequences provided are for the DNA coding strands of the duplex DNA used for the in vitro transcription of the RNA studied. The T7 RNA polymerase promoter in each sequence is underlined.

The first two sequences were used to explore the nanopore sequencer responses for U, Ψ, and the Ψ-(SO_3^-) adducts.

Strand 1

5`- AAGCTAATACGACTCACTATAGGAGCACAGGACCAGACGCTGCACAGAGCCGAAGCACAGCAGACCA
GACCTTATCCAGAAGACGAGACCAAATGACCAGAAGCCGAAGCACAGACGAAATTAGCCAGACGGACA
ACAGCAGAGACCGAAGCGTGGGCAGACACGCAGCGACAGAGCAGCAGGTGAGGACCAGTCAGGACA
ACAGAAAACAAAAAAAAAA

Strand 2

5`-AAGCTAATACGACTCACTATAGGAGCAGCACGAGACGAGGTGACACGACAGAGAGCGGACGCAGTCACGACCG
ACGAACACGCAGCTGCCAGACAAAGAGAACGCAGCACGACGTAGCGACGCAGACGGCGCAGCGAGCATAGCACG
CACGCAGCCACGCACAGACCGTCGCCAGCCGCAGCAGCACGACACATCGCGACGGCACGGAGCGGACGCACGAC
GAGCACAAAACAAAAAAAAAAAA

The third strand studied the expansion of the two k-mers to sequences that include the space up to the helicase.

Strand 3

5`-AAGCTAATACGACTCACTATAGGAGCAGCACGAGACGAGGTGACACGCGAGACAGACACACGGCGGGTGCCG
ACGAACAC AAAACGGCGGCCGTAACGCCAGACAAAGAGCAACGGCGGCCGTAACGACGCAGACGGCGCAGCG
AGCACGCACGCAG CACACGGCGGCCGTAACAGCCGCAGCAGCACGACGACGGCACGGAGCGGACGCACGACG
AGCACAAAACAAAAAAAAAAAA

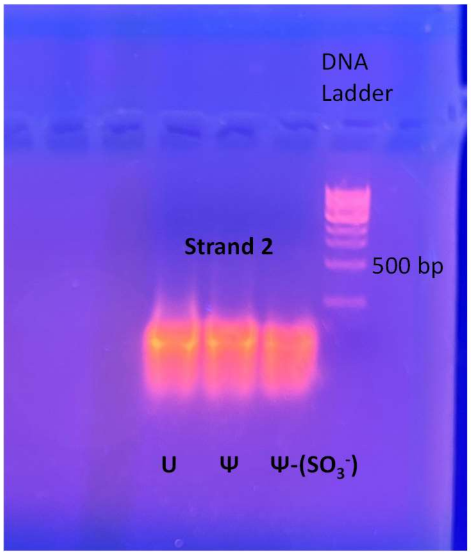
The fourth strand was used to study the nanopore response in the current and dwell time data for two Ψ or Ψ-(SO_3^-) adducts.

Strand 4

5`- AAGCTAATACGACTCACTATAGGAGCAGCACGAGACGAGGCGACACGACAGAGAGCGACGAAAAACGATCAG
CCCCCACGACACGCAGCGCCAGACAAAGAGAACGCAGCACGACGAGCGACGCAGACGGCGCAGCGAGCAAGCAC
GCACGCAGCCACGCAAAAACCGTGCACCCCCAGCCGCAGCAGCACGACACACGCGACGGCACGGAGCGGACGCA
CGACGAGCACAAAACAAAAAAAAAAAA

The fifth strand was used to study the nanopore sequencer responses for C, m^5C, and hm^5C before and after the pH 5 bisulfite reaction.
Strand 5

5`- AAGCTAATACGACTCACTATAGGAGTATAGGATTAGATAGATGGCGGAGTTGAAGTATAGTAGATTAGAGTCAGA
GAAGATGAGATTGAGGTCGGTTAGAAGTTGATGTATAGATGATGCAGTTAGATGGATAGTAATTTTAGTAGAGAT
TGAAGGTCAAGTAGATATGTTAGTAGACCGGTGATGAGGTGATATTGTCGTGGATATTAGATATATGGGGAGATG
ATAGTAGAGGATTGAAAATAAAAAAAAAAAA



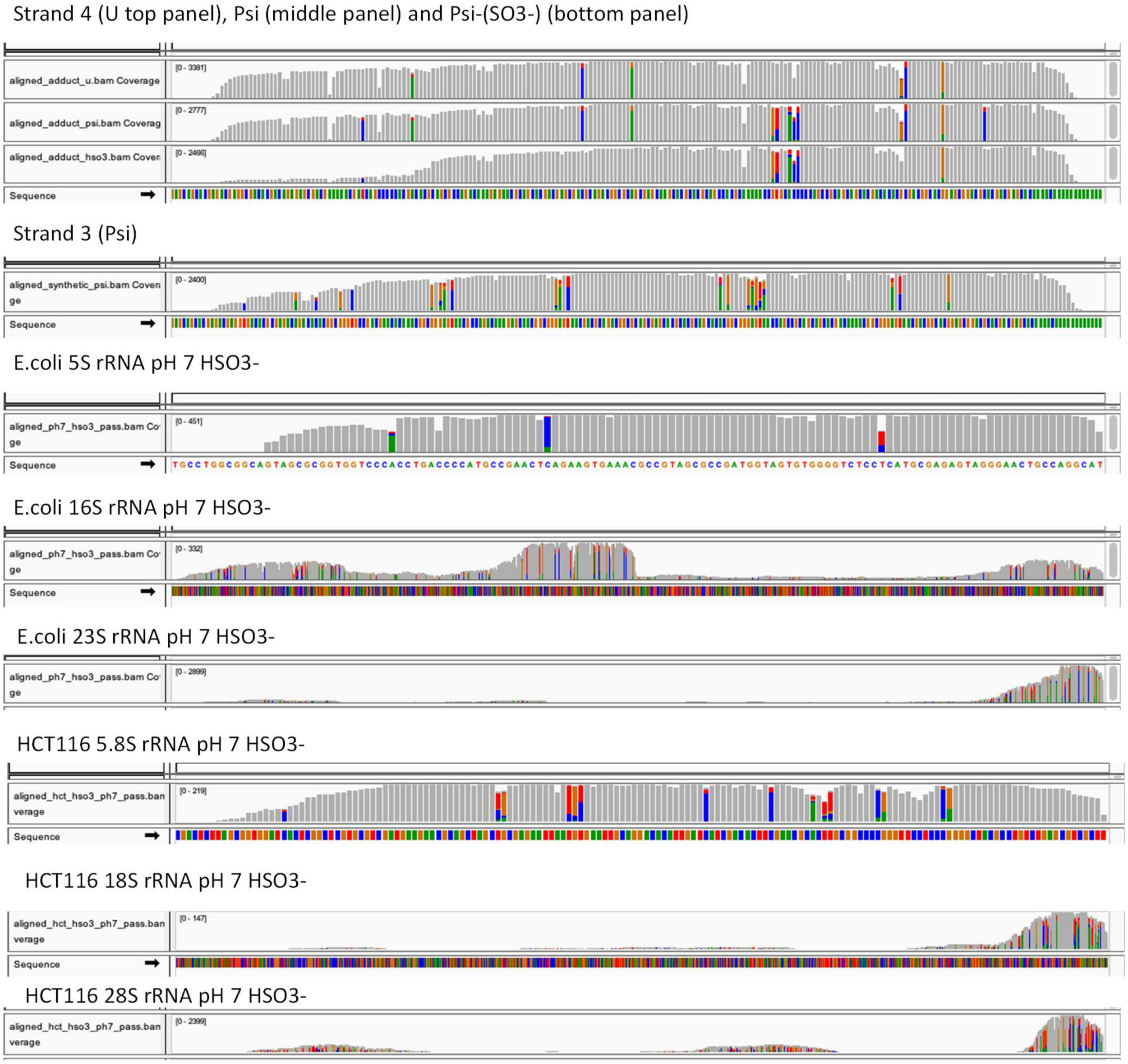
Example 1% agarose gel electrophoresis analyses conducted on strand 2 with U, Ψ, or Ψ-(SO_3^-) adduct. The example gel provided verifies the band profile did not change after the bisulfite reaction for verification the RNA had not undergone significant degradation. The lanes were overloaded to visualize whether there was an increase in the short strands that are less intense from the ethidium bromide stain used for visualization. The commercial ladder used for comparison was a DNA ladder, not an ssRNA ladder. In our hands, by the time the RNA ladders were received in the lab, they had already degraded to a point that rendered them unusable for comparison. The high stability of DNA is far superior for these ladders; moreover, the gel was used to determine if the band profile was the same between the RNAs, which it was, and never used for estimation of the strand length. Information about length was provided by the nanopore sequencing experiment.

Figure S2. Alignment details for the sequencing data.

Sample	Alignment Count	Reads Collected
Strand 5 (C + HSO3-, pH 5)	1305	3819
Strand 5 (m5C + HSO3-, pH 5)	3271	5977
Strand 2 (Psi +HSO3-, pH 7)	1353	13084
Strand 1 (Psi + HSO3-, pH 7)	17326	45848
Strand 4 (U)	3420	6000
Strand 4 (Psi)	2802	6919
Strand 4 (Psi + HSO3-, pH 7)	2482	7242
HCT116	621	3914
HCT116 (+HSO3-, pH 7)	4541	220656
E. coli (+HSO3-, pH 5)	1729	14078
E. coli (+HSO3-, pH 7)	34317	189930
Strand 5 (hm5C)	3885	8913
Strand 5 (HM5C, HSO3-, pH 5	360	3175
Strand 3 (Psi)	2416	6981

The alignment data for the synthetic RNA, *E. coli* and strands 1, 2, and 5 before the bisulfite reaction were previously reported by our lab.[1-3]

Figure S3. The IGV plots for the sequencing experiments.


Strand 4 (U top panel), Psi (middle panel) and Psi-(SO3-) (bottom panel)

Strand 3 (Psi)

E.coli 5S rRNA pH 7 HSO3-

E.coli 16S rRNA pH 7 HSO3-

E.coli 23S rRNA pH 7 HSO3-

HCT116 5.8S rRNA pH 7 HSO3-

HCT116 18S rRNA pH 7 HSO3-

HCT116 28S rRNA pH 7 HSO3-

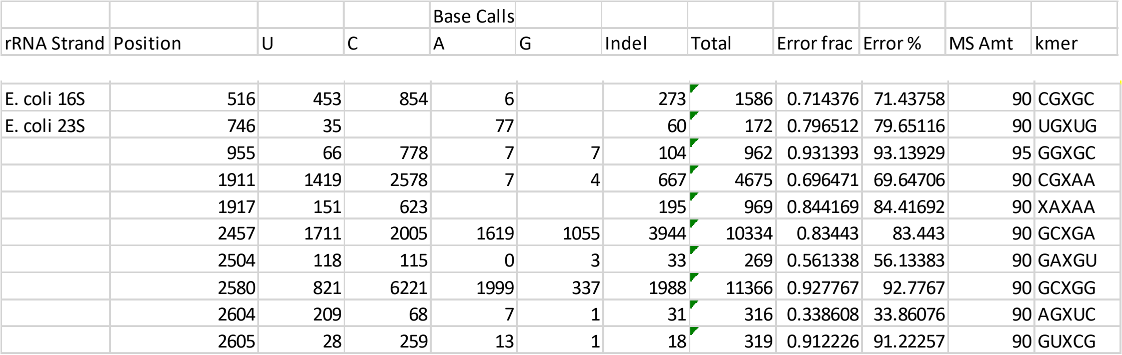
aligned_adduct_u.bam Coverage
[0 - 3381]
aligned_adduct_psi.bam Coverag
[0 - 2777]
aligned_adduct_hso3.bam Cover.
[0 - 2466]
Sequence
aligned_synthetic_psi.bam Coverage
[0 - 2400]
Sequence
aligned_ph7_hso3_pass.bam Coverage
[0 - 451]
Sequence
TGCCTGGCGGCAGTAGCGCGGTGGTCCCACCTGACCCCATGCCGAACTCAGAAGTGAAACGCCGTAGCGCCGATGGTAGTGTGGGGTCTCCTCATGCGAGAGTAGGGAACTGCCAGGCAT
aligned_ph7_hso3_pass.bam Coverage
[0 - 332]
Sequence
aligned_ph7_hso3_pass.bam Coverage
[0 - 2899]
aligned_hct_hso3_ph7_pass.bam Coverage
[0 - 219]
Sequence
aligned_hct_hso3_ph7_pass.bam Coverage
[0 - 147]
Sequence
aligned_hct_hso3_ph7_pass.bam Coverage
[0 - 2399]

The IGV analysis of the aligned sequencing data provides a graphical analysis of the coverage across the reference sequence and gives an indication of the base call errors. These examples IGV plots were constructed with the default settings in IGV. The color code is gray = sequence reads give >70% consensus with the reference base; when the sequence alignment yields a mixture of nucleotides that are <70% consensus, the mixture of bases is color-coded in which green = A, blue = C, yellow = G, red = U, and white space above = indels (for this reason, we have remade the plots in the text and later in the ESI with the indels color-coded black). The reference sequence is color-coded below the bar charts using the same color key. These plots demonstrate the sequence reads have greater coverage at the 3` end that decreases toward the 5` ends. This is an expected result that is found in all nanopore RNA sequencing data because the strand is threaded 3` to 5`. The sites where there exists the greatest mixture of base calls occur at the Ψ sites.

Figure S4. The rRNA nanopore direct RNA sequencing base call analysis data.

				Base Calls							
rRNA Strand	Position	U	C	A	G	Indel	Total	Error frac	Error %	MS Amt	kmer
Human 5.8S	55	89	81	16	14	58	258	0.655039	65.50388	60	GCXGC
	69	77	155	3		12	247	0.688259	68.82591	61	UGXGA
Human 18S	34	5					5	0	0	100	UGXCX
	36		6				6	1	100	82	XCXCA
	93	2	4				6	0.666667	66.66667	87	AAXGG
	105	2	2		2		6	0.666667	66.66667	99	AAXCA
	109	2	4				6	0.666667	66.66667	99	GUXAU
	119		4			4	8	1	100	94	CCXUU
	210	6				2	8	0.25	25	83	GAXGC
	218	2	6				8	0.75	75	100	CAXUU
	296	6			2	2	10	0.4	40	25	UCXAG
	406	6	4				10	0.4	40	87	GGXGA
	572	5				1	6	0.166667	16.66667	97	CUXUA
	609		6				6	1	100	90	UCXGG
	649	2				4	6	0.666667	66.66667	93	UAXUA
	651	1	4			1	6	0.833333	83.33333	93	UUXCX
	681	2		1	2	1	6	0.666667	66.66667	62	XCXGC
	686	2	1	1		2	6	0.666667	66.66667	95	GAXCU
	801	1	6				7	0.857143	85.71429	100	UUXAC
	814	4	3				7	0.428571	42.85714	100	AAXXAG
	815		7				7	1	100	100	AAXXAG
	822	3				7	10	0.7	70	99	UGXUC
	863	4	2	1		1	8	0.5	50	95	AAXAA
	866	3	3	2		1	9	0.666667	66.66667	88	AAXGG
	897	5	1	1		3	10	0.5	50	23	GUXUU
	918	7	1			1	9	0.222222	22.22222	42	AUXAA
	966	1	11				12	0.916667	91.66667	89	AUXCU
	1004	4	9				13	0.692308	69.23077	97	UUXGC
	1045	2	6	1		7	16	0.875	87.5	92	GGXXCG
	1046	1	7	1	1	3	13	0.923077	92.30769	100	GGXXCG
	1056	2	9			2	13	0.846154	84.61538	93	GAXCA
	1081	4				10	14	0.714286	71.42857	94	CAXAA
	1136	9	1	1		5	16	0.4375	43.75	7	AXCUC
	1174	7	7	1		5	20	0.65	65	100	UAXGG
	1177	7	7	4		2	20	0.65	65	100	GUXGC
	1232	6	11	1		1	19	0.684211	68.42105	98	CCXGC
	1238	8	6	2	1	1	18	0.555556	55.55556	97	GCXUA
	1244	11	6			2	19	0.421053	42.10526	100	UUXGA
	1347	11	11		1	2	25	0.56	56	98	GUXGG
	1367	12	5	2	2	3	24	0.5	50	98	GUXAA
	1445	8	7	1		6	22	0.636364	63.63636	90	CUXAG
	1625	10	4	2		9	25	0.6	60	79	AUXCC
	1643	15	6	1		4	26	0.423077	42.30769	96	AUXCC
	1692	19	6			1	26	0.269231	26.92308	98	UUXGU

				Base Calls							
rRNA Strand	Position	U	C	A	G	Indel	Total	Error frac	Error %	MS Amt	kmer
Human 28S	1523	1	3			2	6	0.833333	83.33333	88	ACUAU
	1569	5				1	6	0.166667	16.66667	68	UCXGG
	1664	2	1			3	6	0.666667	66.66667	97	CCXCC
	1670	2	2			1	5	0.6	60	96	GAXAG
	1731	2	2			1	5	0.6	60	100	AAXGA
	1766		4	1			5	1	100	40	CCXAU
	1768	3		1		1	5	0.4	40	100	UAXXCU
	1769	1	4			1	6	0.833333	83.33333	100	AUXCU
	1779	2	1	1		1	5	0.6	60	100	UUXAA
	1847	3	2			1	6	0.5	50	95	ACXUXUG
	1849	4	1	1		2	8	0.5	50	95	ACXUXUG
	2495	5	4			2	11	0.545455	54.54545	92	GAXCG
	2619	5	2			4	11	0.545455	54.54545	90	UUXUC
	2826		11		1	6	18	1	100	20	UGXAG
	2830	10	2				12	0.166667	16.66667	9	GGXAA
	3616	8	6		1	6	21	0.619048	61.90476	89	ACXGXUU
	3618	10	7	1		2	20	0.5	50	95	ACXGXUU
	3674	11	9			1	21	0.47619	47.61905	99	UUXCU
	3709	7	77	1		1	86	0.918605	91.86047	72	AUXCA
	3713	15	4			3	22	0.318182	31.81818	98	AAXGA
	3737	10	8	1		1	20	0.5	50	85	AGXAA
	3741	6	5		2	7	20	0.7	70	100	ACXAX
	3743	5	11			3	19	0.736842	73.68421	100	XAXGA
	3747	9	4	4		4	21	0.571429	57.14286	100	ACXCX
	3749	13				6	19	0.315789	31.57895	100	XCXCU
	3801	15	2		1	4	22	0.318182	31.81818	50	GAXGA
	3823	4	10			7	21	0.809524	80.95238	66	CCXAC
	3830	9	4	1		7	21	0.571429	57.14286	92	ACXAX
	3832	12	7	1		6	26	0.538462	53.84615	100	XAXCC
	3863	12	4	1		6	23	0.478261	47.82609	33	CUXGG
	3899		19			3	22	1	100	100	GCXUG
	3938		16	3		4	23	1	100	93	UGXAG
	4263	7	12	3		12	34	0.794118	79.41176	98	GAXCU
	4266	22	8			1	31	0.290323	29.03226	90	CUXGA
	4269	4	7	1		20	32	0.875	87.5	93	AUXUU
	4282	10	13	1		9	33	0.69697	69.69697	83	AAXAC
	4323	13	10	3		5	31	0.580645	58.06452	95	UUXUG
	4331	5	22	1		4	32	0.84375	84.375	93	UUXAA
	4373	9	22			6	37	0.756757	75.67568	96	GCXUG
	4390	9	17			10	36	0.75	75	99	GUXCA
	4393	13	13		1	12	39	0.666667	66.66667	97	CAXAG
	4401	5	21			6	32	0.84375	84.375	89	CGXCG
	4412	11	8	3		14	36	0.694444	69.44444	100	GAXCC
	4427	4	16	3		8	31	0.870968	87.09677	98	GCXCU
	4441	21	8		1	1	31	0.322581	32.25806	87	UGXGA
	4463	17	2	1	2	10	32	0.46875	46.875	17	GUXGG
	4470	16	5		4	8	33	0.515152	51.51515	100	UGXUC
	4491	21	7	2	1	5	36	0.416667	41.66667	91	CGXGA
	4502	5	22		2	3	32	0.84375	84.375	100	UUXAG
	4522	6	30	1		7	44	0.863636	86.36364	98	GUXAG
	4546	13	15	2	1	10	41	0.682927	68.29268	100	UGXUG
	4549	6	24	4	2	7	43	0.860465	86.04651	100	GUXGC
	4598	27	12	1		9	49	0.44898	44.89796	92	UUXGG
	4606	29	11	1	1	4	46	0.369565	36.95652	42	UAXGU
	4643	22	9	2		13	46	0.521739	52.17391	39	CAXCU
	4659	13	23	3	1	7	47	0.723404	72.34043	87	ACXGA
	4937	12	5	4		13	34	0.647059	64.70588	81	AGXCA
	4966	10	6	2		2	20	0.5	50	86	GGXUU
	4975	7	9	1		5	22	0.681818	68.18182	75	CGXAG

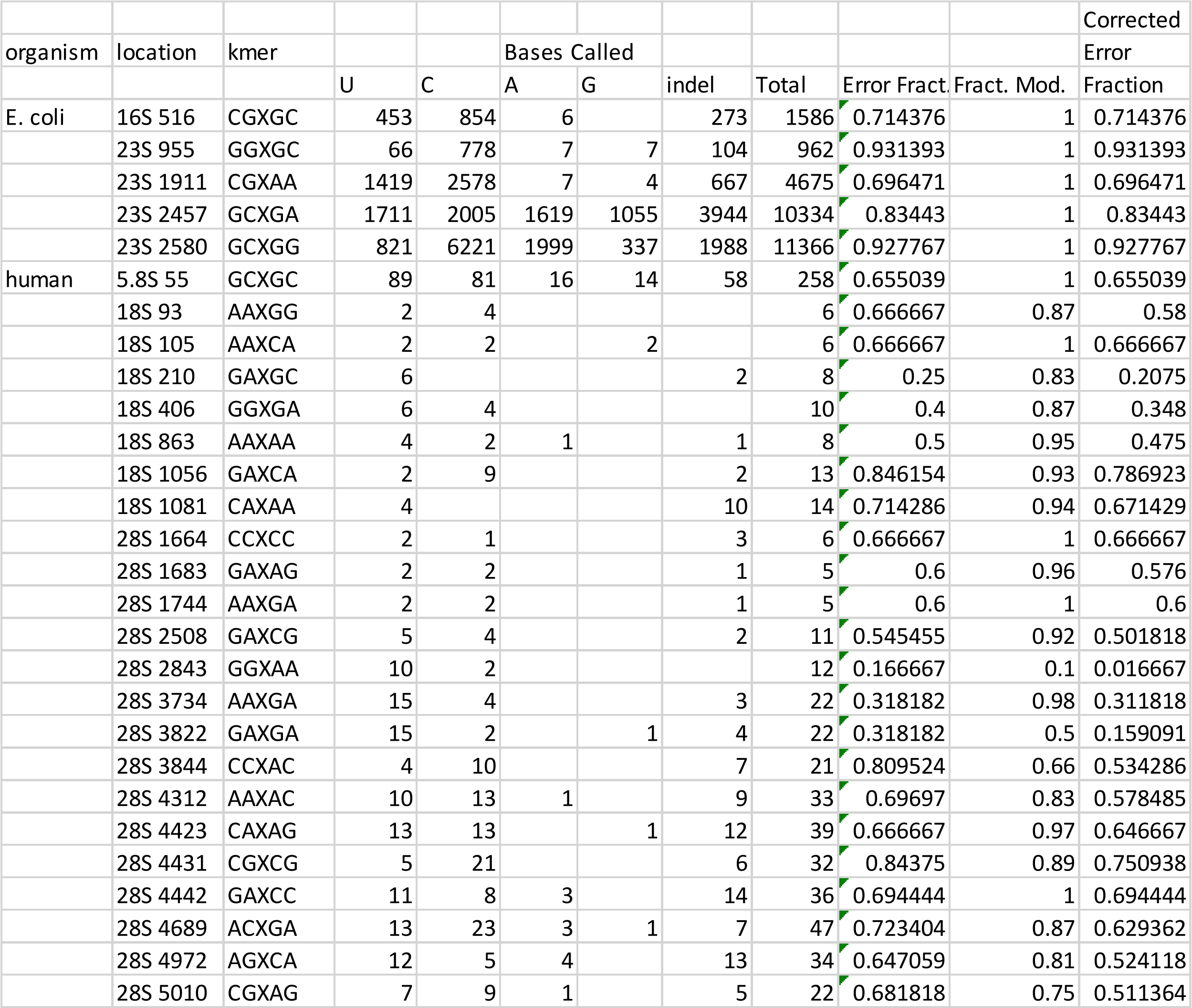


				Base Calls							
rRNA Strand	Position	U	C	A	G	Indel	Total	Error frac	Error %	MS Amt	kmer
E. coli 16S	516	453	854	6		273	1586	0.714376	71.43758	90	CGXGC
E. coli 23S	746	35		77		60	172	0.796512	79.65116	90	UGXUG
	955	66	778	7	7	104	962	0.931393	93.13929	95	GGXGC
	1911	1419	2578	7	4	667	4675	0.696471	69.64706	90	CGXAA
	1917	151	623			195	969	0.844169	84.41692	90	XAXAA
	2457	1711	2005	1619	1055	3944	10334	0.83443	83.443	90	GCXGA
	2504	118	115	0	3	33	269	0.561338	56.13383	90	GAXGU
	2580	821	6221	1999	337	1988	11366	0.927767	92.7767	90	GCXGG
	2604	209	68	7	1	31	316	0.338608	33.86076	90	AGXUC
	2605	28	259	13	1	18	319	0.912226	91.22257	90	GUXCG

The HCT116 RNA was sequenced in the smaller Flongle flow cell, which is why the read count is low. The base calling error analysis used data with a read depth of >5.

The mass spectrometry (MS) values were obtained from the literature.[4,5] The base call values for *E. coli* rRNA were previously reported by our lab.[2]

The base call data for the 28 rRNA Ψ sites in 5-nt k-mer contexts that fit the sequence 5`-VVΨVV-3` (V ≠ U).

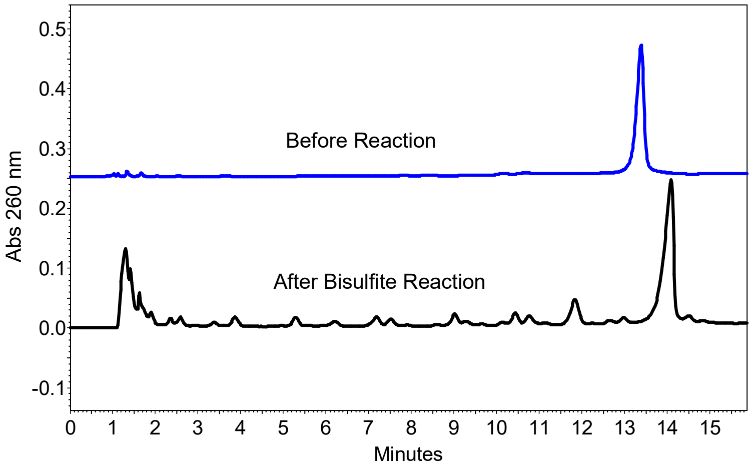


											Corrected
organism	location	kmer			Bases Called						Error
			U	C	A	G	indel	Total	Error Fract.	Fract. Mod.	Fraction
E. coli	16S 516	CGXGC	453	854	6		273	1586	0.714376	1	0.714376
	23S 955	GGXGC	66	778	7	7	104	962	0.931393	1	0.931393
	23S 1911	CGXAA	1419	2578	7	4	667	4675	0.696471	1	0.696471
	23S 2457	GCXGA	1711	2005	1619	1055	3944	10334	0.83443	1	0.83443
	23S 2580	GCXGG	821	6221	1999	337	1988	11366	0.927767	1	0.927767
human	5.8S 55	GCXGC	89	81	16	14	58	258	0.655039	1	0.655039
	18S 93	AAXGG	2	4				6	0.666667	0.87	0.58
	18S 105	AAXCA	2	2		2		6	0.666667	1	0.666667
	18S 210	GAXGC	6				2	8	0.25	0.83	0.2075
	18S 406	GGXGA	6	4				10	0.4	0.87	0.348
	18S 863	AAXAA	4	2	1		1	8	0.5	0.95	0.475
	18S 1056	GAXCA	2	9			2	13	0.846154	0.93	0.786923
	18S 1081	CAXAA	4				10	14	0.714286	0.94	0.671429
	28S 1664	CCXCC	2	1			3	6	0.666667	1	0.666667
	28S 1683	GAXAG	2	2			1	5	0.6	0.96	0.576
	28S 1744	AAXGA	2	2			1	5	0.6	1	0.6
	28S 2508	GAXCG	5	4			2	11	0.545455	0.92	0.501818
	28S 2843	GGXAA	10	2				12	0.166667	0.1	0.016667
	28S 3734	AAXGA	15	4			3	22	0.318182	0.98	0.311818
	28S 3822	GAXGA	15	2		1	4	22	0.318182	0.5	0.159091
	28S 3844	CCXAC	4	10			7	21	0.809524	0.66	0.534286
	28S 4312	AAXAC	10	13	1		9	33	0.69697	0.83	0.578485
	28S 4423	CAXAG	13	13		1	12	39	0.666667	0.97	0.646667
	28S 4431	CGXCG	5	21			6	32	0.84375	0.89	0.750938
	28S 4442	GAXCC	11	8	3		14	36	0.694444	1	0.694444
	28S 4689	ACXGA	13	23	3	1	7	47	0.723404	0.87	0.629362
	28S 4972	AGXCA	12	5	4		13	34	0.647059	0.81	0.524118
	28S 5010	CGXAG	7	9	1		5	22	0.681818	0.75	0.511364

The base call data for the sequence matched synthetic 5-nt k-mer contexts that fit the sequence 5`-VVΨVV-3` (V ≠ U). The base call errors for synthetic RNA were previously reported by our lab.[3]

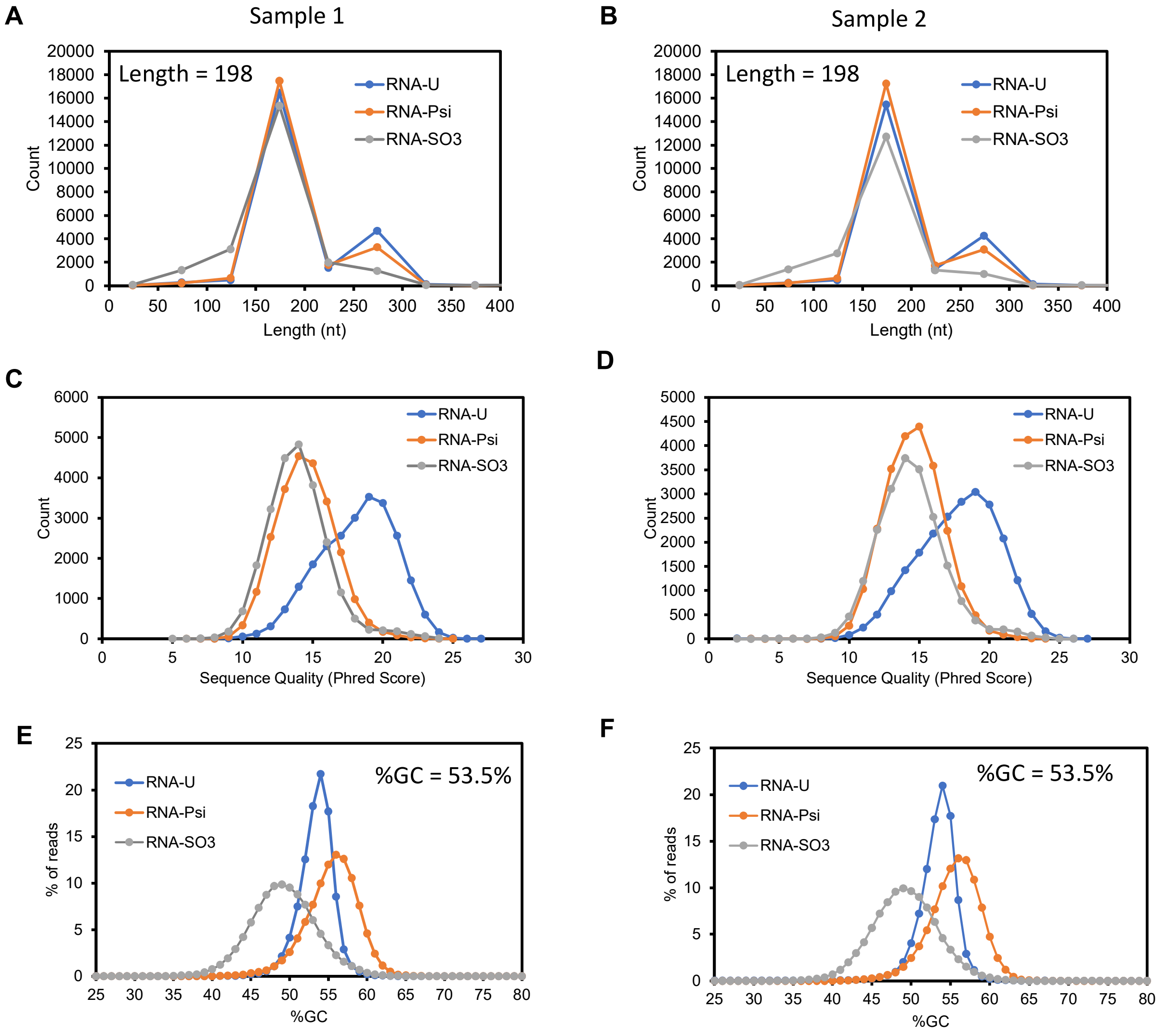
Synthetic RNA		kmer							
			U	C	A	G	indel	Full Total	Error Fract.
		CGXGC	7	15			4	26	0.730769
		GGXGC	3	21	2		7	33	0.909091
		CGXAA	45	21	2		125	193	0.766839
		GCXGA	5	15	1		4	25	0.8
		GCXGG	3	30	9		1	43	0.930233
		GCXGC	1874	3428	87	7	548	5944	0.684724
		AAXGG	31	11	2		22	66	0.530303
		AAXCA	7	19	1	4	6	37	0.810811
		GAXGC	16	4			4	24	0.333333
		GGXGA	671	2777		5	175	3628	0.81505
		AAXAA	4	10			1	15	0.733333
		GAXCA	20	2		1	37	60	0.666667
		CAXAA	40	7	1	1	129	178	0.775281
		CCXCC	5	27	2		31	65	0.923077
		GAXAG	2	7		1	2	12	0.833333
		AAXGA	2	18			7	27	0.925926
		GAXCG	18	27			3	48	0.625
		GGXAA	16	61	2	1	12	92	0.826087
		AAXGA	12	18			7	37	0.675676
		GAXGA	31	6			4	41	0.243902
		CCXAC	18	21	1	1	33	74	0.756757
		AAXAC	9	8	6		23	46	0.804348
		CAXAG	798	4400	82	14	970	6264	0.872605
		CGXCG	468	4734	141	12	666	6021	0.922272
		GAXCC	71	138	16		61	286	0.751748
		ACXGA	8	20			8	36	0.777778
		AGXCA	976	1723	242	1277	1864	6082	0.839526
		CGXAG	1182	4142	91	22	746	6183	0.808831

Figure S5. Characterization of the bisulfite adduct to Ψ-containing RNA.



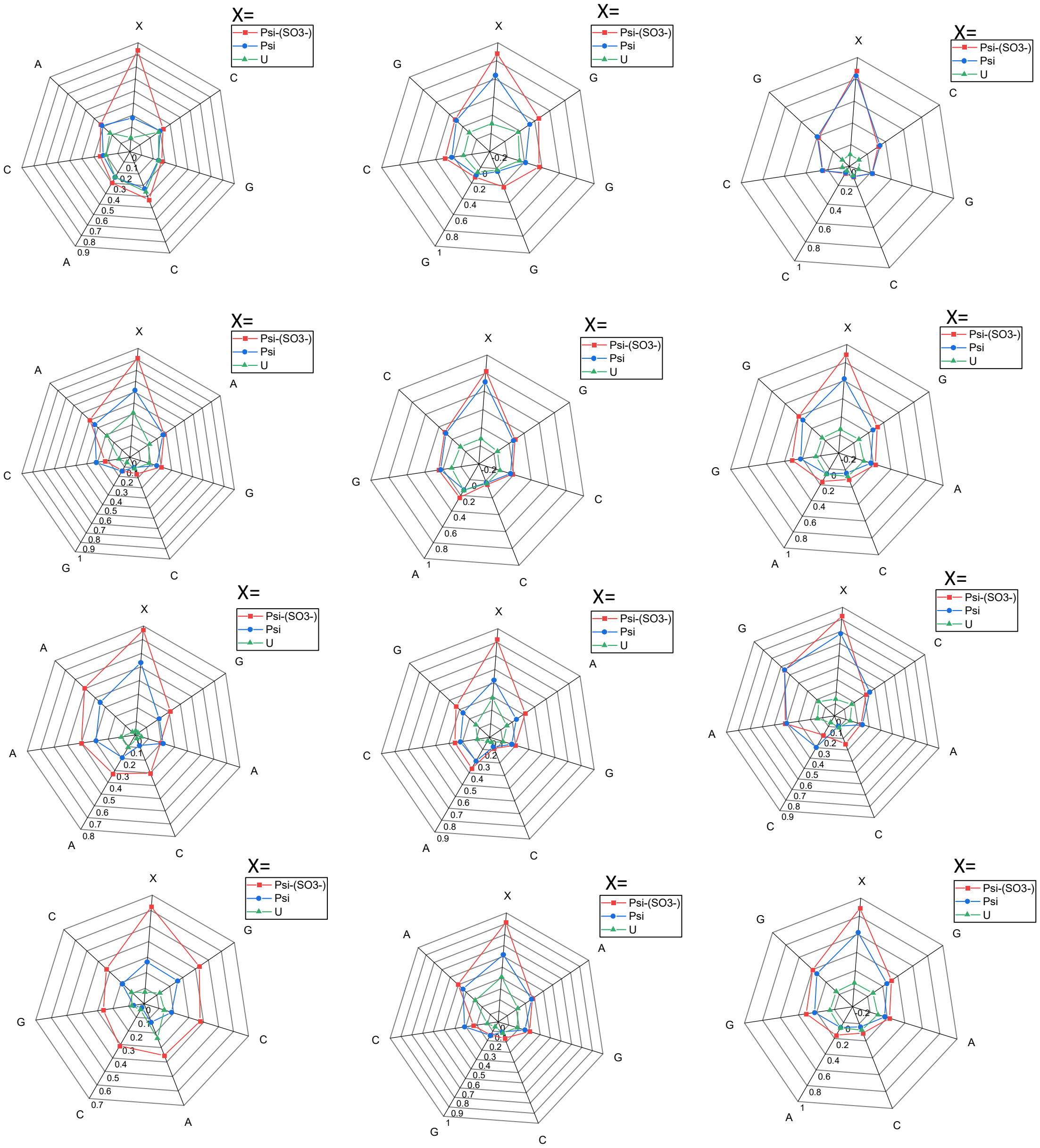
The sequence 5′-UAUU**Ψ**UAAGGUGGAAGUUAGAGGt-3′ was synthesized by established solid-phase synthesis methods using a thymidine-charged column (t) to enhance the synthetic yields; hence, the t nucleotide on the 3′ end. The RNA strand was studied for verification of the bisulfite adduct forming in an RNA strand. The RNA was exposed to $NaHSO_3$ (3 M) at pH 7 and 65 °C for 4 h. The reacting salts were removed using a Nap-25 column (GE Health Sciences) using the manufacturer's protocol. The collected sample was then incubated at pH 8.5 in Tris buffer at 37 °C for 1h. The RNA strand was analyzed by anion-exchange HPLC before (blue trace) and after (black trace). The HPLC method was running a DNAPac PA-100 column with lines A = 1:9 MeCN:ddH$_2$O and B = 1.5 M NaOAc (pH 7) in 1:9 MeCN:ddH$_2$O. The method was initiated at 15% B followed by a linear gradient to 100% B with a flow rate of 1 mL/min while monitoring the elution via the absorbance at 260 nm.

Figure S6. The FastQC analysis of the reads.



Example Fastqc analysis for strand 1 replicates (Sample 1 and Sample 2) with either a U, Ψ (Psi), or Ψ-(SO_3^-) adduct (SO3). The FASTQC analysis allows inspection of the sequencing reads before and after the reaction to look for changes. In plots A and B, read length histograms are provided. The distributions for U (blue), Ψ (orange), and Ψ-(SO_3^-) adduct (gray) are the same. This leads to the conclusion that Ψ and the Ψ-(SO_3^-) adduct go through the pore. In plots C and D are the histograms for the sequencing quality. These plots demonstrate the sequencing quality is best for the U-containing RNA and decreases for the Ψ and Ψ-(SO_3^-) adduct RNAs. Plots E and F provide the percentage of reads vs. the percent GC content. These find the U-containing RNA give a distribution centered around the predicted value provided in the upper right-hand corner of the plots. The Ψ-containing RNA give a slightly higher %GC content because Ψ is miscalled as a C; thus, increasing the average. The Ψ-(SO_3^-) adduct gave a slightly lower %GC that likely results from the increased indel frequency at these sites and adjacent sites.

Figure S7. The ESB radar plots for U, Ψ, and the Ψ-(SO_3^-) adducts in the sequences studied.

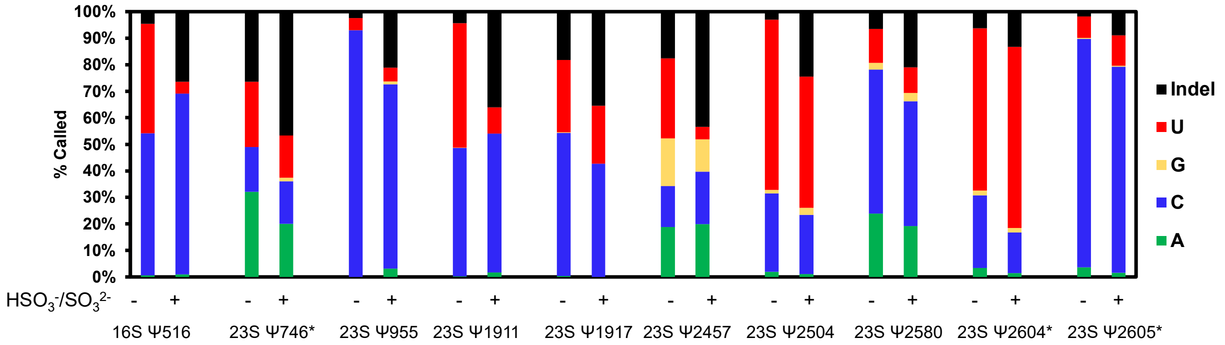


The radar plots illustrate that the ESB values for Ψ (blue line) are greater than the parent base U (green line) in 12 different sequence contexts; however, the ESB values for Ψ are > 0.3, which in some contexts is near the error of U. Further, they show the ESB values for the Ψ-(SO_3^-) adduct (red) line are the greatest and always >0.8. This observation suggests the adduct will always have greater base calling error when they pass through the nanopore and are interpreted by the base caller.

Figure S8. The ELIGOS2 computed *P*-values for the titration of U with Ψ or Ψ-(SO_3^-).

k-mer	100% Psi	100% psi-(so3-)		50% Psi	50% Psi-(SO3-)		33% Psi	33% Psi-(SO3-)		20% Psi	20% Psi-(SO3-)		10% Psi	10% Psi-(SO3-)
GCXGC	256.2	279.7		61.1	192.4		23.4	146.6		5.9	132.7		6.3	11.4
CCXXA	257.5	248.4		83.7	129.6		33.5	146.6		9.0	32.7		5.4	10.3
CXXAX	201.4	270.0		172.6	267.3		66.9	299.2		18.0	70.7		9.7	19.9
XAXCC	222.3	244.4		71.3	111.7		26.0	143.5		7.7	34.8		0.0	9.9
AAXGA	203.2	251.0		132.3	99.5		57.7	159.6		16.2	48.7		8.5	9.5
AAXXA	240.2	236.7		302.8	267.3		118.9	146.6		34.0	231.5		17.9	20.2
AXXAG	244.1	275.6		294.9	296.7		135.1	277.6		43.1	77.7		28.1	28.5
CGXCC	206.1	207.2		87.2	187.0		31.6	134.7		6.9	241.4		7.0	12.8
GGXGA	279.1	271.2		60.8	128.7		20.3	144.6		6.7	148.7		0.0	9.3
AGXCA	233.6	258.1		132.1	179.7		55.7	179.7		18.5	127.3		10.6	18.8
GGXGA	207.2	255.2		65.0	279.7		32.6	180.4		15.0	49.4		11.4	18.1
AGXCA	102.6	257.5		124.0	183.4		100.0	179.5		23.5	65.3		23.7	25.1
GCXGC	107.5	201.4		119.0	244.4		93.4	190.0		37.2	102.4		15.0	44.0
CGXAG	49.1	219.3		199.1	200.0		50.2	99.2		17.7	44.2		6.1	18.0
CAXAG	235.6	203.2		206.2	159.7		73.7	73.7		15.2	71.7		18.4	28.3
CGXCG	279.1	240.2		197.7	199.7		60.0	140.0		17.7	256.2		18.7	27.6
CAXCG	253.1	231.1		123.4	176.2		29.4	74.4		22.5	49.3		24.6	16.6

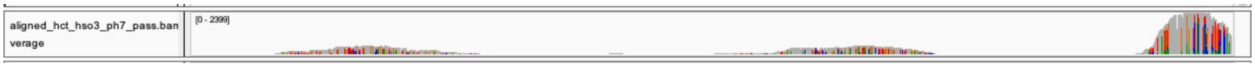
Figure S9. Additional data and discussion regarding the bisulfite reaction on rRNA.



The bar chart above provides the base calling profile for the 10 E. coli rRNA Ψ sites before and after the pH 7 bisulfite reaction. These data provide additional examples of the indel frequency increasing as a result of the bisulfite adduct. *Positions 23S Ψ746, 23S Ψ2604, and 23S Ψ2605 are Ψ sites in the rRNA where other modifications reside, which can impact the signals. For 23S Ψ746 there is a m^1G at 745 and m^5U at 747, and Ψ2604/ Ψ2605 are adjacent to one another.

The section below outlines an attempt to use the pH 7 bisulfite reaction to sequence for Ψ in human rRNA. As described below this did not work in our hands during two attempts. Additional optimizations in the future could get this experiment to work. We did not pursue this further in the present studies.

Example IGV plot for HCT116 28S after the pH 7 bisulfite reaction.

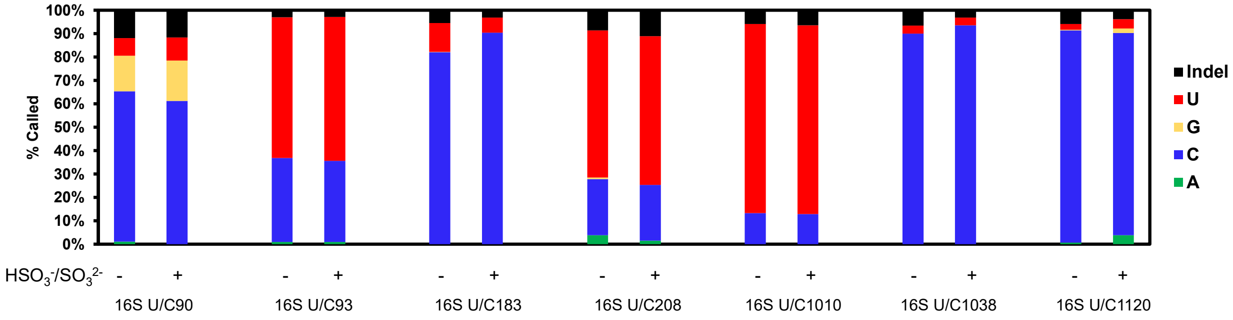


The poor alignment after the bisulfite reaction can come from many sources (see IGV image above). The first is it is well established that the bisulfite reaction causes low-yielding degradation of DNA,[6] and the less stable RNA polymer, likely degrades with higher yields. The degradation has been characterized as strand breaks and abasic site formation; whether this occurs in RNA is not known, and we did not evaluate this chemistry. While conducting the studies to understand the structures of the bisulfite ring-opened sugar adducts to Ψ,[7] we conducted test reactions on C, m^5C, hm^5C, and U to determine whether these other pyrimidines could form sugar adducts. On the basis of HPLC analysis identical to what was conducted with Ψ, when these pyrimidines were treated with bisulfite, low levels (<1%) of sugar adducts were formed; the yield for each of these nucleosides was so low that characterizing these adducts was not successful and this is why we have not reported on this chemistry. In long RNA, low-level reactions can become problematic for sequencing that may be leading to the challenges observed.

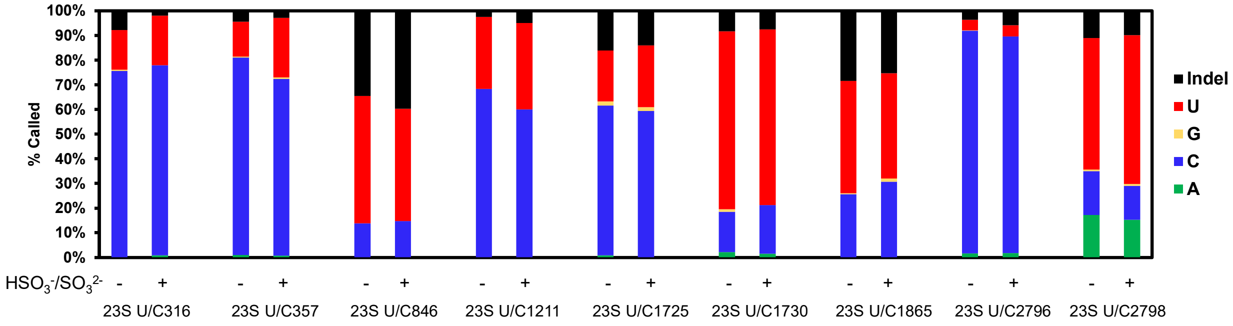
The key difference in the present work compared to other publications using the bisulfite reaction for analysis of Ψ or m^5C in RNA is that we directly sequenced the RNA; in contrast, all other reports convert the RNA to a cDNA via reverse transcription followed by exponential PCR amplification.[8-10] In the present approach, all side reactions on the RNA from the bisulfite treatment will impact the sequencing. In any approach that utilizes reverse transcription and PCR, the polymerases will sanitize the reactions of the side reaction products because they will either not PCR amplify and are lost, or will be bypassed and either remain silent or yield a signature that is omitted from the downstream analysis because it is present in such low levels.

Figure S10. Base call data at known U/C sequence variations in the *E. coli* rRNA strands.

16S rRNA U/C Sequence Variations



23S rRNA U/C Sequence Variations



The 23S rRNA from E. coli has additional U/C sequence variation for which data are not provided at position s542, 1178, and 1229. The bisulfite-treated rRNA when sequenced failed to be read at sufficient depth at these positions to make reliable base call analyses.

Figure S11. Studies on Ψ-(SO_3^-) adducts impacting the raw nanopore data.

We have a history of studying chemical modifications and adducts to DNA with single nanopore systems in which differences in the current vs. time traces were inspected.[11-13] We thought that it would be interesting to follow the Ψ-(SO_3^-) adduct passing through the ONT system; furthermore, this information may help understand why the alignment of these adducted RNA strands was lower in yield than the unreacted RNA (Fig. S2, ESI†). A 200-nt long RNA was designed and studied with two Ψ sites separated by 99 nts in different sequence contexts (5`-GAXCA and 5`-CGXGC; Fig. S1, ESI†) for study of the current vs. time data for these adducts passing through the helicase-nanopore system. Inspection of the raw nanopore data from the ONT system first requires resquiggling to be conducted, which is the process of appending the base calls to the current levels from which they were derived. Two programs are routinely used for resquiggling, Tombo and Nanopolish.[14,15] This strand was sequenced with the ONT system with U, Ψ, or the Ψ-(SO_3^-) adduct. Using Tombo or Nanopolish we found the percentage of reads successfully resquiggled decreased from ~70% for the U-containing RNA, to ~40% for the Ψ-containing RNA, and finally was ~30% for the Ψ-(SO_3^-)-containing RNA (Fig. S11.1A).

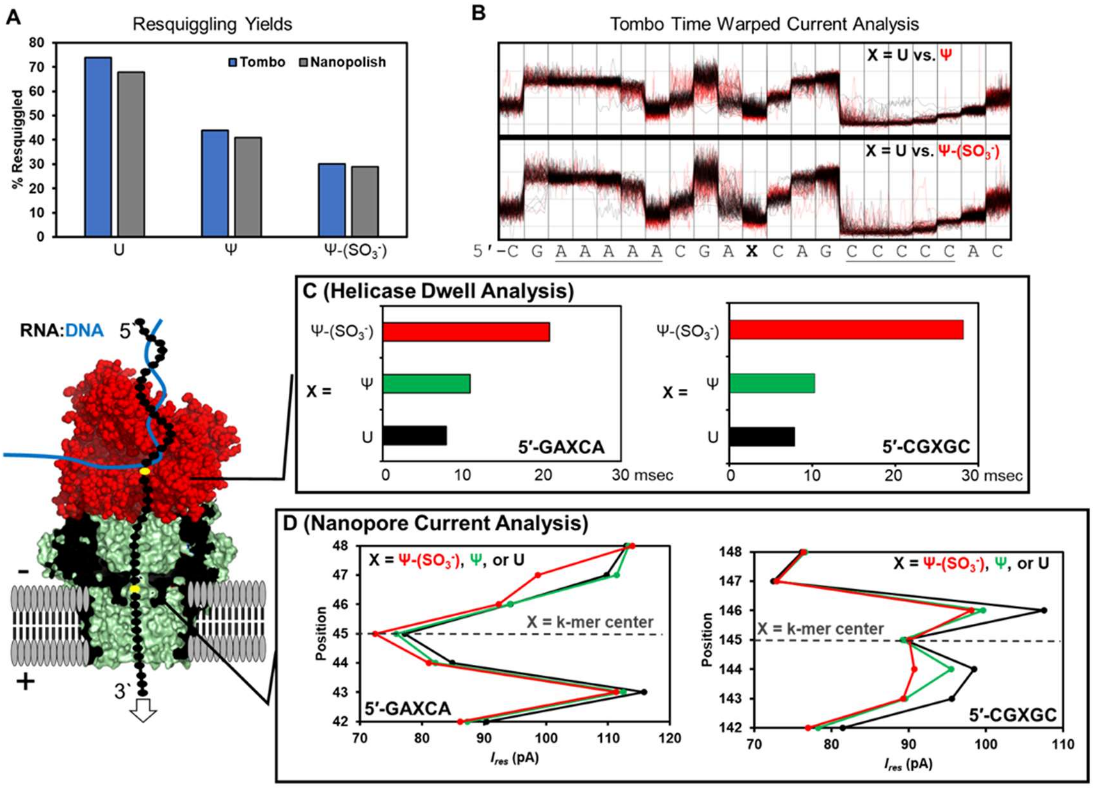
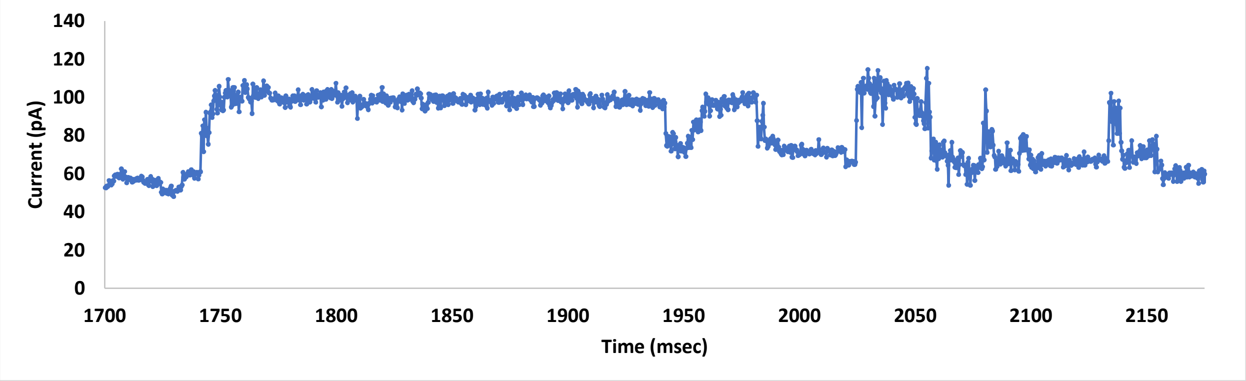


Figure S11.1. Inspection of the ionic current levels and dwell times for U, Ψ, and Ψ-(SO_3^-) adducts as they pass through the dual helicase-nanopore sensors. (A) Percentage of U, Ψ, and Ψ-(SO_3^-) adduct reads resquiggled by Tombo or Nanopolish. (B) Time-warped ionic current levels for the RNA modifications using Tombo. Plots of the (C) dwell times and (D) ionic-current levels as the sites of interest pass through the helicase or nanopore, respectively. The values were obtained from Nanopolish. (E) Example current vs. time traces for U, Ψ, and the Ψ-(SO_3^-) adduct from the ONT nanopore sequencer.

Using the available data, time-warped plots (i.e., the dwell time is scaled to the same value for each event) of the raw data were constructed in Tombo to compare U vs. Ψ (Fig. S11.1B top panel) and U vs. Ψ-(SO_3^-) (Fig. S11.1B bottom panel). The plots for the reads that made it through Tombo did not provide any additional clarity on the passage of the Ψ-(SO_3^-) adduct through the nanopore most likely because they were filtered by the software. Using Nanopolish, the currents and dwell times can be extracted, plotted, and analyzed. Comparison of the helicase dwell times for U, Ψ, and Ψ-(SO_3^-) in the two sequence contexts identified the average dwell time was shortest for U, intermediate for Ψ, and longest for the Ψ-(SO_3^-) adducts (U ~7 msec, Ψ ~10 msec; and the Ψ-(SO_3^-) adduct ~20-30 msec; Fig. S11.1C). As for the residual current levels (I_{res}) in the nanopore protein for U, Ψ, and Ψ-(SO_3^-) adduct in the two sequence contexts, the location of maximal difference between the three nucleotides was sequence dependent (Fig. 11.1D). For the site at position 45 in the sequence 5′-GAXCA, the maximal I_{res} difference occurs when the adduct is in the center of the k-mer and at the 5′ edge in the vestibule toward the helicase (i.e., positions 45 and 47 in Fig. S11.1D). For the site at position 145 in the sequence context 5′-CGXGC, the greatest difference in I_{res} occurred at all positions in the k-mer except when the modifications were in the center (Fig. S11.1D). When a difference in I_{res} was observed between U and Ψ or the Ψ-(SO_3^-) adduct, the modifications were more blocking (i.e., lower I_{res} value) than the parent and the bisulfite adduct blocked the current more than Ψ. These data for sub-populations of the Ψ and Ψ-(SO_3^-) adduct reads (Fig. S11.1A) point to raw data differences that influence the base calling algorithm resulting in the base calling errors observed; additionally, the Ψ-(SO_3^-) adduct is more disruptive to the raw data resulting in greater base calling error compared to Ψ (Fig. S3).

We were not satisfied with only inspecting a subset of the reads with these computational tools; therefore, a small randomly selected population of the reads was extracted from the fast5 data files and inspected manually. The sequences were designed such that a 5-nt poly-A track was on the 5′ side of the inspection site and a 5-nt poly-C track was on the 3′ side to allow finding the position of the U, Ψ, or Ψ-(SO_3^-) adduct visually in the raw data (Fig. S11.1B). This approach of looking at the data turned out to be very challenging because of the large deviation in dwell times from one sample to the next, and the current level differences from one nucleotide to the next were not easily differentiable in some cases; thus, our confidence in the quantification of these data is low. Nevertheless, we learned that the Ψ-(SO_3^-) in many of the events inspected produced very noisy signals that likely challenged Guppy to base call the data, and Tombo and Nanopolish for resquiggling the data (Fig. S11.2). The key point is the data are recorded for highly distorted sites such as the Ψ-(SO_3^-) adduct but the available computational tools impose limitations on studying these events in greater detail.

Example U-containing RNA *i-t* trace



Example Ψ-(SO_3^-)-containing RNA *i-t* trace.

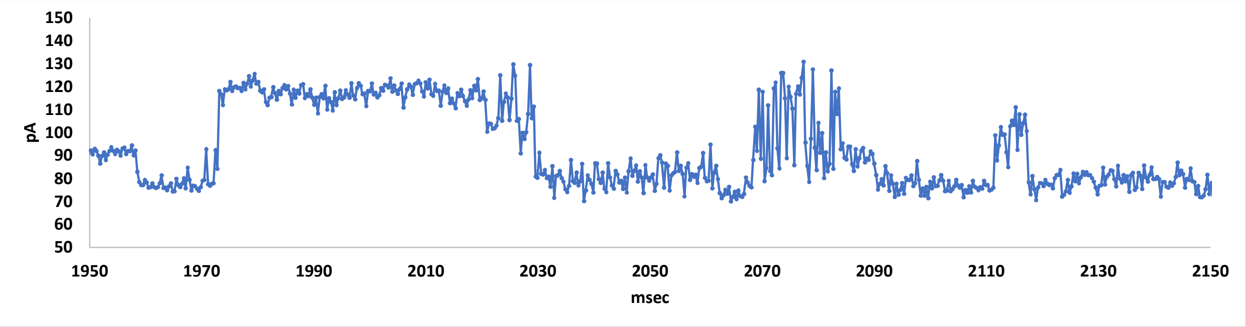


Figure S11.2. Example i-t traces for an RNA with a U (top) or Ψ-(SO_3^-) adduct. The noisy portion of the adduct read starts at 2070 msec.

There are more examples in the data deposited in the public repository. This approach to understand the behavior of an adduct passing through the nanopore turned out to be very difficult to analyze because of the stochastic nature of the data, which is not apparent in plots made from Tombo. The goal was to learn about current levels and dwell times, but this was not achievable to our satisfaction; however, these data confirm the FastQC results that the data are recorded but the downstream programs fail to process the data.

References

(1) Fleming, A. M.; Mathewson, N. J.; Howpay Manage, S. A.; Burrows, C. J. Nanopore dwell time analysis permits sequencing and conformational assignment of pseudouridine in SARS-CoV-2. *ACS Cent. Sci.* **2021**, *7*, 1707-1717.

(2) Fleming, A. M.; Bommisetti, P.; Xiao, S.; Bandarian, V.; Burrows, C. J. Nanopore sequencing for the 17 modification types in 36 locations in *E. coli* ribosomal RNA enables monitoring of stress-dependent changes. *ACS Chem. Biol.* **2023**, doi:10.1021/acschembio.1023c00166.

(3) Fleming, A. M.; Burrows, C. J. Nanopore sequencing for N1-methylpseudouridine in RNA reveals sequence-dependent discrimination of the modified nucleotide triphosphate during transcription. *Nucleic Acids Res.* **2023**, *51*, 1914-1926.

(4) Taoka, M.; Nobe, Y.; Yamaki, Y.; Sato, K.; Ishikawa, H.; Izumikawa, K.; Yamauchi, Y.; Hirota, K.; Nakayama, H.; Takahashi, N.; Isobe, T. Landscape of the complete RNA chemical modifications in the human 80S ribosome. *Nucleic Acids Res.* **2018**, *46*, 9289-9298.

(5) Popova, A. M.; Williamson, J. R. Quantitative analysis of rRNA modifications using stable isotope labeling and mass spectrometry. *J. Am. Chem. Soc.* **2014**, *136*, 2058-2069.

(6) Booth, M. J.; Raiber, E.-A.; Balasubramanian, S. Chemical methods for decoding cytosine modifications in DNA. *Chem. Rev.* **2015**, *115*, 2240-2254.

(7) Fleming, A. M.; Alenko, A.; Kitt, J. P.; Orendt, A. M.; Flynn, P. F.; Harris, J. M.; Burrows, C. J. Structural elucidation of bisulfite adducts to pseudouridine that result in deletion signatures during reverse transcription of RNA. *J. Am. Chem. Soc.* **2019**, *141*, 16450-16460.

(8) Dai, Q.; Zhang, L.-S.; Sun, H.-L.; Pajdzik, K.; Yang, L.; Ye, C.; Ju, C.-W.; Liu, S.; Wang, Y.; Zheng, Z.; Zhang, L.; Harada, B. T.; Dou, X.; Irkliyenko, I.; Feng, X.; Zhang, W.; Pan, T.; He, C. Quantitative sequencing using BID-seq uncovers abundant pseudouridines in mammalian mRNA at base resolution. *Nat. Biotechnol.* **2023**, *41*, 344-354.

(9) Zhang, M.; Jiang, Z.; Ma, Y.; Liu, W.; Zhuang, Y.; Lu, B.; Li, K.; Peng, J.; Yi, C. Quantitative profiling of pseudouridylation landscape in the human transcriptome. *Nat. Chem. Biol.* **2023**, doi:10.1038/s41589-41023-01304-41587.

(10) Edelheit, S.; Schwartz, S.; Mumbach, M. R.; Wurtzel, O.; Sorek, R. Transcriptome-wide mapping of 5-methylcytidine RNA modifications in bacteria, archaea, and yeast reveals m5C within archaeal mRNAs. *PLoS Genet.* **2013**, *9*, e1003602.

(11) An, N.; Fleming, A. M.; White, H. S.; Burrows, C. J. Nanopore detection of 8-oxoguanine in the human telomere repeat sequence. *ACS Nano* **2015**, *9*, 4296-4307.

(12) Johnson, R. P.; Fleming, A. M.; Perera, R. T.; Burrows, C. J.; White, H. S. Dynamics of a DNA mismatch site held in confinement discriminate epigenetic modifications of cytosine. *J. Am. Chem. Soc.* **2017**, *139*, 2750-2756.

(13) Schibel, A. E.; An, N.; Jin, Q.; Fleming, A. M.; Burrows, C. J.; White, H. S. Nanopore detection of 8-oxo-7,8-dihydro-2'-deoxyguanosine in immobilized single-stranded DNA via adduct formation to the DNA damage site. *J. Am. Chem. Soc.* **2010**, *132*, 17992-17995.

(14) Stoiber, M.; Quick, J.; Egan, R.; Eun Lee, J.; Celniker, S.; Neely, R. K.; Loman, N.; Pennacchio, L. A.; Brown, J. De novo identification of DNA modifications enabled by genome-guided nanopore signal processing. *bioRxiv* **2017**, 094672.

(15) Loman, N. J.; Quick, J.; Simpson, J. T. A complete bacterial genome assembled de novo using only nanopore sequencing data. *Nat. Meth.* **2015**, *12*, 733-735.