




ARTICLE

DOI: 10.1038/s41467-018-07675-z

OPEN

# Microbiome characterization by high-throughput transfer RNA sequencing and modification analysis

Michael H. Schwartz<sup>1,2</sup>, Haipeng Wang <sup>1,3,4</sup>, Jessica N. Pan<sup>1</sup>, Wesley C. Clark<sup>1</sup>, Steven Cui<sup>5</sup>, Matthew J. Eckwahl<sup>1</sup>, David W. Pan<sup>1</sup>, Marc Parisien <sup>1</sup>, Sarah M. Owens<sup>5,6</sup>, Brian L. Cheng<sup>2</sup>, Kristina Martinez<sup>5</sup>, Jinbo Xu<sup>4</sup>, Eugene B. Chang<sup>5</sup>, Tao Pan<sup>1,2</sup> & A. Murat Eren <sup>2,5,7</sup>

Advances in high-throughput sequencing have facilitated remarkable insights into the diversity and functioning of naturally occurring microbes; however, current sequencing strategies are insufficient to reveal physiological states of microbial communities associated with protein translation dynamics. Transfer RNAs (tRNAs) are core components of protein synthesis machinery, present in all living cells, and are phylogenetically tractable, which make them ideal targets to gain physiological insights into environmental microbes. Here we report a direct sequencing approach, tRNA-seq, and a software suite, tRNA-seq-tools, to recover sequences, abundance profiles, and post-transcriptional modifications of microbial tRNA transcripts. Our analysis of cecal samples using tRNA-seq distinguishes high-fat- and low-fat-fed mice in a comparable fashion to 16S ribosomal RNA gene amplicons, and reveals taxon- and diet-dependent variations in tRNA modifications. Our results provide taxon-specific in situ insights into the dynamics of tRNA gene expression and post-transcriptional modifications within complex environmental microbiomes.

<sup>1</sup>Department of Biochemistry and Molecular Biology, University of Chicago, Chicago, IL 60637, USA. <sup>2</sup>Committee on Microbiology, University of Chicago, Chicago, IL 60637, USA. <sup>3</sup>School of Computer Science and Technology, Shandong University of Technology, Zibo, Shandong, China. <sup>4</sup>Toyota Technological Institute at Chicago, Chicago, IL 60637, USA. <sup>5</sup>Department of Medicine, University of Chicago, Chicago, IL 60637, USA. <sup>6</sup>Argonne National Laboratory, Lemont, IL 60439, USA. <sup>7</sup>Marine Biological Laboratory, Woods Hole, MA 02543, USA. These authors contributed equally: Michael H. Schwartz, Haipeng Wang. Correspondence and requests for materials should be addressed to T.P. (email: [taopan@uchicago.edu](mailto:taopan@uchicago.edu)) or to A.M.E. (email: [meren@uchicago.edu](mailto:meren@uchicago.edu))

Microbes occupy almost all ecological niches on Earth<sup>1</sup>, and the microbial communities associated with the human body play important roles in health and disease<sup>2,3</sup>. High-throughput sequencing of 16S ribosomal RNA (rRNA) gene amplicons and shotgun metagenomes enabled extensive culture-independent characterizations of microbial diversity and functional potential<sup>4</sup>, however, these approaches are not suitable to gain insights into in situ physiology of microbial populations. An alternative way to study community structures of naturally occurring microbial populations, along with the added benefit of recovering information about their physiological states, could rely on transfer RNA (tRNA) molecules.

A bacterial genome has 30–50 different tRNA gene sequences, and a single cell contains up to 100,000 tRNA transcripts<sup>5</sup>. Similar to the 16S rRNA gene, tRNA genes can be used as phylogenetic markers to resolve relationships between microbial taxa<sup>6</sup>. Yet, in contrast to the sequencing of 16S rRNA gene amplicons, which relies on PCR amplification from genomic DNA, the direct sequencing of tRNA transcripts (hereafter ‘tRNA-seq’) provides access to translational activities of the cell<sup>7–11</sup>. On average, a bacterial tRNA transcript contains 8 modifications, which play important roles in cellular physiology<sup>12–14</sup>. While modifications in the anticodon loop directly tune translational efficiency and fidelity, others tune tRNA stability and physical interactions with cellular proteins and the ribosome<sup>15,16</sup>. Systematical studies of bacterial tRNA modifications have so far been limited to five species in culture, *Escherichia coli*, *Bacillus subtilis*, *Mycoplasma capricolum*, *Lactococcus lactis*, and *Streptomyces griseus*<sup>12,17–19</sup>, and used approaches such as two-dimensional thin-layer chromatography (TLC) and liquid chromatography-mass spectrometry (LC-MS). Insights from these studies revealed that while some tRNA modifications are shared among multiple taxa, others may be more unique. Taxon-dependent modifications may have distinct roles within complex microbial communities, and the identification of post-transcriptional tRNA modifications de novo could offer insights into changing metabolic activities of microbes as a response to changing environmental conditions. However, the presence of these modifications as well as the rigid structure of the tRNA molecule interrupt the cDNA synthesis that is required for sequencing, which limits comprehensive investigations of tRNA modifications and their utility to study microbial physiology in complex communities.

We recently addressed the experimental bottlenecks that limit the high-throughput sequencing of tRNA transcripts by employing two demethylase enzymes to remove tRNA modifications, and a thermophilic reverse transcriptase (TGIRT) to obtain long reads (demethylase tRNA-seq or DM-tRNA-seq<sup>8</sup>). In DM-tRNA-seq, each sample is split in two and processed with and without demethylase treatment prior to cDNA synthesis. We found that the demethylase-treated samples matched the cellular tRNA abundance measured by hybridization in human tRNA studies<sup>8</sup>, whereas the untreated samples were well suited for the analysis of tRNA modifications<sup>20</sup>. TGIRT can read through many Watson–Crick base methylations while leaving behind mutation signatures that allow for the identification of certain modifications at single base resolution and enable comparative analyses of modification levels<sup>20</sup>. We have validated that the mutation fractions for the well-studied N1-methyladenosine (m<sup>1</sup>A) modification determined by tRNA-seq correlated well with modification fractions inferred by a gel-based method<sup>20</sup>.

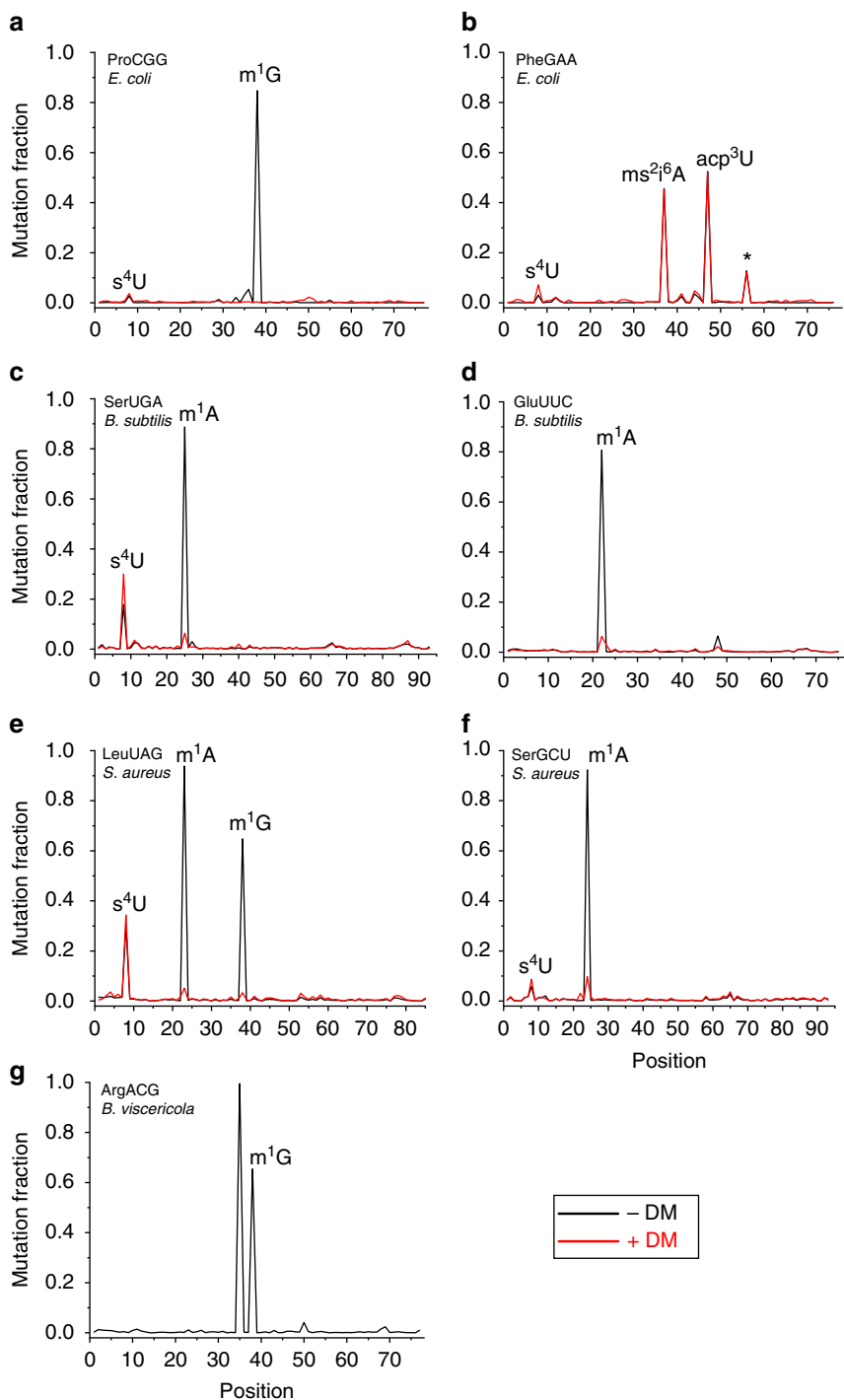
Here, we apply DM-tRNA-seq to gut microbiome of mice fed high-fat (HF) and low-fat (LF) diets and introduce an open-source software suite, tRNA-seq-tools, to identify and characterize tRNA sequences in high-throughput sequencing data that emerge from DM-tRNA-seq experiments. Our workflow enabled

us to infer microbial community structures and taxonomy in our samples using tRNA sequences, investigate the expression of tRNAs with different anticodons across different environmental conditions, and recover tRNA modification levels across different taxa. Our results show class-dependent modification patterns, and diet-dependent changes in modification levels that can be associated with diet-dependent proteome in the microbiome. These findings demonstrate the versatility of tRNA-seq to investigate tRNA expression and modifications within microbiomes.

## Results

**tRNA-seq reveals tRNA modifications in bacterial cultures.** To determine the feasibility of employing tRNA-seq for modification studies in simple model systems, we performed DM-tRNA-seq on bacterial cultures from four species: *Escherichia coli*, *Bacillus subtilis*, *Staphylococcus aureus*, and *Barnesiella viscericola*. Bacterial tRNAs share some modifications with eukaryotic tRNAs, but also possess unique modifications of their own<sup>21</sup>. Since almost all *E. coli* tRNAs and the majority of *B. subtilis* tRNAs have been mapped for modifications by 2D-TLC and LC/MS, these two cultures allowed us to benchmark tRNA-seq results against the known modifications. In contrast, *S. aureus* and *B. viscericola* were suitable to identify new modifications using tRNA-seq.

We focused our analysis on Watson–Crick face base modifications for which DM-tRNA-seq was particularly well suited<sup>20</sup>. We first determined the mutation and the stop fraction at each nucleotide position within tRNA sequences<sup>20</sup>. Here, we calculated the mutation fraction by only including sequencing reads that pass the modification site while excluding shorter reads that stopped before the modification site. Figure 1 and Supplementary Figure 1 display representative mutation and stop fractions for several tRNAs with modification sites annotated using standard tRNA nomenclature<sup>22</sup>. *E. coli* tRNA<sup>Pro</sup>(CGG) has N1-methylguanosine (m<sup>1</sup>G)<sub>37</sub> and 4-thiouridine (s<sup>4</sup>U)<sub>8</sub>, but only m<sup>1</sup>G can be removed by the demethylase treatment. As expected, both mutations and stops showed a reduction to the background levels at the m<sup>1</sup>G<sub>37</sub> position in the demethylase-treated sample (Fig. 1a, Supplementary Fig. 1a). We identified a second mutation peak at position U<sub>8</sub> that was more than 10-folds higher compared to the background level. This peak did not change in the demethylase-treated sample, consistent with the known s<sup>4</sup>U modification. *E. coli* tRNA<sup>Phe</sup>(GAA) has 3-(3-amino-3-carboxypropyl)uridine (acp<sup>3</sup>U)<sub>47</sub>, 2-methylthio-N<sub>6</sub>-isopentenyladenosine (ms<sup>2</sup>i<sup>6</sup>A)<sub>37</sub>, and s<sup>4</sup>U<sub>8</sub>, all were detected by mutation, and none responded to demethylase treatment (Fig. 1b, Supplementary Fig. 1b). Ms<sup>2</sup>i<sup>6</sup>A<sub>37</sub> showed ~99% stops, and ~45% of the remaining 1% read-through were mutations. *B. subtilis* has m<sup>1</sup>A<sub>22</sub> modifications, which are absent in *E. coli*. This methylation can be almost fully removed by the demethylases as reflected by the reduction of mutation fraction upon demethylase treatment for *B. subtilis* tRNA<sup>Ser</sup>(UGA) and tRNA<sup>Glu</sup>(UUC) (Fig. 1c, d). For tRNA<sup>Ser</sup>(UGA), the s<sup>4</sup>U<sub>8</sub> was readily identified through its mutations (Fig. 1c), whereas the N<sub>6</sub>-isopentenyladenosine (i<sup>6</sup>A<sub>37</sub>) modification only showed stops (Supplementary Fig. 1c). Modifications in tRNA<sup>Glu</sup>(UUC) have not been previously mapped; our result is consistent with the presence of m<sup>1</sup>A<sub>22</sub> in this tRNA (Fig. 1d, Supplementary Fig. 1d). No tRNA modifications have been mapped in *S. aureus* and *B. viscericola*. Our results are consistent with the presence of demethylase-sensitive m<sup>1</sup>G<sub>37</sub> and m<sup>1</sup>A<sub>22</sub> and demethylase-insensitive s<sup>4</sup>U<sub>8</sub> in *S. aureus* tRNA<sup>Leu</sup>(UAG) (Fig. 1e, Supplementary Fig. 1e), and m<sup>1</sup>A<sub>22</sub> and s<sup>4</sup>U<sub>8</sub> in tRNA<sup>Ser</sup>(GCU) (Fig. 1f, Supplementary Fig. 1f). Finally, we identified mutation and stop signatures that are consistent with the expected m<sup>1</sup>G<sub>37</sub> and

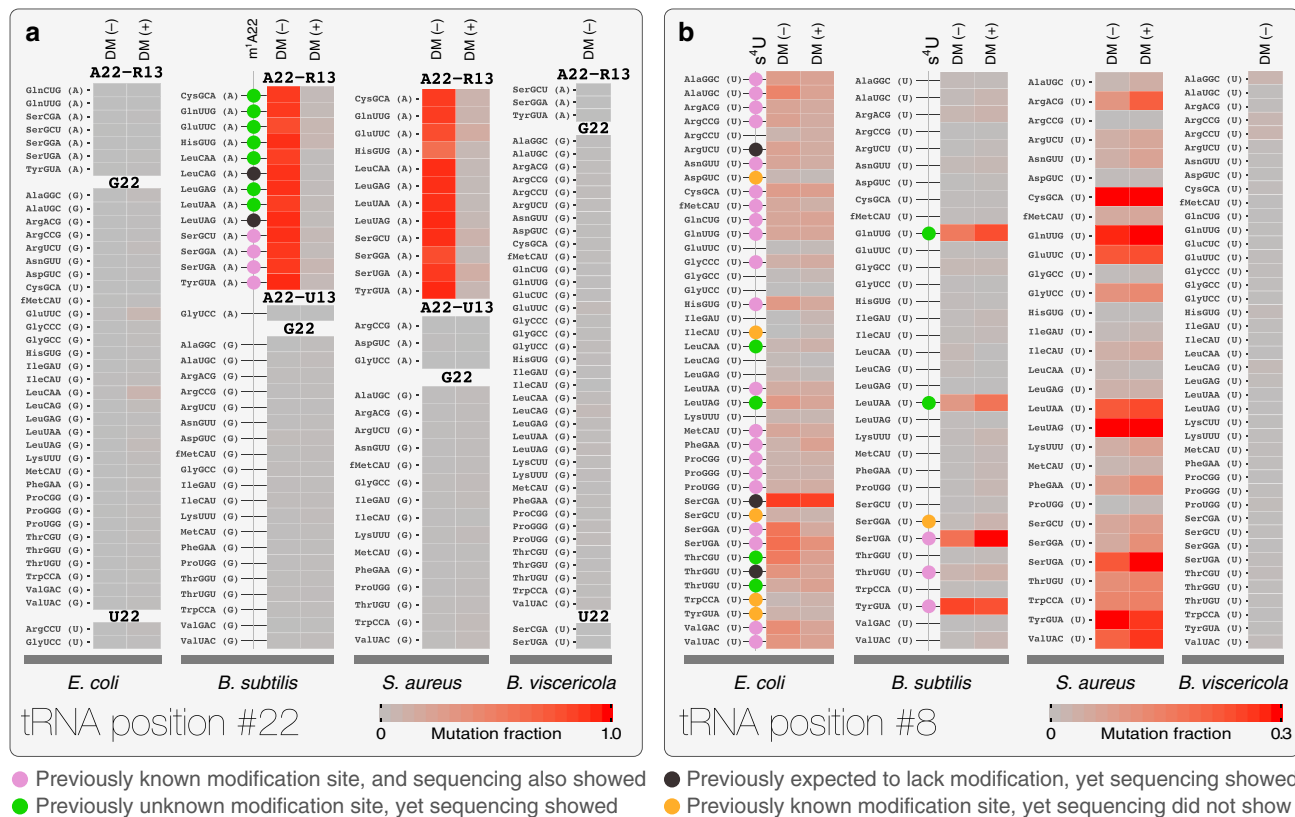


**Fig. 1** tRNA modifications of bacterial cultures. Red and black lines show mutation fractions in representative tRNA sequences with (+DM) and without (-DM) demethylase treatment, respectively. **a** *E. coli* tRNA<sup>Pro</sup>(CGG) shows m<sup>1</sup>G37 and s<sup>4</sup>U8. **b** *E. coli* tRNA<sup>Phe</sup>(GAA) shows acp<sup>3</sup>U47, ms<sup>2</sup>i<sup>6</sup>A37 and s<sup>4</sup>U8. The additional peak denoted by asterisk (\*) may represent an unknown modification. **c** *B. subtilis* tRNA<sup>Ser</sup>(UGA) shows m<sup>1</sup>A22, and s<sup>4</sup>U8. **d** *B. subtilis* tRNA<sup>Glu</sup>(UUC) shows m<sup>1</sup>A22. **e** *S. aureus* tRNA<sup>Leu</sup>(UAG) shows m<sup>1</sup>G37, m<sup>1</sup>A22 and s<sup>4</sup>U8. **f** *S. aureus* tRNA<sup>Ser</sup>(GCU) shows m<sup>1</sup>A22 and s<sup>4</sup>U8. **g** *B. viscericola* tRNA<sup>Arg</sup>(ACG) shows m<sup>1</sup>G37 and I34

inosine (I)34 in *B. viscericola* tRNA<sup>Arg</sup>(ICG) (Fig. 1g, Supplementary Fig. 1g). Using mutation and/or stop signatures, our results are consistent with all known Watson–Crick face base modifications in these four bacterial species (Supplementary Fig. 2, Supplementary Table 1).

We examined the complete mutation profiles that correspond to m<sup>1</sup>A, m<sup>1</sup>G, and s<sup>4</sup>U for tRNAs in the four bacterial cultures (Fig. 2, Supplementary Fig. 3). We only found mutation

signatures that are consistent with m<sup>1</sup>A at position 22 located at the junction of the D stem and D loop in *B. subtilis* and *S. aureus*, but not in *E. coli* and *B. viscericola* (Fig. 2a). We found mutation signatures that are consistent with 13 and 12 tRNAs with m<sup>1</sup>A22 modification in *B. subtilis* and *S. aureus*, respectively; only four have been previously mapped in *B. subtilis* by 2D-TLC<sup>23,24</sup>. Among all tRNAs with A22 in these two bacteria, all with G/A13–A22 pairs were modified, whereas all with U13–A22



**Fig. 2** Mutation fractions of two tRNA sites. Heatmaps of mutation fractions for positions 22 and 8 (using standard, canonical tRNA nomenclature) are shown. tRNAs with different anticodons are grouped by their sequences at the respective position of modification (in parenthesis) and in alphabetical orders. Only *E. coli* and *B. subtilis* tRNA modifications have been mapped previously by 2D-TLC and LC/MS, but the mapping was not done for every tRNA species<sup>12</sup>. Every *E. coli* and *B. subtilis* tRNA species with mutation fraction at 10-times above background is marked with a circle on the right with the following designations: Purples correspond to those known to be present and also identified by sequencing here; blacks correspond to those supposed to be absent but identified by sequencing; greens correspond to those not mapped previously but identified by sequencing; oranges correspond to those known to be present but were not found by sequencing. **a** m<sup>1</sup>A22; R corresponds to A or G. **b** s<sup>4</sup>U8

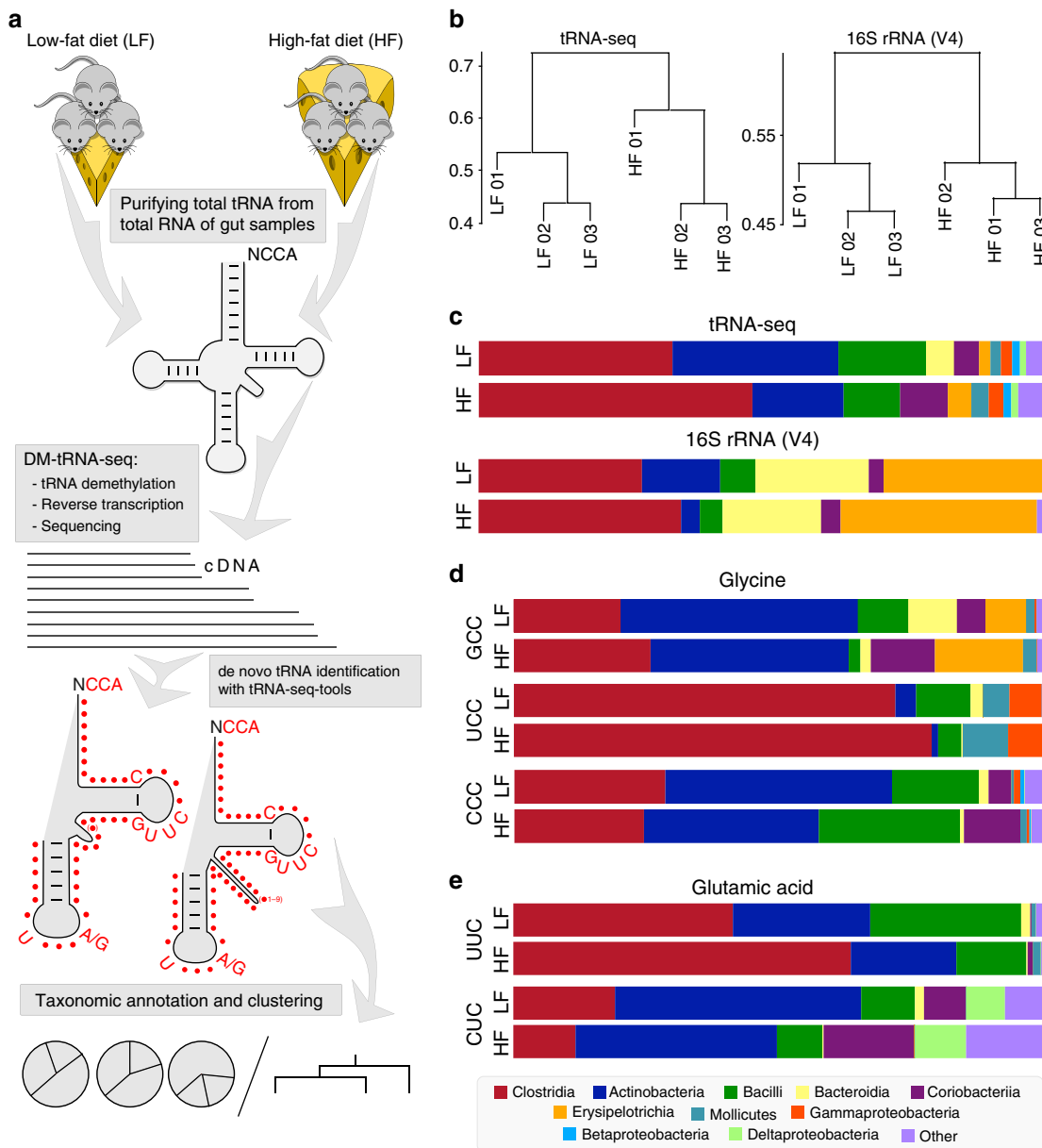
were not modified, consistent with this modification requiring a non-Watson–Crick base paired A22. M<sup>1</sup>A22 mutation fractions ranged from 0.81–0.94 in *B. subtilis* and 0.61–0.94 in *S. aureus*, suggesting that most tRNAs are modified at high fractions at A22 under our culture conditions. M<sup>1</sup>G is located at position 37 immediately 3' to the anticodon nucleotides in bacteria. In all four bacteria, we found that mutation fractions ranged from 0.33–0.90 when the tRNA has G37 in its sequence (Supplementary Fig. 3). Unlike m<sup>1</sup>A22, m<sup>1</sup>G37 was accompanied by highly variable stop fractions in a tRNA-dependent manner (ref. <sup>20</sup>, Supplementary Fig. 1). Since m<sup>1</sup>G37 prevents translational frameshifting<sup>25,26</sup> and is thought to be present at high fractions in tRNA, mutation alone may not correlate well with modification fractions for m<sup>1</sup>G37.

We identified mutation signatures consistent with many s<sup>4</sup>U8 modifications in three of the four bacteria (Fig. 2b). We found 28 sites in the *E. coli* tRNA with mutation fractions of 0.03–0.26 (0.03 corresponded to >10× above background), 20 of which overlapped with the known s<sup>4</sup>U8 modifications by 2D-TLC and LC/MS. The other 8 sites could either be present only under our culture conditions or were missed in the previous mapping studies. We found 5 sites in *B. subtilis* tRNAs with mutation fractions of 0.03–0.25, two of which overlapped with the known s<sup>4</sup>U8 modifications. We found 21 sites with mutation fractions of 0.03–0.39 in *S. aureus* tRNAs, indicating that *S. aureus* tRNAs can be highly modified with s<sup>4</sup>U8. *B. viscericola* showed no mutation significantly above background for any tRNA, suggesting that s<sup>4</sup>U8 is absent in this bacterium. Overall, our results suggest that bacterial tRNA modifications can be highly variable

at the transcriptome level among tRNA species even under mid-log growth conditions. These findings all together demonstrate the utility of tRNA-seq to study modifications in bacterial tRNA.

**Community and anticodon profiles in gut microbiomes.** Next, we evaluated the utility of tRNA-seq to infer tRNA abundance and modification profiles in complex gut microbial ecosystems using mice that were fed either a HF or LF diet. Since HF and LF diet affect microbial community structures in the gut<sup>27</sup>, this model provided an opportunity to also compare community structures inferred by tRNA-seq to those that were identified through 16S rRNA gene amplicons. We collected our samples from the mouse cecum, and processed them both for 16S rRNA gene analysis for taxonomic comparisons, and tRNA-seq with and without demethylase treatment. We developed an open-source software, tRNA-seq-tools (<https://github.com/merenlab/tRNA-seq-tools>) to de novo identify tRNA sequences with conserved sequence motifs and secondary structures from the raw sequencing data (Fig. 3a). We assigned taxonomy to resulting tRNA sequences, and we used only those sequences that were assigned to a unique anticodon for downstream analyses.

DM-tRNA-seq generated all tRNA reads starting from the conserved 3'-CCA<sup>8</sup>. We designed tRNA-seq-tools to use the canonical tRNA sequence motifs (Supplementary Fig. 4) that start with the 7 nucleotides 3' acceptor stem and 17 nucleotides TΨC stem-loop, a variable region of 4–5 for type I and 13–22 nucleotides for type II tRNAs (tRNA<sup>Ser/Leu/Tyr</sup> for bacteria), and



**Fig. 3** Microbiome tRNA-seq workflow and taxonomy analysis. **a** Workflow of tRNA sequencing of gut microbiome samples fed with a high-fat (HF) or low-fat (LF) diet and de novo tRNA assignment. Conserved tRNA residues that were searched for in this work are shown in red. **b** Dendrograms compare relationships between HF and LF samples that were inferred based on community profiles of tRNA transcripts, or 16S rRNA gene amplicons. **c** Class-level taxonomy for averaged HF and LF samples based on tRNA-seq (top) and 16S rRNA gene amplicons (bottom). All bacterial classes at >1% level are shown in distinct colors, all other bacterial classes are grouped together and shown in purple. **d** tRNA<sup>Gly</sup> taxonomy for anticodons GCC, UCC, and CCC. **e** tRNA<sup>Glu</sup> taxonomy for anticodons UUC and CUC. Among the other category for GCC/UCC/CCC and UUC, no class has an abundance of ≥1%; for CUC, other classes with an abundance of ≥1% include Alphaproteobacteria, Gemmatimonadetes, and Ignavibacteria. tRNAs decoding these two amino acids are the most abundant in our tRNA-seq results

finally a 17 nucleotides anticodon stem-loop. We included the conserved, canonical tRNA sequence of GTTC/C in the T<sub>Ψ</sub>C stem-loop and T 5' and A/G 3' to the anticodon nucleotides. We further restricted our assignment to anticodon stems containing only Watson-Crick or G-U wobble base pairs, but allowed for one mismatch among all conserved nucleotide sequences. The minimal length of our assigned tRNA reads corresponded to the 5' end of the anticodon stem to the 3'-CCA which was 48–49 for type I and 57–66 nucleotides for type II tRNAs.

Our analysis on average resulted in ~2.2 million tRNA sequences per demethylase-treated sample containing unambiguous anticodon assignments (Supplementary Table 2). There

were on average ~3.8 million other reads that were either too short to be assigned to a unique tRNA anticodon or did not unambiguously fit to the canonical tRNA signatures. Short sequences are often the result of persistent modifications demethylases fail to remove, or other tRNA structural features that interrupt the cDNA synthesis. An example of the former in our data was the ms<sup>2</sup>i<sup>6</sup>A37 modification in *E. coli* tRNA<sup>Phe</sup> which stopped the RT at ~99% (Supplementary Fig. 1b). Therefore, most tRNA<sup>Phe</sup> reads in *E. coli* did not reach the anticodon and would not have been taken into consideration here. For the tRNA sequences with anticodon assignments, grouping the biological replicates by individual anticodons showed good compatibility

among the HF or LF groups (Supplementary Fig. 5). Our tRNA extraction method from the mouse cecum resulted in less than 1% contamination with mouse tRNA sequences, which was consistent with our microarray studies<sup>28</sup>. We randomly selected one type I and one type II tRNA and aligned matching quality-filtered short reads to visualize the coverage and homogeneity of individual nucleotide positions (Supplementary Fig. 6). Consistent with our assignment rules, all reads were beyond the 5' stem of the anticodon stem-loop with gradual drop-offs toward the 5' end of tRNA, which we had also observed with human tRNA reads<sup>8</sup>.

We performed de novo analysis of tRNA sequences and 16S rRNA gene amplicons using Minimum Entropy Decomposition<sup>29</sup> to infer community structures, which revealed nearly identical relationships between samples (Fig. 3b). We also compared the average taxonomic distribution of HF/LF samples from tRNA-seq and 16S rRNA gene amplicons (Fig. 3c). While we used the SILVA database<sup>30</sup> with >350,000 entries to assign taxonomy for 16S rRNA gene amplicons, we used 4235 gold-standard bacterial genomes obtained from the Ensembl database<sup>31</sup> to assign taxonomy for tRNA sequences (Supplementary Table 3). For the six most abundant bacterial classes, the full tRNA taxonomy qualitatively matched the 16S rRNA gene-based taxonomy (Fig. 3c), but proportions of taxa differed between the two approaches. For instance, tRNA-seq showed higher fractions of Clostridia and Actinobacteria, and lower fractions of Bacteroidia and Erysipelotrichia classes. Multiple factors could result in differences between abundance estimates of the two approaches, one among them being the utilization of RNA transcripts in the tRNA-seq workflow in contrast to PCR amplification of genomic DNA in the 16S rRNA gene-based workflow (see Discussion).

We also analyzed the taxonomic make up of tRNA sequences at the anticodon-level (Fig. 3d, e, Supplementary Table 4). The two most abundant tRNA sequences in our dataset decoded amino acids of glycine and glutamic acid. Previous analyses based on available genomes show that for tRNA<sup>Gly</sup> (Fig. 3d), UCC is present in only all, GCC in most, and CCC in only a small number of bacterial taxa<sup>32</sup>. The taxonomy of tRNA<sup>Gly</sup>(GCC) was similar to the taxonomic make up of all tRNA sequences. The tRNA<sup>Gly</sup>(UCC) taxonomy has a larger representation of Clostridia, and the tRNA<sup>Gly</sup>(CCC) taxonomy diverged from the all-tRNA taxonomic profiles such as it showed more Bacilli in the HF sample. Similarly, for tRNA<sup>Glu</sup> (Fig. 3e), UUC is present in all, while CUC is found only in some bacterial genomes. Our results showed that the tRNA<sup>Glu</sup>(UUC) taxonomy was more similar to the overall taxonomy of all tRNA sequences, whereas the tRNA<sup>Glu</sup>(CUC) taxonomy had a different profile with a larger representation of Actinobacteria in the LF sample.

To gain further insights into tRNA anticodon-based taxonomy, we performed additional analyses of the anticodons CUC and UUC of sequences that encode for glutamic acid (Supplementary Fig. 7). tRNA<sup>Glu</sup>(UUC) can read both GAA and GAG codons and is essential in all cells, whereas tRNA<sup>Glu</sup>(CUC) can only read GAG codon and is optional for life. We found that the taxonomic proportion of tRNA<sup>Glu</sup>(UUC) varies widely among the six major bacterial classes compared to the taxonomic profiles based on all tRNA sequences as well as 16S rRNA gene amplicons (Supplementary Fig. 7a). Zooming into the class Bacilli further, we found that 75% of all tRNA<sup>Glu</sup>(CUC) genes (78/104) are concentrated in the Lactobacillaceae family (Supplementary Fig. 7b), suggesting that this family can be uniquely assessed using tRNA<sup>Glu</sup>(CUC) reads. Lactobacillaceae could be resolved both through tRNA sequences and 16S rRNA gene amplicons which showed similar abundance patterns between HF and LF microbiome samples (Supplementary Fig. 7c).

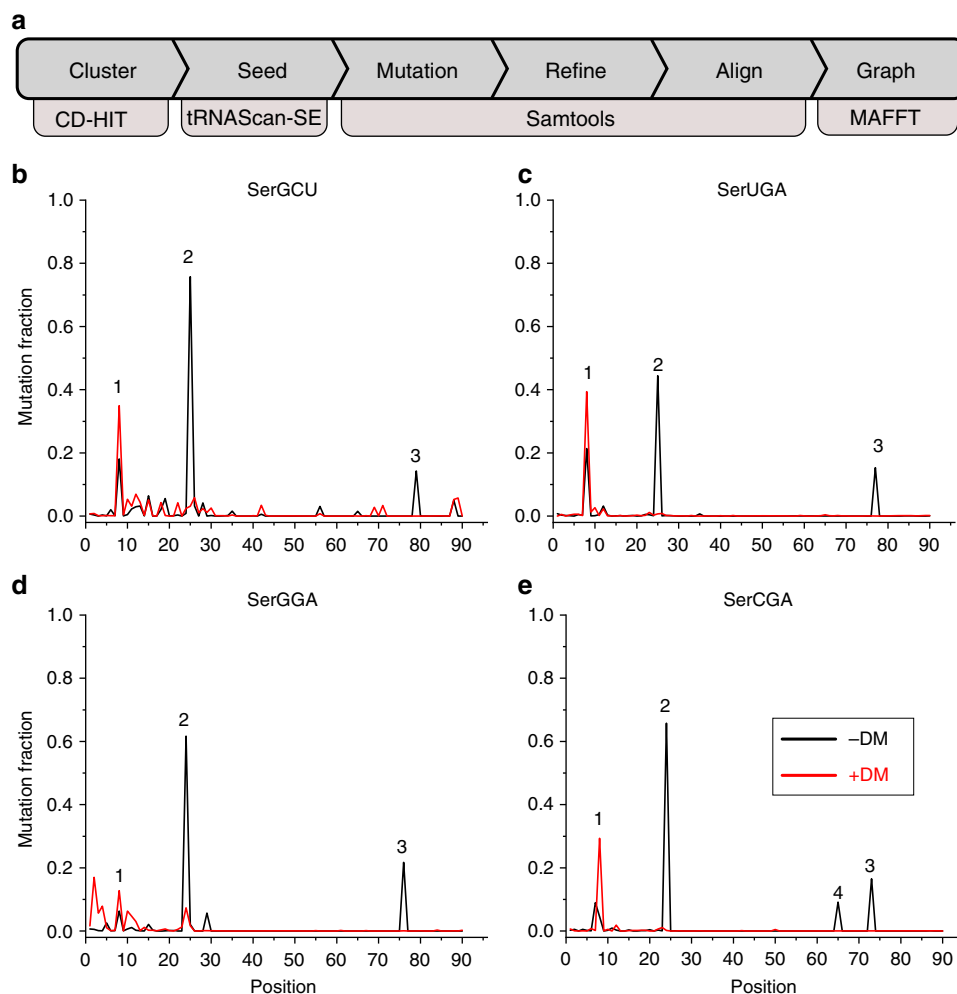
Focusing further on tRNA<sup>Glu</sup>(CUC) and tRNA<sup>Glu</sup>(UUC) reads from the family of Lactobacillaceae, we found that the ratio of tRNA<sup>Glu</sup>(CUC) to tRNA<sup>Glu</sup>(UUC) was higher in the LF samples than the HF samples (Supplementary Fig. 7d). This could be due to either higher tRNA<sup>Glu</sup>(CUC) expression or a higher proportion of Lactobacillaceae organisms that contain tRNA<sup>Glu</sup>(CUC) in the LF samples. In the case of the former, the increased tRNA<sup>Glu</sup>(CUC) expression could potentially add an additional capacity for the Lactobacillaceae in LF samples to specifically decode the GAG codon in translation. Finally, we found that all tRNA<sup>Glu</sup>(CUC) reads in our samples are from the *Lactobacillus* genus in the Lactobacillaceae family. This genus is represented by 31 genomes in our tRNA database; 18 different tRNA<sup>Glu</sup>(CUC) gene sequences are present among the 31 genomes, which allowed us to investigate the species-level distribution of reads that resolved to tRNA<sup>Glu</sup>(CUC) (Supplementary Fig. 7e). Overall, these results indicate that anticodon-level taxonomic analyses of tRNA transcripts can provide additional insights into physiological differences between individual branches of bacteria.

**Diet dependent differences in tRNA modification levels.** A unique feature of tRNA-seq is the capability to assess in situ tRNA modifications. Compared to bacterial cultures, the analysis of modifications in tRNA transcripts in complex microbial communities presents additional challenges. For instance, even though both mutations and stops are useful for tRNA modification studies in pure cultures, in microbiome studies we can only rely on mutation information since stops yield short reads that often do not assign to specific tRNA seed sequences unambiguously. Mutations in the sequencing data can be identified through the alignment of individual reads to reference seed sequences. For bacterial cultures, the seed sequences are simply the annotated tRNA genes. In contrast, complex environments may require de novo-identified tRNA sequences to serve as seed sequences for modification analyses due to the lack of comprehensive reference genomes (Fig. 4a, Supplementary Fig. 8).

To study modification fractions we selected our seed sequences from demethylase-treated samples. Our clustering-based approach (see Materials and Methods) resulted in an average of ~10,700 seed sequences for each sample, which was approximately proportional to the total number of tRNA reads per sample (Supplementary Table 2).

We ran a multiple sequence alignment for the same anticodon while allowing flexible lengths in the variable loop and the  $\alpha$  and  $\beta$  regions of the D loop to visualize our results. Figure 4 shows our workflow to study modification fractions, and the alignment results for the tRNA<sup>Ser</sup> transcripts in a HF sample. We found three major peaks for tRNA<sup>Ser</sup>(GCU) (Fig. 4b) without demethylase treatment. Peaks 2 and 3 disappeared upon demethylase treatment, therefore assigning them to m<sup>1</sup>A. Peak 1 remained upon demethylase treatment, thereby assigning it to s<sup>4</sup>U. We found the same 3 peaks for the other three tRNA<sup>Ser</sup>s with the same responses to demethylase treatment (Fig. 4c–e). We found a demethylase-removable fourth peak for tRNA<sup>Ser</sup>(CGA) (Fig. 4e) which may represent m<sup>1</sup>A in tRNA species with shorter variable arms.

We then analyzed bacterial taxon-dependent modifications. Examples in Fig. 5a were tRNA<sup>Ser</sup>(UGA) in the same microbiome sample. For this tRNA from *Lactobacillus*, class Bacilli, we readily identified mutation signatures consistent with m<sup>1</sup>A22 and s<sup>4</sup>U8 (Fig. 5a, Supplementary Fig. 9a), both were also known in tRNA<sup>Ser</sup>(UGA) of *B. subtilis*, class Bacilli. For *Bifidobacterium*, class Actinobacteria, mutation signature consistent with an m<sup>1</sup>A in the T loop (m<sup>1</sup>A59 in standard tRNA nomenclature) can be readily identified, but this tRNA did not have s<sup>4</sup>U8 (Fig. 5a,



**Fig. 4** Microbiome tRNA modification analysis. **a** Workflow for modification assignment using mutation signatures. **b–e** Representative positional plots showing m<sup>1</sup>A and s<sup>4</sup>U modifications for transcripts of tRNA<sup>Ser</sup>(GCU) (**b**), tRNA<sup>Ser</sup>(UGA) (**c**), tRNA<sup>Ser</sup>(GGA) (**d**), and tRNA<sup>Ser</sup>(CGA) (**e**), HF-fed mouse sample. The peak numbers correspond to those described in the text with peak 1 called for s<sup>4</sup>U8 and peak 2 for m<sup>1</sup>A22. Peak 3 is located around nucleotide 73–79 in the type II tRNA<sup>Ser</sup>s, but is m<sup>1</sup>A59 in the standard tRNA nomenclature. Red and black lines show mutation fractions in tRNA<sup>Ser</sup> sequences with (+DM) and without (–DM) demethylase treatment, respectively

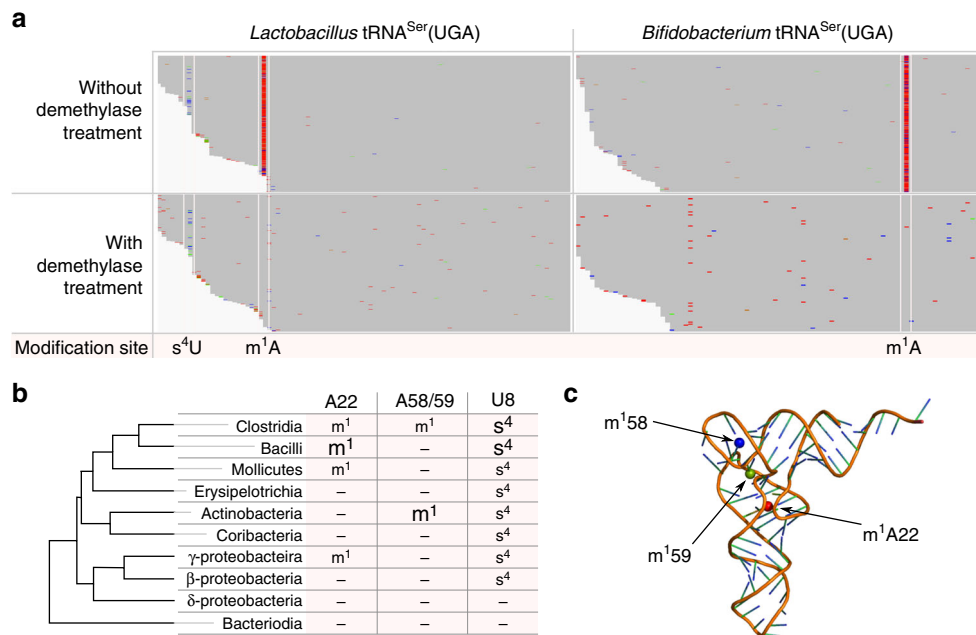
Supplementary Fig. 9b). M<sup>1</sup>A59 modification has also been found in *Streptomyces griseus*<sup>12</sup> in class Actinobacteria, although *S. griseus* is in the order Actinomycetales, whereas *Bifidobacterium* belongs to the order Bifidobacteriales. We also identified the more common m<sup>1</sup>A58 modifications, e.g. in tRNA<sup>Phe</sup> from *Bifidobacterium* (Supplementary Fig. 9c).

We identified a class-dependent modification pattern at three positions (Fig. 5b). M<sup>1</sup>A22 was predominantly present in the closely related Clostridia and Bacilli classes. M<sup>1</sup>A58 and m<sup>1</sup>A59 were predominantly present in Actinobacteria. S<sup>4</sup>U8 is widely spread in most of the bacterial classes, but absent in Bacteroidia and Deltaproteobacteria. The absence of m<sup>1</sup>A and s<sup>4</sup>U in Bacteroidia is unlikely due to the lack of sequencing depth as this result is consistent with our bacterial culture study that did not find s<sup>4</sup>U8 for *B. vissericola* from class Bacteroidia (Fig. 2b). We superimposed the three m<sup>1</sup>A modifications onto the three-dimensional tRNA structure (Fig. 5c). M<sup>1</sup>A22, m<sup>1</sup>A58, and m<sup>1</sup>A59 all introduce a positive charge in the tRNA, and are all located in the elbow region of the tRNA structure, suggesting that they may serve a common function.

To investigate whether modification fractions changed between HF and LF group, we analyzed the mutation fractions of m<sup>1</sup>A and s<sup>4</sup>U sites in all tRNAs (Fig. 6). The mutation fraction is not fully

quantitative for absolute modification fraction due to tRNA sequence context-dependent differences. However, comparative analyses of modification levels at each site between two samples can be interpreted with high confidence in regard to relative changes in modification fractions, as each site has the same sequence context. For Lachnospiraceae, m<sup>1</sup>A22 was present in eight tRNA anticodon groups, but in seven of the eight, the median level was higher in the HF than the LF samples (Fig. 6a). For Bifidobacteriaceae, m<sup>1</sup>A58 and m<sup>1</sup>A59 were present in 12 tRNA anticodon groups; in all 12 groups, the median level was higher in the HF than the LF samples (Fig. 6b). The m<sup>1</sup>A level change under dietary conditions of the mouse gut microbiome may be related to translation activity. In a human cell study, m<sup>1</sup>A modified tRNA had an increased affinity to the translation elongation factor and increased the expression of a reporter gene<sup>33</sup>. On the other hand, m<sup>1</sup>A-hypomodified tRNAs may be better tuned for interaction with other proteins in cells to perform extra-translational functions<sup>34</sup>.

For Lachnospiraceae, s<sup>4</sup>U8 was present in 24 tRNA anticodon groups; no clear preference for HF versus LF samples could be discerned (Fig. 6c), indicating that the s<sup>4</sup>U levels did not respond to the dietary conditions of the mouse gut microbiome. S<sup>4</sup>U8 is known to serve as a sensor to elicit cellular responses to UV;<sup>35</sup> gut



**Fig. 5** Taxonomic differences of modification sites. **a** Examples of aligning tRNA sequencing reads to two seed sequences of tRNA<sup>Ser</sup>(UGA) from *Lactobacillus*, class Bacilli, and *Bifidobacterium*, class Actinobacteria without and with demethylase treatment. Modification sites identified (s<sup>4</sup>U and m<sup>1</sup>A) are highlighted between the white lines. **b** m<sup>1</sup>A22, m<sup>1</sup>A58/59, and s<sup>4</sup>U8 identified in the abundant bacterial classes from Fig. 3c in the context of their phylogenetic relationship. Large fonts indicate bacterial classes in which the majority of the modifications are found (m<sup>1</sup>A22 in Clostridia and Bacilli, m<sup>1</sup>A58/59 in Actinobacteria, and s<sup>4</sup>U8 in Clostridia and Bacilli). **c** Proximal location of the m<sup>1</sup>A22 (red), m<sup>1</sup>58 (blue), and m<sup>1</sup>59 (green) modifications in a tRNA three-dimensional structure

microbiomes are not exposed to UV light and therefore may not adjust its s<sup>4</sup>U8 levels to dietary differences.

#### tRNA modification fractions and protein expression levels.

While differences across dietary conditions suggest tRNA modification fractions could reflect environmental responses, these results by themselves do not communicate whether modification fractions are associated with protein synthesis. To investigate associations between the m<sup>1</sup>A modification fractions observed in our study and protein synthesis, we used the gut metaproteomics data generated by Figeys, Stintzi, Mack, and co-workers<sup>36</sup> also from mice fed with HF and LF diet. In their study, the authors describe a total of 849 proteins with significantly different expression levels between the HF and LF samples collected over a course of 43 days. The day 29 and 43 experimental conditions most closely resemble the experimental condition of our tRNA-seq experiment. The metaproteomics data from both day 29 and 43 showed a marked difference between HF and LF conditions (e.g., Figure 7a). Our taxonomic analysis assigned 77% (655/849) of these proteins to the class Clostridia (Fig. 7b), and we focused our subsequent analysis specifically on this group (Fig. 7c). To examine a possible relationship between these differentially expressed proteins and tRNA m<sup>1</sup>A modifications, we compared the amino acid and codon compositions of these proteins that are highly expressed in HF samples (log >1) and depleted in HF samples (log < -1). In Clostridia, m<sup>1</sup>A modification is present in tRNAs that decode amino acids Cys, Glu, Gln, and Ser (Fig. 6a). We combined amino acid residues or codons based on protein expression levels in HF over LF, and subtracted the compositions of HF over-expressed proteins minus the HF under-expressed proteins (Fig. 7d, e). In both cases, Cys and Glu, but not Gln and Ser are over-represented among the highly over-expressed proteins, although this over-representation was not unique to Cys and Glu.

Since ribosome decoding includes an adjacent pair of mRNA codons in the A and P site, we also combined amino acid and codon pairs in a similar fashion (Fig. 7f, g, Supplementary Fig. 10). Excitingly, the top five most over-represented amino acid pairs in both day 29 and 43 HF over-expressed proteins relative to the HF depleted proteins were E-E (Glu-Glu), K-E (Lys-Glu), L-E (Leu-Glu), E-K (Glu-Lys), and R-E (Arg-Glu). For codon pairs, the top five most over-represented were GAG-GAA (Glu-Glu), AAG-GAA (Lys-Glu), GCG-GAG (Ala-Glu), CTG-GAA (Leu-Glu), and CCG-GAG (Arg-Glu) which closely matched the over-represented amino acid pairs. Decoding the top five over-represented amino acid and codon pairs all involves the m<sup>1</sup>A-containing tRNA<sup>Glu</sup> (anticodon CUC, UUC) in HF samples (Fig. 6a). These proteomic results are therefore consistent with our hypothesis that tRNAs with higher m<sup>1</sup>A levels enhance the decoding of their respective amino acid/codon pairs. Overall, this analysis indicates that tRNA m<sup>1</sup>A modification level differences are consistent with m<sup>1</sup>A modified tRNAs facilitating increased expression of microbiome proteins with specific amino acid and codon contexts.

#### Discussion

Here we described an approach to perform and characterize the high-throughput sequencing of tRNA transcripts that enables insights into the translational states of naturally occurring microbial communities. Although we previously have demonstrated the efficacy of tRNA-seq for studying human cells, the application of this strategy to study environmental microbes comes with unique experimental and computational challenges that have not been addressed previously.

Our study establishes a computational workflow for the recovery of tRNA sequences and the characterization of their abundance and modification patterns in environmental microbiomes. However, there is substantial room for further



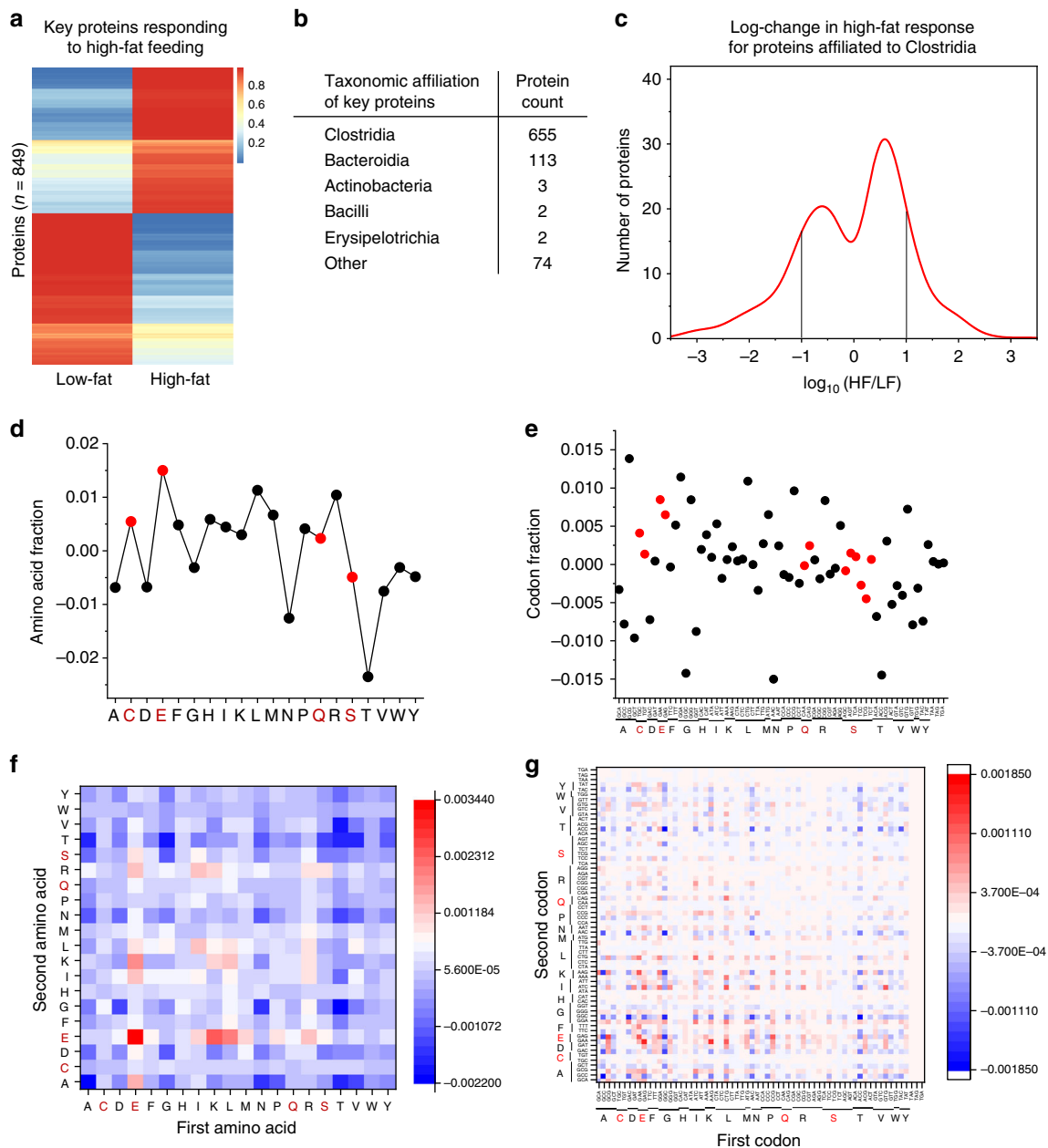


**Fig. 6** Comparisons of mutation fractions of HF versus LF samples. Bacterial families with the highest numbers of modifications at the respective nucleotides are shown. Each pair shows HF and LF samples with distinct anticodons marked on top. The amino acid whose codons are read by the corresponding tRNA with designated anticodon can be found in Supplementary Table 4. Box-and-whisker plots show median as a line, upper and lower quartiles in the box, and outliers outside of the line. **a**  $m^1A22$  from Lachnospiraceae, class Clostridia. **b**  $m^1A58$  and  $m^1A59$  from Bifidobacteriaceae, class Actinobacteria. **c**  $s^4U8$  from Lachnospiraceae, class Clostridia

improvements. For example, the tRNA gene database we used for taxonomic annotations contained only gold-standard isolate genomes of bacteria. tRNA gene annotation from the large number of metagenome-assembled and single-amplified genomes<sup>37</sup> could dramatically expand the database of reference tRNA sequences. This would be particularly important for applying tRNA-seq to ecological samples for which fewer isolate genomes are available. Our strategy to de novo identify tRNA sequences of identical origin to infer microbial community structures relied on an algorithm which has originally been developed for marker gene amplicon surveys<sup>29,38</sup>. However, unlike the sequencing products of most marker gene amplicon surveys, tRNA sequences contain more length variation. Hence, potential improvements of recovering abundance and distribution patterns through more specific algorithms warrant further study. Finally, strategies to select seed sequences for modification analyses directly influence the de novo identification of tRNA modification sites and mutation rates in environmental samples. While our current workflow only relied on seed sequences recovered from tRNA-seq data itself, this critical step could benefit from more comprehensive approaches that combine de novo identified sequences from tRNA-seq with the expanded reference database of tRNA gene sequences.

On the experimental side, bacterial tRNA modifications at the Watson–Crick face that cannot be removed by the demethylases can bias the quantitative analysis of tRNA-based taxonomy. This is particularly difficult for certain tRNA families with modifications 3' to the anticodon that stop the thermophilic reverse transcriptase we used here. For example, the  $ms^2i^6A37$  modification seems to be present in tRNA<sup>Phe</sup> of three cultured bacteria we studied here. Because our current microbiome analysis assigns each tRNA to a specific anticodon, any read that stopped at this position would not have an anticodon assignment. Such modification-dependent RT stops can explain in part why on average only ~33% of tRNA reads were assigned to a unique anticodon in this work. One possible improvement is to chemically or enzymatically remove this modification from microbiome tRNA prior to cDNA synthesis. The relatively high biomass requirement, which is in the range of 0.1–10  $\mu$ g total RNA, of the current experimental workflow poses another bottleneck. While this limitation is not necessarily a concern for gut microbiome studies, most applications of tRNA-seq to terrestrial and marine habitats with low microbial biomass will require improvements in molecular biology techniques.

Our analyses that used the same set of samples both with 16S rRNA gene amplicons and tRNA-seq revealed quantitative



**Fig. 7** Analysis of differential protein expression and tRNA modification. Proteomics data from HF- and LF-fed mice were from reference<sup>36</sup>. **a** Average differential expression of 849 proteins between HF- and LF-fed mouse gut microbiome from day 43 mice that most mimics the experimental condition of our tRNA-seq experiment. **b** BLASTp protein sequence analysis shows that most of these proteins are from class Clostridia. **c** Quantitative difference between the clostridia proteins from day 43 mice. Lines show the boundaries of the proteins used for downstream analysis that are highly enriched ( $\log > 1$ , 88 proteins) or depleted ( $\log < -1$ , 105 proteins) in HF over LF samples. The difference in amino acid (**d**) or codon content (**e**) determined by subtracting the compositions of HF over-expressed proteins minus the HF under-expressed proteins. The amino acids or codons for which their decoding tRNAs were found to contain m<sup>1</sup>A modifications are in red: Cys, Glu, Gln, Ser. The difference in amino acid pair (**f**) and codon pair content (**g**) determined by subtracting the pair compositions of HF over-expressed proteins minus the HF under-expressed proteins. The first amino acid represents the N-terminal residue and the first codon represents the 5' codon

differences in estimates of taxonomic composition, which may be due to several factors. For instance, the taxonomic analysis of tRNA sequences in our study relied on a database of ~4000 isolate genomes with biases towards certain bacterial groups. In contrast, the same analysis for 16S rRNA gene amplicons used a database of >350,000 rRNA gene sequences. In addition, RT-stopping modifications can also bias tRNA abundances in a taxa-dependent manner. It is also worth noting that in contrast to the tRNA-seq workflow, the 16S rRNA gene amplicons originate from the genomic DNA through the use of 'universal' PCR

primers, the efficiency of which may differ among microbial branches<sup>39,40</sup>, and can amplify environmental DNA from inactive or dead cells<sup>41</sup>. To what extent these factors are responsible for the observed differences and how they impact our understanding of microbial ecosystems need further investigation.

Besides these challenges, tRNA-seq comes with several distinct strengths as it offers access to an overall physiological assessment of all cells in a complex environmental sample through an actively expressed gene product via high-throughput sequencing. As it does not require any prior knowledge, tRNA-seq enables

investigations of tRNA modifications associated with uncharacterized bacterial branches or tRNAs with non-canonical sequence and structural features. For instance, our application of tRNA-seq to mouse gut microbiome revealed previously unknown class-dependent tRNA modification patterns in bacteria. While m<sup>1</sup>A22 and s<sup>4</sup>U8 in Bacilli are the same as those in *B. subtilis*, here we were able to map the presence or absence of m<sup>1</sup>A22, m<sup>1</sup>A58, m<sup>1</sup>A59, and s<sup>4</sup>U8 in other bacterial classes. The class-dependent modification patterns at these locations suggest a non-identical use of tRNA modifications. Unexpectedly and excitingly, the m<sup>1</sup>A sites showed significant differences in their modification levels for the gut microbiome fed with different diets. Our analysis of amino acid and codon contents of the metaproteomes under the same dietary conditions lend support to our hypothesis that increased m<sup>1</sup>A modification levels enhance translation of specific proteins in the microbiome in an amino acid and codon context-dependent manner. These results suggest a potential use of modification fractions as molecular markers to investigate microbial responses to changing environments.

In summary, we described an approach that makes use of high-throughput sequencing of tRNA transcripts to gain insights into physiological states of naturally occurring microbial communities. tRNA-seq represents another direction in microbiome research as it relies on an actively expressed gene product that is core to protein synthesis machinery, universally abundant, and phylogenetically informative. It offers a complementary approach to the sequencing of 16S rRNA gene amplicons and genome-resolved metagenomics for broader characterization of environmental microbiomes.

## Methods

**Bacterial culture growth.** We grew *E. coli* BW25113, *B. subtilis* 168, and *S. aureus* RN4220 cultures at 37 °C in Erlenmeyer flasks with a 10:1 flask medium ratio at 225 rpm in a shaker. We grew *E. coli* and *B. subtilis* in LB Lennox and *S. aureus* in tryptic soy broth. We obtained *Barnesiella viscericola* C46 from the Deutsche Sammlung von Mikroorganismen und Zellkulturen (DSMZ). We grew *B. viscericola* in serum vials under strict anaerobic conditions at 37 °C in chopped meat medium<sup>42</sup>. We harvested all cultures at midlog by centrifugation.

**Animals.** All murine experimental procedures were approved by the University of Chicago Institutional Animal Care and Use Committee (IACUC). We followed all relevant ethical regulations. We bred C57BL/6 mice and maintained them under standard 12:12 h light/dark conditions at the University of Chicago. We fed age and litter-matched specific pathogen free (SPF) male C57BL/6 mice between 8 and 12 weeks old either a purified low-fat (LF; Harlan Teklad TD.00102) diet, or high saturated milk fat (HF; Harlan Teklad TD.97222) diet for four weeks ad libitum. Diets were formulated by Harlan Teklad and compositions<sup>43</sup>. We  $\gamma$ -irradiated gnotobiotic diets and tested them for sterility prior to use. We anesthetized mice using 10 mg/ml ketamine/xylazine followed by exsanguination and cervical dislocation. We snap-froze cecal contents in liquid nitrogen and stored at -80 °C. Three mice each from LF or HF-fed group were used in this study.

**tRNA extraction, purification, and sequencing.** We lysed cecal contents or pelleted bacteria in 400  $\mu$ L of 0.3 M NaOAc/HOAc, 10 mM EDTA pH 4.8 with an equal volume acetate saturated phenol chloroform pH 4.8. We placed samples in a reciprocating bead beater for 45 s on maximum intensity using FastPrep lysing matrix B. We centrifuged samples at 21,000 rcf for 15 min at 4 °C before re-extraction and ethanol precipitation of total RNA. We purified tRNA by running RNA pellets on a 10% denaturing polyacrylamide gel and cutting type I and type II tRNA bands identified through UV shadowing. We obtained about 2  $\mu$ g total tRNA from 25  $\mu$ g total RNA. We eluted tRNA in 200 mM KCl/50 mM KOAc, ethanol precipitated, resuspended in water.

We performed tRNA sequencing as carried out for human tRNA studies<sup>8</sup>. Briefly, starting from about 2  $\mu$ g total tRNA, we left 1  $\mu$ g tRNA untreated, and subjected the remaining 1  $\mu$ g tRNA to demethylase treatment to remove Watson-Crick face methylations on tRNAs known to be present in bacterial tRNAs that can impede the processivity of reverse transcription (such as m<sup>1</sup>A or m<sup>1</sup>G). We performed reverse transcription of 100 ng of demethylase treated and untreated tRNA with a thermophilic reverse transcriptase (TGIRT), which synthesizes cDNA with high processivity. TGIRT only requires base pairing with the 3' terminal nucleotide, which is adenosine in all mature tRNAs, hence no sequence-specific primer is needed. We purified and eluted cDNAs on a 10% polyacrylamide denaturing gel, circularized cDNA and followed by PCR amplification of 12 cycles

for Illumina library construction. We sequenced each tRNA sample twice: with demethylase treatment to generate data for stringent abundance measurement and without demethylase treatment to generate data for modification analysis.

**16s rRNA gene sequencing.** We used a fraction of total cecal contents to extract genomic DNA by lysing samples in 100  $\mu$ L water with an equal volume phenol/chloroform and 100  $\mu$ L of 0.5 mm glass beads. We then performed five 1 min vortexing cycles followed by 1 min on ice in between each cycle. We then centrifuged samples for 15 min at 21,000  $\times$  g at 4 °C, re-extracted, and used ethanol precipitation to yield purified genomic DNA. We then produced amplicon libraries targeting the V4 region of the 16S rRNA gene using barcoded PCR primers 515F (5'-GTGCCAGCMGCGCGGTAA-3') and 806R (5'-GGACTACHVGGGTWTC TAAT-3') for Illumina MiSeq<sup>44</sup>. Sequencing of these libraries at the Environmental Sample Preparation and Sequencing Facility (ESPSE) at Argonne National Laboratory generated 151 bp paired-end amplicon sequences.

**Analysis of the 16S rRNA gene amplicons.** We removed low-quality reads from the raw sequencing results using illumina-utils<sup>45</sup> (available from <https://github.com/merenlab/illumina-utils>). For the analysis of 16S rRNA gene amplicons data, we used the program 'iu-merge-pairs' with default parameters, which merged partially overlapping paired-end Illumina reads while simultaneously removing any pair with more than 3 mismatches at the overlapped region. For pairs with three or less mismatches, we picked the base to be used in the final merged sequence from the read with the higher Q-score. However, if neither of the reads had a Q-score higher than Q10 at a position of mismatch, we also discarded the pair to diminish the impact of noise to the final results. We used Minimum Entropy Decomposition<sup>29</sup> with default parameters to cluster high-quality 16S rRNA gene amplicons at 1-nt resolution, and GAST<sup>46</sup> to assign taxonomy to each read individually.

**Analysis of bacterial culture tRNA-seq data.** We performed tRNA-seq analysis for bacterial cultures, each with a single reference genome, in the same way as described in detail for human tRNA-seq analysis<sup>8,20</sup>. A brief description follows. We aligned sequencing reads using Bowtie 1.0<sup>47</sup> to a modified tRNA genome file containing annotated tRNA genes for each of the four bacterial species from Genomic tRNA database<sup>32</sup>. When needed, we appended 3'-CCA to tRNA genes so that all tRNA genes end with 3'-CCA. Prior to mapping, we processed reads using Trimmomatic v0.32<sup>48</sup> using default parameters to remove de-multiplexing and to remove primers, adapters, or any other low-quality sequences. We aligned sequences longer than 15 bp to the library using Bowtie 1.0 with sensitive options using the highest allowed mismatch settings for Bowtie 1.0. We mapped reads to all references simultaneously, and we reported only one alignment declared as valid by the respective mapping software for each read. For each position in the reference file, we calculated the following: at each position, how many total counts were present, how many mutations, and how many aligned reads stopped at the position.

**Analysis of microbiome tRNA-seq data.** Because tRNA molecules (74–96 nucleotides) are shorter than 100 bases, the short insert size results in sequencing of the same molecule twice and the additional sequencing of the Illumina adapters. As a result, the overlapping region in the alignment of the first read and the reverse-complement of the second read in a given pair represents the sequence of the tRNA molecule, and the tailing ends represent the Illumina adapters. To recover tRNA sequences in this dataset we also used the program 'iu-merge-pairs', but with two additional flags and an additional parameter: --retain-only-overlap (to keep only the overlapping region while trimming tailing ends from both reads in a pair), --marker-gene-stringent (to align the reverse-complement of the second read even if the alignment occurs between the end of the second read and the beginning of the first read --which is a unique case for short inserts), and --max-num-mismatches 0 (to remove any pair if there was a disagreement between the aligned portions to increase the quality dramatically as described in ref. 45). We developed a software tool to identify tRNA sequences in the raw sequencing results: tRNA-seq-tools (available from <https://github.com/merenlab/tRNA-seq-tools>). We used the tRNA-seq-tools program 'trna-profile' with default parameters to identify all sequences that matched our criteria for standard tRNA (Supplementary Fig. 3). Standard tRNAs contain several secondary structural regions: the acceptor stem, T $\psi$ C stem-loop, anticodon stem-loop, D stem-loop, and a short variable loop of 4–5 residues (type I tRNA) or a long variable stem-loop of 13–22 residues (type II tRNA). Conserved sequences at defined positions include the 3'-CCA, a C and GTTC in the T stem-loop, and TXXX(A/G) in the anticodon loop where XXX is the unique identifier for the anticodon residues. All reads in our sequencing method start from the 3'-CCA, so that each read can be assigned to a unique anticodon (Fig. 3a, Supplementary Fig. 3). From the self-contained database of tRNA sequences 'trna-profile' generated for each sample, we recovered tRNA sequences (full-length or non-full-length but with/without anticodon), and sequences matching to a specific anticodon using the programs 'trna-get-sequences' and 'trna-gen-anticodon-profile'. We used the full-length tRNA sequences to characterize the community composition de novo using Minimum Entropy Decomposition<sup>29</sup> with default parameters. For taxonomy assignment, we generated a database of tRNA sequences tRNAscan-SE (v1.3.1)<sup>49,50</sup> identified from 4,235

gold-standard bacterial genomes (non-endosymbiont genomes with an assembly level of ‘chromosome’) stored in the Ensembl Genomes 2016 database<sup>31</sup>. We then queried our database of tRNA sequences with GAST (queried in March 2017)<sup>46</sup> to individually assign taxonomy to each high-quality read in our tRNA sequencing results. We generated hierarchical clustering dendrograms with Bray-Curtis distance metric and Ward clustering algorithm to compare sample community profiles for both tRNA sequences and 16S rRNA gene amplicons.

**Analysis of microbiome tRNA modifications.** (a) Derivation of seed sequences from tRNA reads: To identify modified nucleotides in tRNA by read mapping, we first constructed a reference database consisting of seed sequences to which raw tRNA reads were then mapped back. For each FASTA file containing tRNA reads from each demethylase-treated sample, we grouped identical reads and collapsed them into a single sequence and recorded its count. We then clustered the resulting sequences to further reduce redundancy using CD-HIT (v4.6.8)<sup>51</sup> with sequence identity threshold set to 0.98. To reduce the impact of noise, we discarded any cluster with 5 or less reads. For each remaining cluster, we chose a representative read to generate a seed sequence by performing the following procedure. First, to obtain as many high-quality full-length tRNAs as possible, we used tRNAscan-SE (v2.0)<sup>50</sup> to evaluate the longest read in each cluster and kept only sequences with tRNAscan-SE score  $\geq 50$  to ensure that the resulting seed sequence had a high likelihood to fold like a tRNA for subsequent processing. The longest read could be truncated due to the RT stops at the 5' end. In >93% of all clusters, the seed sequence derived from the longest read was consistent with the most abundant read (with 100% sequence identity). Second, truncation of 5' ends in some of the representative reads from different clusters may lead to generation of identical seed sequences. After removing these redundant sequences we obtained the reference database from each demethylase-treated sample for subsequent alignment of tRNA reads.

(b) *Mapping of tRNA reads and identification of mutations:* To obtain mutation signatures we ran the alignment program Bowtie<sup>47</sup> on each seed for both libraries with and without demethylase treatment allowing two mismatches. To minimize sequencing artifacts where mutations might be present because of low coverage for a given seed we kept only the alignment results with total coverage of >50 at each nucleotide position. We then generated an output file for each nucleotide position, with the mutation fraction between 0.05–0.98. We used this range because mutation fractions <0.05 had a higher likelihood to be false positives derived from sequencing errors, whereas mutation fractions >0.98 might be derived from a highly abundant but redundant seed sequence that was removed in the last step of seed sequence identification. We further filtered the result by removing those positions where their mutation was only to one of the three possible sequences. These sites likely corresponded to single nucleotide polymorphisms in tRNA sequences. This step was justified because all bacterial culture mutations (Figs 1, 2) had two or three sequence changes, albeit at different ratios. For example, s<sup>4</sup>U produced both U-to-C and U-to-A mutations, but U-to-C is present at higher frequency than U-to-A. We performed one last filter to remove redundant entries from the seeds that differed by two nucleotides because these generated identical mutation results once we allowed two mismatches in Bowtie alignment.

We analyzed the alignment results by converting to binary BAM format, sorted and indexed using SAMtools (v1.5)<sup>52</sup>. We calculated the overall mutation rate and fractions of mutated nucleotide for each position in each seed sequence and kept only positions with depth of coverage >50 and mutation rate in the range of 0.05–0.98 for subsequent filtering. We defined positions with 2 zero nucleotide fractions as SNPs (single nucleotide polymorphisms) and removed them from the list of mutation entries. We also discarded the redundant mutation entries with the same nucleotide fractions provided that their seed sequences contained the same anticodon and differed by no more than 2 bases. These seed sequences were originated from the same tRNA sequence but were not grouped into one cluster in the previous steps.

(c) Graphing overall mutation rates for seed sequences in each anticodon family: We performed multiple sequence alignments for all seed sequences in each anticodon family by using MAFFT (v7.369b)<sup>53</sup>. For each aligned position we calculated a weighted sum of mutation rates from all seed sequences while excluding gaps and SNPs from the summation.

As expected, the degree of identification of mutation signatures depends on the sequencing depths of the individual sample. As an example, the m<sup>1</sup>A and s<sup>4</sup>U signatures in *Lactobacillus* and *Bifidobacterium* tRNA<sup>Ser</sup>(UGA) (Fig. 5a) can be found in all three LF samples and two of the three HF samples (HF02 and HF03), consistent with the sequencing coverage of these samples (Supplementary Table 2).

**Analysis of the microbiome metaproteomics data.** In the microbiome metaproteomics study from Mack, Stintzi, Figeys, and co-workers (MetaPro-IQ)<sup>36</sup>, 849 proteins were identified that have significantly different expression levels between high-fat and low-fat-fed mouse gut microbiome. We obtained the individual expression level of these proteins from Table S5 of the MetaPro-IQ paper. The MetaPro-IQ authors performed an iterative database search strategy in the MeaPro-IQ workflow to identify the protein sequences using the gut microbial gene catalog from reference<sup>54</sup> as the reference database. We then requested this catalog of mouse gut metagenome comprising ~2.6 million nonredundant genes

from the authors and extracted the 849 coding sequences corresponding to the above 849 proteins according to their entry IDs.

We obtained the taxonomic assignment for each of the 849 proteins using BLASTp, followed by converting the GI# of the top hit to the taxon ID. Among the hits, 98.9% had  $e$ -value  $< e^{-14}$ , these were used for subsequent analysis. Among those, 96.8% had >50% amino acid sequence identity and the remaining had an identity between 26.7 and 49.8%. We found that 77% of these (655/849) most closely match or resemble proteins in class Clostridia. These 655 proteins were analyzed further by counting the frequency of the amino acid residues or codons, either singly or in adjacent pairs, and comparing the protein groups that are highly over-expressed or depleted in HF versus LF samples.

**Code availability.** The source code of tRNA-seq-tools is accessible at <https://github.com/merenlab/tRNA-seq-tools>.

## Data availability

The raw sequencing data have been deposited in NCBI GEO database, accession number GSE100263. Figures 1–6 have associated raw data from the deposited raw sequencing data above. Figure 7 raw data was from the published work in ref. <sup>36</sup>. A Reporting Summary for this Article is available as a Supplementary Information file.

Received: 9 May 2018 Accepted: 9 November 2018

Published online: 17 December 2018

## References

- Pace, N. R. A molecular view of microbial diversity and the biosphere. *Science* **276**, 734–740 (1997).
- Human Microbiome Project Consortium. Structure, function and diversity of the healthy human microbiome. *Nature* **486**, 207–214 (2012).
- Gensollen, T., Iyer, S. S., Kasper, D. L. & Blumberg, R. S. How colonization by microbiota in early life shapes the immune system. *Science* **352**, 539–544 (2016).
- Turnbaugh, P. J. et al. The human microbiome project. *Nature* **449**, 804–810 (2007).
- Kirchner, S. & Ignatova, Z. Emerging roles of tRNA in adaptive translation, signalling dynamics and disease. *Nat. Rev. Genet.* **16**, 98–112 (2015).
- Widmann, J., Harris, J. K., Lozupone, C., Wolfson, A. & Knight, R. Stable tRNA-based phylogenies using only 76 nucleotides. *RNA* **16**, 1469–1477 (2010).
- Pang, Y. L., Abo, R., Levine, S. S. & Dedon, P. C. Diverse cell stresses induce unique patterns of tRNA up- and down-regulation: tRNA-seq for quantifying changes in tRNA copy number. *Nucleic Acids Res.* **42**, e170 (2014).
- Zheng, G. et al. Efficient and quantitative high-throughput tRNA sequencing. *Nat. Methods* **12**, 835–837 (2015).
- Cozen, A. E. et al. ARM-seq: AlkB-facilitated RNA methylation sequencing reveals a complex landscape of modified tRNA fragments. *Nat. Methods* **12**, 879–884 (2015).
- Schimmel, P. The emerging complexity of the tRNA world: mammalian tRNAs beyond protein synthesis. *Nat. Rev. Mol. Cell Biol.* **19**, 45–58 (2018).
- Gogakos, T. et al. Characterizing expression and processing of precursor and mature human tRNAs by hydro-tRNAseq and PAR-CLIP. *Cell Rep.* **20**, 1463–1475 (2017).
- Machnicka, M. A. et al. MODOMICS: a database of RNA modification pathways--2013 update. *Nucleic Acids Res.* **41**, D262–D267 (2013).
- Sprinzl, M. & Vassilenko, K. S. Compilation of tRNA sequences and sequences of tRNA genes. *Nucleic Acids Res.* **33**, D139–D140 (2005).
- Roundtree, I. A., Evans, M. E., Pan, T. & He, C. Dynamic RNA Modifications in gene expression regulation. *Cell* **169**, 1187–1200 (2017).
- El Yacoubi, B., Bailly, M. & de Crécy-Lagard, V. Biosynthesis and function of posttranscriptional modifications of transfer RNAs. *Annu. Rev. Genet.* **46**, 69–95 (2012).
- Grosjean, H., de Crécy-Lagard, V. & Marck, C. Deciphering synonymous codons in the three domains of life: co-evolution with specific tRNA modification enzymes. *FEBS Lett.* **584**, 252–264 (2009).
- Andachi, Y., Yamao, F., Muto, A. & Osawa, S. Codon recognition patterns as deduced from sequences of the complete set of transfer RNA species in *Mycoplasma capricolum*. Resemblance to mitochondria. *J. Mol. Biol.* **209**, 37–54 (1989).
- Puri, P. et al. Systematic identification of tRNAome and its dynamics in *Lactococcus lactis*. *Mol. Microbiol.* **93**, 944–956 (2014).
- Cao, X. & Limbach, P. A. Enhanced detection of post-transcriptional modifications using a mass-exclusion list strategy for RNA modification mapping by LC-MS/MS. *Anal. Chem.* **87**, 8433–8440 (2015).

20. Clark, W. C., Evans, M. E., Dominissini, D., Zheng, G. & Pan, T. tRNA base methylation identification and quantification via high-throughput sequencing. *RNA* **22**, 1771–1784 (2016).
21. Cantara, W. A. et al. The RNA modification database, RNAMDB: 2011 update. *Nucleic Acids Res.* **39**, D195–D201 (2011).
22. Rich, A. & Rajbhandary, U. L. Transfer RNA: molecular structure, sequence, and properties. *Annu. Rev. Biochem.* **45**, 805–860 (1976).
23. Menichi, B., Arnold, H. H., Heyman, T., Dirheimer, G. & Keith, G. Primary structure of *Bacillus subtilis* tRNAs<sup>Tyr</sup>. *Biochem. Biophys. Res. Commun.* **95**, 461–467 (1980).
24. Matsugi, J., Jia, H. T., Mura, K. & Ishikura, H. Nucleotide sequences of serine tRNAs from *Bacillus subtilis*. *Biochim. Biophys. Acta* **1130**, 333–335 (1992).
25. Farabaugh, P. J. & Bjork, G. R. How translational accuracy influences reading frame maintenance. *EMBO J.* **18**, 1427–1434 (1999).
26. Bjork, G. R., Wikstrom, P. M. & Bystrom, A. S. Prevention of translational frameshifting by the modified nucleoside 1-methylguanosine. *Science* **244**, 986–989 (1989).
27. Howe, A. et al. Divergent responses of viral and bacterial communities in the gut microbiome to dietary disturbances in mice. *ISME J.* **10**, 1217–1227 (2016).
28. Schwartz, M. H. & Pan, T. tRNA misacylation with methionine in the mouse gut microbiome in situ. *Microb. Ecol.* **74**, 10–14 (2017).
29. Eren, A. M. et al. Minimum entropy decomposition: unsupervised oligotyping for sensitive partitioning of high-throughput marker gene sequences. *ISME J.* **9**, 968–979 (2015).
30. Quast, C. et al. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res.* **41**, D590–D596 (2013).
31. Kersey, P. J. et al. Ensembl Genomes 2016: more genomes, more complexity. *Nucleic Acids Res.* **44**, D574–D580 (2016).
32. Chan, P. P. & Lowe, T. M. GtRNAdb 2.0: an expanded database of transfer RNA genes identified in complete and draft genomes. *Nucleic Acids Res.* **44**, D184–D189 (2016).
33. Liu, F. et al. ALKBH1-mediated trna demethylation regulates translation. *Cell* **167**, 816–828 (2016).
34. Parisien, M. et al. Discovering RNA-protein interactome by using chemical context profiling of the RNA-protein interface. *Cell Rep.* **3**, 1703–1713 (2013).
35. Kramer, G. F., Baker, J. C. & Ames, B. N. Near-UV stress in *Salmonella typhimurium*: 4-thiouridine in tRNA, ppGpp, and ApppGpp as components of an adaptive response. *J. Bacteriol.* **170**, 2344–2351 (1988).
36. Zhang, X. et al. MetaPro-IQ: a universal metaproteomic approach to studying human and mouse gut microbiota. *Microbiome* **4**, 31 (2016).
37. Bowers, R. M. et al. Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea. *Nat. Biotechnol.* **35**, 725–731 (2017).
38. Eren, A. M. et al. Oligotyping: Differentiating between closely related microbial taxa using 16S rRNA gene data. *Methods Ecol. Evol.* **4**, 1111–1119 (2013).
39. Polz, M. F. & Cavanaugh, C. M. Bias in template-to-product ratios in multitemplate PCR. *Appl. Environ. Microbiol.* **64**, 3724–3730 (1998).
40. Brown, C. T. et al. Unusual biology across a group comprising more than 15% of domain Bacteria. *Nature* **523**, 208–211 (2015).
41. Carini, P. et al. Relic DNA is abundant in soil and obscures estimates of soil microbial diversity. *Nat. Microbiol.* **2**, 16242 (2016).
42. Bacic, M. K. & Smith, C. J. Laboratory maintenance and cultivation of *Bacteroides species*. *Curr. Protoc. Microbiol.* **9**, 13C.1.1–13C.1.21 (2008).
43. Devkota, S. et al. Dietary-fat-induced taurocholic acid promotes pathobiont expansion and colitis in Il10<sup>-/-</sup> mice. *Nature* **487**, 104–108 (2012).
44. Caporaso, J. G. et al. Ultra-high-throughput microbial community analysis on the Illumina HiSeq and MiSeq platforms. *ISME J.* **6**, 1621–1624 (2012).
45. Eren, A. M., Vineis, J. H., Morrison, H. G. & Sogin, M. L. A filtering method to generate high quality short reads using illumina paired-end technology. *PLoS ONE* **8**, e66643 (2013).
46. Huse, S. M. et al. Exploring microbial diversity and taxonomy using SSU rRNA hypervariable tag sequencing. *PLoS Genet.* **4**, e1000255 (2008).
47. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* **10**, R25 (2009).
48. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).
49. Lowe, T. M. & Eddy, S. R. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res.* **25**, 955–964 (1997).
50. Lowe, T. M. & Chan, P. P. tRNAscan-SE On-line: integrating search and context for analysis of transfer RNA genes. *Nucleic Acids Res.* **44**, W54–W57 (2016).
51. Li, W. & Godzik, A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **22**, 1658–1659 (2006).
52. Li, H. et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
53. Katoh, K., Misawa, K., Kuma, K. & Miyata, T. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res.* **30**, 3059–3066 (2002).
54. Xiao, L. et al. A catalog of the mouse gut metagenome. *Nat. Biotechnol.* **33**, 1103–1108 (2015).

## Acknowledgements

The authors are grateful to Vanessa Leone for help with the germ free mouse colony, Rich Fox for help with the computer systems administration, and Mitch Sogin for comments on the manuscript. We also thank Drs. Xu Zhang and Daniel Figeyes for providing the 849 protein sequences of the mouse gut microbiome metaproteomics study, and Drs. Karsten Kristiansen, Liang Xiao, and Zhongkui Xia for the mouse gut metagenome catalog. Haipeng Wang acknowledges the support from the National Natural Science Foundation of China (31500669), Shandong Provincial Natural Science Foundation (ZR2014FQ024), and China Scholarship Council (201708370053). This work was supported by grants from the NIH P30DK042086 (pilot to T.P. and A.M.E.), P30DK020595 (pilot to T.P.) and RM1HG008935 (to T.P.), the University of Chicago startup funds (to A.M.E.), and the Keck Foundation (to T.P., A.M.E., and E.B.C.).

## Author contributions

M.H.S. performed all tRNA-seq experiments. H.W. performed microbiome tRNA modification analyses. J.N.P. and H.W. performed meta-proteomics analysis. M.J.E., D.W.P., M.P., and J.X. contributed to bioinformatics analyses. W.C.C. and T.P. performed tRNA-seq analysis of bacterial cultures. S.M.O. performed library preparation and sequencing for 16S rRNA gene amplicons. B.L.C., K.M., and E.B.C. provided mouse cecum samples. S.C. and A.M.E. developed the tRNA-seq-tools. M.H.S., T.P., and A.M.E. conceived the project, analyzed data, generated figures, and wrote the paper.

## Additional information

**Supplementary Information** accompanies this paper at <https://doi.org/10.1038/s41467-018-07675-z>.

**Competing interests:** The authors declare no competing interests.

**Reprints and permission** information is available online at <http://npg.nature.com/reprintsandpermissions/>

**Publisher's note:** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2018