# Supplementary Figures and Tables

## The extensive $m^5C$ epitranscriptome of *Thermococcus kodakarensis* is generated by a suite of RNA methyltransferases that support thermophily

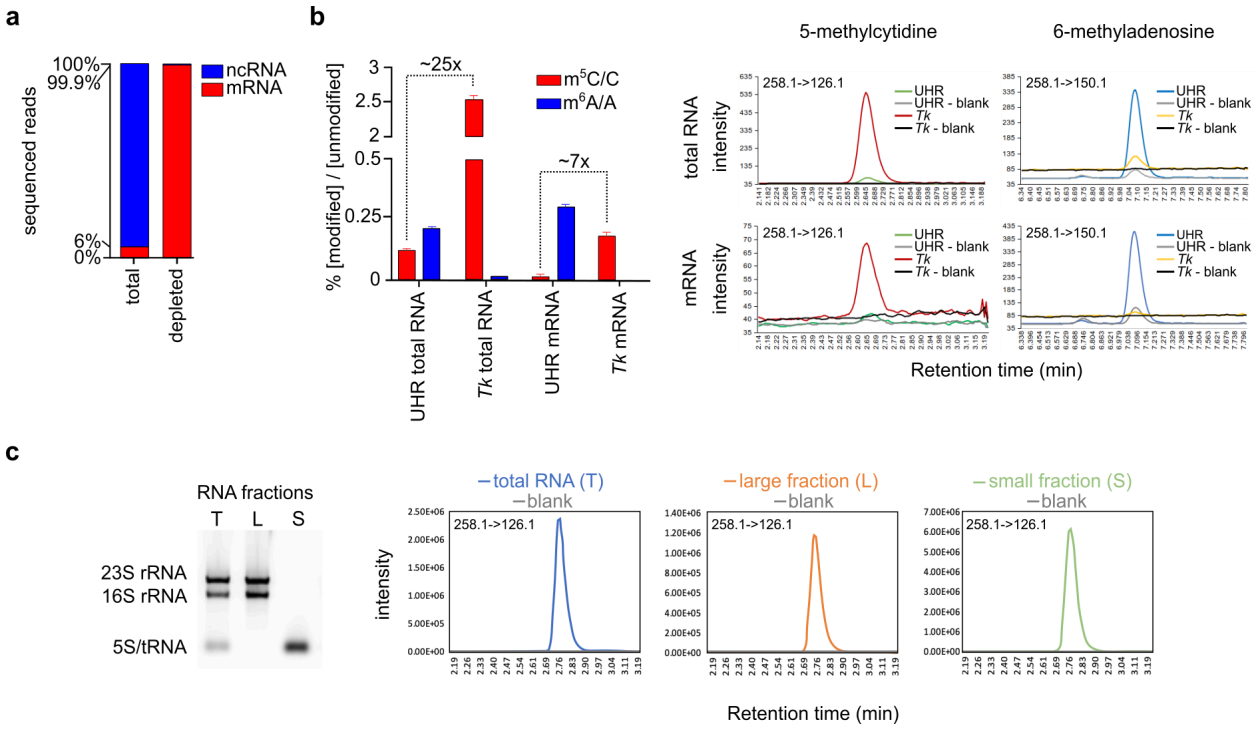
Kristin A. Fluke[1], Ryan T. Fuchs[2], Yueh-Lin Tsai[2], Victoria Talbott[1], Liam Elkins[3], Hallie P. Febvre[3], Nan Dai[2], Eric J. Wolf[2], Brett W. Burkhart[3], Jackson Schiltz[3], G. Brett Robb[2], Ivan R. Corrêa Jr.[2], Thomas J. Santangelo[1,3,4]

[1] Cell and Molecular Biology Graduate Program, Colorado State University, Fort Collins, CO, United States, 80523

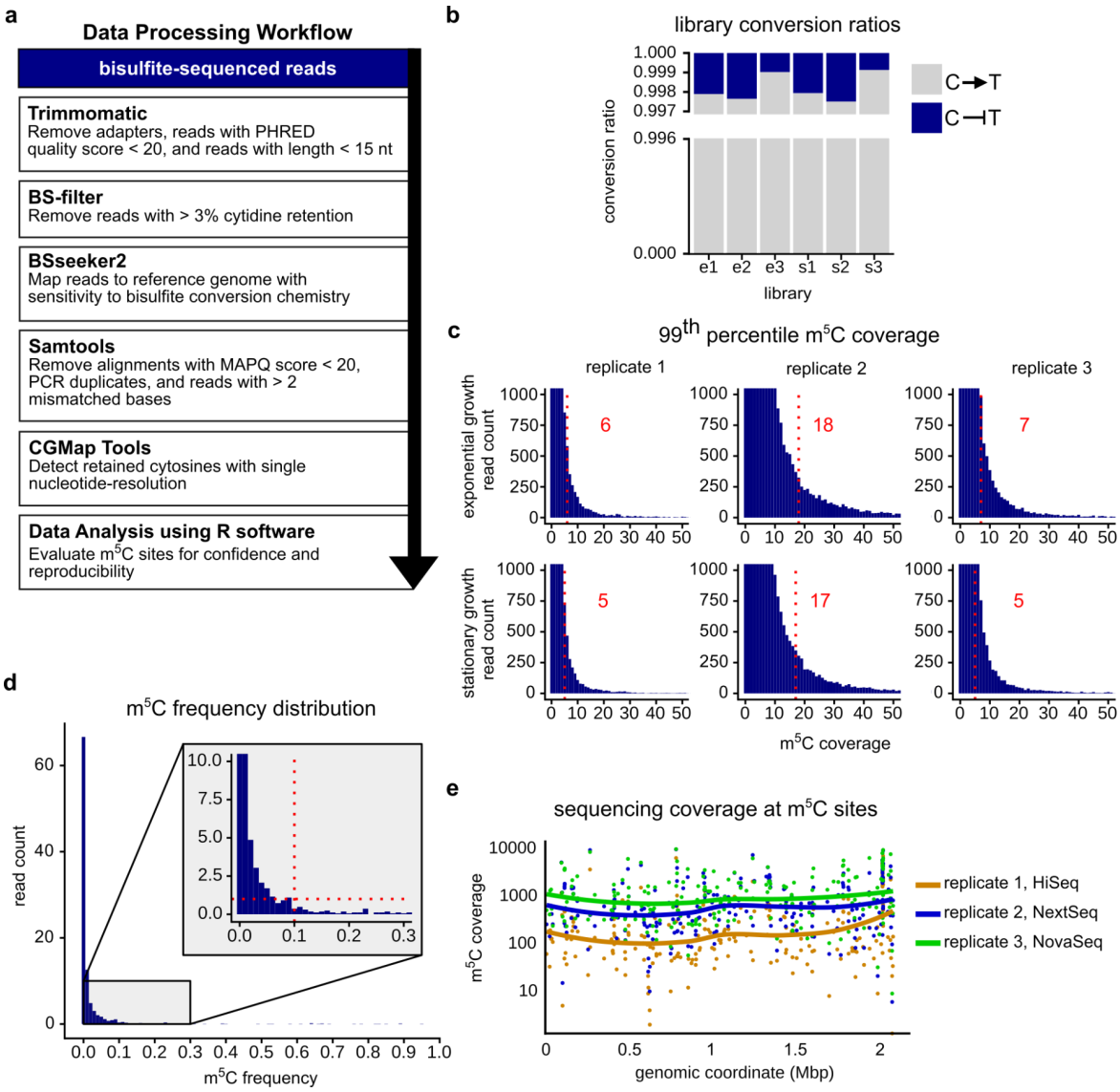
[2] New England Biolabs Inc., Beverly, Massachusetts, United States, 01915

[3] Department of Biochemistry and Molecular Biology, Colorado State University, Fort Collins, CO, United States, 80523

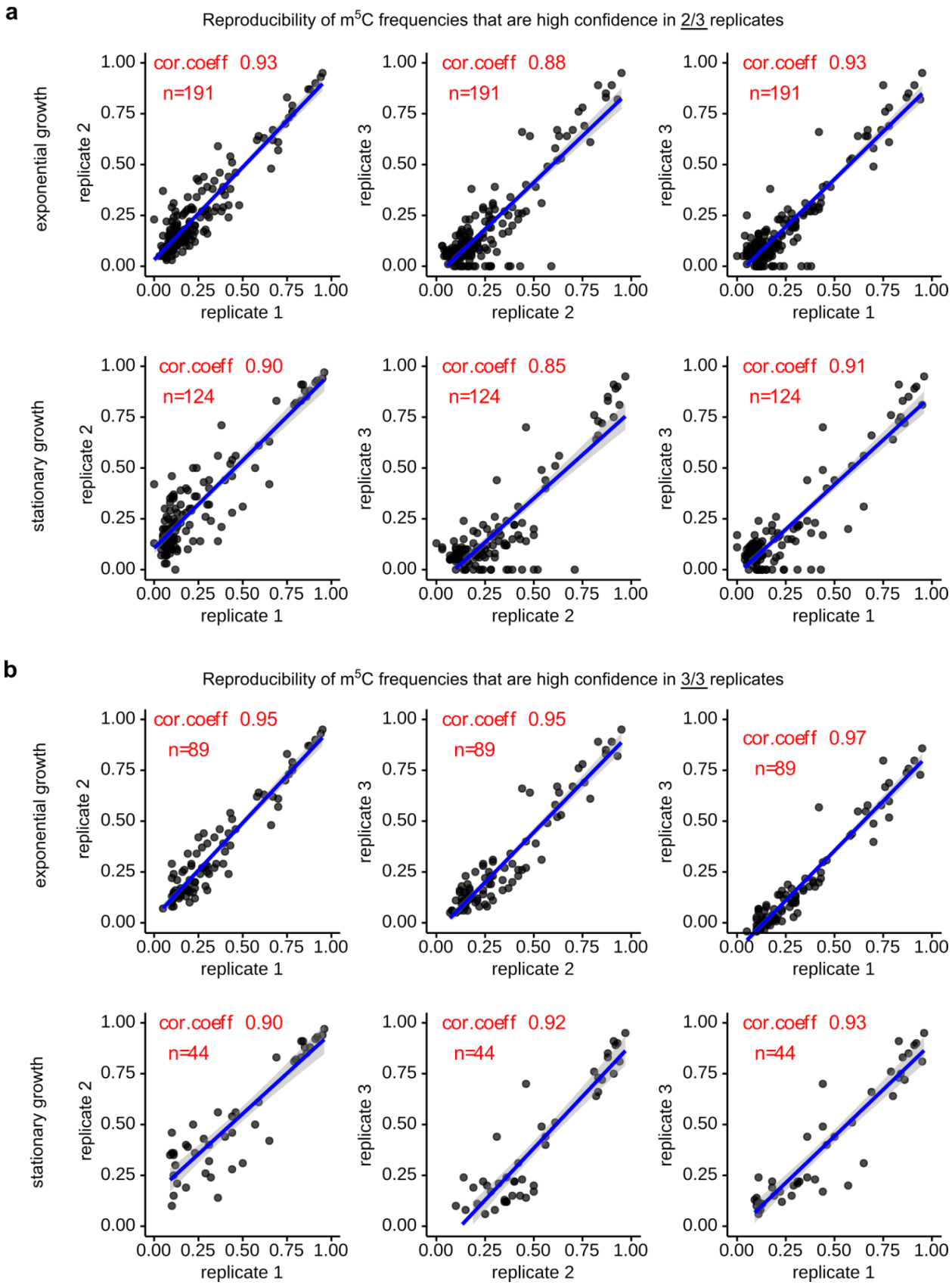
[4] Correspondence: thomas.santangelo@colostate.edu; ORCiD: 0000-0003-4559-3244.



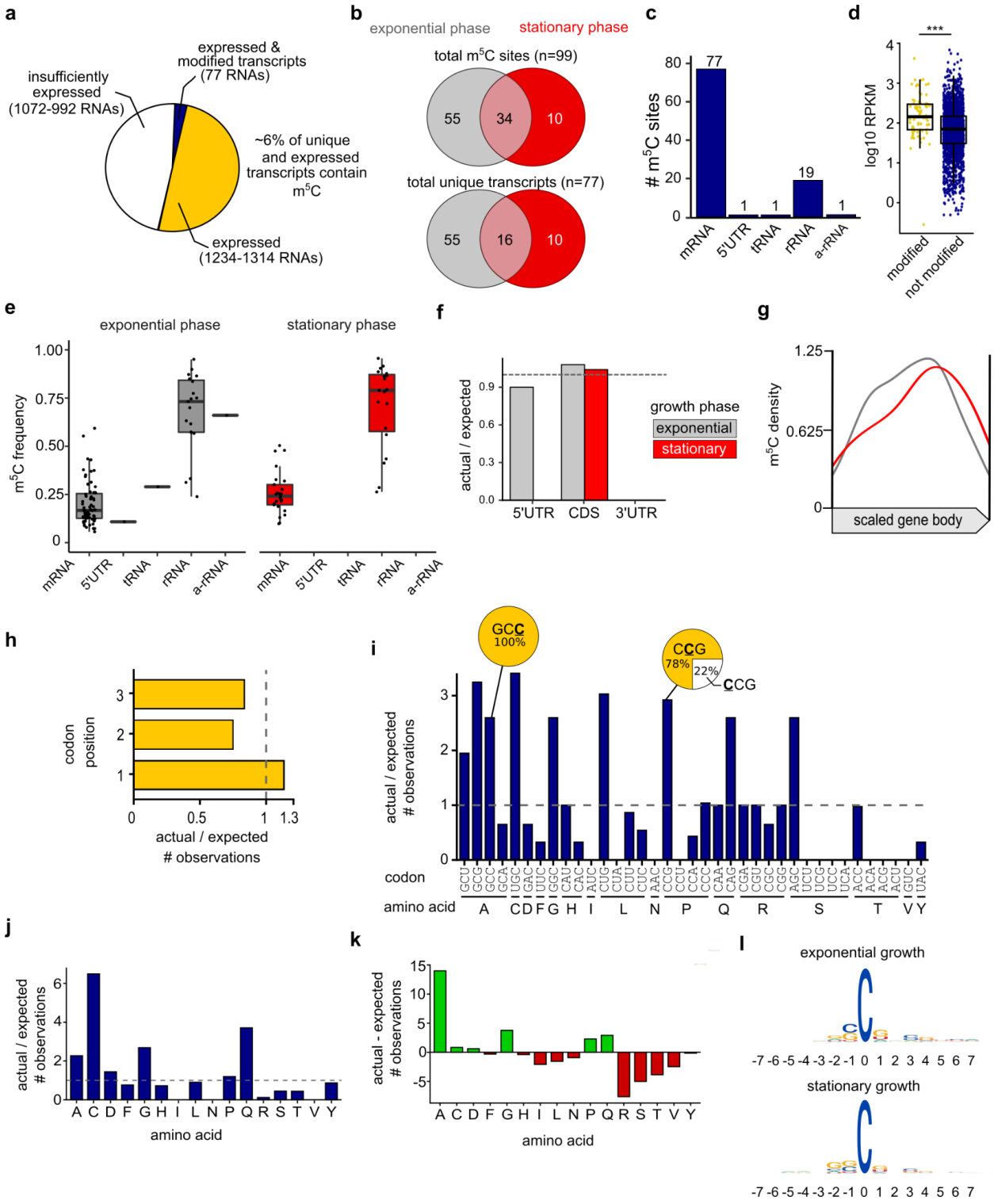

**Supplementary Fig. 1 | $m^5C$ dominates the epitranscriptomic landscape in *T. kodakarensis*.** **a** RNA was collected from cells and the rRNA and tRNAs depleted. Total RNA includes ~6% mRNAs whereas virtually all post-depletion RNA is mRNA. **b** Methyl-5 cytidine without phosphate groups has a retention time of ~2.65 min. Mass spectrometry analysis of total and depleted RNA from a Universal Human Reference (UHR) and *T. kodakarensis* indicate $m^5C$. The bar graph displays mean ± 1 standard deviation where n = 3 technical replicates. **c** Total, large and small fractions of RNA from *T. kodakarensis* indicate that $m^5C$ is highly abundant in rRNA and tRNAs. An uncropped gel image is provided in Source Data.




**Supplementary Fig. 2 | Bisulfite sequencing of RNA permits comprehensive detection of high confidence and reproducible m$^5$C sites.** RNA was collected from cells grown to exponential or stationary growth phase and bisulfite-sequenced on an Illumina platform. **a** Reads sequenced with low quality were removed and adapters trimmed using Trimmomatic. A custom python script was used to remove reads where cytidines constitute >3% of the nucleotides. The reads were then mapped to the reference genome using BSseeker2 where C to T conversions are not counted as mis-matched bases. The aligned reads were further filtered where low MAPQ scores, PCR duplicates, and alignments with > 2 non-C-to-T mismatches were removed. Cytidine retention at single nucleotide resolution was quantified using CGmaptools. Finally, CGmaps were analyzed using R software. **b** The library-wide cytidine conversion ratio was quantified for three replicates from each biological condition. **c** For each sequenced library, we calculated the 99$^{th}$ percentile for m$^5$C coverage (listed in red text) for all genomically encoded cytidines. **d** A histogram illustrates the m$^5$C frequency of all candidate m$^5$C sites. A sharp decline followed by a steady pattern of m$^5$C frequencies are observed at ~10% modification frequency. **e** The m$^5$C coverage at all high-confidence and reproducible m$^5$C sites is extraordinarily high in all sequenced libraries, ranging from 100-1000X coverage on average.


a
Reproducibility of m5C frequencies that are high confidence in 2/3 replicates
exponential growth
stationary growth
replicate 1
replicate 2
replicate 3
cor.coeff 0.93
n=191
cor.coeff 0.88
n=191
cor.coeff 0.93
n=191
cor.coeff 0.90
n=124
cor.coeff 0.85
n=124
cor.coeff 0.91
n=124
b
Reproducibility of m5C frequencies that are high confidence in 3/3 replicates
cor.coeff 0.95
n=89
cor.coeff 0.95
n=89
cor.coeff 0.97
n=89
cor.coeff 0.90
n=44
cor.coeff 0.92
n=44
cor.coeff 0.93
n=44
0.00
0.25
0.50
0.75
1.00

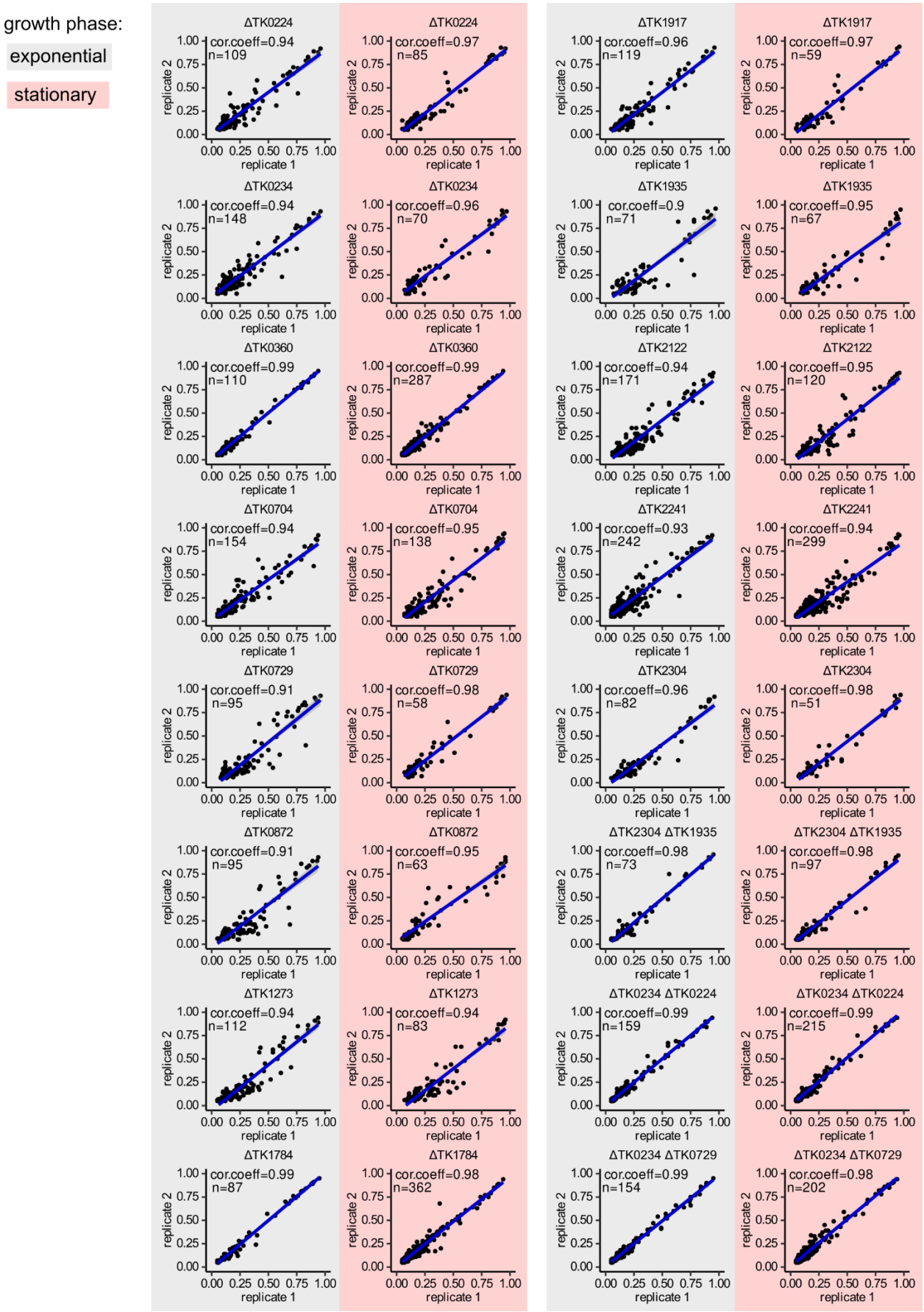
**Supplementary Fig. 3 | Linear regression of $m^5C$ frequencies indicated highly reproducible modification frequencies.** Linear regression and Pearson's correlation coefficients (cor. coeff) were used to compare the site-specific modification frequencies of high confidence and reproducible $m^5C$ sites in each replicate. The analysis was performed for sites reproducible in **(a)** 2/3 and **(b**) 3/3 replicates. The analysis was performed using R software.



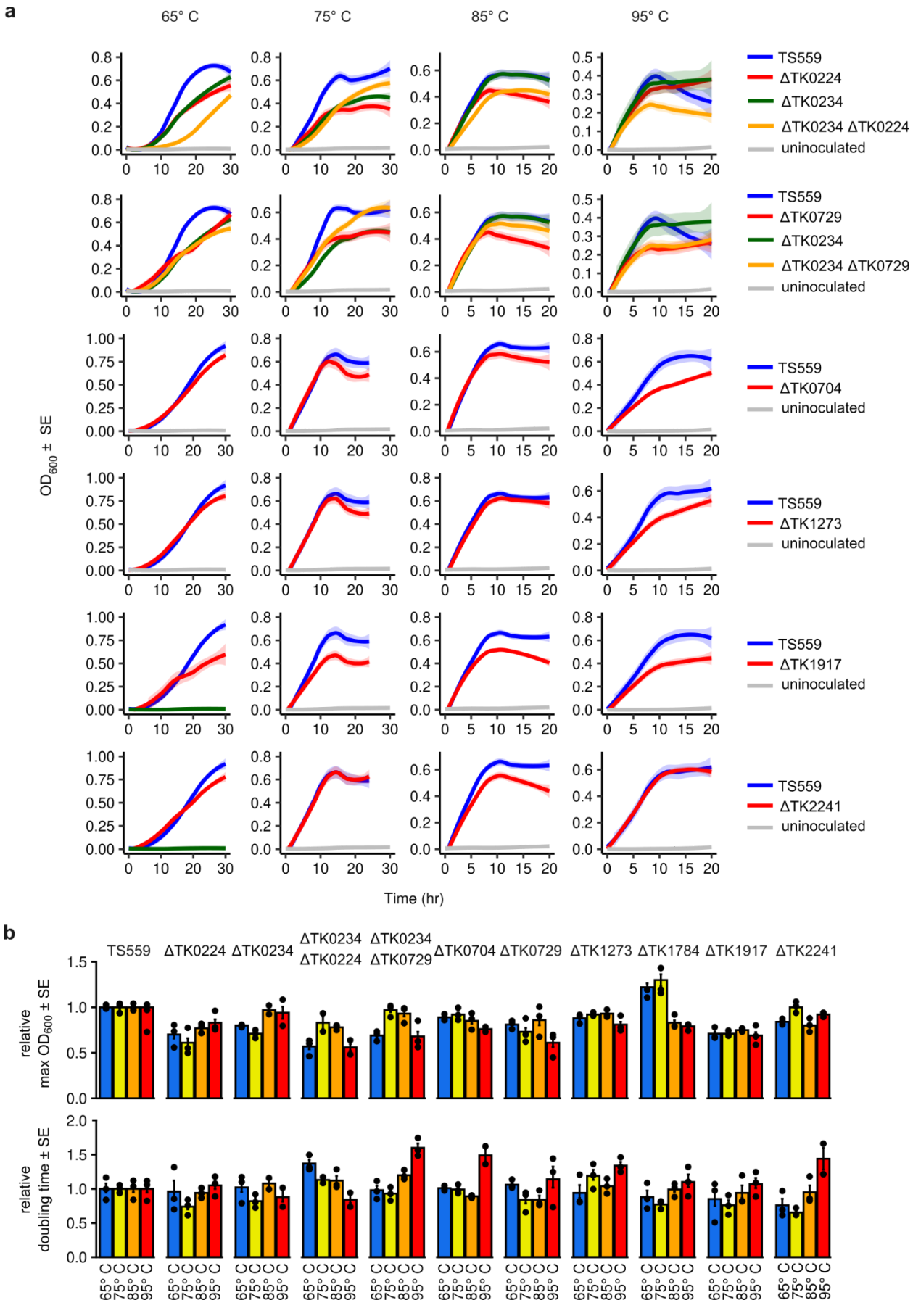

**Supplementary. Fig. 4 | Analysis of m$^5$C sites present in 3/3 replicates reveal largely similar conclusions compared to m$^5$C sites present in at least 2/3 replicates.** Parallel analysis of the context in which m$^5$C is found throughout the transcriptome was performed for m$^5$C sites found in all three replicates. **a** We detected 1234-1314 RNAs with expression at or above the thresholds set for the

minimum coverage required to call $m^5C$ sites in all three replicates. Between two biological conditions, 6% of expressed transcripts contained at least one $m^5C$ site. **b** The number of $m^5C$ sites and unique transcripts were calculated in two biological conditions, and distinct $m^5C$-epitranscriptomic profiles were observed between exponential and stationary growth phases. **c** $m^5C$ are similarly distributed throughout a diverse set of RNAs. **d** Gene expression (log 10 mean RPKM) of modified (yellow) and unmodified (blue) mRNAs. **e** The average modification frequencies are nearly identical when compared to analysis of $m^5C$ sites present in at least 2/3 replicates. **f** We compared the actual number observed and the expected number of $m^5C$ sites in different regions of an mRNA. We expect the actual / expected number of $m^5C$ sites to be equal to 1 if the $m^5C$ site is distributed throughout the transcriptome randomly. Unlike $m^5C$ sites present in at least 2/3 replicates, there were no sites mapped to 3'UTRs, so we did not observe an enrichment in this region when only analyzing sites present in all 3 replicates. **g** When $m^5C$ mapped to coding sequences, there was no apparent positional bias, **(h)** nor did we observe a strong bias in codon position. **i** The number of $m^5C$ sites mapped to each codon was compared to the number expected by random chance. Obvious deviations from expectation (actual / expected = 1) were similarly observed between sites present in 2/3 or 3/3 replicates. (**j**) The actual / expected and **(k)** difference between actual and expected (actual - expected = 0 where no bias is detected) number of $m^5C$ sites mapping to particular amino acid codons was calculated and indicate clear bias in $m^5C$ positioning in select amino acid contexts. **l** Logos sequence analysis of nucleotides adjacent to $m^5C$ sites do not reveal a strong RNA sequence context.

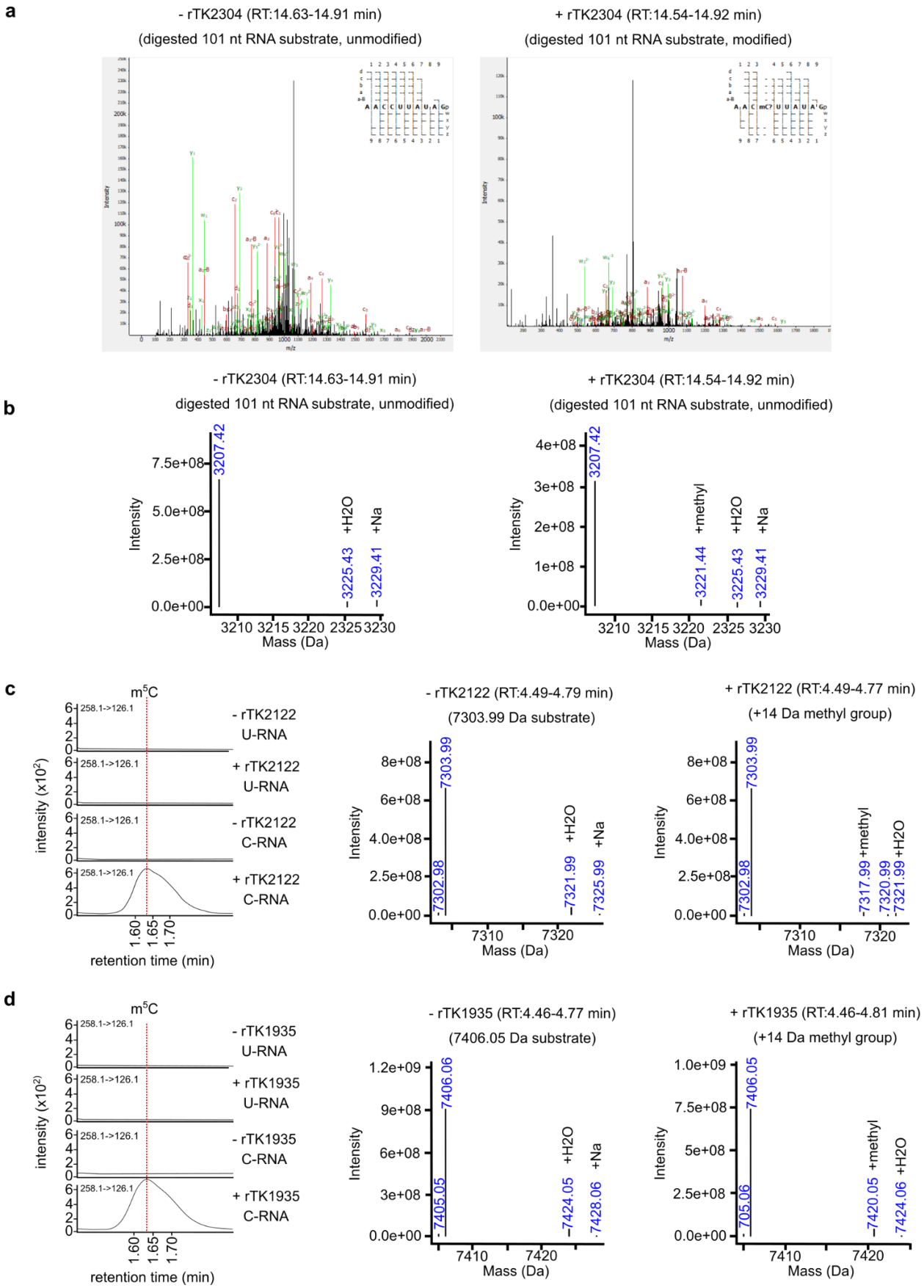

**Supplementary Fig. 5 | Genes encoding putative RMTases were genomically deleted.** Genomic deletions were initially screened using PCR for preliminary confirmation (data not shown), and final confirmation was performed using Minion whole genome sequencing (DNA-seq) and Illumina RNA bisulfite sequencing (RNA-seq). Visual inspection of DNA and RNA sequences aligned to the *T. kodakarensis* reference genome using Integrative Genomics Viewer. In each window, the coverage and read tracks are stacked. The RNA-seq alignments are displayed using bisulfite-mode. Reads absent from the deleted loci indicate markerless removal of the loci encoding the target RMTase. In the DNA-seq windows, black lines connect contiguous reads and indicate a gapped alignment, further confirming the target DNA region is removed. The boundaries of each gene and their gene IDs are represented by blue bars, and white arrows illustrate the direction in which the gene is transcribed. The coverage range for each window is listed in brackets.


growth phase:
exponential
stationary
ΔTK0224
cor.coeff=0.94
n=109
ΔTK0224
cor.coeff=0.97
n=85
ΔTK1917
cor.coeff=0.96
n=119
ΔTK1917
cor.coeff=0.97
n=59
ΔTK0234
cor.coeff=0.94
n=148
ΔTK0234
cor.coeff=0.96
n=70
ΔTK1935
cor.coeff=0.9
n=71
ΔTK1935
cor.coeff=0.95
n=67
ΔTK0360
cor.coeff=0.99
n=110
ΔTK0360
cor.coeff=0.99
n=287
ΔTK2122
cor.coeff=0.94
n=171
ΔTK2122
cor.coeff=0.95
n=120
ΔTK0704
cor.coeff=0.94
n=154
ΔTK0704
cor.coeff=0.95
n=138
ΔTK2241
cor.coeff=0.93
n=242
ΔTK2241
cor.coeff=0.94
n=299
ΔTK0729
cor.coeff=0.91
n=95
ΔTK0729
cor.coeff=0.98
n=58
ΔTK2304
cor.coeff=0.96
n=82
ΔTK2304
cor.coeff=0.98
n=51
ΔTK0872
cor.coeff=0.91
n=95
ΔTK0872
cor.coeff=0.95
n=63
ΔTK2304 ΔTK1935
cor.coeff=0.98
n=73
ΔTK2304 ΔTK1935
cor.coeff=0.98
n=97
ΔTK1273
cor.coeff=0.94
n=112
ΔTK1273
cor.coeff=0.94
n=83
ΔTK0234 ΔTK0224
cor.coeff=0.99
n=159
ΔTK0234 ΔTK0224
cor.coeff=0.99
n=215
ΔTK1784
cor.coeff=0.99
n=87
ΔTK1784
cor.coeff=0.98
n=362
ΔTK0234 ΔTK0729
cor.coeff=0.99
n=154
ΔTK0234 ΔTK0729
cor.coeff=0.98
n=202
replicate 1
replicate 2
0.00 0.25 0.50 0.75 1.00

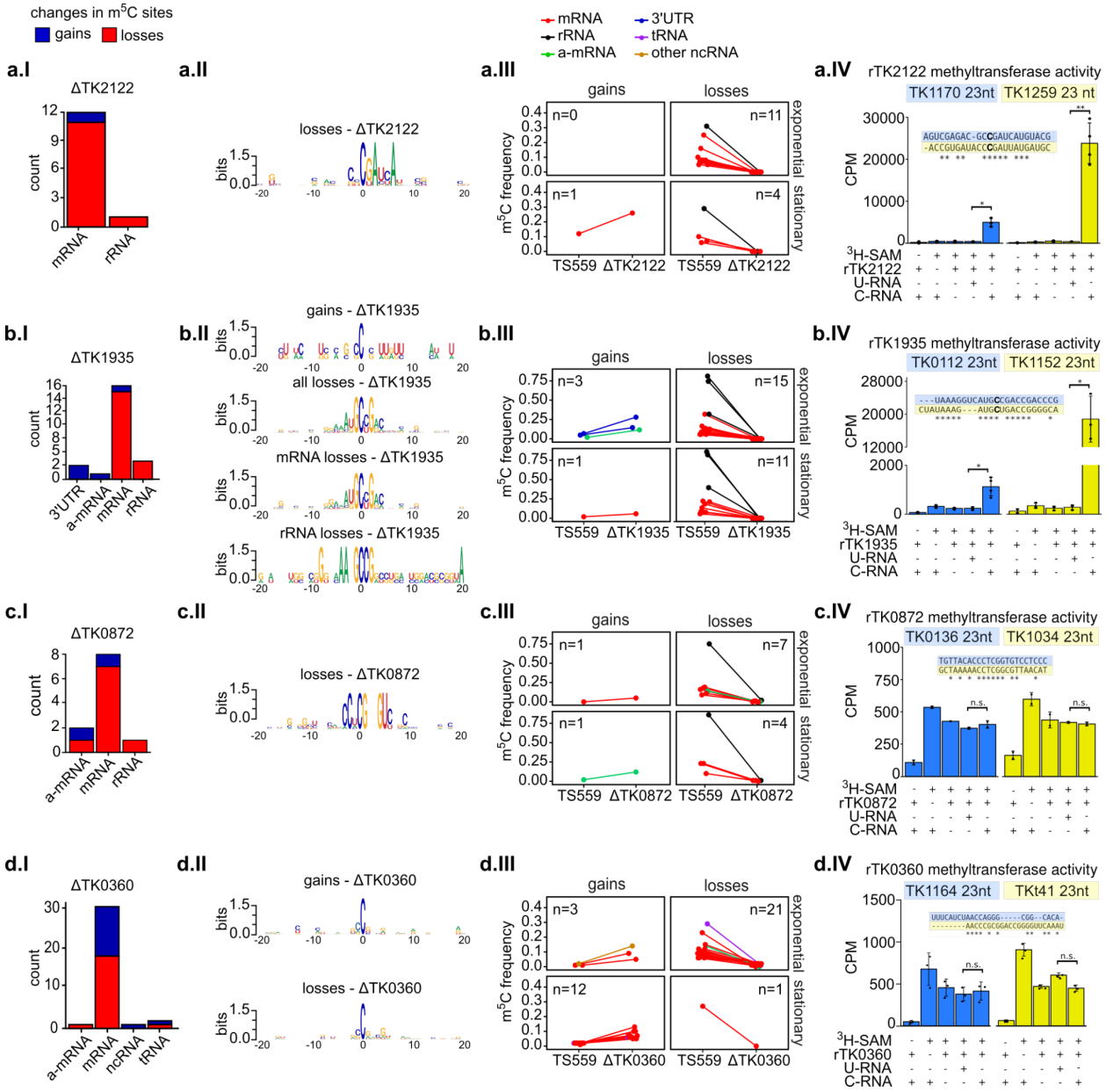
**Supplementary Fig. 6 | Linear regression of modification frequencies between replicates in strains deleted for putative RMTases.** Linear regression and Pearson's correlation coefficients (cor. coeff) were used to compare the site-specific modification frequencies of high confidence and reproducible $m^5C$ sites in each strain across replicates and growth phases. The analysis was performed for sites reproducible. The analysis was performed using R software.


a
65° C
75° C
85° C
95° C
TS559
ΔTK0224
ΔTK0234
ΔTK0234 ΔTK0224
uninoculated
TS559
ΔTK0729
ΔTK0234
ΔTK0234 ΔTK0729
uninoculated
TS559
ΔTK0704
uninoculated
TS559
ΔTK1273
uninoculated
TS559
ΔTK1917
uninoculated
TS559
ΔTK2241
uninoculated
OD600 ± SE
Time (hr)
b
TS559
ΔTK0224
ΔTK0234
ΔTK0234 ΔTK0224
ΔTK0234 ΔTK0729
ΔTK0704
ΔTK0729
ΔTK1273
ΔTK1784
ΔTK1917
ΔTK2241
relative max OD600 ± SE
relative doubling time ± SE
65° C
75° C
85° C
95° C

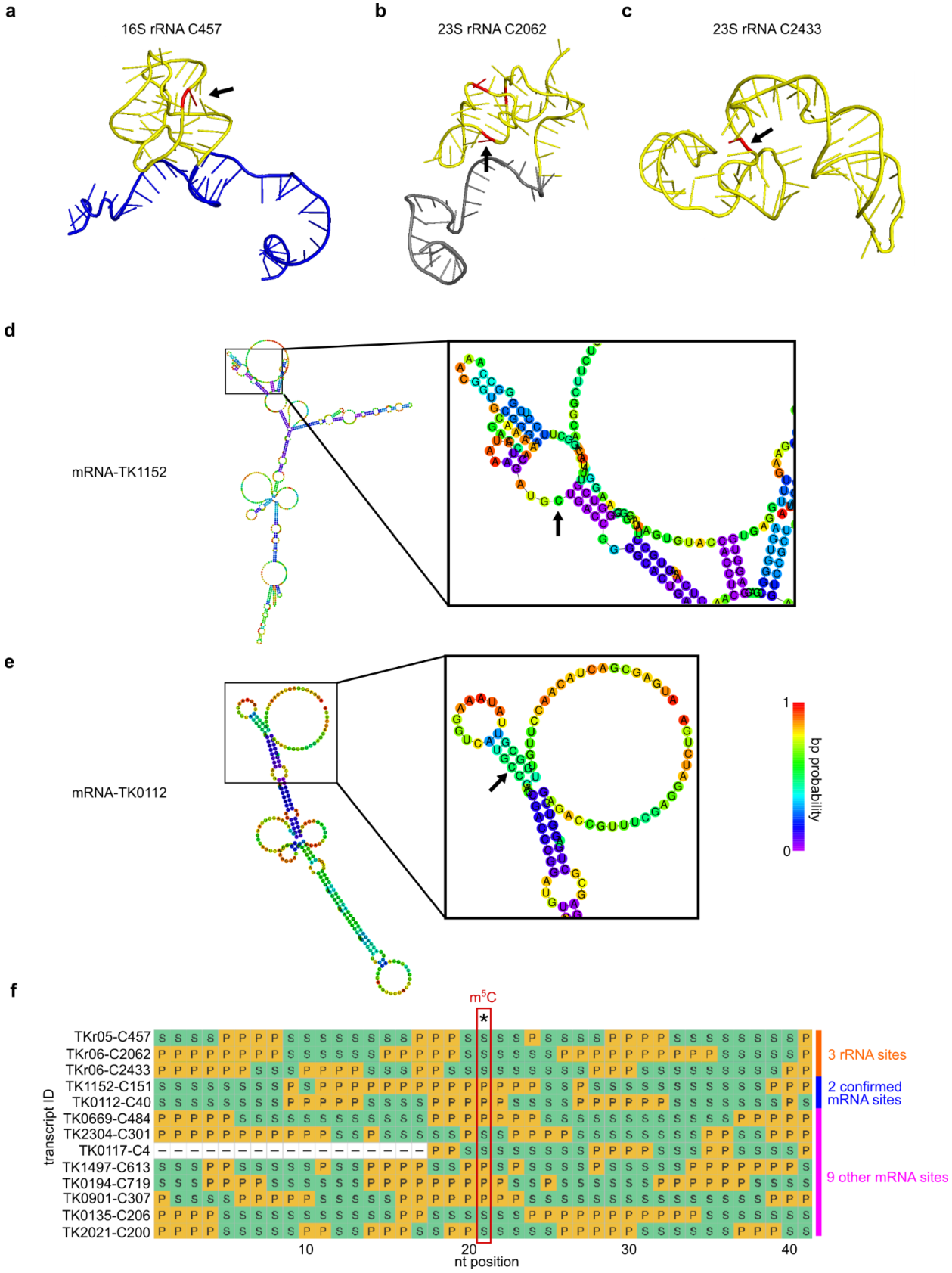
**Supplementary Fig. 7 | Phenotypic analysis of 10 putative RMTases indicate a role for the epitranscriptome in thermophily. a** Head-to-head growth competitions were performed at 65°, 75°, 85°, and 95° C for each strain deleted for a single of 2 putative RMTase and parent strain TS559. **b** The maximum optical density (Max $OD_{600}$) and the rate of growth (doubling time) for each culture is illustrated relative to parent strain TS559. ± 1 standard error (SE) is represented by error bars for n=3 biological replicates. Raw data is provided in Source Data.


a
- rTK2304 (RT:14.63-14.91 min)
(digested 101 nt RNA substrate, unmodified)
+ rTK2304 (RT:14.54-14.92 min)
(digested 101 nt RNA substrate, modified)
b
- rTK2304 (RT:14.63-14.91 min)
digested 101 nt RNA substrate, unmodified)
+ rTK2304 (RT:14.54-14.92 min)
(digested 101 nt RNA substrate, unmodified)
Intensity
Mass (Da)
3207.42
+H2O
+Na
+methyl
3225.43
3229.41
3221.44
c
m5C
258.1->126.1
- rTK2122
U-RNA
+ rTK2122
U-RNA
- rTK2122
C-RNA
+ rTK2122
C-RNA
intensity (x10²)
retention time (min)
- rTK2122 (RT:4.49-4.79 min)
(7303.99 Da substrate)
+ rTK2122 (RT:4.49-4.77 min)
(+14 Da methyl group)
7303.99
7302.98
7321.99
7325.99
7317.99
7320.99
d
- rTK1935
U-RNA
+ rTK1935
U-RNA
- rTK1935
C-RNA
+ rTK1935
C-RNA
- rTK1935 (RT:4.46-4.77 min)
(7406.05 Da substrate)
+ rTK1935 (RT:4.46-4.81 min)
(+14 Da methyl group)
7406.06
7405.05
7424.05
7428.06
7420.05
7424.06
7406.05

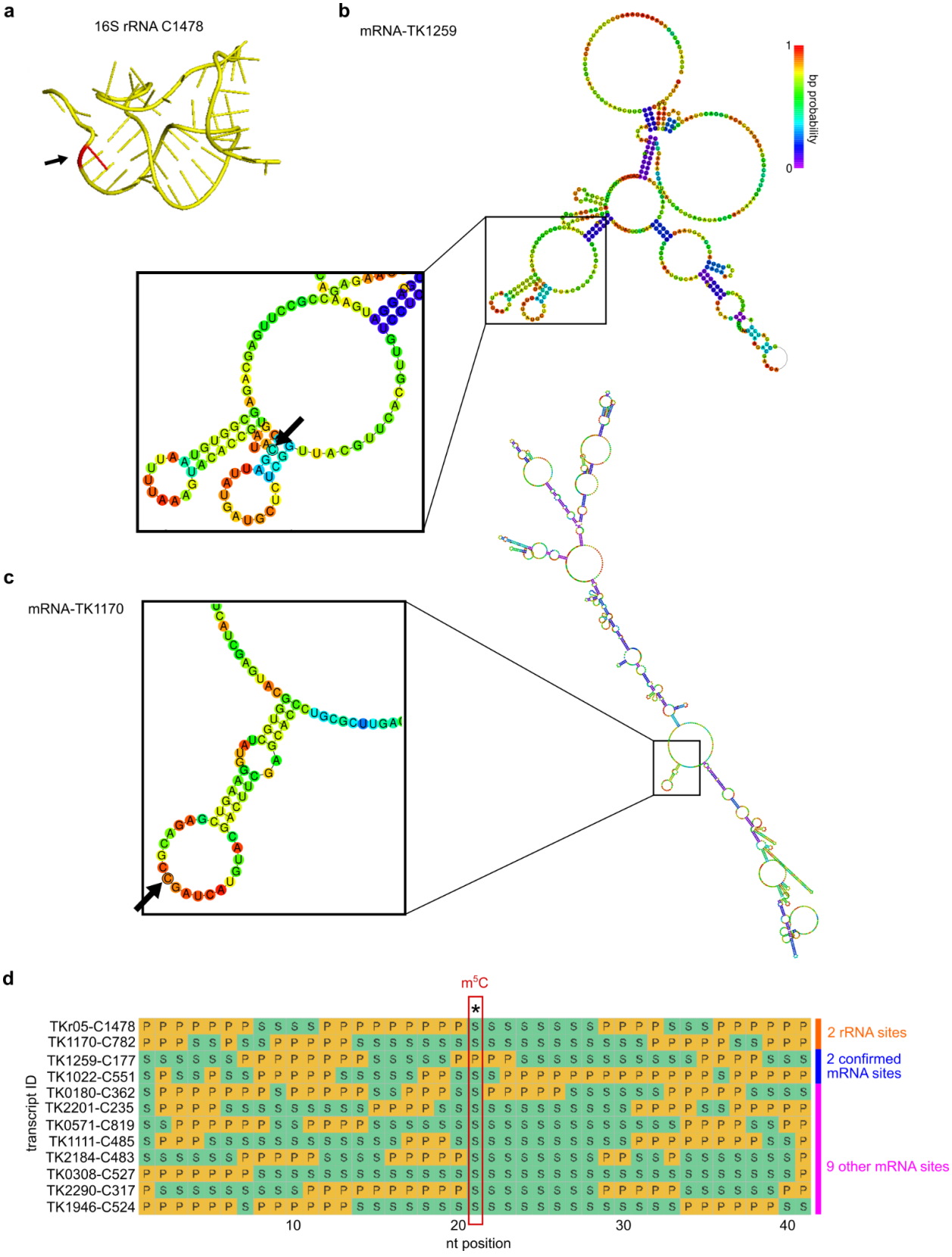
**Supplementary Fig. 8 | Mass spectrometry analysis R5CMTs demonstrate *bonafide* MTase activity.** **a** RNase T1 digestion results in the modified fragment AAC**C**UUAUAG being released from the 101 nt oligonucleotide. The underlined C is the anticipated target cytidine for modification by rTK2304. The mass spectra of the substrate fragments incubated with or without rTK2304 enzyme are shown. n = 1 biological replicate. **b** Mass deconvolution of TK2304-treated 101nt substrate RNaseT1 Product (AACCUUAUAGp, 3207.4265 Da) indicates a mass shift consistent with a methyl group (~14 Da) at the expected cytidine. The relative methylated product abundance is 5.85%. Analysis of a 23 nt RNA substrate modified by **(c)** rTK2122 or **(d)** rTK1935 indicate a mass shift consistent with $m^5C$. n = 2 biological replicates.



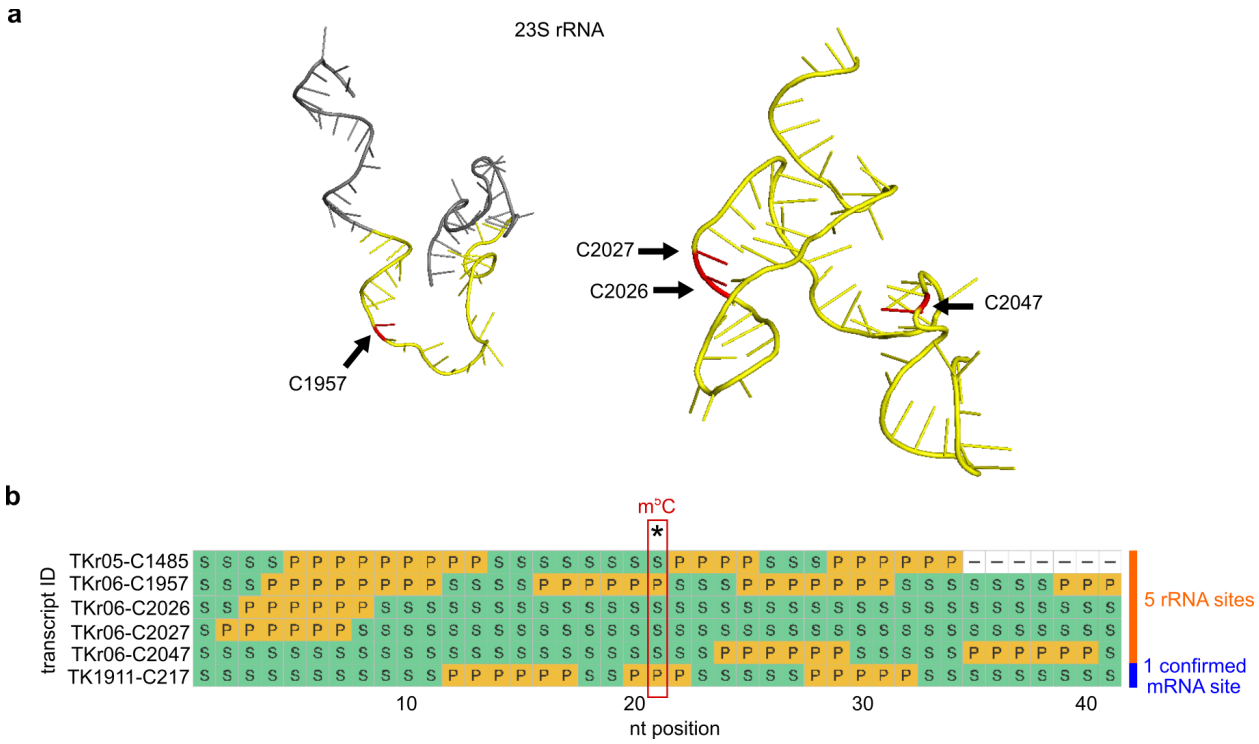

**Supplementary Fig. 9 | Four additional R5CMTs contribute to the maintenance of the m$^5$C epitranscriptome in *T. kodakarensis*.** Genes encoding **(a)** TK2122, **(b)** TK1935, **(c)** TK0872, and **(d)** TK0360 exhibit both gains and losses (subfigures I) in m$^5$C sites upon gene deletion. Sites where losses or gains were detected are defined within sequence contexts (subfigures II). The change in modification frequencies of individual sites between the parent and deletion strains are illustrated (subfigures III). Methyltransferase activity assays were completed for each enzyme using RNA substrates identified *in vivo* to be methylated by the select enzyme (subfigures IV). Methyltransferase activity is measured in counts per minute (CPM). The substrates used are 23 nt oligonucleotides with the target cytidine at the center in the C-RNA. Sequence alignments between the two 23 nt substrates are displayed with stars denoting a match. n = 3 biological replicates. * $p < 0.05$, ** $p < 0.01$, not significant (n.s.). Raw data represented in part IV is provided in Source Data.


a
16S rRNA C457
b
23S rRNA C2062
c
23S rRNA C2433
d
mRNA-TK1152
e
mRNA-TK0112
1
0
bp probability
f
m5C
*
TKr05-C457
TKr06-C2062
TKr06-C2433
TK1152-C151
TK0112-C40
TK0669-C484
TK2304-C301
TK0117-C4
TK1497-C613
TK0194-C719
TK0901-C307
TK0135-C206
TK2021-C200
transcript ID
10
20
30
40
nt position
3 rRNA sites
2 confirmed mRNA sites
9 other mRNA sites

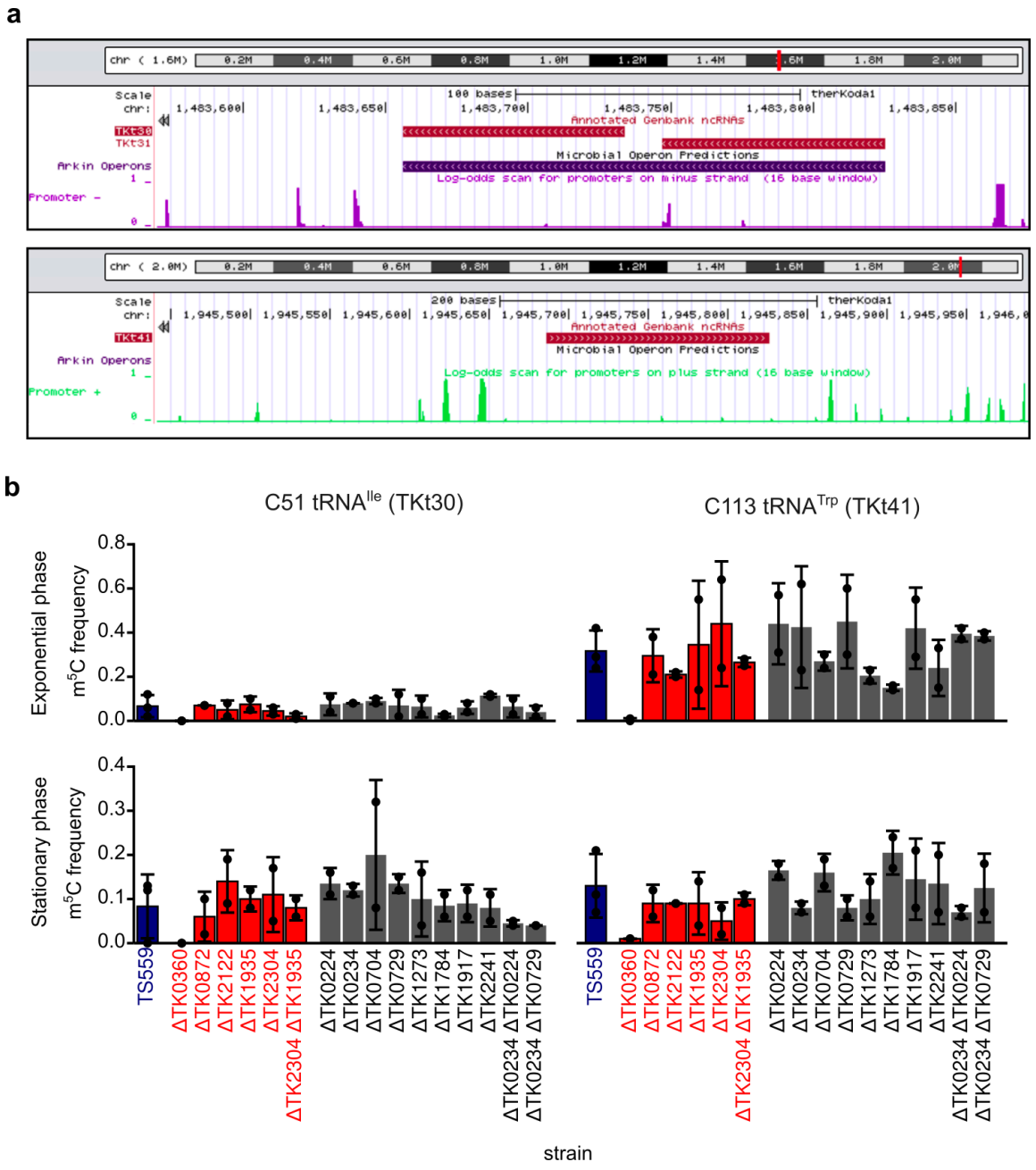
**Supplementary Fig. 10 | Structural analysis of RNAs targeted for methylation by the protein encoded by TK1935. a-c** The tertiary structure of mature rRNAs (CryoEM) where a loss in cytidine retention is detected (red, arrow) after RNA BS-seq of the strain deleted for gene TK1935. **(a)** C456 in the 16S rRNA (TKr05) and **(b)** C2062 and **(c)** C2433 in the 23S rRNA (TKr06) are double stranded. **d, e** The predicted structure of mRNAs that TK1935 methylate as demonstrated in vitro. The cytidine targeted for methylation is indicated by an arrow. **f** For each candidate $m^5C$ site that was lost in the TK1935 deletion strain, the predicted secondary structure is aligned. "P" indicates paired and "S" indicates single stranded nucleotides. The cytidine targeted for methylation is at the center (position 21) and surrounded by 20 nt of up and downstream sequence. In the case of mRNA-TK0117, the $m^5C$ site is the 4th nucleotide, and therefore only 3 nt upstream are included. The transcript ID described the transcript name and transcript coordinate of the cytidine targeted for methylation.


a
16S rRNA C1478
b
mRNA-TK1259
1
bp probability
0
c
mRNA-TK1170
d
m5C
transcript ID
TKr05-C1478
TK1170-C782
TK1259-C177
TK1022-C551
TK0180-C362
TK2201-C235
TK0571-C819
TK1111-C485
TK2184-C483
TK0308-C527
TK2290-C317
TK1946-C524
10
20
30
40
nt position
2 rRNA sites
2 confirmed mRNA sites
9 other mRNA sites

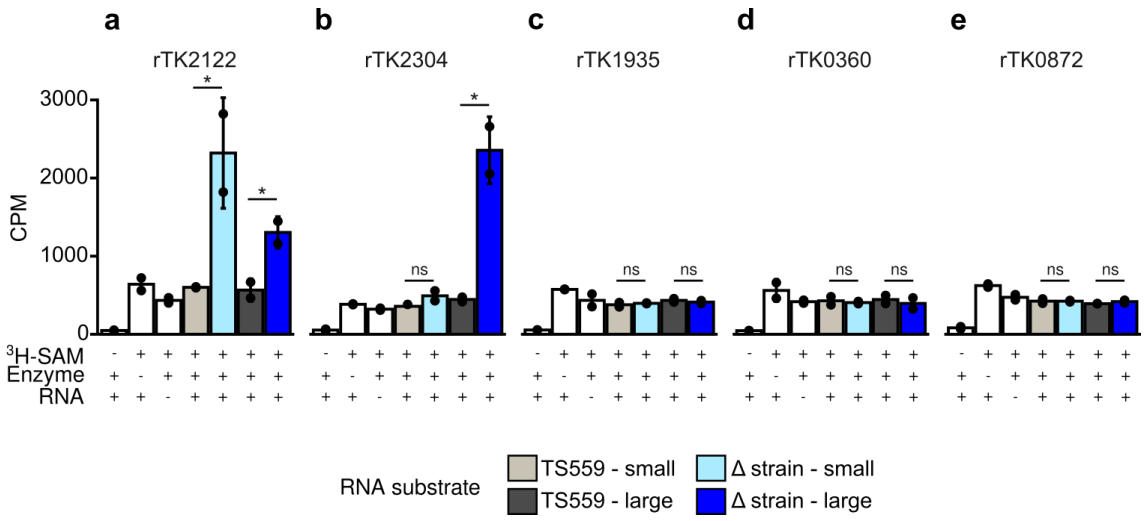
**Supplementary Fig. 11 | Structural analysis of RNAs targeted for methylation by the protein encoded by TK2122. a** The tertiary structure of mature rRNAs (CryoEM) where a loss in cytidine retention is detected (red, arrow) after RNA BS-seq of the strain deleted for gene TK2122. **b, c** The predicted structure of mRNAs that TK2122 methylates as demonstrated in vitro. The cytidine targeted for methylation is indicated by an arrow. **d** For each candidate $m^5C$ site that was lost in the TK2122 deletion strain, the predicted secondary structure is aligned. "P" indicates paired and "S" indicates single stranded nucleotides. The cytidine targeted for methylation is at the center (position 21) and surrounded by 20 nt of up and downstream sequence. The transcript ID described the transcript name and transcript coordinate of the cytidine targeted for methylation.




**Supplementary Fig. 12 | Structural analysis of RNAs targeted for methylation by the protein product encoded by TK2304. a** The tertiary structure of rRNAs (CryoEM) where a loss in cytidine retention is detected (red, arrow) after RNA BS-seq of the strain deleted for gene TK2304. C1957, C2026, C2027 and C2047 in the 23S rRNA (TKr06) are either single or double stranded in the mature ribosome. C1485 in the 16s rRNA (TKr05) is not resolved in the structure. **b** For each candidate $m^5C$ site that was lost in the TK2304 deletion strain, the predicted secondary structure is aligned. "P" indicates paired and "S" indicates single stranded nucleotides. The cytidine targeted for methylation is at the center (position 21) and surrounded by 20 nt of up and downstream sequence. That the predicted secondary structure of the rRNA is not entirely consistent with the CryoEM structure of the ribosome. The transcript ID described the transcript name and transcript coordinate of the cytidine targeted for methylation.




**Supplementary Fig. 13 | TK0360 likely encodes a tRNAs methyltransferase. a** UCSC Archaeal Genome Browser depicting the genomic region of TKt30 (top) and TKt41 (bottom), inclusive of operon predictions and transcription promoter locations. **b** The $m^5C$ modification frequency at C51 and C113 in TKt30 and TKt41, respectively, across strains and growth phases. Strains colored red are *bona fide* (TK1935, TK2122, and TK2304) or putative (TK0360 and TK0872) genes that encode enzymes that generate $m^5C$. Strains colored in gray are putative methyltransferases. Parent strain TS559 is colored in dark blue.




**Supplementary Fig. 14 | Methyltransferase activity on small and large RNA fractions. (a)** rTK2122, **(b)** rTK2304, **(c)** rTK1935, **(d)** rTK0360, and **(e)** rTK0872 where challenged to methylate large (> ~200 nt, dark gray or dark blue) and small (< ~200 nt, light gray or light blue) RNA fractions. Size fractions were isolated from total RNA purified from parent strain TS559 (grays) or the deletion strain (blues). n=2 biological replicates. * $p < 0.05$, not significant (ns). Raw data is provided in Source Data.

**Supplementary Table 1: Differential $m^5C$ sites detected in strains deleted for putative RMTase, not thought to install $m^5C$.**

| | | exponential phase[b] | | stationary phase[c] | | # losses/ gains |
|---|---|---|---|---|---|---|
| Gene ID | Description[a] | rel. gains | rel. losses | rel. gains | rel. losses | |
| TK0224 | class I SAM-dependent MTase, UbiE/COQ5 family | 3 | 0 | 0 | 0 | 0 |
| TK0234 | MTase domain-containing protein | 3 | 0 | 0 | 0 | 5 |
| TK0704 | class I SAM-dependent MTase, UbiE/COQ5 family | 0 | 0 | 0 | 0 | 10 |
| TK0729 | class I SAM-dependent MTase, UbiE/COQ5 family | 1 | 0 | 0 | 0 | 20 |
| TK1273 | class I SAM-dependent MTase, UbiE/COQ5 family | 0 | 0 | 0 | 0 | 30 |
| TK1784 | predicted SAM-dependent MTase, DUF890 family | 1 | 10 | 34 | 0 | |
| TK1917 | class I SAM-dependent MTase | 1 | 0 | 0 | 0 | |
| TK2241 | class I SAM-dependent MTase, UbiE/COQ5 family | 2 | 0 | 20 | 0 | |
| TK0234,TK0224 | double deletion | 10 | 5 | 25 | 1 | |
| TK0234,TK0729 | double deletion | 7 | 3 | 12 | 0 | |

[a] The NCBI description and protein family (Uniprot and InterPro predictions) are provided.

[b,c] BS-seq of individual or double deletion of non-essential RMTases, we identified relative losses and gains (≥ 2X fold change) in $m^5C$ sites between exponential and stationary growth phase cells. The number of $m^5C$ sites is represented by color scale.

**Supplementary Table 2. RNA substrates analyzed by LC-MS/MS**

| Recombinant Enzyme | RNA substrate name | RNA substrate sequence[a] |
|---|---|---|
| rTK2304 | TK1911_23nt_C | UAGCGAAGAAC**C**UUAUAGUUGCU |
| rTK2304 | TK1911_23nt_U | UAGCGAAGAAC**U**UUAUAGUUGCU |
| rTK2304 | TK1911_101nt_C | AGGGCAAGGAGCAGGGCCUUCGUGACAGGUCGGAGAUGAUAGCGAAG AAC**C**UUAUAGUUGCUGGAAAAAUUGAGCCUGAGGGGACAUCUAAAAA GGUUGAG |
| rTK2304 | TK1911_101nt_U | AGGGCAAGGAGCAGGGCCUUCGUGACAGGUCGGAGAUGAUAGCGAAG AAC**U**UUAUAGUUGCUGGAAAAAUUGAGCCUGAGGGGACAUCUAAAAA GGUUGAG |
| rTK2122 | TK1259_23nt _C | ACCGUGAUACC**C**GAUUAUGAUGC |
| rTK2122 | TK1259_23nt _U | ACCGUGAUACC**U**GAUUAUGAUGC |
| rTK1935 | TK1152_23nt_C | CUAUAAAGAUG**C**UGACCGGGGCA |
| rTK1935 | TK1152_23nt_U | CUAUAAAGAUG**U**UGACCGGGGCA |

[a] The cytidine targeted for methylation or the uridine substitution are underlined and bolded.