Supplementary Figures

Supplementary Figure 1. Example of a multi-modification (m6A+I) NanoSpeech FASTA file output. Header lines containing positional information for multi-modification basecalling are highlighted in red. If detected, modification indices are stored as lists according to the following pattern: [alternative_modified_base_symbol]_[canonical_base_symbol]=INDEX1, INDEX2, …, INDEXn (e.g., M_A for m6A and I_A for Inosine in this example). Modified bases (in blue) are replaced by canonical versions to facilitate downstream mapping procedures.
A NanoSpeech accessory script (see Data Availability) enables the indexing of header lines, allowing random access to modified bases. Additionally, it retrieves per-read mapping coordinates for each modification, aggregating modification ratios in genome space.

Supplementary Figure 2. Analysis of NanoSpeech basecalled reads from synthetic samples. (A) Statistics of raw sequences for pure*G* (blue), pure*I* (orange), and CC (green) IVT samples showing: len – the total length of the basecalled reads; Acount – the number of basecalled As in each read, including inosines; Acount_noI – the same count excluding Is; Icount_norm_len – the number of identified inosines in each read, normalized by its length; Iper_on_As – the percentage of Is relative to the total number of As. (B) An extended scatterplot version displaying the inosine content (normalized on length) for the same samples. (C) The inosine content of IVTs from A549 cells and two replicates of HeLa cell lines (blue, orange, and green, respectively). Reads were basecalled using the NanoSpeech model trained on multiple species for enhanced generalization capacity.

Supplementary Figure 3. Alignment screenshot of NanoSpeech and Guppy basecalled reads from pure*G* and pure*I* samples based on a subset of 100 IVT reads. Alignment profiles were strongly improved in modified synthetic sequences.

Supplementary Figure 4. The percentage of NanoSpeech (blue) and Guppy (red) aligned reads from different samples including synthetic (pure*G*, pure*I*, CC, coli-IVT-mod and coli-IVT-unmod) and real (*E. coli*, *S. cerevisiae*, HEK293T - hKO, hWT, and hOE) data.

Supplementary Figure 5. Alignment screenshot of NanoSpeech and Guppy basecalled reads from the highly modified *E. coli* IVT sample. A random and independent subset of reads (N=1000) were mapped using minimap2 (with k=14) by NanoSpeech and Guppy. The mapping of Guppy reads did not produce alignments. On the contrary, NanoSpeech reads showed low noise and high mappability.

Supplementary Figure 6. Analysis of NanoSpeech basecalled reads from synthetic coli-IVT samples. (A) Statistics of raw sequences for modified (blue) and unmodified (orange) coli-IVT samples, showing: len – the total length of the basecalled reads; Acount – the number of basecalled As in each read, including inosines; Acount_noI – the same count excluding Is; Icount_norm_len – the number of identified inosines in each read, normalized by its length; Iper_on_As – the percentage of Is relative to the total number of As. (B) An extended scatterplot version displaying the inosine content (normalized on length) for the same samples. Reads were basecalled using the NanoSpeech model trained on multiple species for enhanced generalization capacity.

Supplementary Figure 7. Per-site detection of m6A and I in reads from IVT synthetic constructs basecalled by NanoSpeech implementing the m6A_I model for the concomitant identification of m6A and I. The model was generated by NanoListener and included reads produced using the R9.4.1 flowcells from (A) modified and (B) unmodified CC (with or without m6As) or (C) pure*I* and (D) pure*G*

IVTs (with or without inosines). Reads were basecalled by NanoSpeech using the multi-modification model with the following extended vocabulary {A, C, G, U, I, m6A}. The modification frequencies Ifreq (inosine) and Mfreq (m6A) per site were computed by aggregating aligned reads. Ifreq is in blue, Mfreq is in red.

Supplementary Figure 8. Analysis of current intensity profiles of re-squiggled events by f5c eventalign on DRACH motifs in CC IVT molecules. Unmodified CC IVT reads were compared, position-wise, with their corresponding versions containing either m1A or m6A modifications to identify differences in their raw signal signatures.

Supplementary Figure 9. Statistical analysis of the current distributions depicted in Supplementary Figure 8. Re-squiggled current events were extracted from unmodified, m1A, and m6A-containing CC IVTs, and their distributions on representative DRACH contexts were summarized as mean, standard deviation, and sample size. Mann-Whitney U tests were used to compare the three groups pairwise, and p-values are reported in the last three columns.

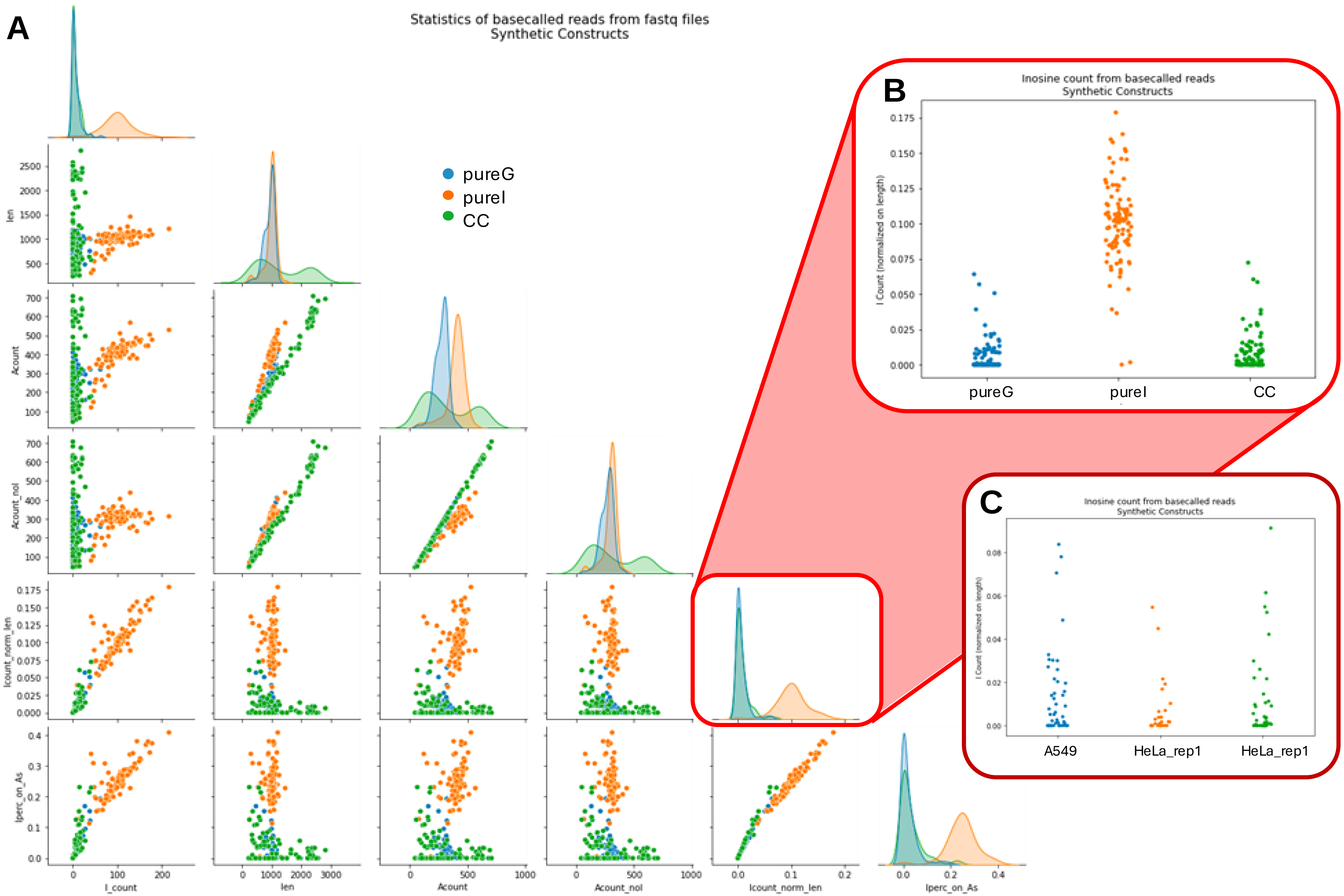
Supplementary Figure 10. Alignment screenshot of NanoSpeech (red) and Guppy (blue) basecalled reads from a subset of unmodified and modified (with pU) sequences of the CC IVT samples. Alignment profiles were strongly improved in modified sequences.

Supplementary Figure 11. Correlation between the length of electric signal chunks and the corresponding output kmers. The “Simulated Randomers Strategy,” implemented by NanoListener, was tested using different combinations of parameters to define the expected ranges of raw signal chunk lengths to extract and annotate with nucleotide sequences. In (A), for the general multi-species dataset, the correlation between extracted chunks and output kmers is shown, with a Spearman’s correlation coefficient of $\rho = 0.794$, $p < 10^{-15}$ (N = 50,000). In (B), for the R9 unmodified curlcakes dataset, two sets of chunks with variable lengths (1) from 900 to 1500 and (2) from 1300 to 1900 current measurements, were produced. When merged, these two datasets showed a positive Spearman’s correlation between electric signals and output kmers, with $\rho = 0.374$, $p < 10^{-15}$ (N = 100,000 chunks). In (C), the same analysis was performed on NanoListener datasets of chunk/output-kmer pairs generated using data from unmodified curlcakes sequenced with the RNA004 kit. In this case, the evaluated chunk length ranges were (1) 800–1300, (2) 1100–1700, (3) 1500–2100, (4) 1900–2500, and (5) 2500–3200 samples. The correlation was even stronger in this global dataset, with $\rho = 0.841$, $p < 10^{-15}$ (N = 250,000 chunks). In (D) and (E), the learning curves for training on CCs and pure*I*/pure*G* datasets using shorter and longer chunk strategies, respectively, are shown. In (F), the learning curves for five parallel training sessions on an RNA004 CCs mini-dataset using different NanoListener configurations are displayed. In (G), the validation losses for the same dataset are shown.

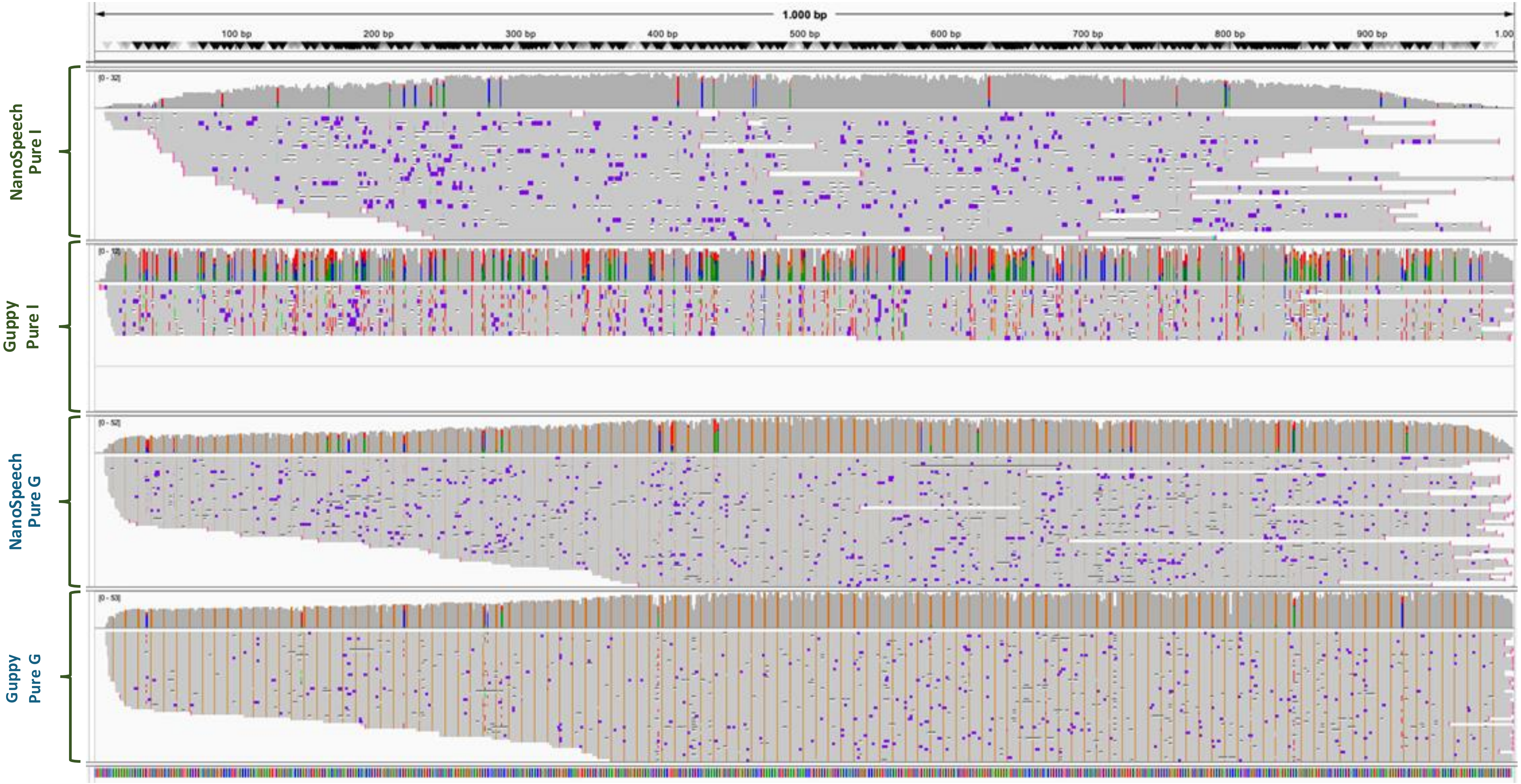
Supplementary Figure 1

```
>77128b6d-c6da-486c-a44d-8fa5cab03597 I_A=2452,2461,2470,2479,2489,2498,2504,2513,2522
M_A=0,12,15,18,22,28,30,35,36,42,44,45,46,51,54,59,63,64,70,74,87,88,92,93,102,110,114,124,130,132,133,142,171,172,181,184,192,199,204,205,206,207,210,218,223,240,241,245,257,26
4,266,267,269,274,280,282,284,285,287,293,298,299,301,303,308,314,316,318,319,321,327,328,334,338,347,348,353,357,358,359,368 ,369,374,376,378,379,389,398,399,405,414,419,420,422,4
30,435,439,442,443,446,447,453,454,455,456,459,461,472,477,478,488,489,492,493,494,498,501,505,509,517,519,520,522,526,528,53 7,539,545,551,552,553,555,556,557,559,562,566,567,568,
575,577,585,587,589,651,653,656,659,661,662,663,664,665,666,679,685,692,696,698,710,719,724,727,739,741,752,756,762,768,769,7 76,784,793,806,810,812,816,821,829,834,835,836,837,840
,841,842,843,844,846,847,851,854,857,864,871,877,881,887,893,897,901,913,914,916,925,933,936,939,944,949,951,952,954,961,965, 971,983,984,985,986,989,993,995,997,1003,1005,1006,101
1,1043,1045,1048,1051,1054,1055,1057,1058,1059,1075,1076,1081,1082,1084,1085,1102,1106,1111,1118,1122,1126,1136,1137,1139,114 0,1151,1154,1158,1168,1170,1175,1182,1183,1184,11
87,1197,1201,1207,1209,1210,1215,1216,1217,1219,1226,1230,1233,1236,1238,1240,1241,1245,1246,1250,1276,1289,1309,1313,1319,13 24,1328,1329,1334,1339,1342,1345,1349,1353,1355,1
366,1377,1382,1385,1390,1394,1407,1413,1414,1415,1421,1427,1428,1431,1432,1436,1449,1450,1454,1456,1465,1466,1469,1475,1480,1 481,1483,1485,1486,1490,1495,1497,1499,1501,1505,
1509,1511,1513,1515,1516,1517,1519,1528,1541,1542,1549,1552,1554,1560,1564,1569,1582,1585,1588,1595,1598,1601,1609,1612,1613, 1620,1625,1631,1635,1641,1643,1644,1655,1656,1663
,1664,1666,1669,1676,1678,1696,1709,1719,1726,1732,1734,1739,1749,1752,1753,1755,1764,1765,1768,1770,1772,1785,1787,1789,1808 ,1818,1819,1822,1826,1829,1835,1836,1838,1845,184
6,1848,1850,1852,1853,1855,1859,1861,1866,1867,1868,1893,1911,1913,1914,1923,2017,2019,2020,2021,2024,2025,2026,2029,2036,204 6,2051,2052,2053,2068,2078,2079,2082,2088,2089,21
04,2108,2112,2113,2116,2118,2121,2131,2134,2135,2150,2152,2155,2159,2161,2163,2164,2166,2168,2173,2176,2177,2178,2182,2185,21 90,2193,2194,2206,2207,2213,2216,2217,2219,2220,2
222,2228,2233,2236,2238,2243,2244,2246,2248,2249,2252,2253,2257,2268,2270,2277,2279,2290,2291,2292,2297,2301,2305,2311,2315,2 319,2324,2325,2328,2330,2345,2346,2348,2351,2353,
2357,2358,2370,2371,2373,2376,2378,2382,2383,2528,2531,2532,2533
AGGGCTCTCGGCAGGATCACTTATTCTGAGACGCGAAGTTCCATAAATTTTAGGATTCGACGCAATTGGTACGCACGGGGGGTTTTGAATGTAACTGTCCGGACTGGTGGAGTGAGTTGGCGTCAG
TCGCACAAGGGCGCGGACCTTTTTGTTTCCTTTTTTGTTTCCTTTAACGGCGTCTACTATGGCGTGACCCCTTAGCCTAAAACCATGCGCCCATGTCAGCGCTCTTGTGTGGGCAACGGACGCCCG
CCTCTACGCCTTACAACAGTTCAGCCGTACATAACAGCTGGACCTTAACACAGTTCAGCCGTACATAACAGCTGGAACGGGTACCTATCCTGTGCAAGCGCATTGAAACTTGGTCGAACTTCAGATA
ACCCCGTGGCAGTGCCGGGAATGGCGACTGCGCGCACCTCAACATTGTTGTACTGCACGTACCAATGAACCGTGAAAAGTACAGTCCTGGGTCACCCCAACGCTTTTGTAAGGAAACCGACCACG
GATCGAGTTCTTGATAATACTGACATGTGTTCTAGAGGTGTATCTTCAAAGAAAGACCATCCAAATGTGGCACACTGTTCCATAGAGTACTCCGCGGGATCTGGACCCTACATGGATGTGACTTCCT
AGCGCGTGTCCAGGCGGGTTAGAGCATTAGAAAAAACTCGTGCGGGCGATTTGCAGTTTGCAGTCACATCTGCTTGGTTATTCGGCGCAGTTTACCACGGTTCTGCGGATAGTCGGTTGCCATCTA
TCGTCATCTCCAAGGCCCGATTGCGGCATTCGTCCCATTTGGGCGTTGGATGGAGACCCATCGCATGGCTGCAGCTTAAAATTAAAAACAACGGAGGATGATGCCGTAGGCTGTAGCGCGAGCGA
TCGTGATCGCTACCGATCCAGGGTTTCTCCTAATAGTGCTTGGATTCTGCTAGTAGTAGCGTAGCGGATAATATGCGGCAGGCACGTGCAGCGTGGGCGTGAAAATCATCGAGAGAGTTCCATAAT
TTTAGTTGGGACGTTGACTCCAGCTTATGAGAGGAATATTAGTAGCAATAAAGCGCGGCGCGCGGGCAACCGGAACAAGGCCTGCCGTGGTCGTATCGAGTCCAGGGGGTACGGACTGAGTGCG
GGTTAAGAACTGTCGTCTTACCACGTATCTCCTCCTAGATCCTATTTTGTAAATTAGTTTGCGCTACGGACCCTTACAACCTGAAAGAGGTGTTAGGGATTATGAGATAATCGAACTGATCTCGGGTG
GTCCGGGCGTCCCGCTAGCCCGGTTTCGTACCGGGTTTGAAATTTGTTGACTTATGTCCATTGTATTCAACGGTATCCCACTACGACGTAGCGACATCTGTTTGGTATGGGTTTGGTATGGCATCAC
TTGATCCAGCTGGTTCTGTGACTCCGAAATGTTGACTGGGAATGAATGGAGGTTTGTCTGCGAACCCAGATGGCTTGCAAGCACGCGCAGCCGAAGACAATGTACCGTAGACAGACTGACGCACA
CACAAAGATTTTGCTCACTCGTGGCTTTGAATTCGCGAGGACAGCGGGACGGAGGCGACGGGCCCTGTCGACGACGAGCTGCGACGACGAGCTGCCGACTAAGGGCTCACTTTACTTTCAGGGA
CTCCGATAATGGTTCCCTTAAGCGTCTAAGAGTAGCCCGTACAGTCTGCGCCCTTGCTCCACTTGGTTTCTGGATGCCCCTTTACGCGGGACGTCCAGATGCCACTCGGGGTGACGAAGAGTTTG
GCGAATGACATACGTCTCGTTCTGACATAGCGTCGCTCGCTTGTCTCAGTGCTGCTGAAGGACCTACCACCCTCAAGATCTCCTAATACACAAGACTCAGAGCGCAAAGGTCCGGTTGCTGCTCCC
CCTTTCACCCGGTCCCCTCGTGGCATAAGGTTGTCGACCGCCCTTGGTTTAATAATGGTCGCCACGTGAGCTTCTACCTTCCCCAGAATGGGCTAACCCTAGCTGATACATCGCTGGGAGCGGGC
TGCCGAGAAACTAAATCAGGGTTCATGGTGTCCGATTTGAAATCCGTGCTCCCCTTAGCCTGCGGGAACGACCCGTAATGGCCGTTCTGGGGAGTTAGTTAACCAGACCAGCCGTTCGCACTAATT
TGCTCGCCGTTGACATTATGCAGAGAACAGAGCGCAGGAAAGGCAGTACGCCAGGAATTGCGCCCGTCAACTTTGAGCAATAATATGCCTACCTTATTATAGTCTAACAGAAGTAACCTAGCGTGG
TGGCAGACGGCCGATAGGCGGCCTGTAAAGGCCATGGAGGCAGGTGTACCCATCTACGGGAAGTATACGTTTTTCGTGGGGAATATCACATGGAATTTTCGTGGGGAATATCACATGGAACGTGT
ACGTCACCAGGTATAGATTAGCTAGAAGCCTTTCCAGCAACAAGAGCCATGGACTGTCCCTCAACAACACCTATATATCCCATCATTCCCACCTTCTCACATCATTCTCATCCATATAACTCTAAAAC
CTTTAACCAAAATCAAACTGATGTA
```

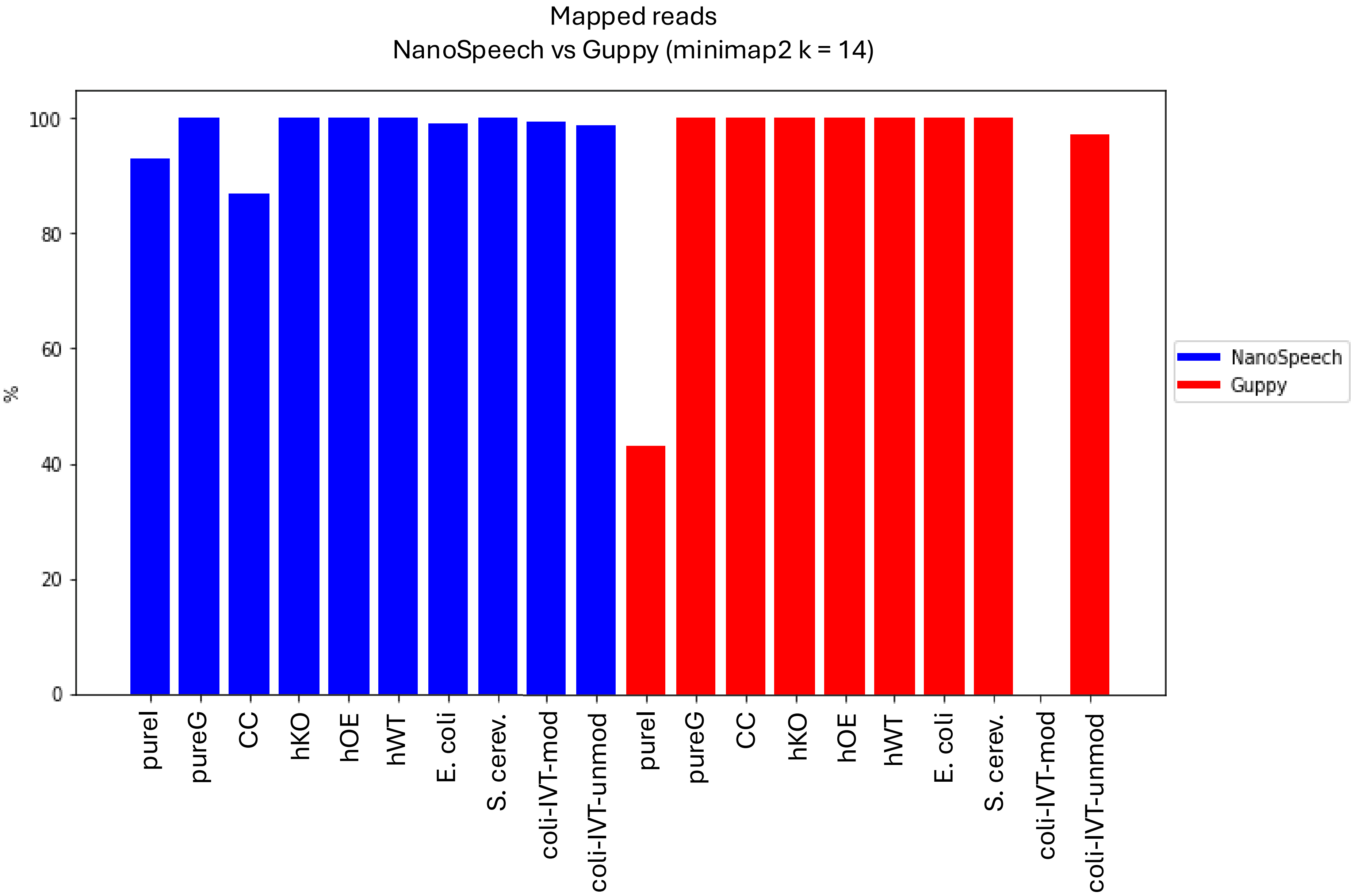
Supplementary Figure 2



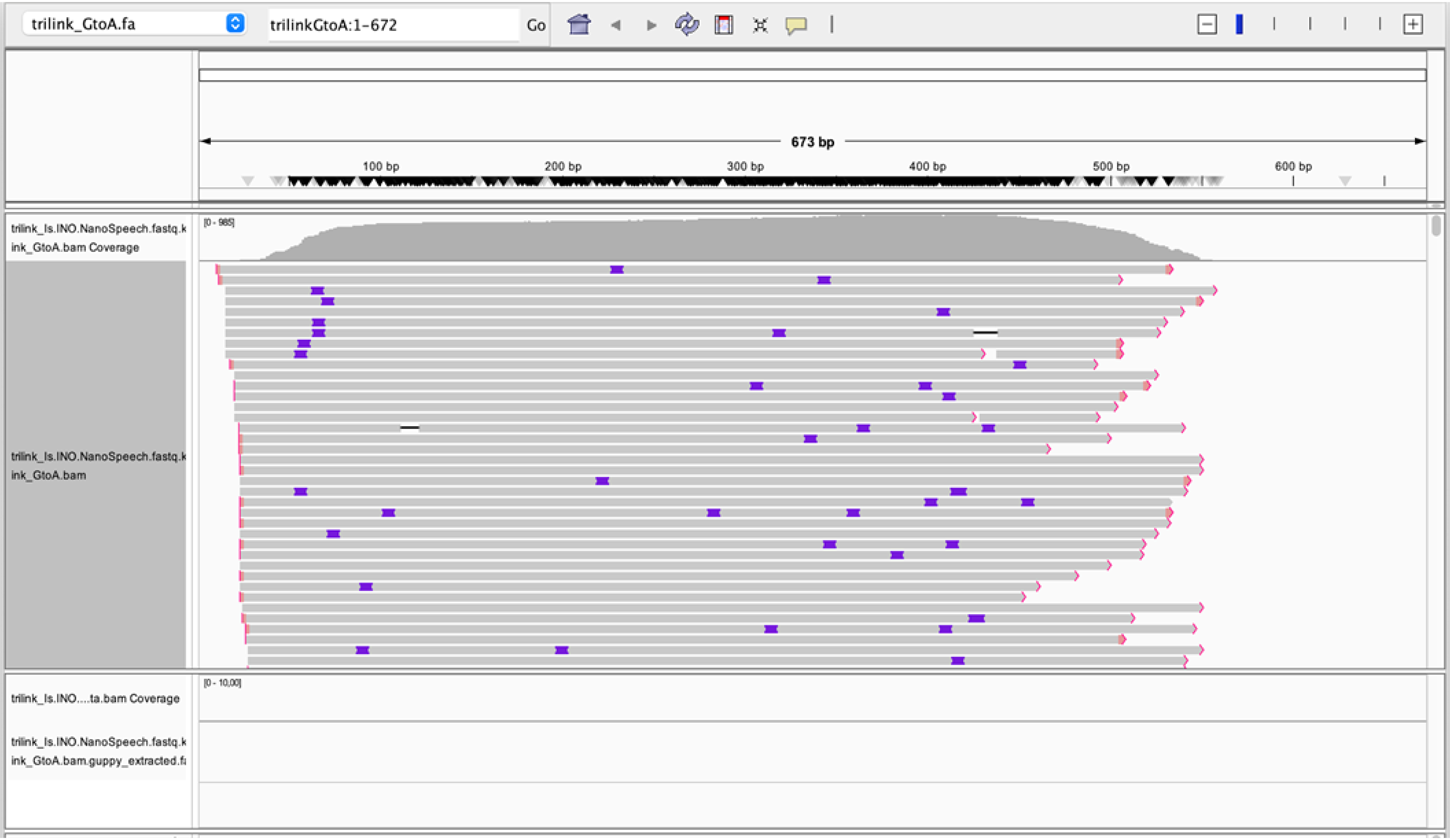
Supplementary Figure 3



Supplementary Figure 4

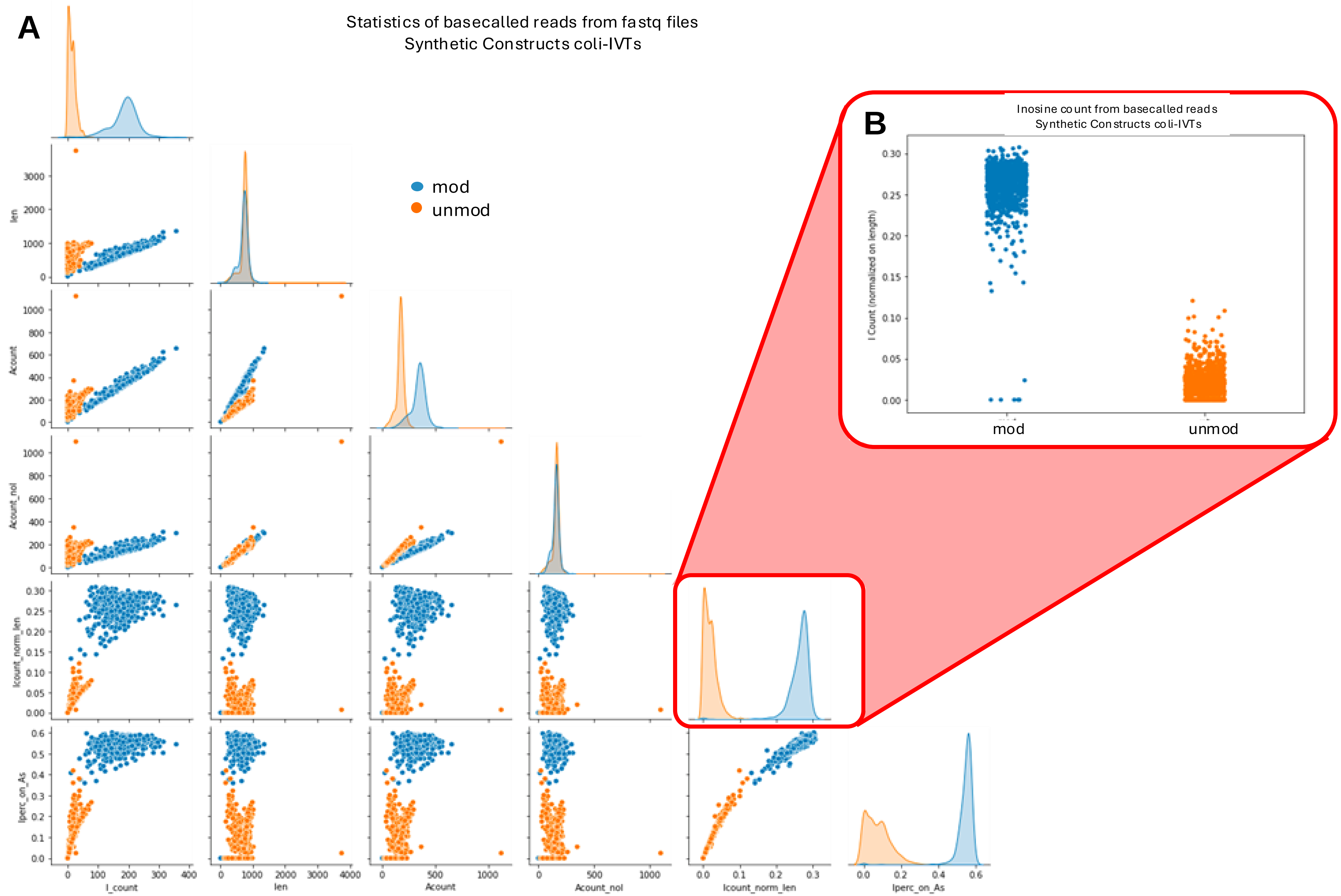


Supplementary Figure 5



trilink_GtoA.fa
trilinkGtoA:1-672
Go
673 bp
100 bp
200 bp
300 bp
400 bp
500 bp
600 bp
trilink_ls.INO.NanoSpeech.fastq.kink_GtoA.bam Coverage
[0 - 985]
trilink_ls.INO.NanoSpeech.fastq.kink_GtoA.bam
trilink_ls.INO....ta.bam Coverage
[0 - 10,00]
trilink_ls.INO.NanoSpeech.fastq.kink_GtoA.bam.guppy_extracted.fa

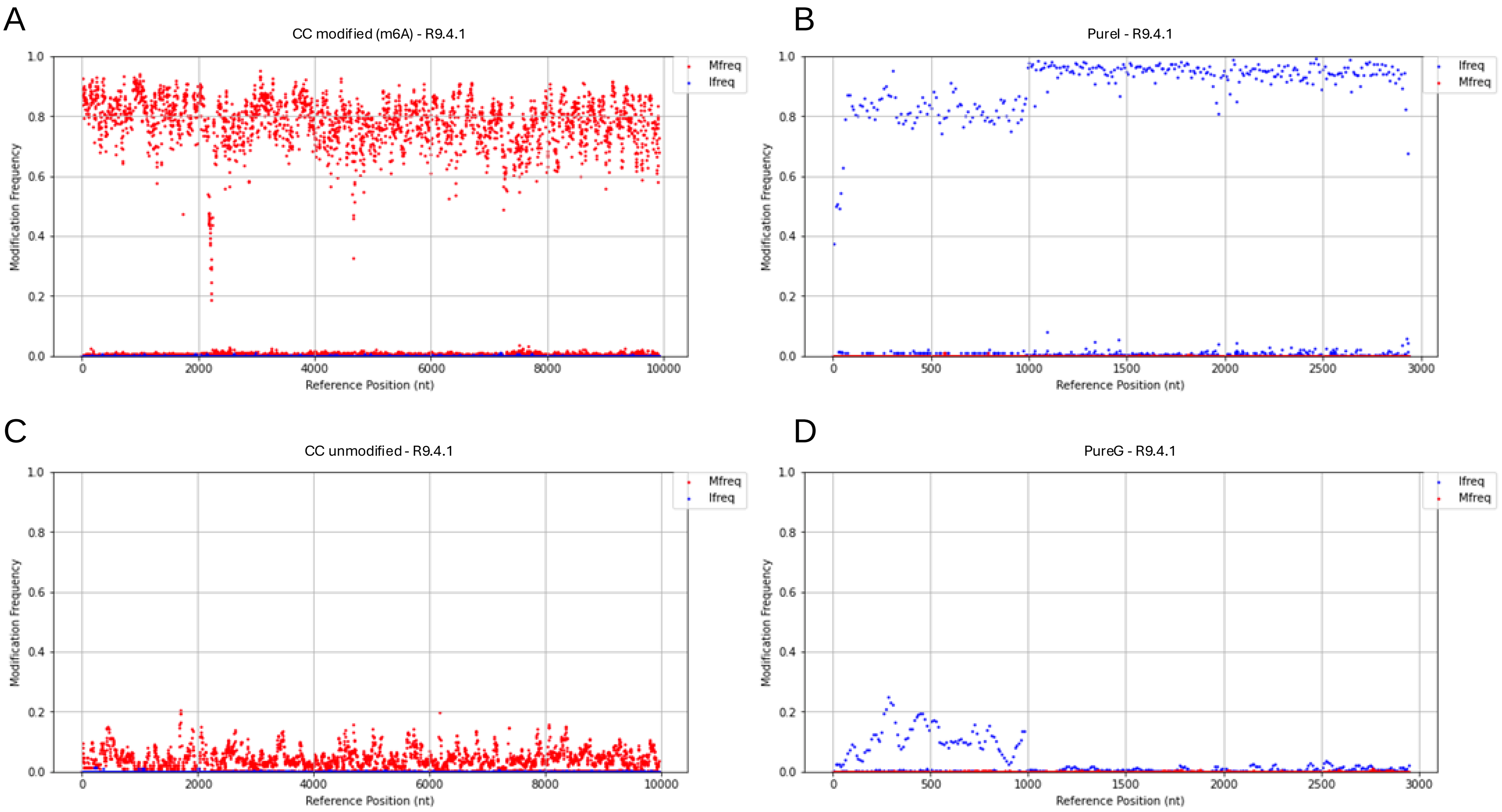
Supplementary Figure 6



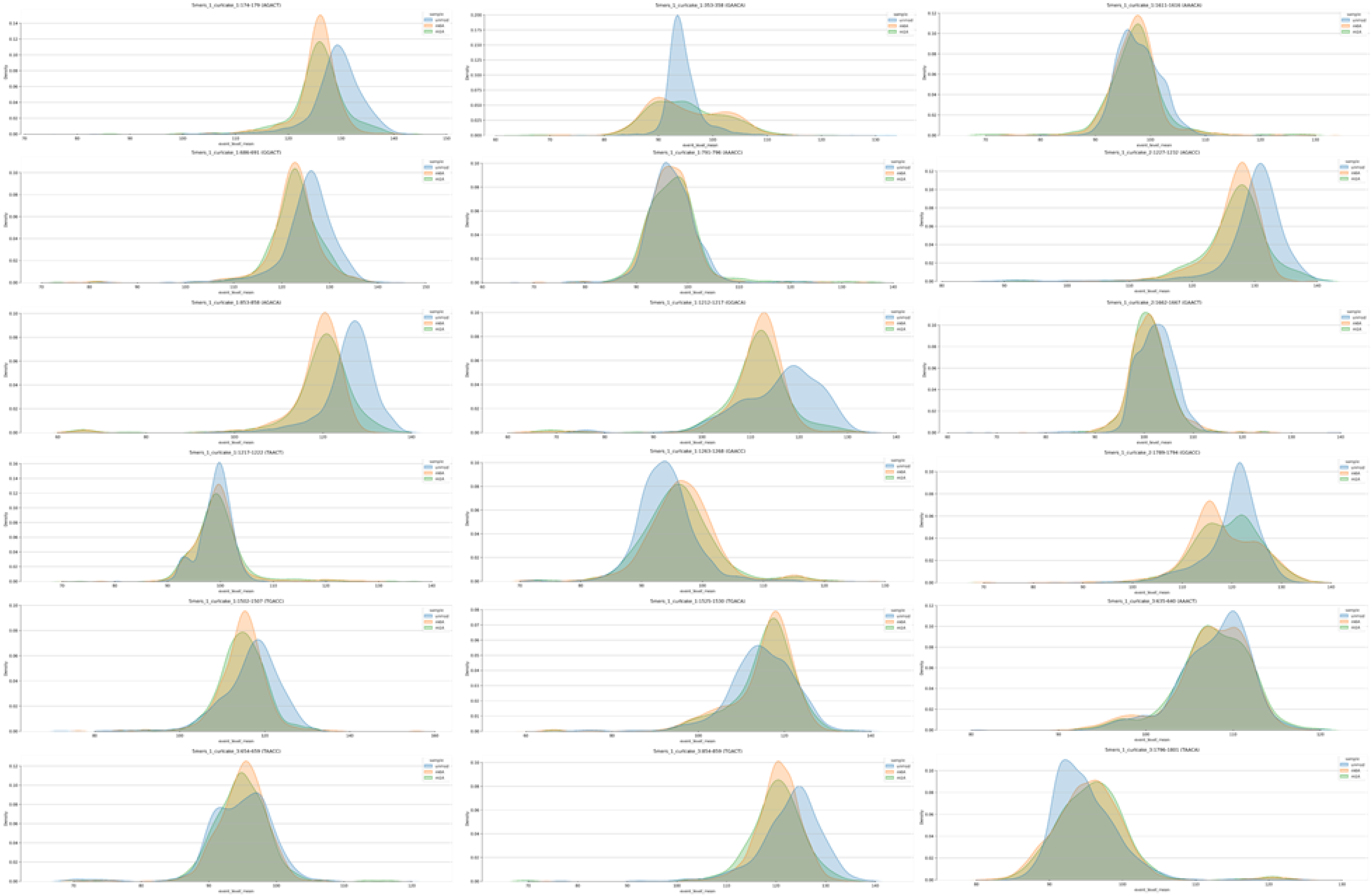
Statistics of basecalled reads from fastq files
Synthetic Constructs coli-IVTs

Inosine count from basecalled reads
Synthetic Constructs coli-IVTs

Supplementary Figure 7



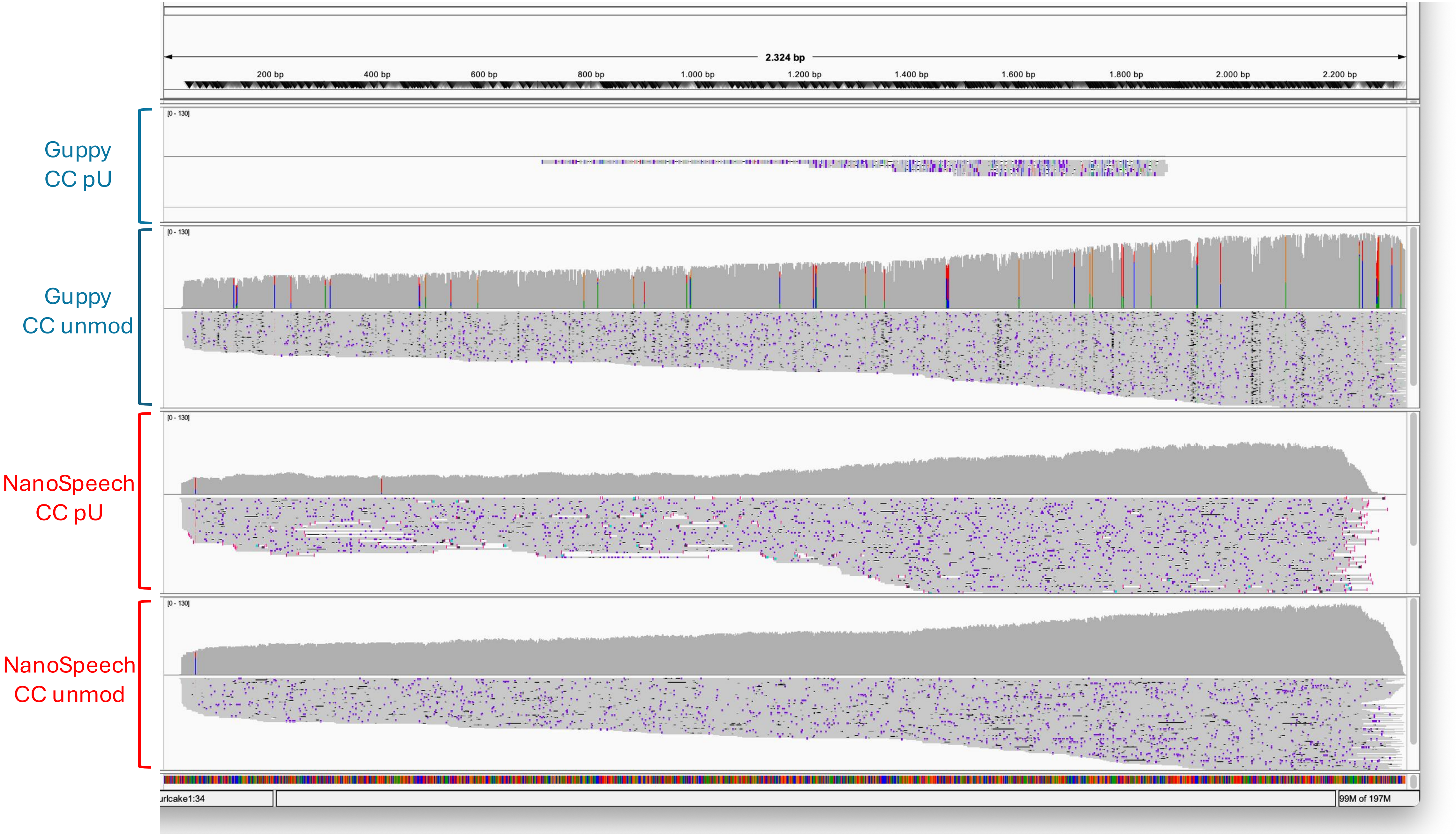
Supplementary Figure 8



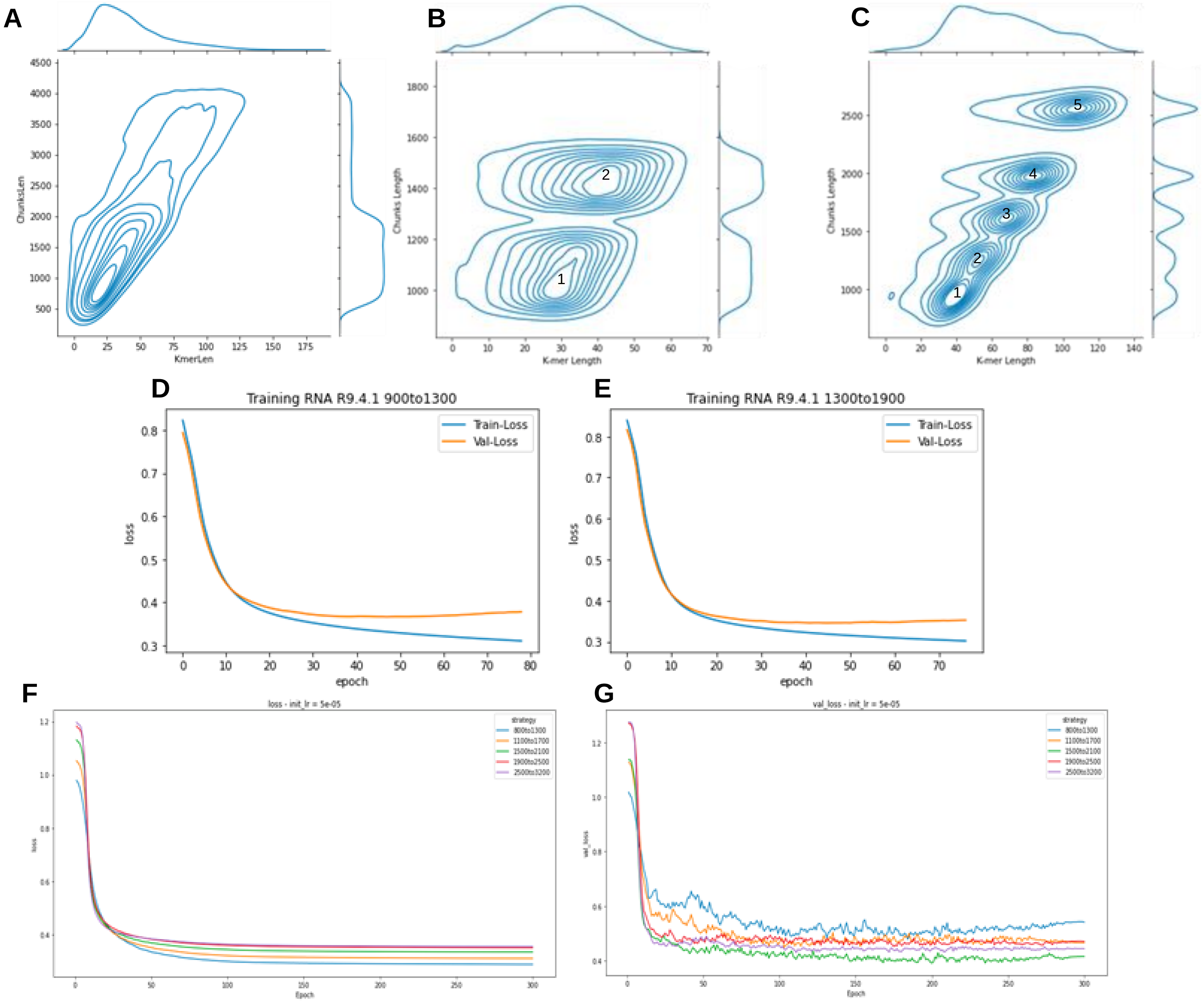
Supplementary Figure 9

contig	start	stop	kmer	unmod_mean	m6A_mean	m1A_mean	unmod_std	m6A_std	m1A_std	unmod_N	m6A_N	m1A_N	unmod_vs_m6A_p	unmod_vs_m1A_p	m6A_vs_m1A_p
5mers_1_curlcake_1	174	179	AGACT	129.3160096	125.4521572	125.7468703	4.842301883	3.936408679	5.108871509	4987	1539	802	3.41E-254	5.69E-109	0.141139944
5mers_1_curlcake_1	353	358	GAACA	94.56351801	95.13601852	95.19900433	3.091411901	6.796281227	6.734308996	4693	1404	693	0.190994377	0.475968127	0.627357182
5mers_1_curlcake_1	686	691	GGACT	125.662737	122.533716	122.99695	5.287202687	5.5678547	5.182279596	7932	2979	1341	3.24E-241	3.60E-101	0.039886435
5mers_1_curlcake_1	791	796	AAACC	96.93390074	96.66319843	97.18749035	4.448026568	4.248481202	5.482185278	4030	1779	777	0.074847252	0.891635704	0.181896054
5mers_1_curlcake_1	853	858	AGACA	125.9631049	118.6015679	119.4769284	5.777571013	7.437109984	8.801781405	8844	3999	1732	0	0	1.41E-11
5mers_1_curlcake_1	1212	1217	GGACA	116.1320482	111.43486	111.2699292	8.577071838	6.501248794	7.661389773	4028	1214	565	2.04E-107	2.14E-53	0.307259305
5mers_1_curlcake_1	1217	1222	TAACT	99.06345587	99.30972041	99.43301943	2.967564179	4.537248584	4.95402042	6991	2611	1338	0.130256421	0.227508626	0.956786857
5mers_1_curlcake_1	1263	1268	GAACC	94.01860146	96.71914384	96.11230051	4.499890082	5.64735252	5.934968924	4376	3761	1178	1.99E-152	3.24E-39	3.11E-05
5mers_1_curlcake_1	1502	1507	TGACC	116.8889692	114.7054948	114.7512685	6.719624621	5.5864745	5.466713272	4152	1718	607	1.46E-57	1.56E-24	0.674218483
5mers_1_curlcake_1	1525	1530	TGACA	114.7106122	114.8937581	115.2754132	7.740102773	7.846028278	7.300890062	4067	4191	1428	7.91E-07	3.59E-05	0.845950311
5mers_1_curlcake_1	1611	1616	AAACA	97.86604754	97.18281369	97.30011923	3.790751619	4.274527451	5.301301106	4964	2630	671	1.29E-12	0.000932802	0.499572993
5mers_1_curlcake_2	1227	1232	AGACC	129.8810762	126.63275	127.0771746	5.724572409	4.293156257	4.979920284	5194	3360	899	0	4.67E-102	0.08435981
5mers_1_curlcake_2	1662	1667	GAACT	102.4463327	101.2824951	101.1797086	3.882664061	4.07053175	3.906544556	4322	3603	755	9.68E-51	2.61E-20	0.64310733
5mers_1_curlcake_2	1789	1794	GGACC	120.2448649	118.221537	119.1715504	5.462500994	7.079771543	6.640495069	6366	4242	903	3.36E-84	5.52E-09	1.08E-05
5mers_1_curlcake_3	635	640	AAACT	107.8301829	107.4942456	107.6829912	4.048101914	4.227820981	4.137106358	2733	2280	565	0.019579169	0.261909923	0.782250221
5mers_1_curlcake_3	654	659	TAACC	94.70363692	94.94123561	94.73068429	4.722900773	3.686237301	3.850658781	3239	2258	643	0.234496184	0.287300474	0.038187415
5mers_1_curlcake_3	854	859	TGACT	123.0103914	120.480202	120.1216729	6.172653075	4.455537619	5.644448033	2606	2921	807	5.27E-99	8.93E-47	0.101151864
5mers_1_curlcake_3	1796	1801	TAACA	94.85080637	95.52580708	95.88136433	4.161809658	5.046435668	4.993260206	5655	11269	1312	3.39E-29	7.41E-17	0.015057534

Supplementary Figure 10



Supplementary Figure 11


A
ChunksLen
KmerLen
B
Chunks Length
K-mer Length
C
Chunks Length
K-mer Length
D
Training RNA R9.4.1 900to1300
Train-Loss
Val-Loss
loss
epoch
E
Training RNA R9.4.1 1300to1900
Train-Loss
Val-Loss
loss
epoch
F
loss - init_lr = 5e-05
strategy
800to1300
1100to1700
1500to2100
1900to2500
2500to3200
loss
Epoch
G
val_loss - init_lr = 5e-05
strategy
800to1300
1100to1700
1500to2100
1900to2500
2500to3200
val_loss
Epoch