nature

https://doi.org/10.1038/s41586-020-2418-2

Supplementary information

Dynamic RNA acetylation revealed by quantitative cross-evolutionary mapping

In the format provided by the authors and unedited

Aldema Sas-Chen, Justin M. Thomas, Donna Matzov, Masato Taoka, Kellie D. Nance, Ronit Nir, Keri M. Bryson, Ran Shachar, Geraldy L. S. Liman, Brett W. Burkhart, Supuni Thalalla Gamage, Yuko Nobe, Chloe A. Briney, Michaella J. Levy, Ryan T. Fuchs, G. Brett Robb, Jesse Hartmann, Sunny Sharma, Qishan Lin, Laurence Florens, Michael P. Washburn, Toshiaki Isobe, Thomas J. Santangelo, Moran Shalev-Benami✉, Jordan L. Meier✉ & Schraga Schwartz✉

Supplementary Notes:

Supplementary Note 1. Optimization of a chemical reaction for sequencing N4-acetylcytidine in RNA. To further optimize our previously reported ac^{4}C-sequencing method, we hypothesized that protonation of ac^{4}C may increase its susceptibility to borohydride addition and enable quantitative reduction (**Extended Data Fig. 1a**). Screening several borane reagents, we found that sodium cyanoborohydride ($NaCNBH_3$) rapidly reduced ac^{4}C under acidic, but not neutral, conditions (**Extended Data Fig. 1b**). Reduction of ac^{4}C using $NaCNBH_3$ occurred at rates ~10-fold faster than that of unbuffered $NaBH_4$ used in previous studies, while hydrolysis was relatively unaffected (**Extended Data Fig. 1c,d**). Mass spectrometry (MS) demonstrated that a major product of this reaction is reduced ac^{4}C (**Extended Data Fig. 1e,f**). Reverse transcription of a model RNA containing a single site of ac^{4}C revealed the processive TGIRT reverse transcriptase produced full length cDNA products (**Extended Data Fig. 1g**), which upon Sanger sequencing exhibited higher rates of ac^{4}C-dependent C->T misincorporations using the second generation acidic $NaCNBH_3$ chemistry (**Extended Data Fig. 1h**). Our results are consistent with a model in which improved ac^{4}C reduction kinetics results in increased levels of misincorporation. Together with previously established reactions for chemical deacetylation[12], these experiments provide an optimized set of chemistries for the mutational sequencing of ac^{4}C in RNA.

Supplementary Note 2. Statistical power of ac^{4}C-seq and comparisons to previously published data.

Supplementary Note 2a. Statistical power of ac^{4}C-seq

The ability of ac^{4}C-seq to achieve statistical power to detect an acetylation site is dependent both on sequencing depth and on the stoichiometry of modification. The theoretical relationship between the two is explored in **Extended Data Fig 2h,** in which we assess our theoretical power to detect a modified site on the basis of simulated data, under the assumption that ac^{4}C invariably results in a C->T mismatch. Under this model, to be able to achieve >80% statistical power for calling a site as being modified at a ratio of 20%, a coverage of 40 reads is required; Such coverage is available in our data for 46,512 (28,131 in wild-type) of 433,719 (10.7%, or 6.5% for wild-type) CCG sites in the human transcriptomes studied here (**Extended Data Fig 2h**).

To further explore whether a subset of the sites previously detected by antibody-based mapping[10] may reflect low level acetylation events, we assessed overlapping data between the two experiments. For this analysis, we analyzed C->T misincorporations at sites originally detected by Arango et al, which also have very high coverage in our ac^{4}C-seq data. Pooling together all data generated in either wild-type or Nat10/Thumpd1 overexpression cells, we identified 57 'peak regions' originally identified by Arango et al that also had high coverage by ac^{4}C-seq, with >80% of the cytosines in these peak regions covered by a minimum of 400 reads in our data. Of note, at a coverage of 400 reads, we have 99% power to detect ac^{4}C at a stoichiometry causing misincorporation rates of 5%, and 86% power to detect ac^{4}C at a stoichiometry causing a misincorporation rate of 2%. In other words, assuming these 57 sites are ac^{4}C modified at stoichiometries causing misincorporation rates as low as 5% or 2%, respectively, we expect that in all of them, or 47 of them (respectively) we will identify one cytosine passing our detection thresholds. In practice only 2 sites passed these thresholds, both at sites harboring a 5'-CCG-3' motif. To further increase our sensitivity, we employed a threshold-free approach in which we merely identified, for each of these 57 sites, the cytosine harboring the highest levels of C->T misincorporations, without demanding statistical significance. Using these highly relaxed detection criteria, we again identified only the above 2 sites with C->T misincorporation percentages exceeding 2% **(Extended Data Fig. 2f)**. Inspection of the raw data revealed these

two sites originated from the overexpression sample, suggesting ac4C is absent at these sites under physiological conditions. Finally, we focused on a set of 3 precisely identified sites highlighted in Figure 7E of the previous study which harbor a 'CXX' motif and have very high coverage in our data (>400 reads; **Extended Data Fig. 2g**). Performing a focused analysis of all of the C's at these precisely defined sites, we again did not identify any cytidine residues with C->T misincorporation rates exceeding ~0.4% **(Extended Data Fig. 2g)**.

This analysis of deeply sequenced sites finds no evidence for low levels of endogenous ac4C within these previously identified antibody-defined peaks and, together with literature precedent and the comprehensive validation of our method using multiple orthogonal approaches (see **Supplementary Note 2b**), strongly disfavors the hypothesis that the discrepancy between the two studies reflects a technical limitation caused by the sequencing depth of our ac4C-seq data. Nonetheless, several limitations should be acknowledged: First, the above-noted statistical considerations, which are a general feature of all modification detection approaches relying on sequencing, mean that we are particularly limited in detecting low stoichiometry sites in mRNAs expressed at low levels (see also **Supplementary Note 2b**). Second, our study focused on those sites most deeply sampled by ac4C-seq, rather than those most strongly enriched by acRIP. Finally, the above analyses were conducted in HEK-293T cells (for which we had collected deeper sequencing data) whereas the original ac4C peaks were identified in HeLa cells, and it is possible there is biological variability in ac4C topologies between these models.

Supplementary Note 2b. Comparing validation of ac4C sites in the two studies and technical challenges of antibody and LC-MS analyses of mRNA modifications. In ac4C-seq, acetylated sites are called 'de novo' using a statistical test that evaluates C->T misincorporations in treated samples relative to mock-treated controls. Newly identified sites were then subjected to a comprehensive set of orthogonal, site-selective validations including 1) chemical validation: ac4C sites are deacetylated by mild alkali pre-treatment, 2) genetic validation: ac4C sites are responsive to Nat10/Kre33 deletion or overexpression, 3) biochemical validation: ac4C sites occur in a 5'-C**C**G-3' consensus sequence conserved in eukaryotes and archaea, 4) mass spectral (MS) validation: ac4C sites can be identified in endogenous rRNA sequences by LC-MS, 6) conservation: ac4C sites and stoichiometry are conserved between related species, and 7) structural validation: ribosomal ac4C sites can be visualized in density maps generated via high resolution cryo-EM analysis. These combined orthogonal lines of evidence provide high confidence that the new candidate ac4C sites identified are, indeed, acetylated.

The need for orthogonal validation is particularly crucial, as it has been repeatedly demonstrated that genomic mapping efforts can be highly misleading when relying exclusively on antibodies raised against RNA modifications. For example, we have previously shown that >1300 highly specific and reproducible 'peaks' are detected when applying an anti-m6A antibody to yeast mRNA, even in strains completely devoid of m6A [79]. Also the original reports pertaining to widespread presence of N1-methyladenosine (m1A) on mRNA are meanwhile understood to have been due to cross-reactivity of the antibody with the 5' cap of mRNA [24,26,80]. It should be emphasized that the antibodies raised against m1A, m6A and ac4C are all quite specific and sensitive, as can be evaluated based on classical characterization assays. However, as a modification becomes more rare, even subtle, non-specific antibody binding at a low *fraction* of sites in the transcriptome has the potential to result in a huge *number* of false positive sites, due to the huge size of eukaryotic transcriptomes.

In addition to antibody-based enrichment, previously published studies provide two lines of evidence supporting the presence of acetylation in eukaryotic mRNA: First, ac4C was detected in

bulk mRNA based on mass-spectrometry.[10,13] Second, 'ac4C peaks' identified by antibody-based mapping were sensitive to Nat10 depletion.[10] The following alternative interpretations of these data are, in our view, possible:

- With respect to the bulk LC/MS data, previous experience in the field has revealed that precisely measuring levels of rare modifications in mRNA can be challenging due to "contaminants" in the form of tRNA and rRNA (Legrand et al. 2017). Specifically, when a modification is sufficiently abundant in highly expressed tRNA and rRNA, and rare in more modestly expressed mRNA, the signal:noise issue becomes very difficult to overcome. m1A provides an important antecedent for this. This modification was originally estimated based on mass-spectrometry to be relatively abundant on mRNA at levels of 0.04%[81], yet based on genomic approaches this modification - even under the most optimistic scenarios - is far more rare[24–26,80,82,83] and the previously reported sites were shown to be due to antibody cross-reactivity of the anti-m1A antibody[26]. In retrospect the original estimations are likely to reflect contamination by rRNA and tRNA fractions, forming ~95% of the total RNA, and that are hence very difficult to completely eliminate. With regards to LC-MS-based quantification of ac4C specifically, we have observed that ac4C is prone to non-enzymatic chemical deacetylation in aqueous solution, even when stored at reduced temperatures (e.g. -20 °C). In our experience, this can cause degradation of ac4C internal standards, hindering calibration of ac4C standard curves, and lead to overestimation of the absolute levels of ac4C present in bulk or poly(A)-enriched RNA (note that this does not affect comparisons of relative levels of ac4C between experimental conditions, as reported in this manuscript). To the best of our knowledge this issue was not explored in previous studies, and highlights an additional challenge presented by the absolute quantification of this hydrolytically-labile modified RNA nucleobase by LC/MS.

- With respect to the genetic data, it is critical to point out that while the previous study *did* make use of a Nat10 knockdown cell line, this cannot be interpreted as a source of orthogonal validation, as the reduced levels in the Nat10 depleted cells were a *prerequisite* for sites to be *defined* as putative ac4C sites. Thus, the reduced levels observed in the knockout strain is a direct result of the fact that these sites were computationally filtered to have such reduced levels. In contrast, in our study we identified ac4C sites *exclusively* on the basis of our measurements of ac4C levels in WT cells (comparison of treated to untreated controls), and we could subsequently obtain measurements in Nat10 knockdown/knockout cells and confirm their elimination. In the latter case - but not in the former - elimination in the Nat10 knockout cells provides orthogonal, genetic validation.

Supplementary Note 3. Temperature-dependent patterns of ac4C in rRNA, tRNA, ncRNA and mRNA.

Below we characterize the distinct, temperature-dependent patterns of modifications observed for different classes of RNA molecules in *T. kodakarensis*:

- **rRNA**: We identified 172 ac4C sites at CCG motifs spread across the 16S, 23S, and 5S ribosomal RNAs. The stoichiometry of the majority of these sites was highly regulated in a temperature-dependent manner across both subunits (**Fig. 3a**). The three-dimensional arrangement and association of these sites with functional locations in the ribosome is characterized further in the main text via cryo-EM structural analysis.

- **tRNA**: We identified 76 ac^{4}C sites at CCG motifs distributed across nearly every tRNA in *T. kodakarensis*. The number of sites and their distribution by far exceeds all of the previous reports of ac^{4}C, in which ac^{4}C was restricted to a very limited number of sites within a limited number of tRNAs (**Fig. 3c** and **Extended Data Fig. 6d,e**). Overall 13 positions were found to be modified in 37 tRNA molecules. Each tRNA molecule was modified in 1-4 distinct positions. All modifications occurred at duplex RNA in stem structures, and displayed elevated stoichiometry as growth temperature increased.

- **ncRNA**: We identified 35 sites distributed across two highly structured ncRNA molecules: RNAse P noncoding RNA and SRP RNA. Ac^{4}C stoichiometries were relatively high in both, reaching up to 68% and increased with culture temperature (**Extended Data Fig. 6f**). The high density of ac^{4}C in these molecules, the relatively high stoichiometry, and the conservation of ac^{4}C in these two ncRNAs also in additional *Thermococcales* species (**Fig. 2f,g** and **Extended Data Fig. 6h**) are all reminiscent of the modification profiles in tRNA and rRNA in these species, and thus suggest that ac^{4}C may have acquired specific functions within these ncRNAs.

mRNA: Overall 119 sites of ac^{4}C-modification were identified in mRNA. These sites were typically modified at very low stoichiometries at 55˚C, but increased with temperature and peaked at 95˚C (8.3% median) (**Fig. 3a**). While ac^{4}C-modifications were almost exclusively (99%) restricted to C**C**G motifs, these sites were not enriched within any particular codon (**Extended Data Fig. 6g**), nor within a particular reading frame (**Extended Data Fig. 6h**).

Supplementary Data 1. The extracted mass chromatograms for the quantitation of acetyl-4-cytidine (ac4C) in T. kodakarensis rRNAs. The RNase T1 digests of T. kodakarensis rRNAs (450 ng) were analyzed by the nanoflow LC-MS performed as described in Methods. The major peaks in the chromatogram are assigned to the fragments containing ac4C (black arrows) and the corresponding fragments without ac4C (blue arrows) (see Supplementary Table 3). The positions of ac4C in the sequences of each RNase T1 fragment are indicated on the left with the m/z values (±5 ppm mass tolerance). The strains used and the growth temperatures are indicated on the top. Y axis shows the peak intensity relative to the most intensive peak in each set of panels, whose intensity is labeled in the top panel. The peak areas in Supplementary Data 1 were used for quantitation of ac4C as described in Supplementary Table 3.

Supplementary Data 1. The extracted mass chromatograms for the quantitation of acetyl-4-cytidine (ac4C) in T. kodakarensis rRNAs. Continued.

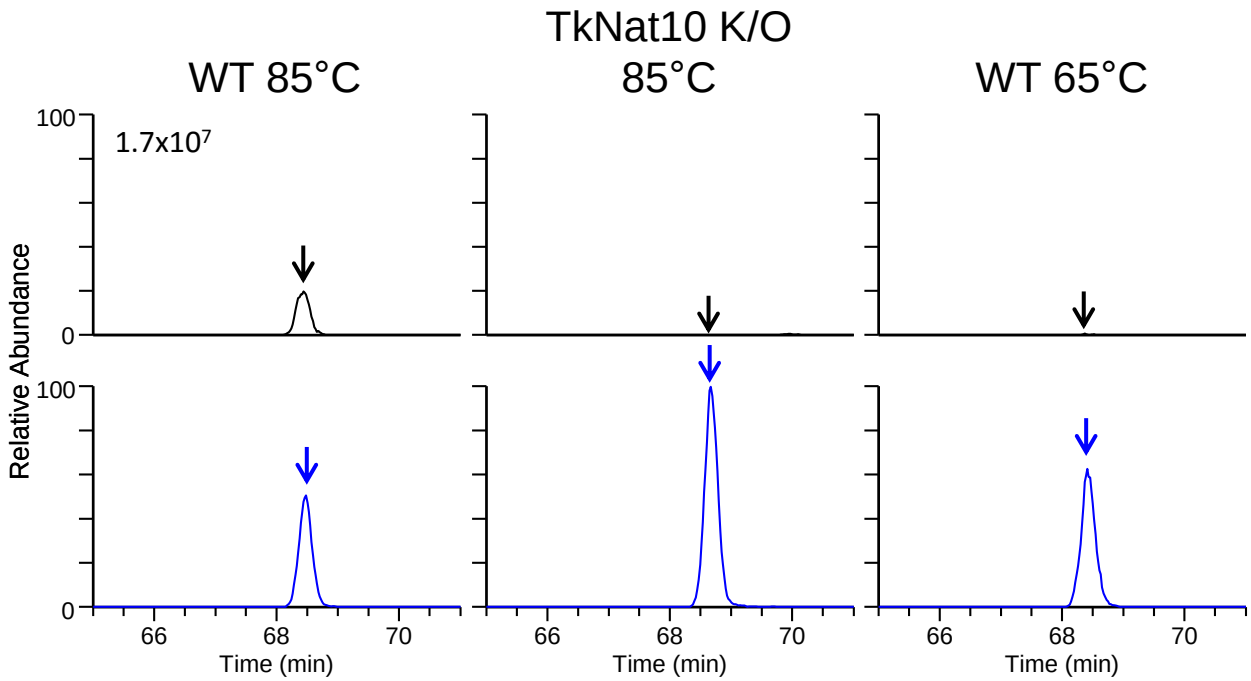
ac^{4}C32 in 5S rRNA

CCACACC(ac^{4}C)Gp

m/z 963.467, 3-

CCACACCCGp

m/z 949.463, 3-



Supplementary Data 1. The extracted mass chromatograms for the quantitation of acetyl-4-cytidine (ac4C) in T. kodakarensis rRNAs. Continued.

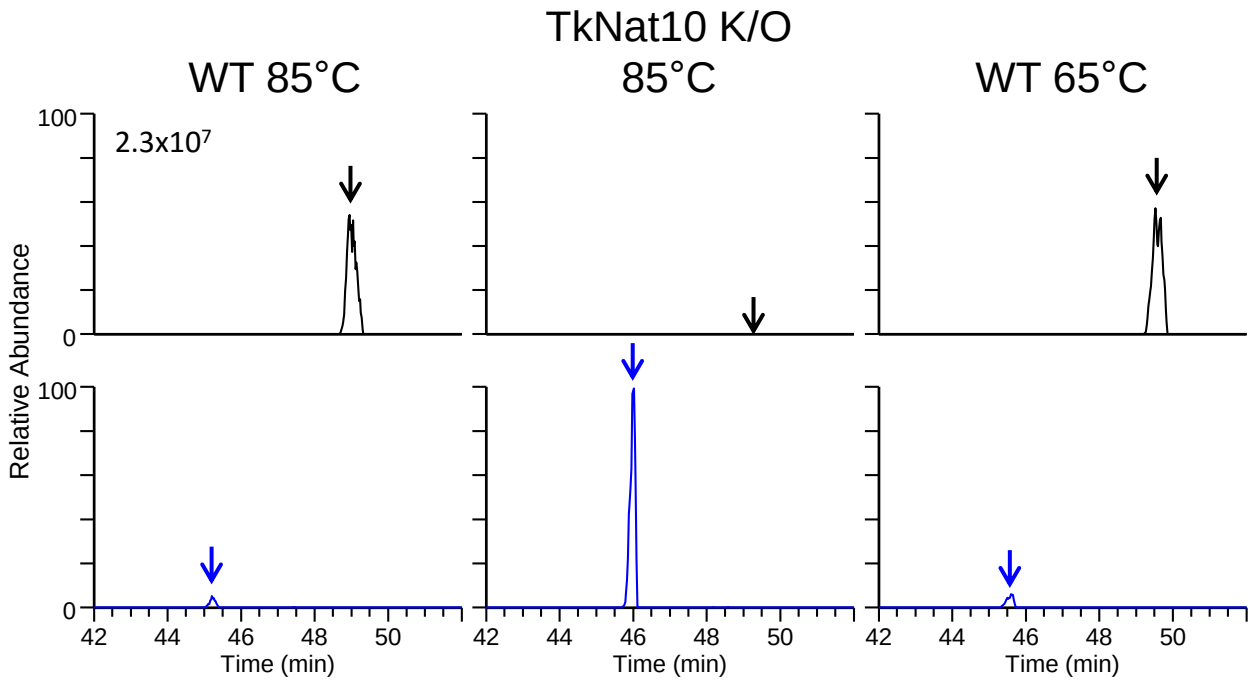
ac^{4}C5 in 16S rRNA

pAUUC(ac^{4}C)Gp

m/z 1017.103, 2-

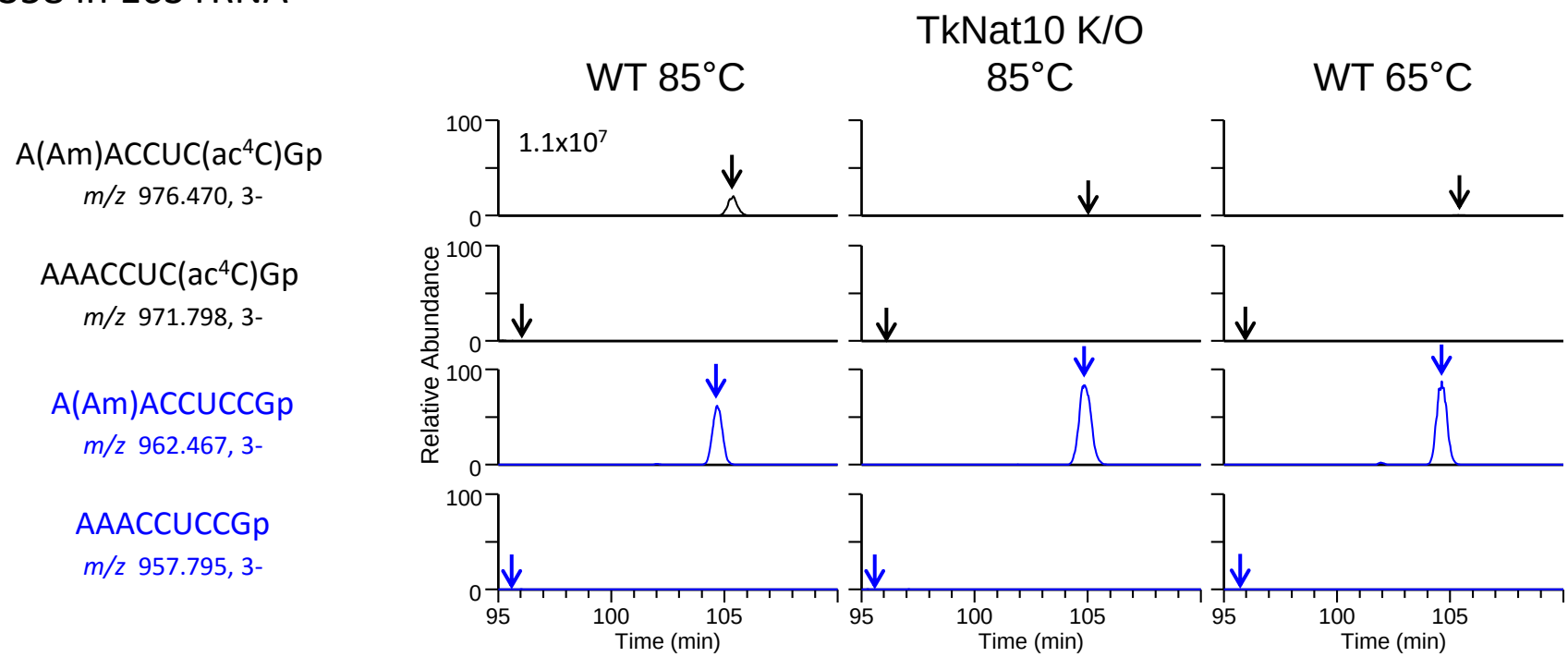
pAUUCCGp

m/z 996.098, 2-



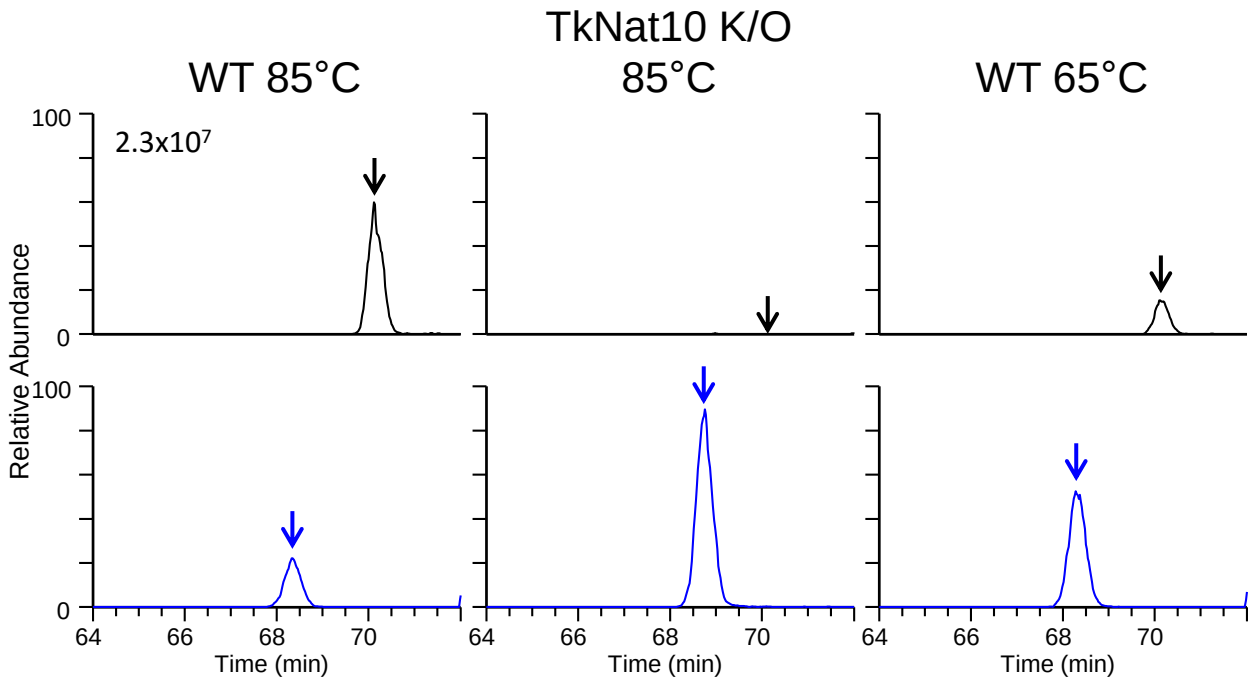
Supplementary Data 1. The extracted mass chromatograms for the quantitation of acetyl-4-cytidine (ac4C) in T. kodakarensis rRNAs. Continued.

ac^{4}C358 in 16S rRNA



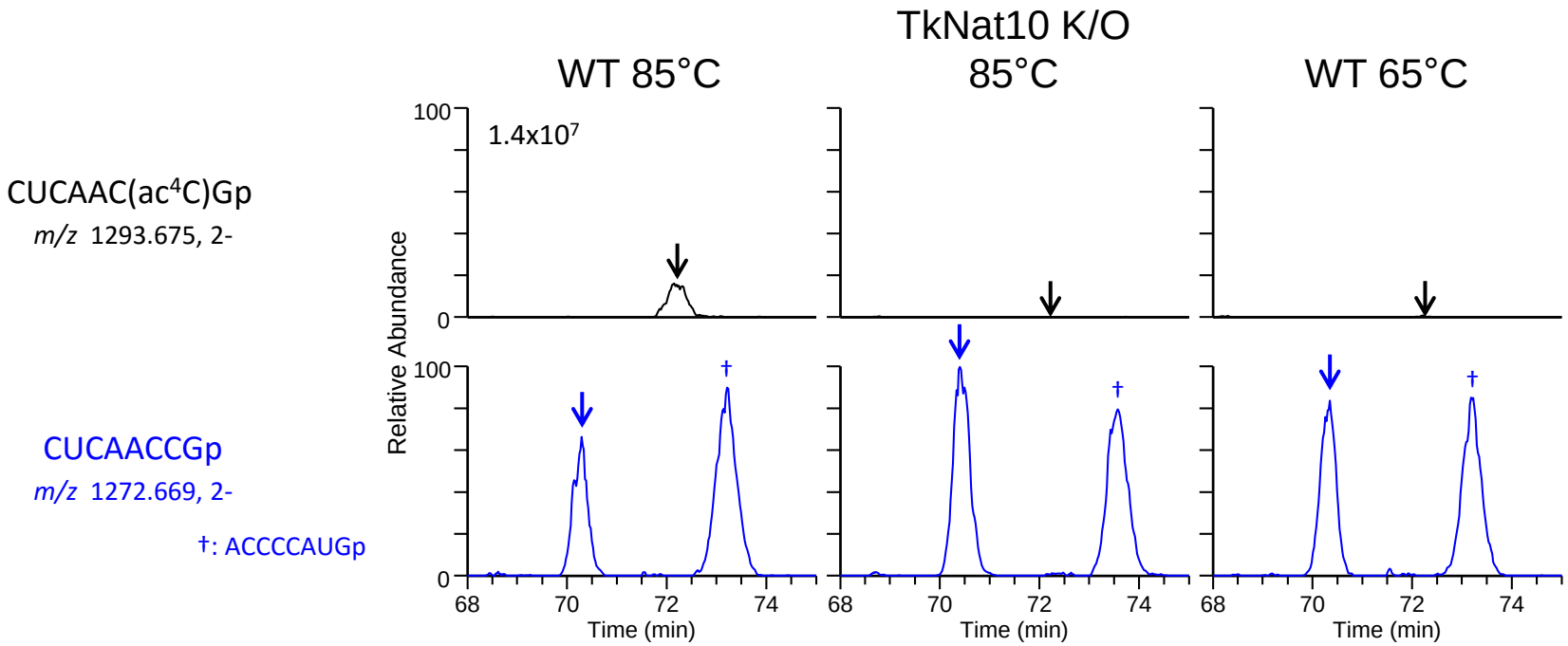
Supplementary Data 1. The extracted mass chromatograms for the quantitation of acetyl-4-cytidine (ac4C) in T. kodakarensis rRNAs. Continued.

ac^{4}C376 in 16S rRNA



Supplementary Data 1. The extracted mass chromatograms for the quantitation of acetyl-4-cytidine (ac4C) in T. kodakarensis rRNAs. Continued.

ac^{4}C569 in 16S rRNA



Supplementary Data 1. The extracted mass chromatograms for the quantitation of acetyl-4-cytidine (ac4C) in T. kodakarensis rRNAs. Continued.

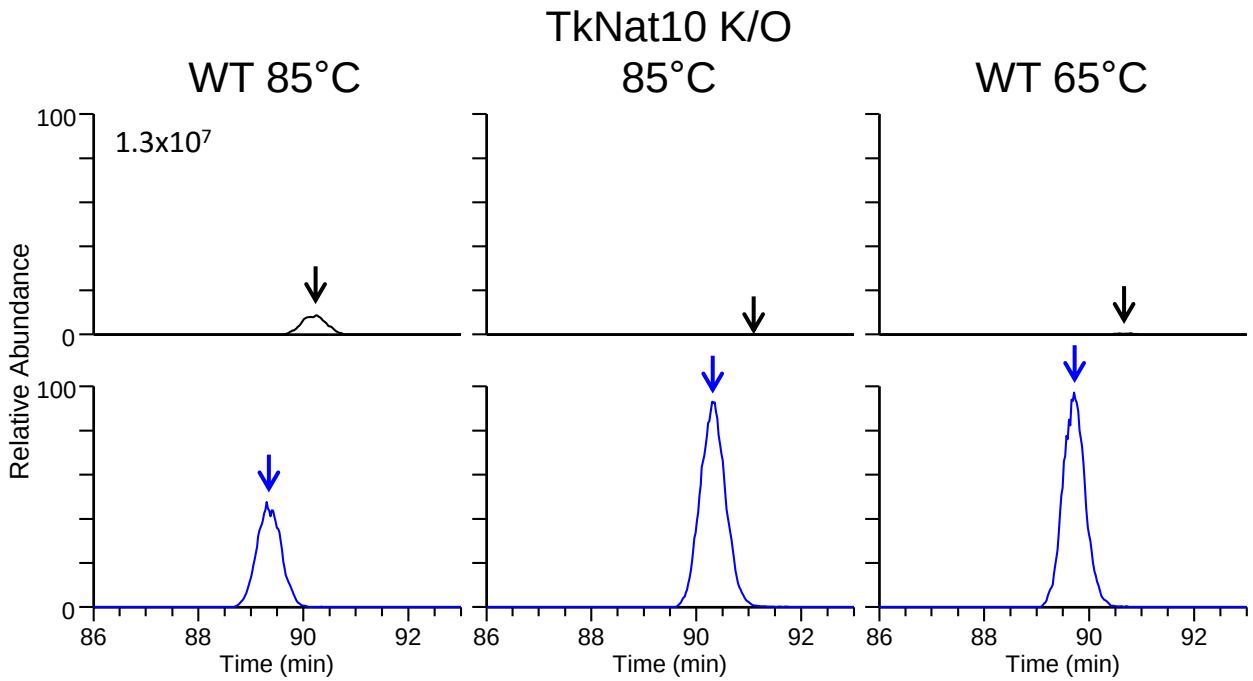
ac^{4}C936 in 16S rRNA

AACCUCAC(ac^{4}C)Gp

m/z 1073.479, 3-

AACCUCACCGp

m/z 1059.475, 3-



Supplementary Data 1. The extracted mass chromatograms for the quantitation of acetyl-4-cytidine (ac4C) in T. kodakarensis rRNAs. Continued.

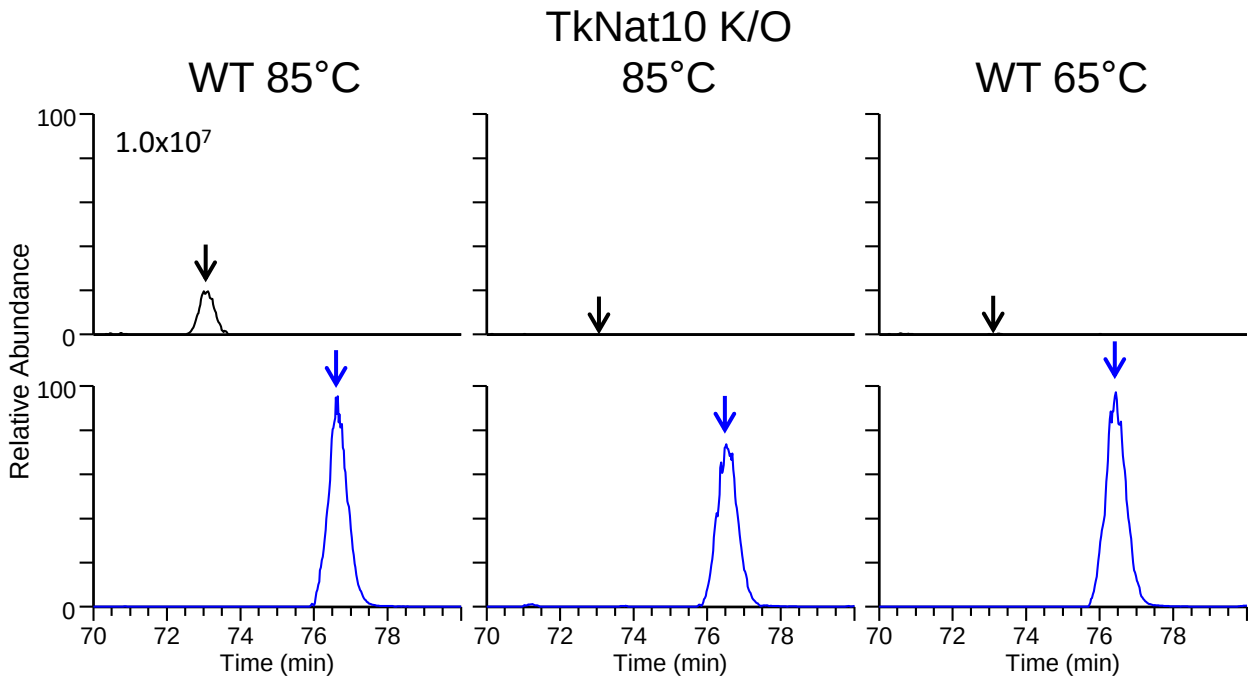
ac^{4}C138 in 23S rRNA

ACCUCC(ac^{4}C)Gp

m/z 1281.669, 2-

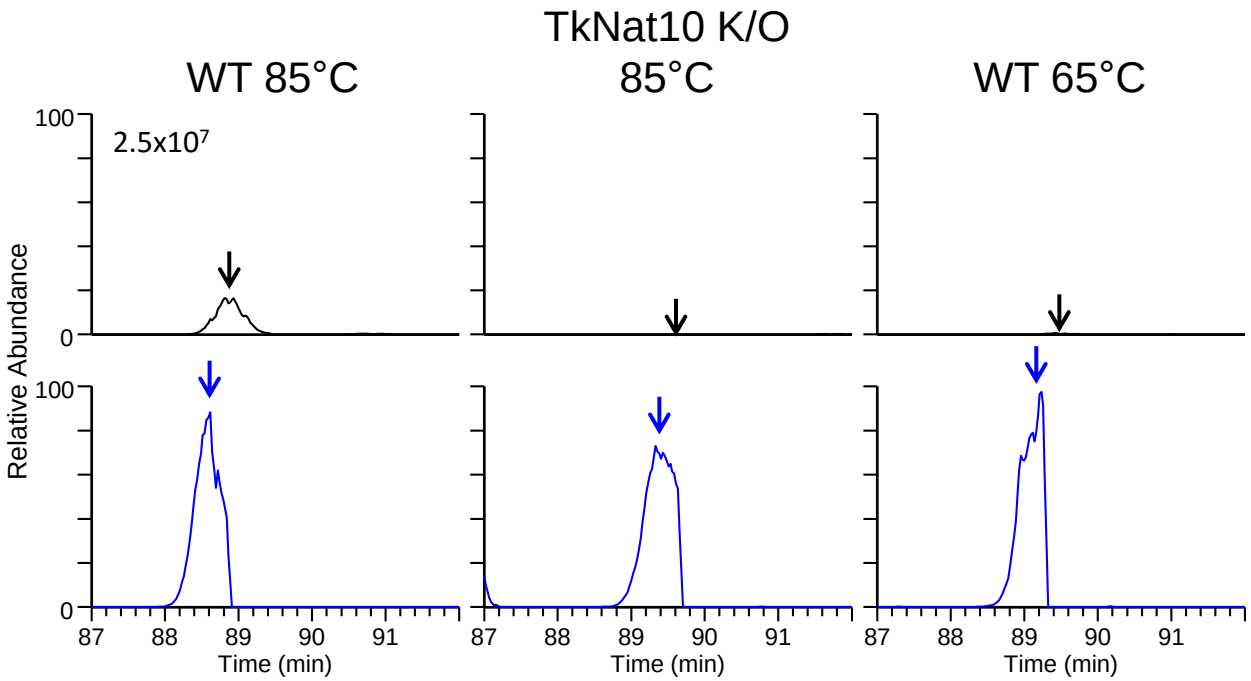
ACCUCCCGp

m/z 1260.664, 2-



Supplementary Data 1. The extracted mass chromatograms for the quantitation of acetyl-4-cytidine (ac4C) in T. kodakarensis rRNAs. Continued.

ac^{4}C433 in 23S rRNA



Supplementary Data 1. The extracted mass chromatograms for the quantitation of acetyl-4-cytidine (ac4C) in T. kodakarensis rRNAs. Continued.

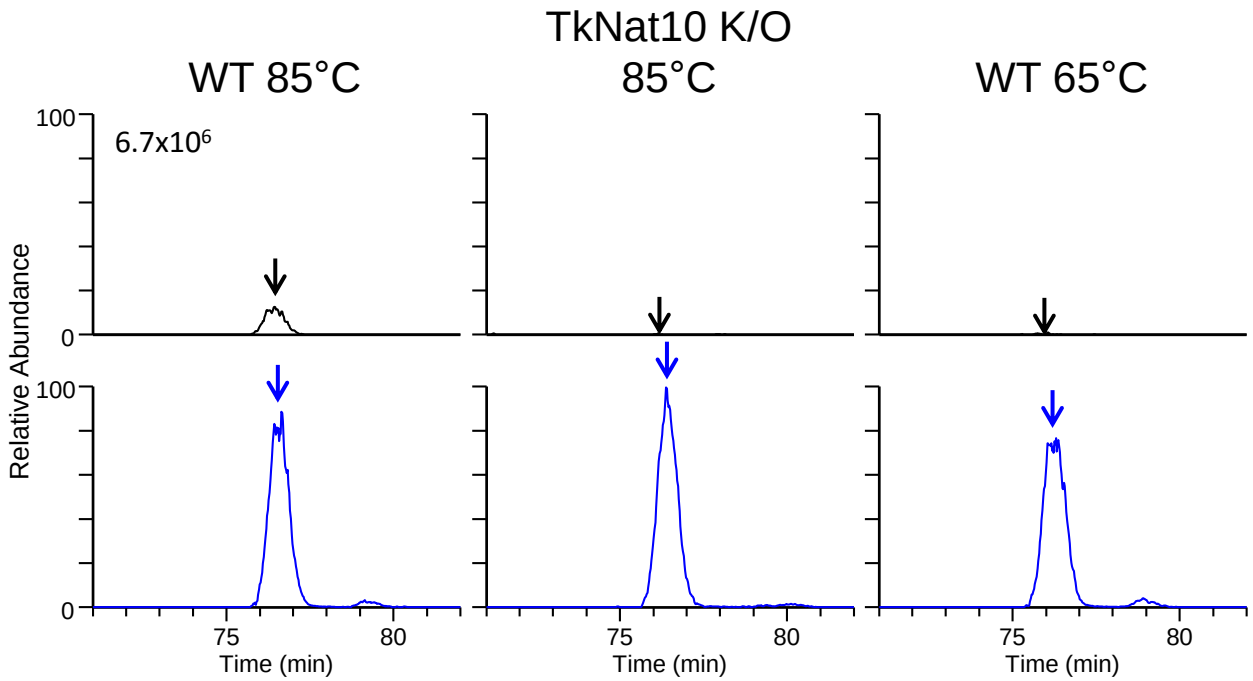
ac^{4}C713 in 23S rRNA

UCAACCC(ac^{4}C)Gp

m/z 963.795, 3-

UCAACCCCGp

m/z 949.791, 3-



Supplementary Data 1. The extracted mass chromatograms for the quantitation of acetyl-4-cytidine (ac4C) in T. kodakarensis rRNAs. Continued.

ac^{4}C1026 in 23S rRNA

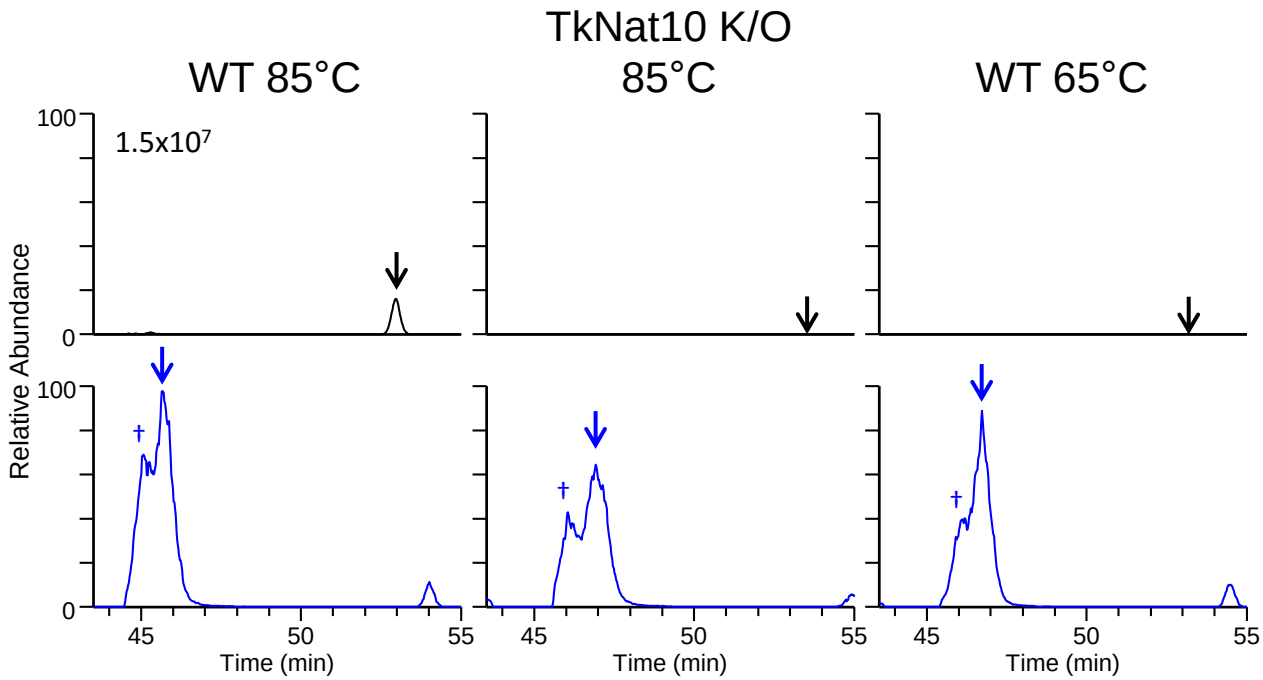
CCUC(ac^{4}C)Gp

m/z 964.622, 2-

CCUCCGp

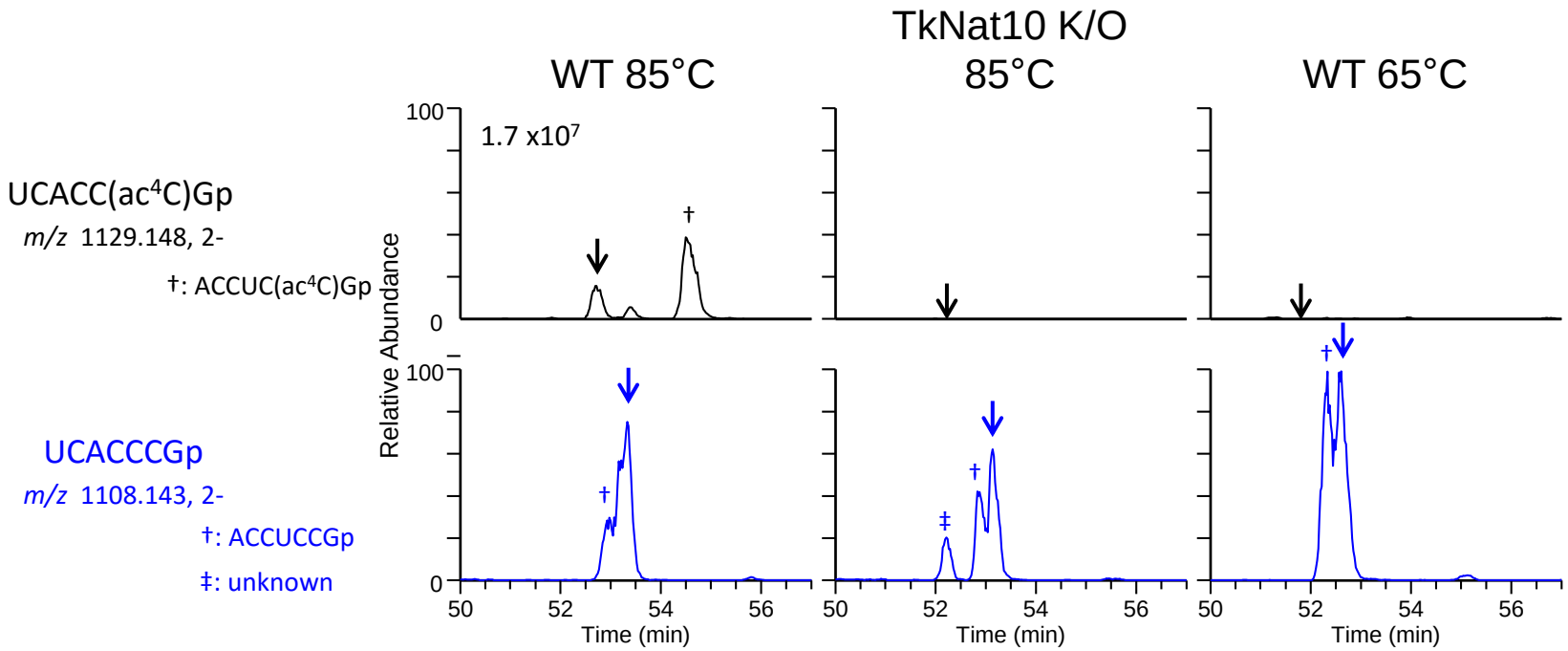
m/z 943.617, 2-

†: CUCCCGp



Supplementary Data 1. The extracted mass chromatograms for the quantitation of acetyl-4-cytidine (ac4C) in T. kodakarensis rRNAs. Continued.

ac^{4}C1099 in 23S rRNA

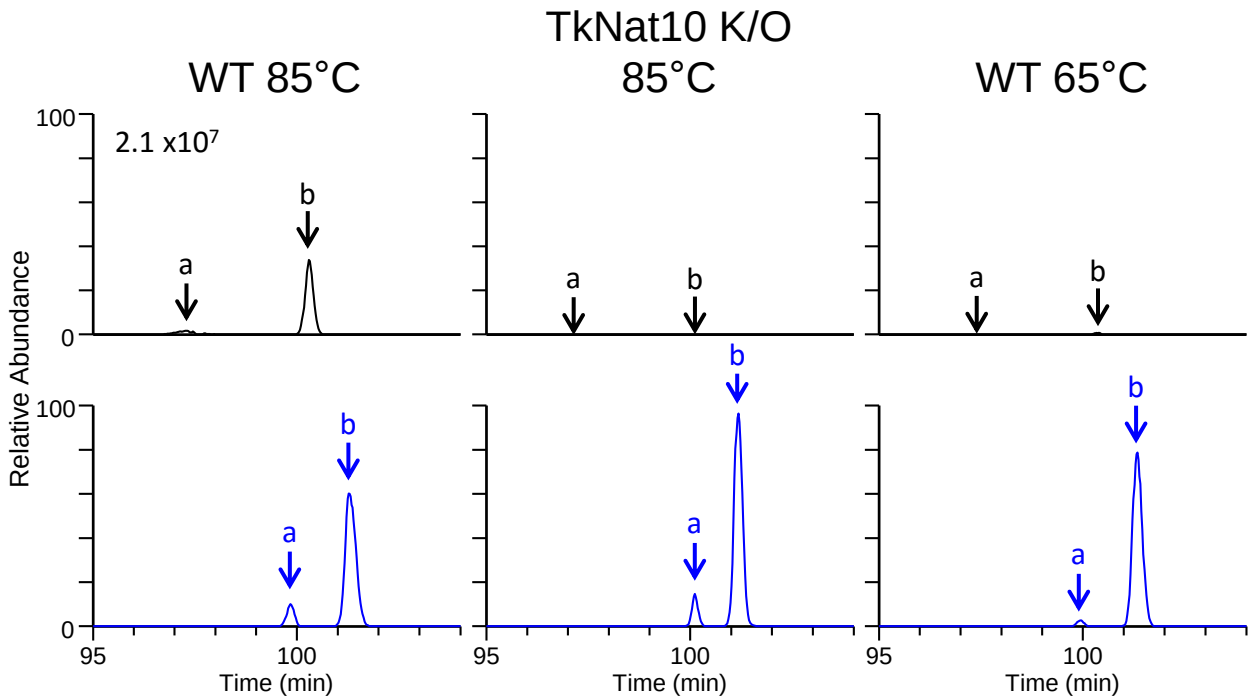


Supplementary Data 1. The extracted mass chromatograms for the quantitation of acetyl-4-cytidine (ac4C) in T. kodakarensis rRNAs. Continued.

ac^{4}C1538 in 23S rRNA

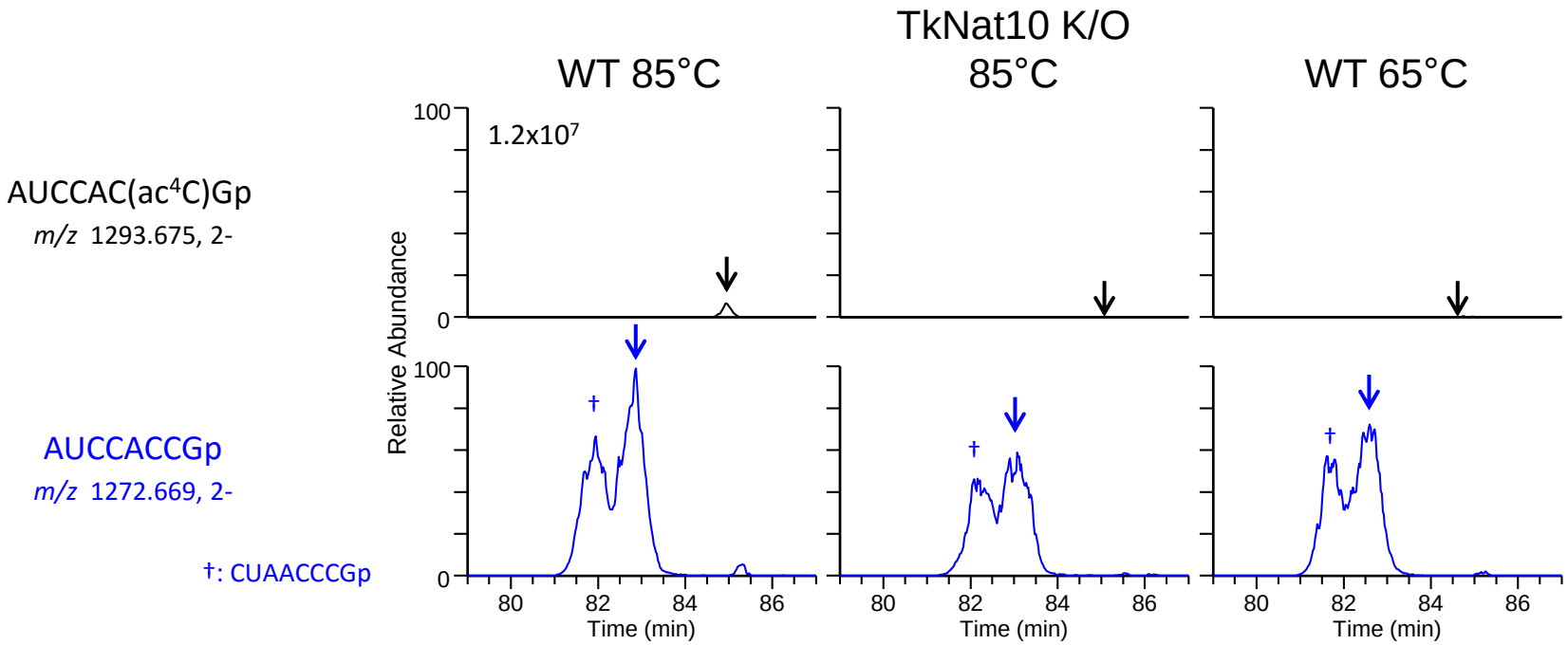
a. UUUAUAUUCC(ac^{4}C)Gp
(with Ψ signature)
b. UUUAUAUUCC(ac^{4}C)Gp
m/z 1270.476, 3-

a. UUUAUAUUCCCGp
(with Ψ signature)
b. UUUAUAUUCCCGp
m/z 1256.472, 3-



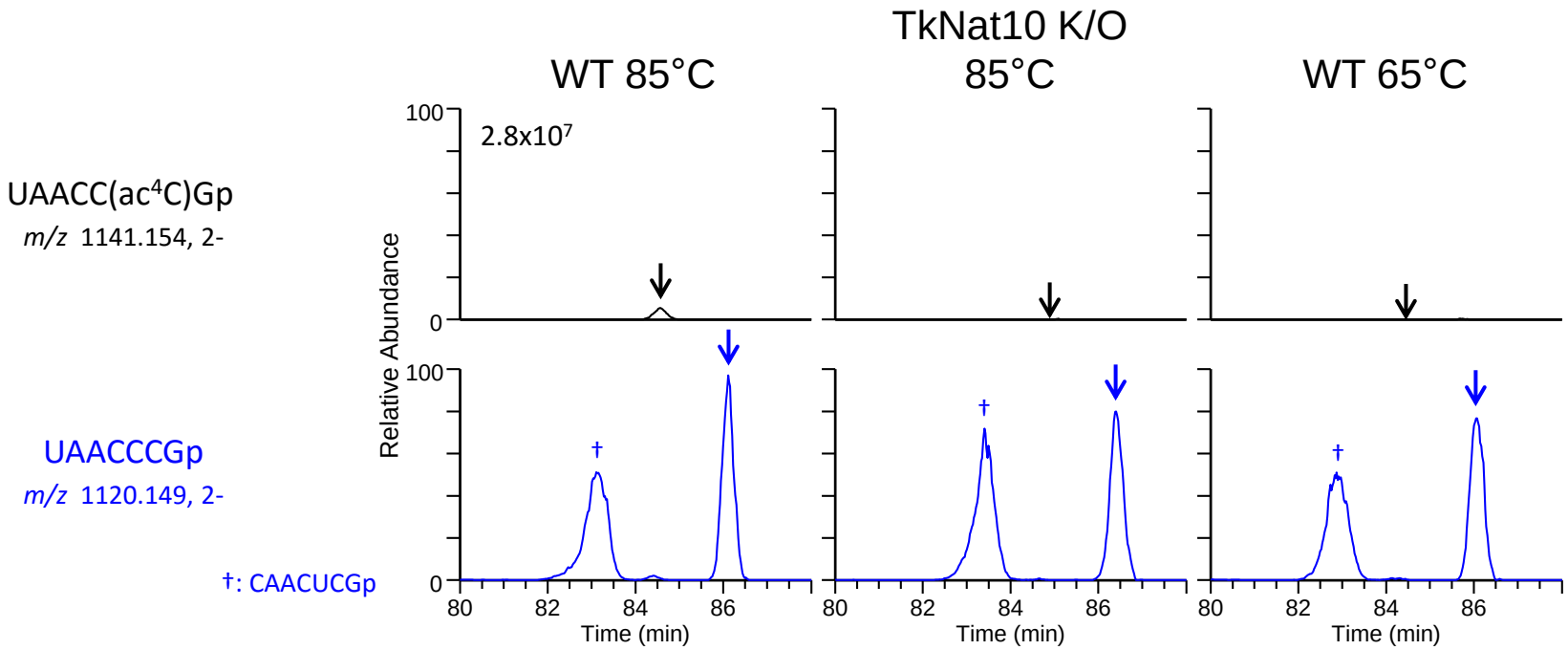
Supplementary Data 1. The extracted mass chromatograms for the quantitation of acetyl-4-cytidine (ac4C) in T. kodakarensis rRNAs. Continued.

ac^{4}C1739 in 23S rRNA



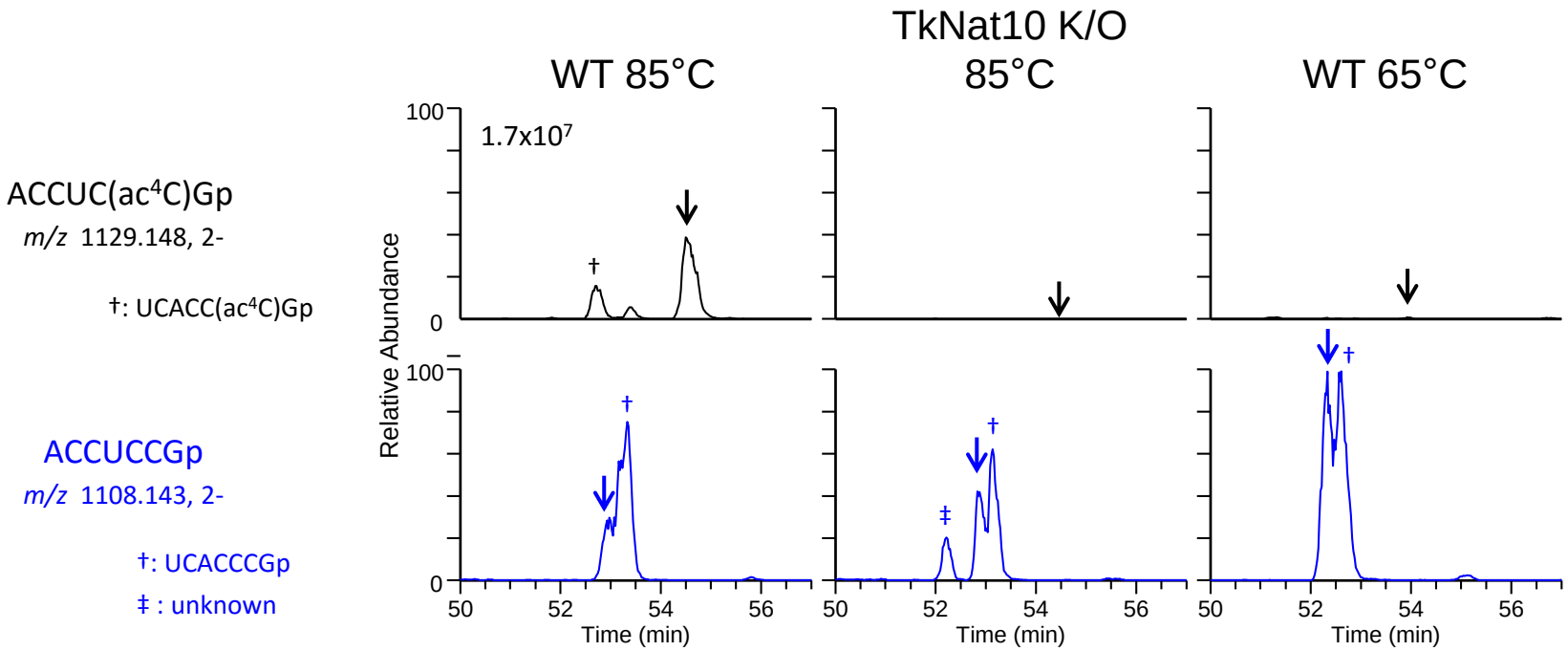
Supplementary Data 1. The extracted mass chromatograms for the quantitation of acetyl-4-cytidine (ac4C) in T. kodakarensis rRNAs. Continued.

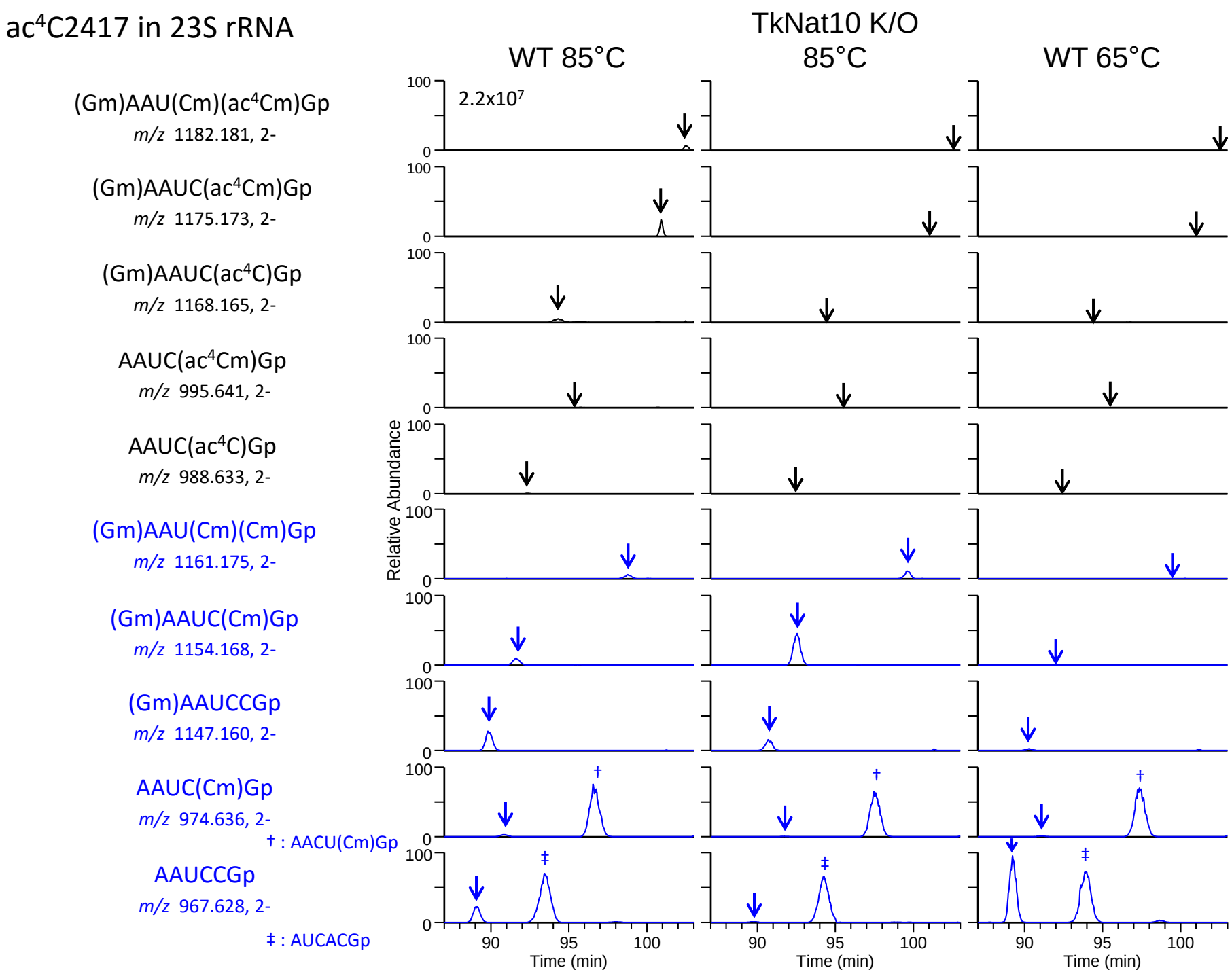
ac^{4}C1873 in 23S rRNA



Supplementary Data 1. The extracted mass chromatograms for the quantitation of acetyl-4-cytidine (ac4C) in T. kodakarensis rRNAs. Continued.

ac^{4}C2159 in 23S rRNA




ac4C2417 in 23S rRNA
TkNat10 K/O
WT 85°C
85°C
WT 65°C
(Gm)AAU(Cm)(ac4Cm)Gp
m/z 1182.181, 2-
(Gm)AAUC(ac4Cm)Gp
m/z 1175.173, 2-
(Gm)AAUC(ac4C)Gp
m/z 1168.165, 2-
AAUC(ac4Cm)Gp
m/z 995.641, 2-
AAUC(ac4C)Gp
m/z 988.633, 2-
(Gm)AAU(Cm)(Cm)Gp
m/z 1161.175, 2-
(Gm)AAUC(Cm)Gp
m/z 1154.168, 2-
(Gm)AAUCCGp
m/z 1147.160, 2-
AAUC(Cm)Gp
m/z 974.636, 2-
† : AACU(Cm)Gp
AAUCCGp
m/z 967.628, 2-
‡ : AUCACGp
2.2x10^7
Relative Abundance
100
0
90
95
100
Time (min)

Supplementary Data 1. The extracted mass chromatograms for the quantitation of acetyl-4-cytidine (ac4C) in T. kodakarensis rRNAs. Continued.

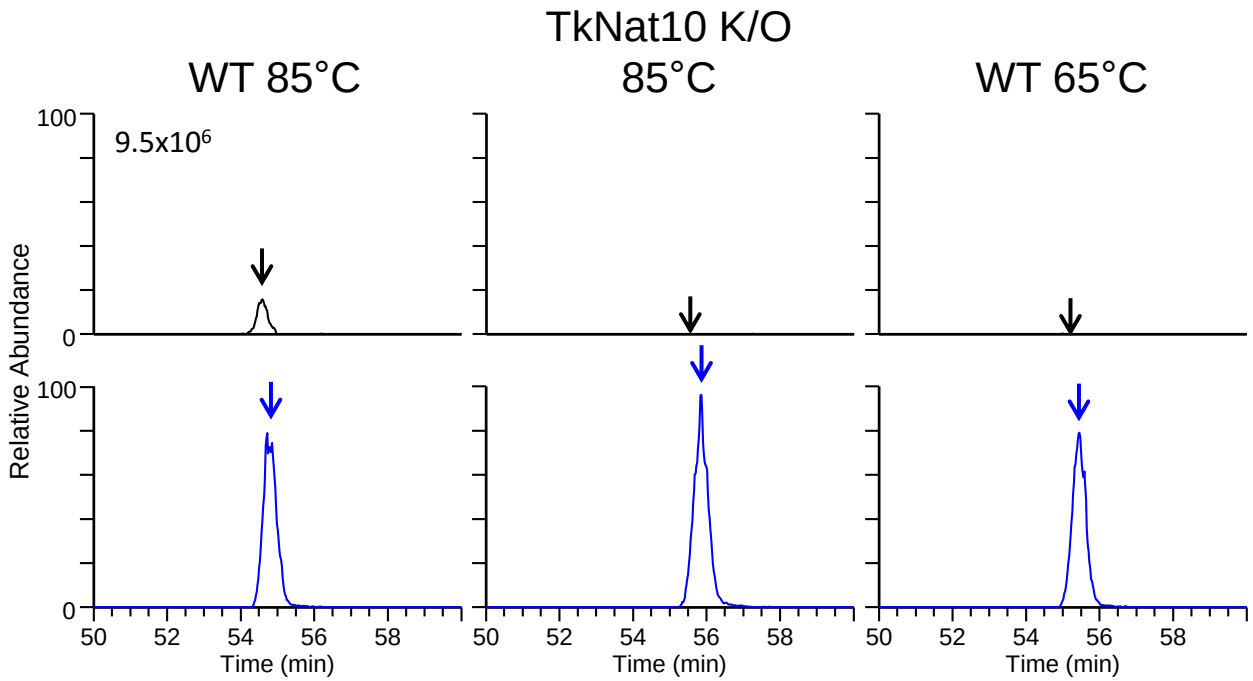
ac^{4}C2514 in 23S rRNA

CCUCCC(ac^{4}C)Gp

m/z 1269.664, 2-

CCUCCCCGp

m/z 1248.658, 2-



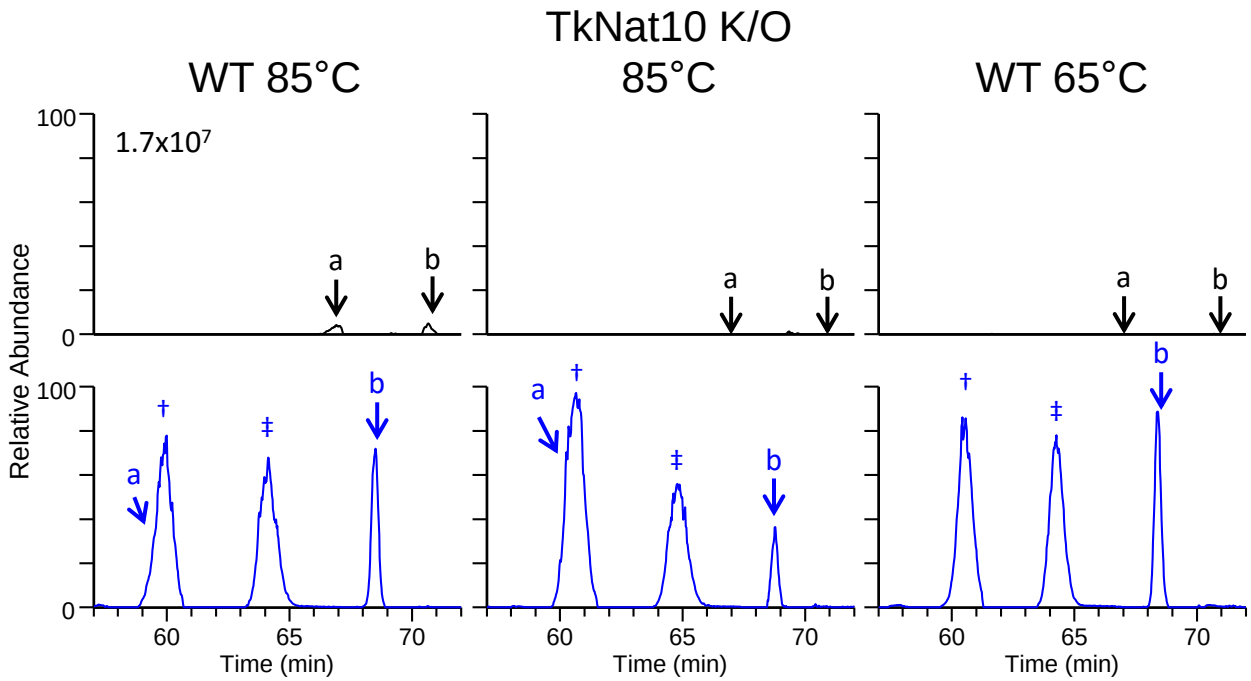
Supplementary Data 1. The extracted mass chromatograms for the quantitation of acetyl-4-cytidine (ac4C) in T. kodakarensis rRNAs. Continued.

ac^{4}C2797 in 23S rRNA

a. UCUAC(ac^{4}C)Gp
(with Ψ signature)
b. UCUAC(ac^{4}C)Gp
m/z 1129.641, 2-

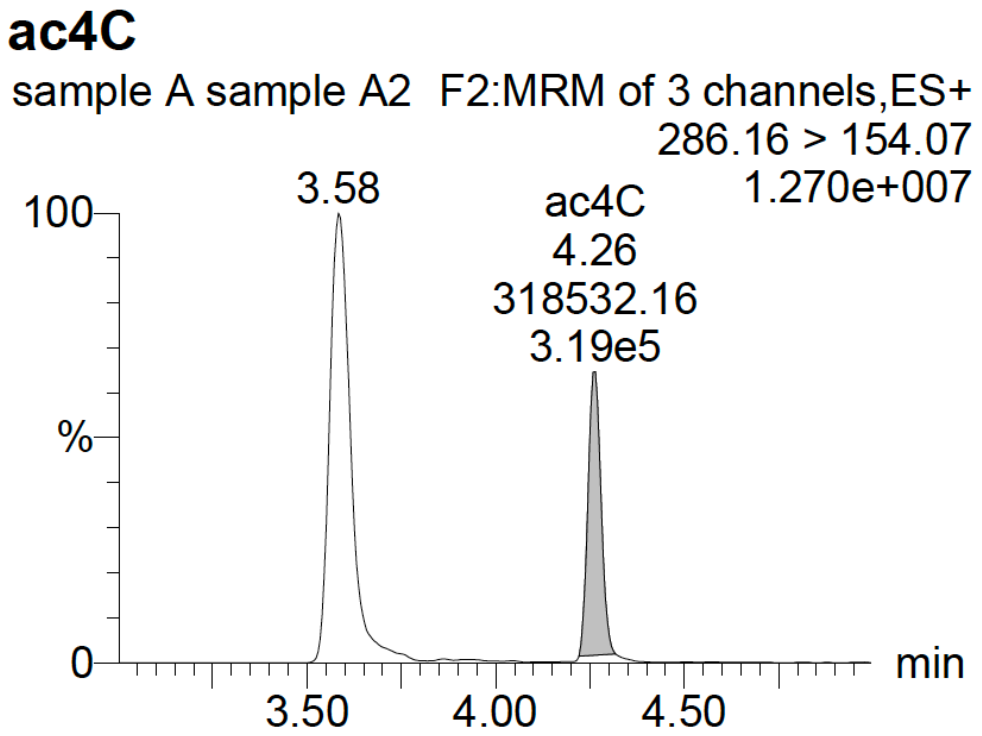
a. UCUACCGp
(with Ψ signature)
b. UCUACCGp
m/z 1108.635, 2-

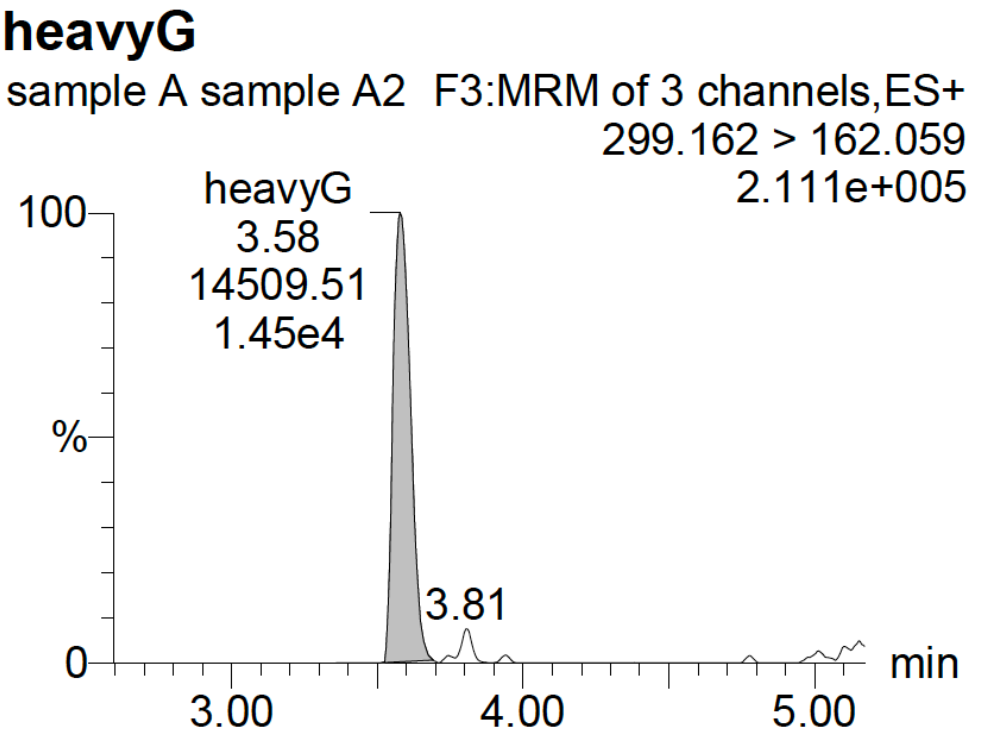
† : CACUCUGp
‡ : UCCUCAGp



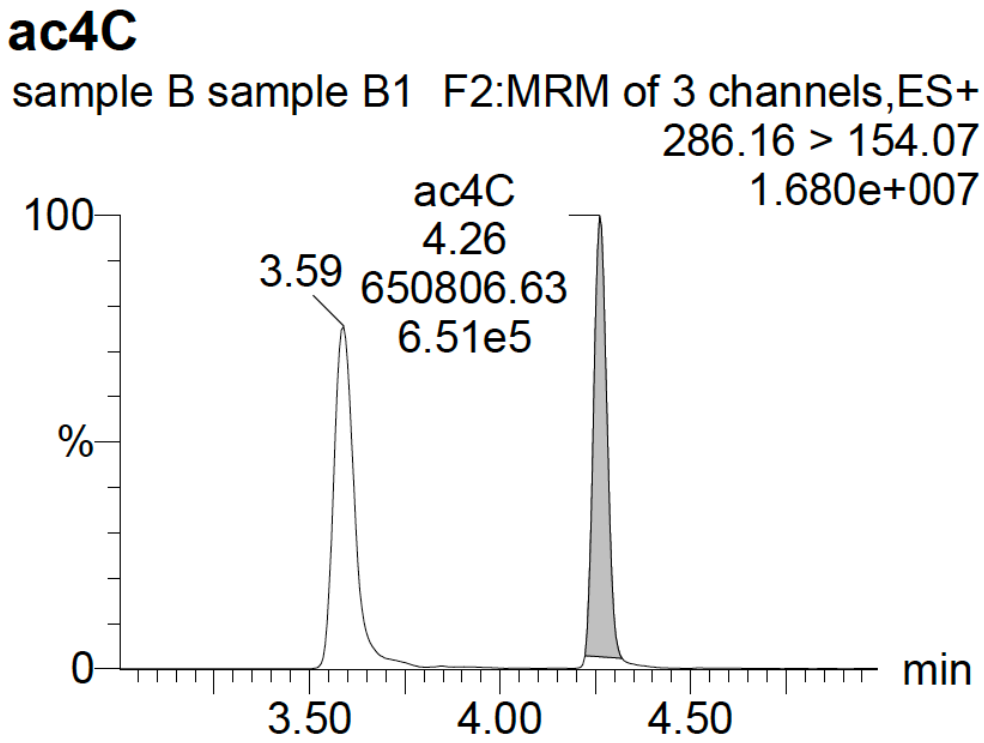
Supplementaru Data 2. LC-MS traces of ac4C in human, yeast and archaea.

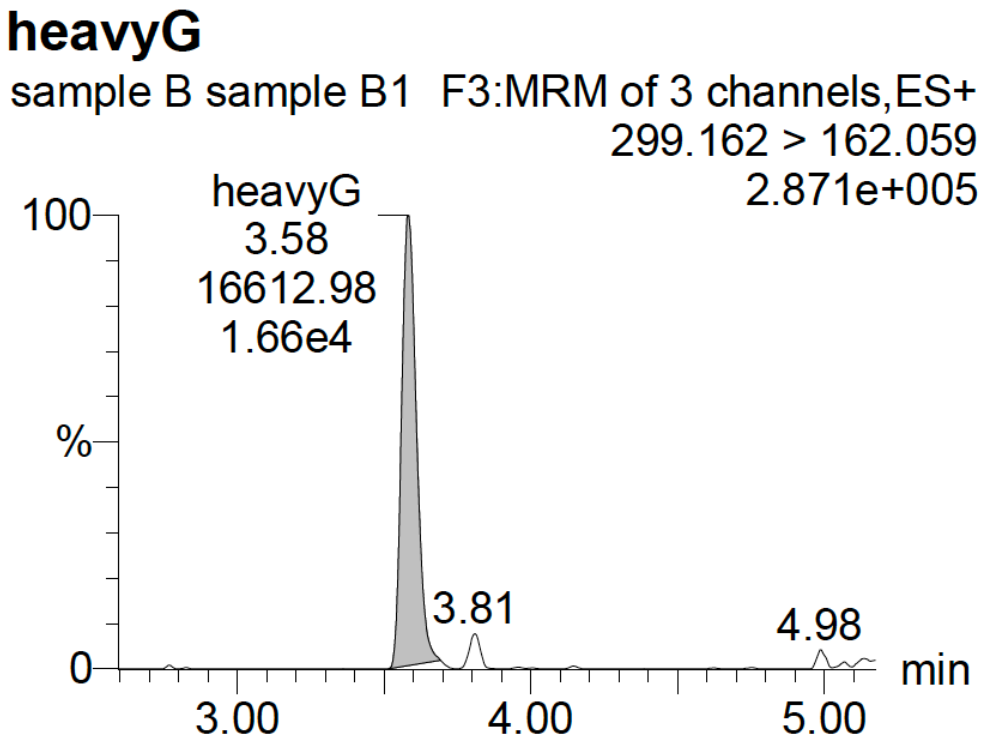
H. sapiens (HeLa)



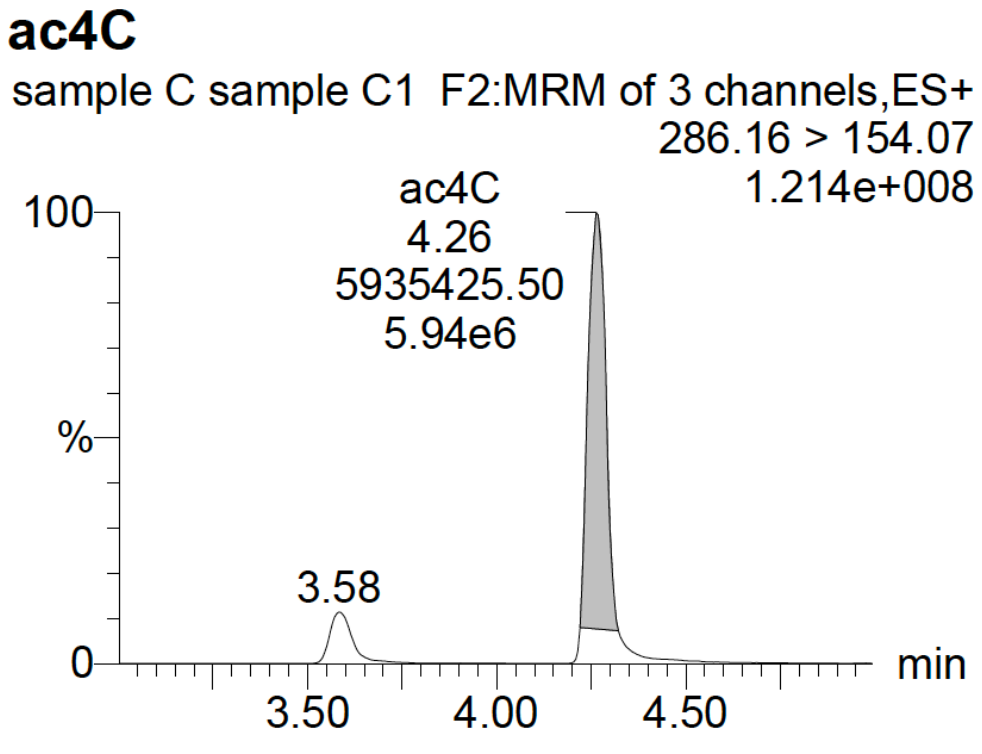


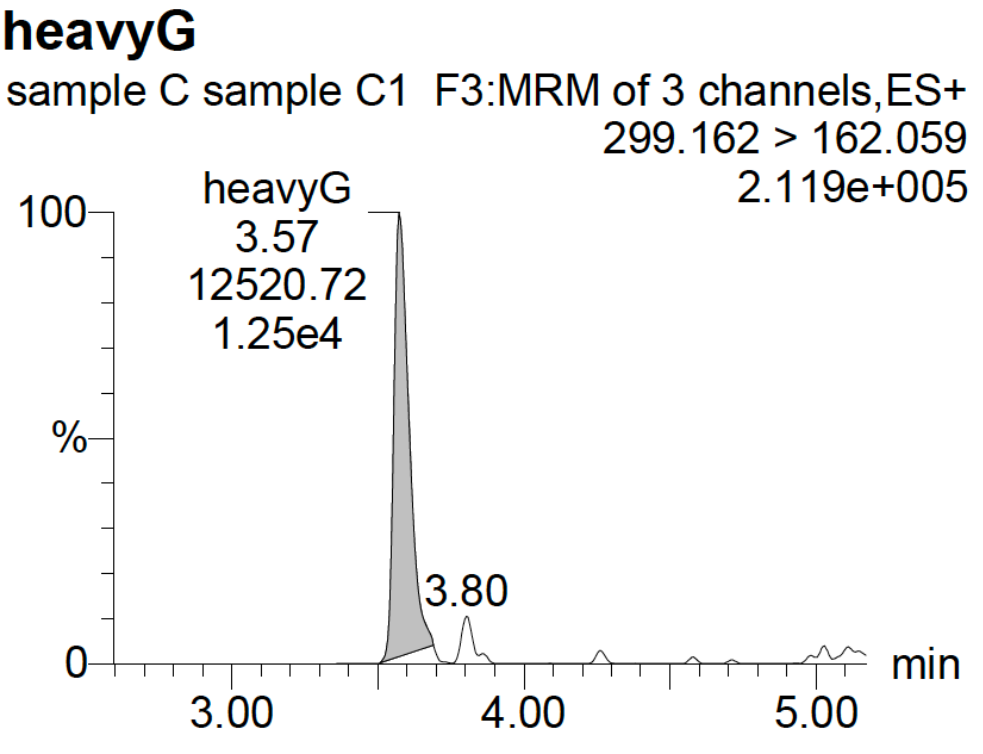
S. cerevisiae



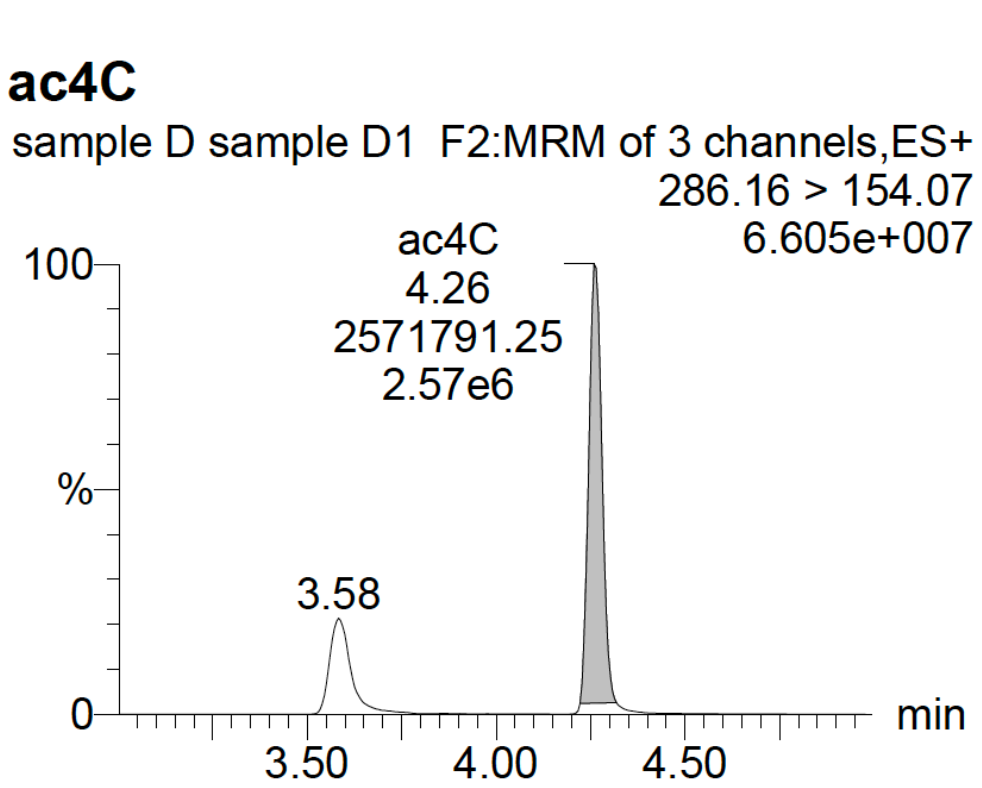


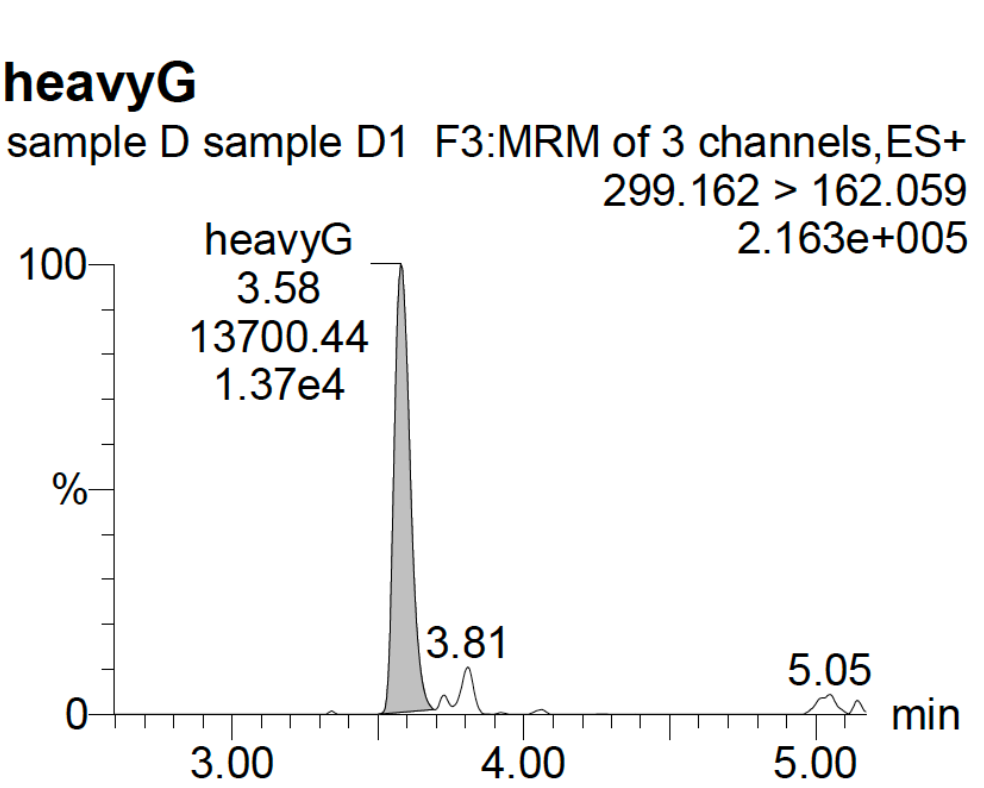
T. kodakarensis



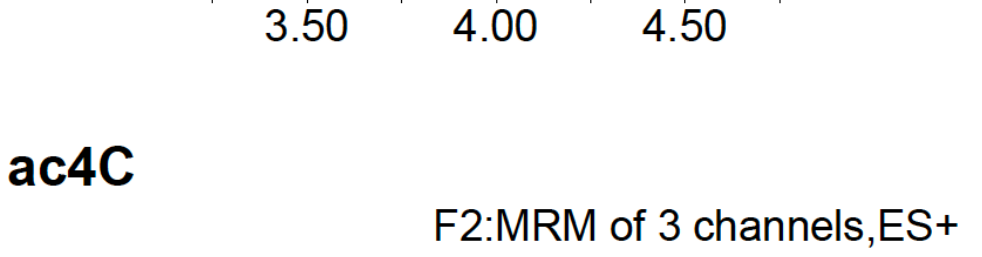


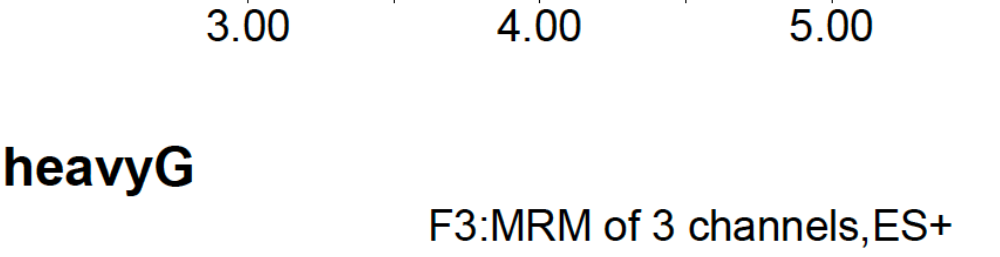
S. solfataricus





ac4C standard



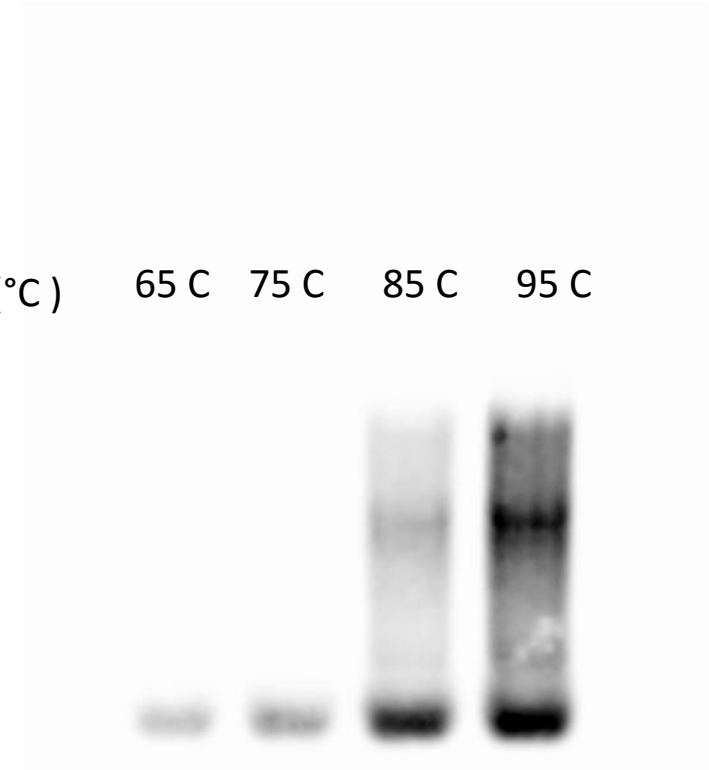


Supplementary Data 3. Raw gels corresponding to data shown in main Figure 3b and in Extended Data Figures 3b, 5e and 6i.

Full Gel for Figure 3b

1. T. kodakarensis temperature-dependent acetylation

EtBr stained 1% agarose- glyoxal gel image

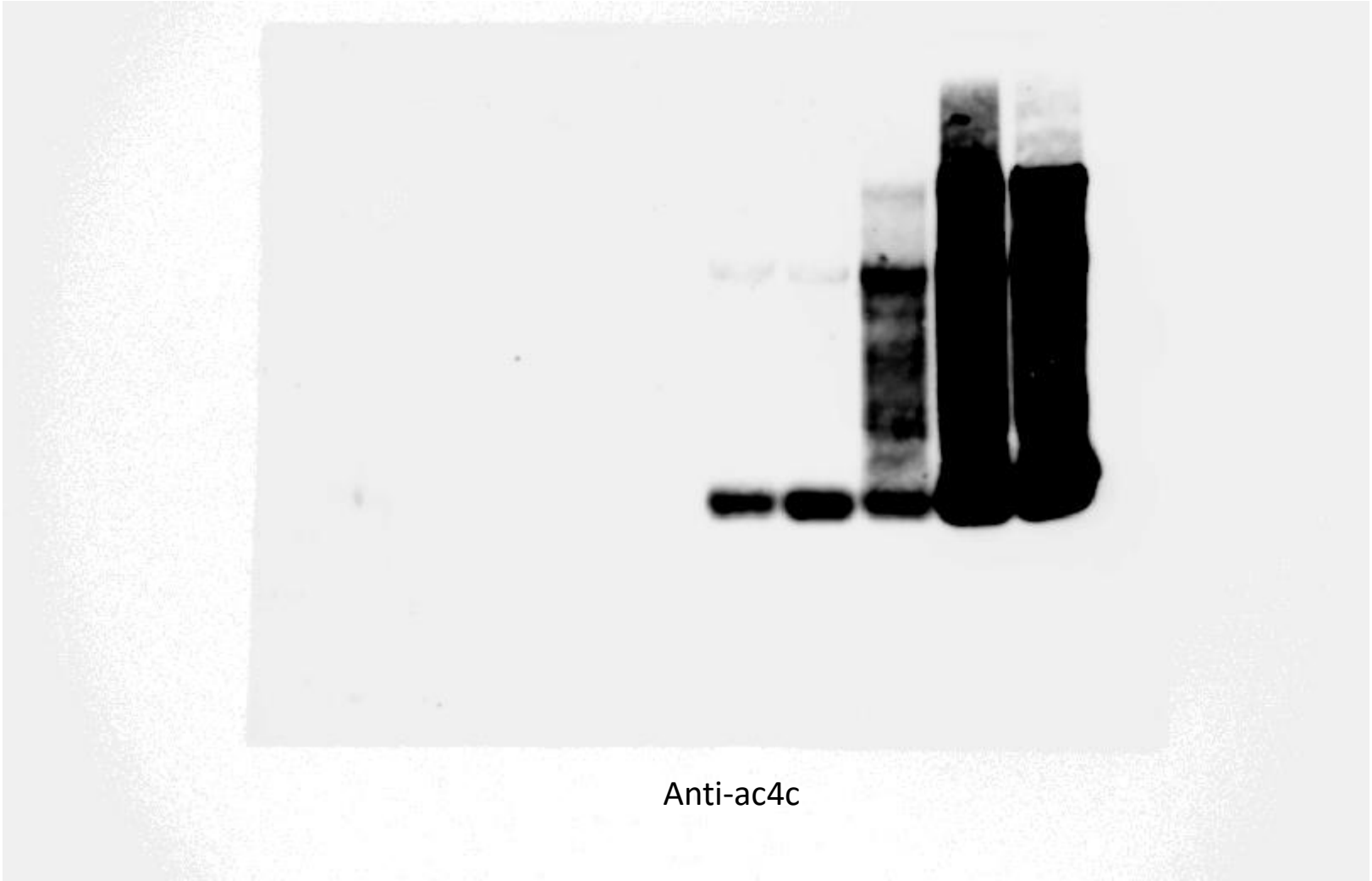


Anti-Ac4C

Full gel for Figure S5e
2. Delta TkNat10 and 55-85

T. kodakarensis

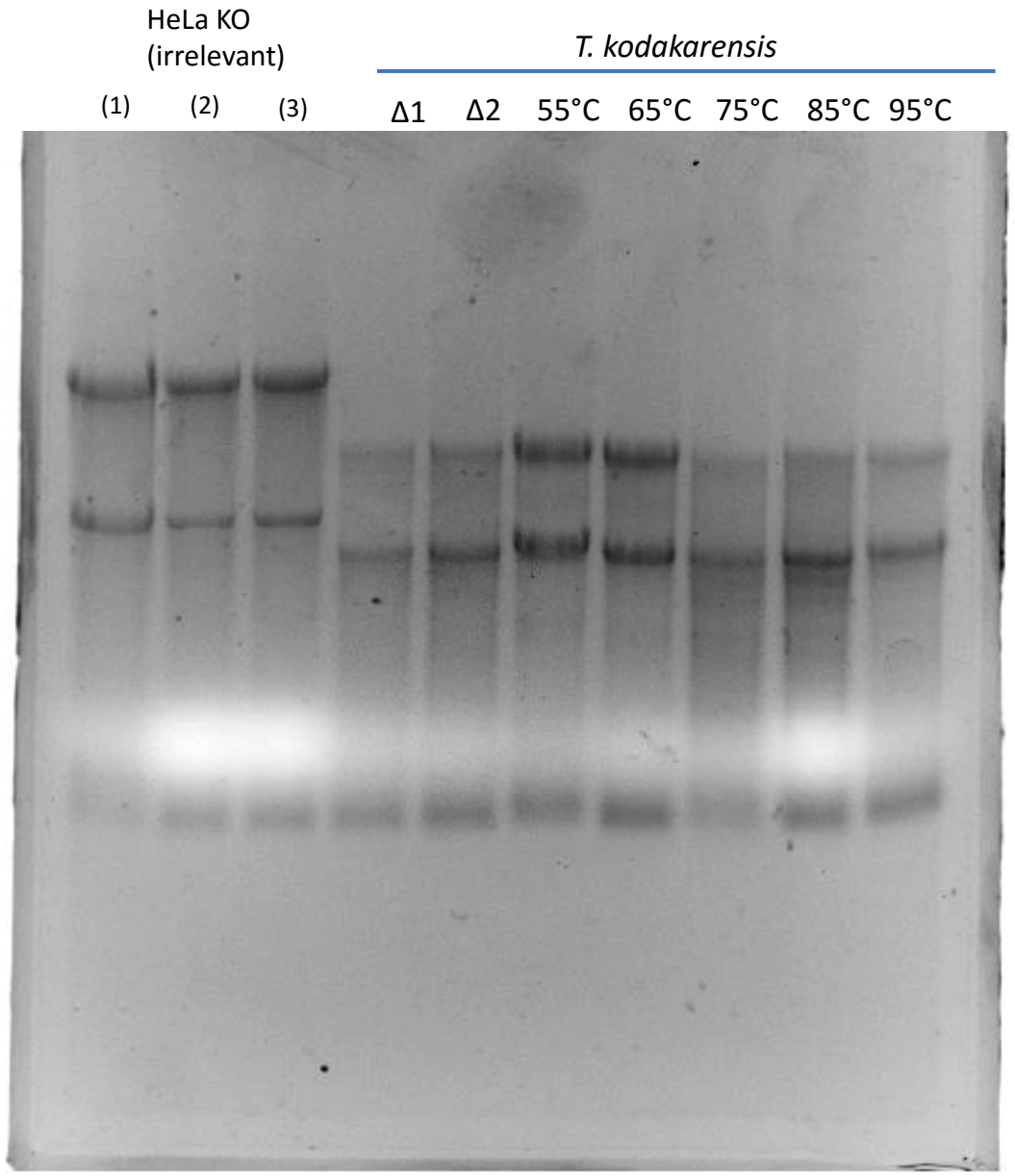
Δ1 Δ2 55°C 65°C 75°C 85°C 95°C



Anti-ac4c

Full gel for Figure S5e

2. Delta TkNat10 and 55-85



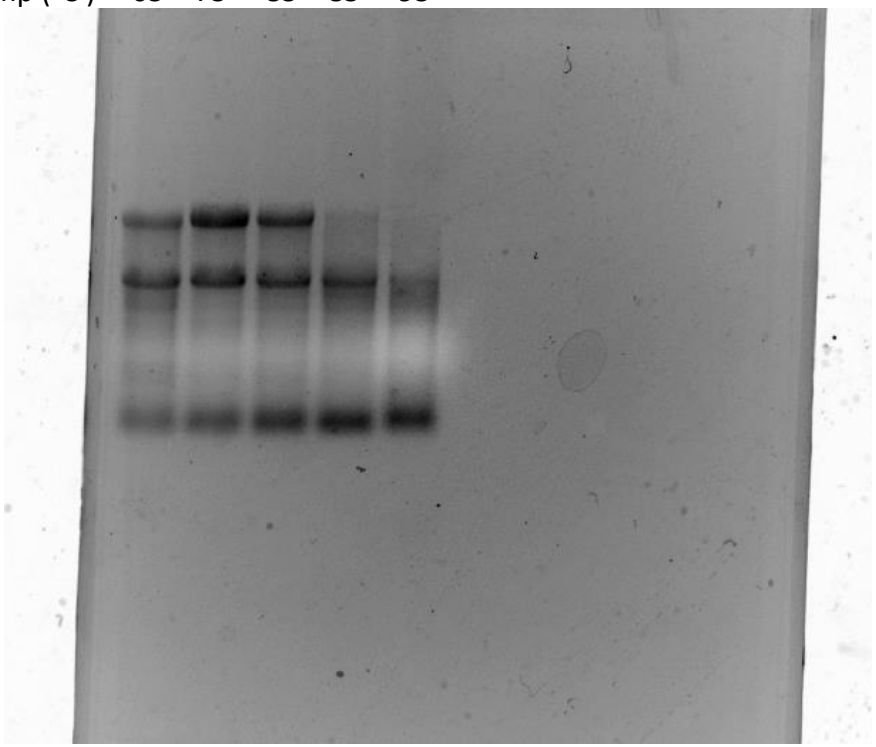
Full gel for Figure S6i

2. P. furiosis and T. AM4 temperature-dependent acetylation

Thermococcus

sp. AM44 *p. furiosis*

Temp (°C) 65 75 85 85 95

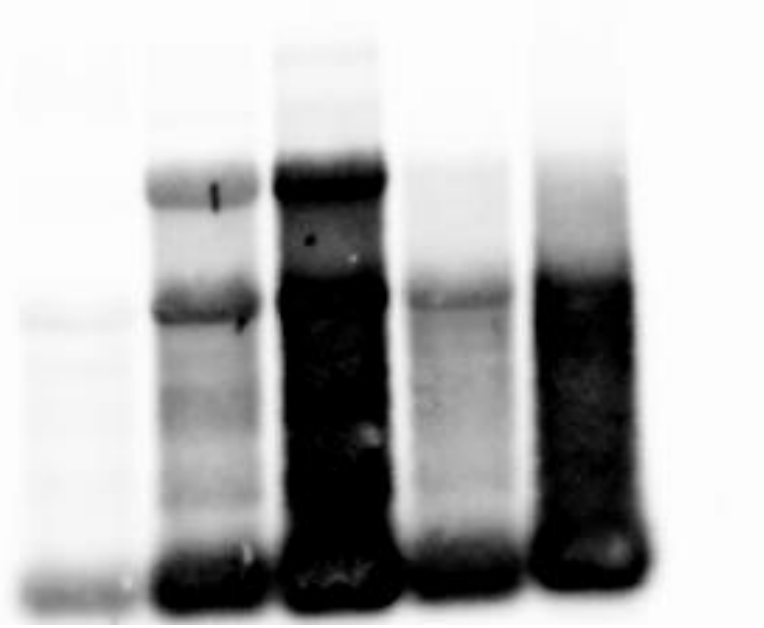


ethidium bromide

Thermococcus

sp. AM44 *p. furiosis*

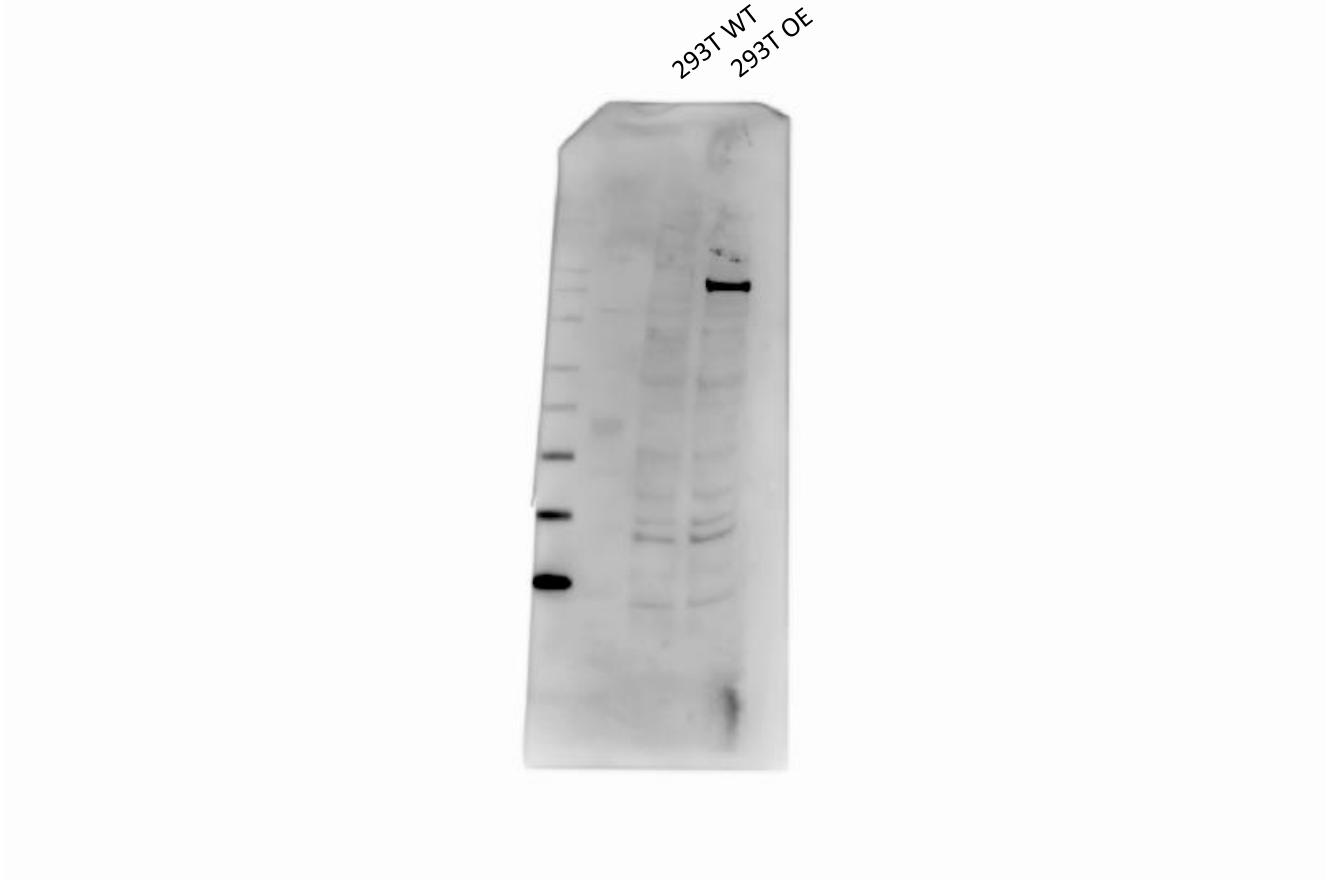
Temp (°C) 65 75 85 85 95



Anti-ac4c

Figure S3b
3. Assessment of Nat10/Thumpd1 overexpression

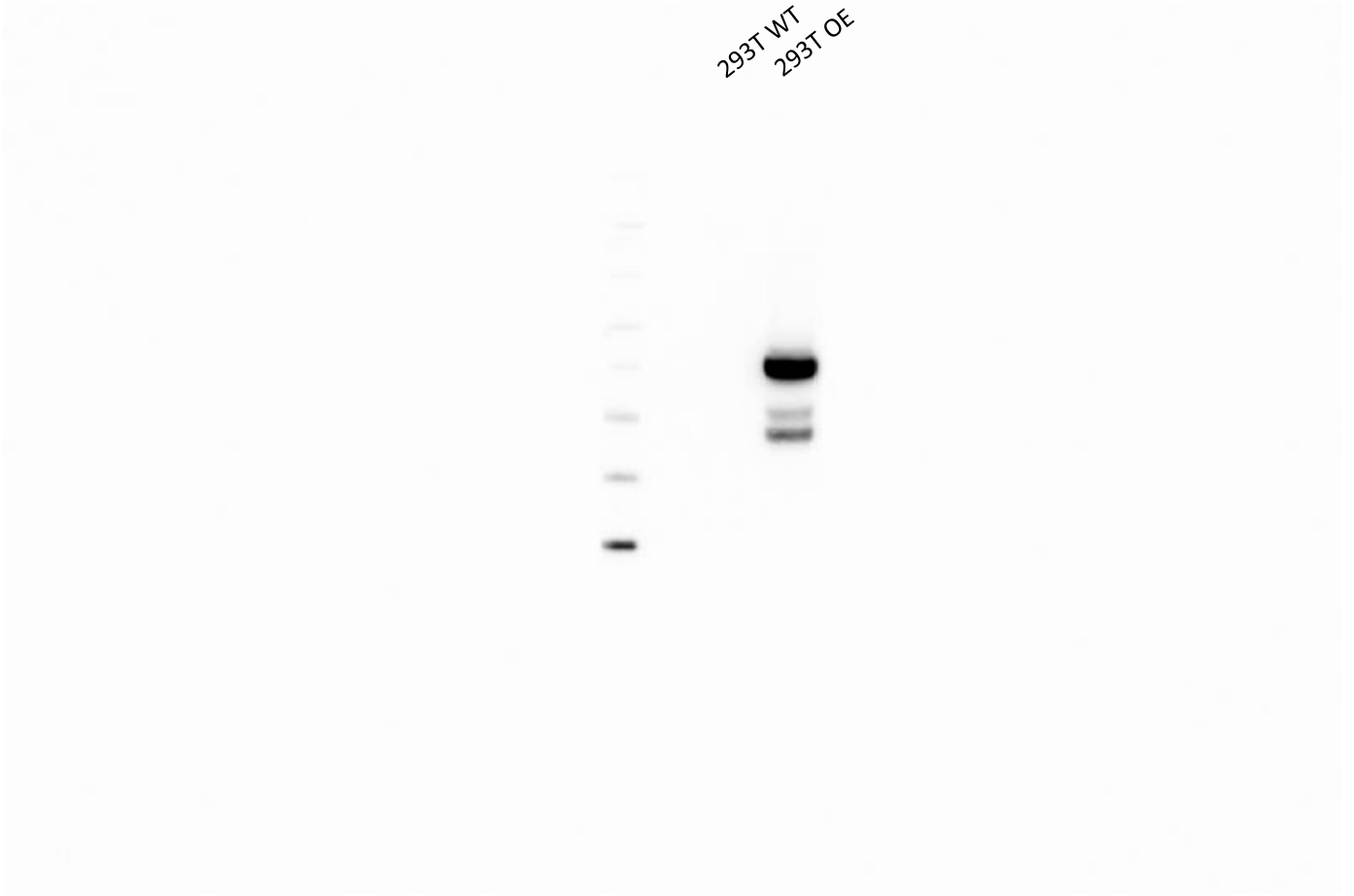
Full gel for Figure S3b
HEK293T 3xFLAG-NAT10/myc-THUMPD1 Transient Overexpression
Anti-FLAG (Nat10)



Full gel for Figure S3b

HEK293T 3xFLAG-NAT10/myc-THUMPD1 Transient Overexpression

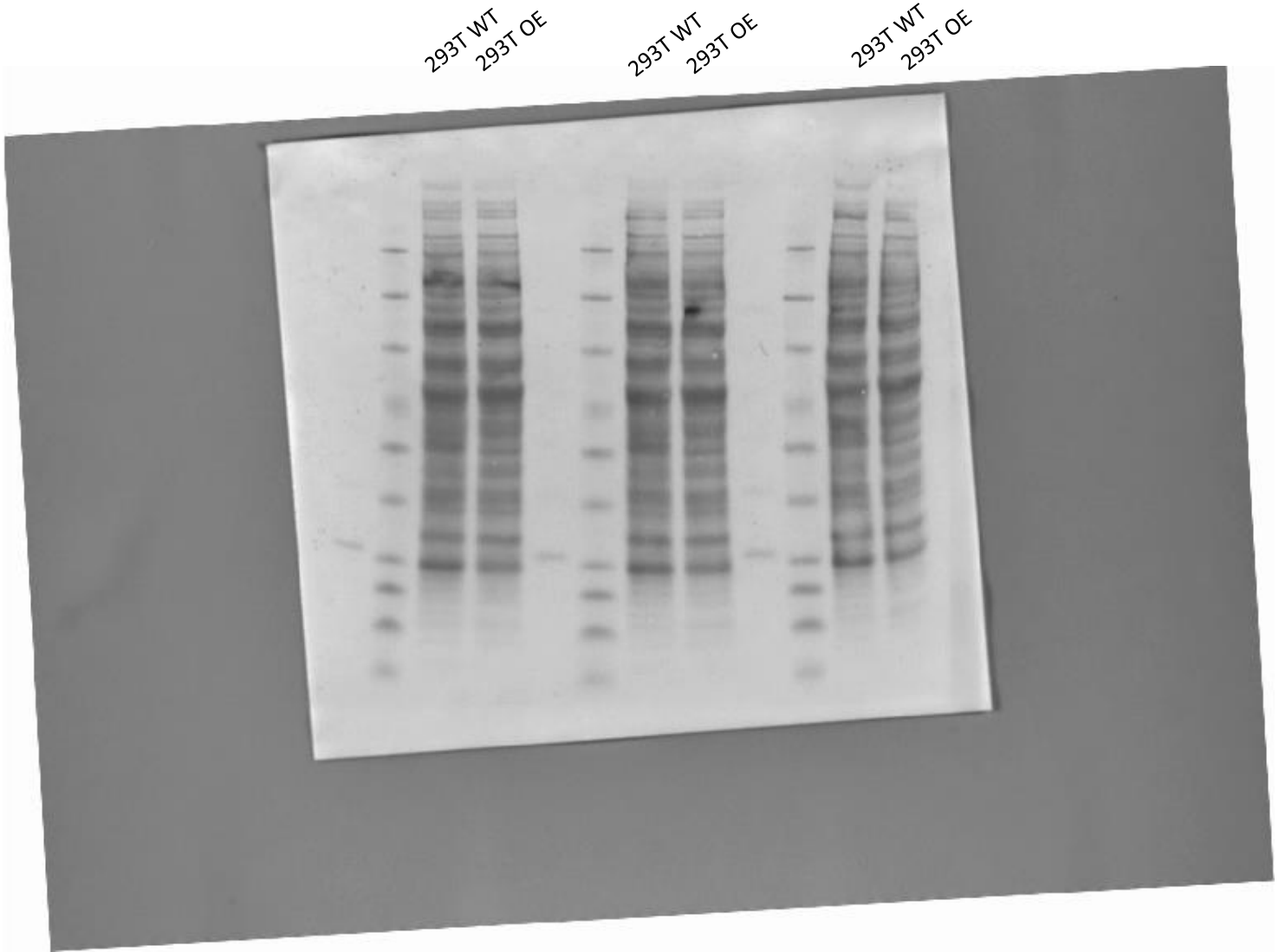
Anti-MYC (Thumpd1)



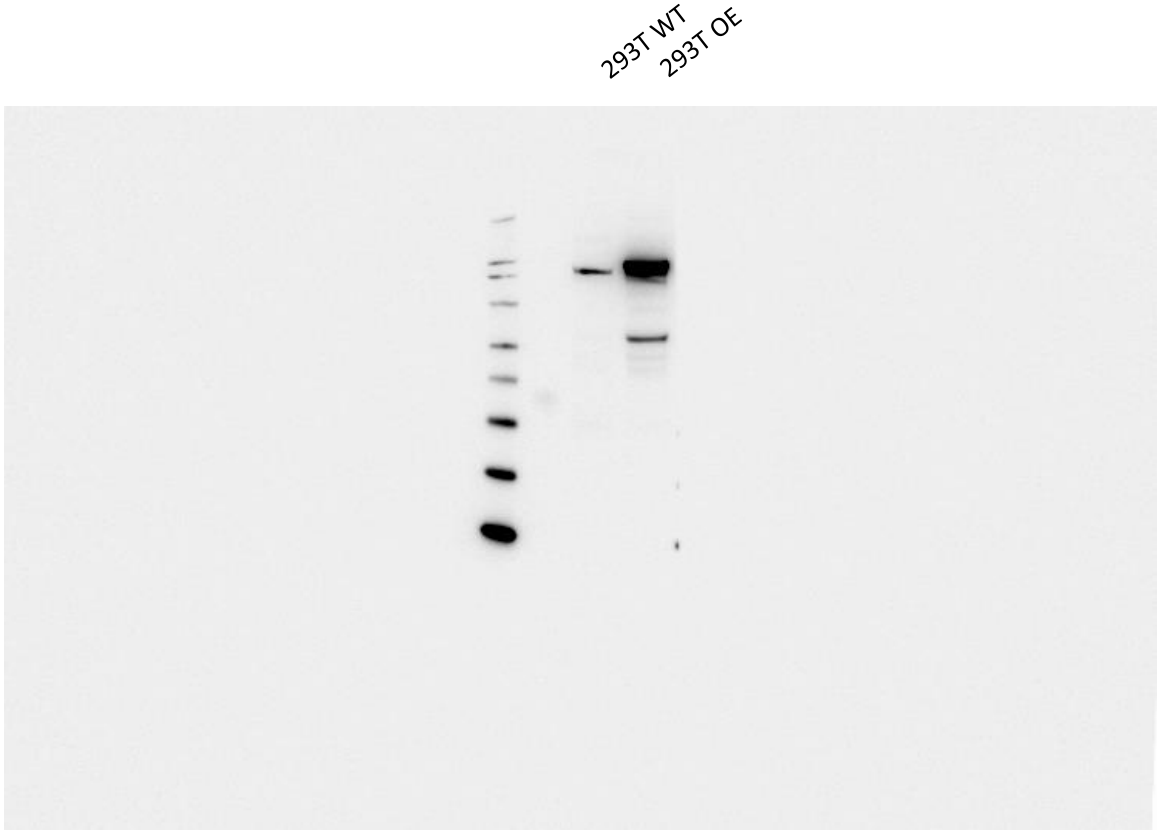
Full gel for Figure S3b

HEK293T 3xFLAG-NAT10/myc-THUMPD1 Transient Overexpression

Ponceau



Full gel for Figure S3b
HEK293T 3xFLAG-NAT10/myc-THUMPD1 Transient Overexpression
Anti-NAT10



Full gel for Figure S3b

HEK293T 3xFLAG-NAT10/myc-THUMPD1 Transient Overexpression

Anti-THUMPD1