Quantitative mapping of the cellular small RNA landscape with AQRNA-seq

Published in final edited form as:

Nat Biotechnol. 2021 August ; 39(8): 978–988. doi:10.1038/s41587-021-00874-y.

Quantitative mapping of the cellular small RNA landscape with AQRNA-seq

Jennifer F. Hu[1,†], **Daniel Yim**[2,¶], **Duanduan Ma**[3], **Sabrina M. Huber**[2,§], **Nick Davis**[2], **Jo Marie Bacusmo**[4], **Sidney Vermeulen**[2], **Jieliang Zhou**[5], **Thomas J. Begley**[6], **Michael S. DeMott**[2,7], **Stuart S. Levine**[3], **Valerie de Crécy-Lagard**[4], **Peter C. Dedon**[2,7,8,*], **Bo Cao**[2,8,9,*]

[1]Department of Chemistry, Massachusetts Institute of Technology, Cambridge, MA, USA.

[2]Department of Biological Engineering, Massachusetts Institute of Technology, Cambridge, MA, USA.

[3]BioMicro Center, Massachusetts Institute of Technology, Cambridge, MA, USA.

[4]Department of Microbiology & Cell Science, University of Florida, Gainsville, FL, USA.

[5]KK Research Center, KK Women's and Children's Hospital, 229899, Singapore

[6]The RNA Institute and Department of Biology, University at Albany, Albany, NY, USA

[*] Corresponding authors: B.C. (caobo@qfnu.edu.cn) and P.C.D. (pcdedon@mit.edu).
[†] Present address: Bristol Myers Squibb
[¶] Present address: A*STAR Genome Institute of Singapore
[§] Present address: Laboratory of Toxicology, ETH Zürich, Switzerland

Author Contributions

P.C.D., B.C., J.F.H. conceived of AQRNA-seq, designed the experiments, and wrote the first draft of the manuscript. P.C.D., J.F.H., B.C., and D.Y. developed the method and performed the sequencing experiments. J.F.H., B.C., D.M., and S.S.L. developed, implemented and interpreted the data processing workflows and computational analyses. D.Y., S.V. and J.F.H. performed mycobacterial culturing and RNA isolation. T.J.B. analyzed proteomics data for codon usage patterns. J.M.B. performed *E. coli* culturing and RNA isolation. N.D. performed proteomics analyses. S.M.H. optimized experimental conditions and characterized demethylation efficiency by LC-MS. S.M.H. and J.F.H. performed northern blot analyses. M.S.D. contributed reagents and analyzed miRNA data. J.Z. analyzed miRNA data. V.C.L. supervised *E. coli* experiments and contributed insights and analysis. All authors participated in the writing of the manuscript.

Competing Interests

B.C., J.F.H., D.Y., S.M.H., M.S.D., and P.C.D are co-inventors on two patents (PCT/US2019/013714, US 2019/0284624 A1) relating to the published work.

Data Availability

All custom scripts have been made available at https://github.com/dedonlab/. All sequencing and proteomics data that support the findings of this study have been variously deposited in public databases: RNA sequencing studies reported in Figures 3–5, Supplementary Figures 2, 4, and 8 and the proteomics studies reported in Figure 4 and Supplementary Figure 6 have been deposited in the NCBI Sequence Read Archive (SRA) under the BioProject ID PRJNA579244; miRNA and standards data shown in Figure 2 have been deposited in the Gene Expression Omnibus (GEO) under accession number GSE139936; data for miRNA studies in HMEC cells shown in Figure 6 have been deposited in GEO as accession number GSE159434.

Code Availability

The software used in the studies presented here is publicly available as follows. Blast version 2.6.0 (nucleotide BLAST) available at https://blast.ncbi.nlm.nih.gov/Blast.cgi?PAGE_TYPE=BlastDocs&DOC_TYPE=Download. Peakfit.m version 9.0 available at Tom O'Haver, MATLAB Central File Exchange - https://terpconnect.umd.edu/~toh/spectrum/. fgrep (Linux command) available at https://unix.stackexchange.com/questions/17949/what-is-the-difference-between-grep-egrep-and-fgrep. fastxtoolkit version 0.013 available at http://hannonlab.cshl.edu/fastx_toolkit/. Custom python scripts are available at GitHub https://github.com/dedonlab/ (https://github.com/dedonlab/aqrnaseq for prokaryotic process scripts and https://github.com/dedonlab/general_aqrnaseq for eukaryotic/general pipeline scripts).

[7]Center for Environmental Health Sciences, Massachusetts Institute of Technology, Cambridge, MA, USA

[8]Singapore-MIT Alliance for Research and Technology Antimicrobial Resistance IRG, Singapore

[9]College of Life Sciences, Qufu Normal University, Qufu, Shandong 273165, China.

Abstract

Current next-generation RNA sequencing methods do not provide accurate quantification of small RNAs within a sample due to sequence-dependent biases in capture, ligation, and amplification during library preparation. We present a method, Absolute Quantification (AQ) RNA-seq, that minimizes biases and provides a direct, linear correlation between sequencing read count and copy number for all small RNAs in a sample. Library preparation and data processing were optimized and validated using a 963-member miRNA reference library, oligonucleotide standards of varying lengths, and northern blots. Application of AQRNA-seq to a panel of human cancer cells revealed >800 detectable miRNAs that varied during cancer progression, while application to bacterial tRNA pools, with the challenges of secondary structure and abundant modifications, revealed 80-fold variation in tRNA isoacceptor levels, stress-induced site-specific tRNA fragmentation, quantitative modification maps, and evidence for stress-induced tRNA-driven codon-biased translation. AQRNA-seq thus provides a versatile means to quantitatively map the small RNA landscape in cells.

While greatly advancing functional genomics,[1, 2] current next-generation RNA sequencing (RNA-seq) methods provide precise and accurate analysis of changes in transcript abundance *between* samples, they cannot accurately quantify small RNA species *within* a sample. This is partly rooted in biased ligation of sequencing linkers to the 3’- and 5’-ends of RNAs, with sequence-dependent 10^3-fold variation in efficiency[3–7] causing 10^6-fold artifacts in sequencing read counts.[5, 8, 9] Highly structured and modified RNA molecules, such as tRNAs, further challenge the quantitative accuracy of RNA-seq[10, 11] by causing polymerase fall-off during cDNA synthesis.[10–13]

Many specialized RNA-seq methods, often limited to miRNA or tRNA,[14–16] attempt to minimize ligation bias using linkers with randomized ends, enhance ligation efficiency with molecular crowding agents,[4] or reduce reverse transcription (RT) polymerase fall-off with two-step ligation[8] and AlkB removal of methyl modifications.[17–19] Even with these changes, residual ligation biases lead to “jackpot” sequences,[8] such as biases caused by the overhanging nucleotide in the adapter strand with the template-switching TGIRT used in DM-tRNA-seq and TGIRT-seq.[18, 2021]

Few RNA-seq techniques have been systematically engineered to optimize ligation and amplification efficiencies or validated for quantitative accuracy and lack of bias artifacts. Furthermore, none are broadly applicable to all RNA species. Here we describe AQRNA-seq, a method that enables absolute quantification of all small RNA species in a sample by providing a direct, linear correlation between sequencing read count and RNA abundance. Library preparation and data mining algorithms were validated by multiple orthogonal approaches. Application of AQRNA-seq to stress-induced mycobacterial persistence revealed large variations in tRNA copy numbers, tRNA fragmentation, and tRNA modification location and abundance within and among samples. In a human mammary epithelial tumor model, AQRNA-seq quantified 875 miRNAs over a 10^5-fold range and revealed that the majority of so-called isomiRs are artifacts of library preparation.

Results

AQRNA-seq design and optimization.

The AQRNA-seq workflow (Fig. 1a) maximizes ligation capture of RNAs using novel adapters (linkers) and minimizes RT fall-off with two-step linker ligation and optional AlkB treatment. Adapter ligation begins at the 3'-end, with two randomized nucleotides at the 5' end of linker 1 to maximize the T4 ligase efficiency.[22] Linker 1 is DNA to facilitate removal of unligated linker with RecJ, a single-stranded-DNA-specific 5'→3' exonuclease, leaving the hybrid RNA-DNA product intact.[23] A 50:1 excess of linker 1 resulted in >90% ligation efficiency (Fig. 1b, Supplementary Fig. 1).

Ligated RNA can then be treated with AlkB to reduce levels of RT-blocking methyl modifications.[17] Though not essential for capturing all RNA sequences, it can provide information about the identities of polymerase-blocking modifications. The buffer provided with the commercial kit caused RNA degradation (Supplementary Fig. 1k,l), so we optimized buffer conditions and AlkB concentration to reduce m^1A (90%), m^1G (48%), and m^1I (96%). Contrary to previous observations,[18] only 12% of m^3C was removed and $m^{2,2}G$ was reduced by only 35%, apparently demethylated to AlkB-resistant m^2G (Fig. 1c).[17] After demethylation, optimized RecJ digestion removed >99% of unligated linker 1 (Fig. 1b), obviating HPLC purification of the RNA-DNA product.

RT is accomplished with a DNA primer complementary to linker 1 and the resulting cDNA is 3'-ligated to a custom DNA adapter (linker 2) using T4 DNA ligase. Linker 2 possesses a hairpin, a random hexamer sequence (splint to enhance cDNA ligation, Fig. 1d), and a downstream primer binding site for subsequent amplification, with ligation optimized to >97% at 50:1 linker excess (Fig. 1d, Supplementary Fig. 1). Excess linker 2 is removed with RecJ and PCR amplification is performed with primers complementing linker 2 and incorporating a standard Illumina anchor and barcode for subsequent sequencing.

AQRNA-seq data processing workflow.

We developed a custom workflow for optimal processing of AQRNA-seq data for bacterial tRNAs (Fig. 1e) and human miRNAs (Fig. 1f), but the generalized workflow can be adapted for any organism. The workflow (Online Methods) allows mapping of reads to highly repetitive targets or genes with similar sequences and makes it possible to map sequences with modifications and high level of mutations. As a result, the pipeline allows accurate quantification of all expressed small RNAs, as well as detection of RNA modifications, sequence alterations, and RNA structural changes that traditional RNAs-seq methods do not capture.

The AQRNA-seq pipeline (Fig.1f) begins with paired end sequence assembly that integrates read1 and read2 sequences, obtains high quality insert sequences by cross validating read pairs, and removes artificial linkers. The abundance of each unique insert sequence is counted in every sample and annotated with the corresponding RNA. Normalization of read counts, differential expression analyses, and analyses of sequence variations, chemical modifications, and structural changes occur after reads are mapped to RNAs.

To reduce sequencing and computational costs, we have also devised a workflow that is customized for prokaryotes, which makes it suitable for single-end sequencing (Fig.1e). The workflow begins with curation of reference sequences used for aligning reads. Reference sequences are first culled of duplicate genes and pseudogenes that lead to ambiguous assignments. Similar consideration must be given to post-transcriptional processing, such as trimming and processing of 5'- and 3'-termini of primary tRNA transcripts as well as tRNA modifications.[24] For mycobacteria, the 3'-CCA of each tRNA is variably genomically encoded or added post-transcriptionally.[25]

The resulting non-redundant reference library is then used to align forward and reverse sequencing reads, first separating uniquely aligned reads from reads matching multiple sequences. Even with careful library curation, high sequence similarity or insufficient read length can result in multiply-mapped reads. We arbitrarily set a 10 nt read length filter to maximize alignments. For closely related reference sequences, multiply-mapped reads are resolved by collapsing ambiguous read assignments into separate groups (Fig. 1e), with subsequent determination of the proportion of multiply-mapped reads (i.e., do they significantly alter the final read count for each RNA) and the cause of multiple mapping (e.g., highly similar sequences such as tRNA isodecoders). These considerations rationalize a decision to discard, average, or sum the read counts from multiple, closely related reference sequences.

Finally, the read count for full-length sequences and fragments is tabulated from the curated set of mapped reads. Data are normalized either to the total number of reads in each barcoded sample or to an internal RNA standard to account for sample-to-sample variation in input RNA as well as variable sample pooling prior to sequencing.

Validation of a linear relationship between read count and RNA abundance.

Four different approaches were used to test the precision and accuracy of AQRNA-seq. First, five RNA oligonucleotides with varying lengths (25–80 nt) were mixed at varying molar ratios and used as input RNA for library preparation (Supplementary Tables 2,3). After sequencing, we found that read counts for each oligonucleotide varied directly and linearly with input RNA abundance (r^2 0.92–1; Fig. 2a), with an average sequencing response (slope) of ~300 ± 50 reads per fmol of input RNA. This demonstrates minimal sequencing bias for quantity or length of input RNA.

To assess library preparation biases, we prepared libraries from the Miltenyi miRXplore Universal Reference consisting of 963 equimolar miRNA sequences from miRBase[26] (16–28 nt), which possessed all 16 possible dinucleotide combinations at 3' and 5' termini. The expected and measured frequencies of the terminal nucleotides in the Miltenyi miRNAs were nearly identical (Fig. 2b), which demonstrates minimal sequence bias in library preparation. We also used the miRXplore reference to assess the quantitative accuracy of AQRNA-seq. Here we calculated a read ratio by dividing "normalized read count" (miRNA reads divided by total counts for all detected miRNAs) by "expected read count" (total counts divided by 963, the number of detected miRNAs), assuming all species are equimolar. A plot of all 963 read ratios ranked from lowest to highest showed that ~75% fell within two-fold of expected abundance (Fig. 2c). The number of jackpot and dropout miRNAs (normalized read ratios >10-fold higher or lower than expected) was <3% of the total mixture. A direct comparison of AQRNA-seq to six commercial small RNA-seq kits (Fig. 2d),[27] as well as additional reports using the miRXplore reference,[5, 21, 28] established AQRNA-seq as the most quantitatively accurate RNA-seq workflow.

Analysis of sequencing reads for miRXplore reference set revealed an unexpected correlation between quantitative accuracy and sequence variants induced during library preparation. Here we defined seven classes of sequence variants representing additions and deletions at the ends of the miRNA inserts according to the nomenclature used in the IsomiR Bank.[29] These sequence variants are not present in the miRXplore set and could only arise during library preparation or sequencing. The proportions of the sequence variants for the six small RNA-seq methods (green background) and AQRNA-seq (yellow background) are depicted in Supplementary Figure 2a. Here it is apparent that all of the small RNA-seq methods produce all of the sequence variants to varying degrees. Similar analysis of the miRNA sequencing data from Kim et al.,[16] using their "AQ-seq" method to quantify miRNA isoforms (i.e., isomiRs) in cells, revealed a higher proportion of 3'-addition variants in human cells compared to analyses performed with the miRXplore reference (Supplementary Fig. 2b). Furthermore, AQ-seq produced all of the sequence variants noted with the other methods (Supplementary Fig. 2a, c). These observations raise concerns about the biological relevance of many isomiRs noted in the literature and databases.[29] Among small RNA-seq methods, a positive correlation was observed between the average number of sequence variants detected and the percentage of miRNAs quantified within 2-fold of expected (Fig. 2e, Supplementary 2c). The minor variation in the sequences of the inserts does not affect the alignment step during the datamining, with AQRNA-seq producing the most sequence variants and the most accurate quantification (Fig. 2d,e).

The fourth validation study tested AQRNA-seq performance against the analysis of the *E. coli* K12 tRNA pool on two-dimensional gels with northern blotting by Dong et al.[30] Applying AQRNA-seq to small RNAs from *E. coli* K12 strain BW25113 (derived from W1485 used by Dong et al.[30]), the total expressed levels of 45 tRNAs (summed counts for full-length and truncated reads) were compared to the 46 tRNAs identified by Dong et al.[30] Excluding one outlier, there was strong agreement ($r^2 = 0.81$) between the two approaches (Fig. 2f).

tRNA dynamics in bacterial persistence.

AQRNA-seq was applied to quantify the dynamics of a challenging set of targets: the 45 tRNAs in the *Mycobacterium bovis* BCG model for the stress-induced, non-replicative, antibiotic-resistant state of persistence in tuberculosis.[31–33] Total small RNA was isolated along the time course of BCG persistence caused by nutrient deprivation (Supplementary Fig. 3a,b), with ~1% of the bacteria surviving as persisters after 20 days in PBS and restoration of growth in nutrient-rich medium (Supplementary Fig. 3a). After size-selection and adapter trimming for the 5 million raw sequencing reads for each sample (Supplementary Fig. 3c), the majority (75%) of the remaining reads consisted of uniquely mapped, paired reads for the full set of mature BCG RNA sequences (Supplementary Fig. 3c). Of these, another 75% mapped to an 80-nt internal standard added in large excess in this experiment, while remaining reads mapped to reference library tRNA sequences (Supplementary Fig. 3c). A lower level of internal standard allows detection of rare RNA species, such as tRNA fragments (Fig. 3e). To account for variation introduced by input RNA and sample processing, reads originating from a single sample can be normalized to a spiked-in standard (80 nt here). For any RNA species of interest, comparison between samples is facilitated by expressing RNA read counts as either a percentage of total aligned reads or total aligned tRNA reads within each sample.

The resulting BCG dataset was mined for information about starvation-induced changes in tRNA expression and fragmentation, and the locations of modified nucleosides in individual tRNAs. These features are best visualized graphically in horizontally stacked alignment plots (Fig. 3a), in which the start and end positions of each read are aligned along an X-axis annotated using the Sprinzl tRNA coordinate system (Fig. 3b),[34] with the 3’ end defining the location of linker 1 (Fig. 3c). Alignment plots for the 45 tRNA species in BCG are shown in Supplementary Figure 4, with stack height directly proportional to the total number of expressed transcripts. The graph can be split into sections: ***bottom,*** “Type 3” or full length reads that span the entire tRNA sequence (Fig. 3b); ***top***, reads not reaching the 3’ end of the tRNA (“Type 1”; Fig. 3b) and corresponding to 5’ tRNA fragments, with the 3’ linker ligated to the 5’ end of the break (Fig. 3c); and ***middle***, reads that start at the tRNA 3’ end but do not reach the 5’ end (“Type 2”; Fig. 3b) and represent tRNA fragments missing a 3’ portion or full-length tRNAs for which cDNA synthesis was prematurely truncated by RT fall-off.

Analysis of 3 tRNA isoacceptors by northern blotting (Supplementary Fig. 5) suggests that the majority of the Type 2 reads are in fact full-length tRNAs. This is consistent with the relatively low level of 5’ tRNA fragments in stack plots for all expressed tRNAs in BCG (Supplementary Fig. 4).

Starvation remodels the tRNA landscape in BCG.

AQRNA-seq provides a global view of changes in the RNA landscape. In starved BCG, the abundance of individual tRNAs – defined as the sum of all reads that aligned to a particular tRNA – spanned a large range. In most samples, tRNA Lys-CTT and tRNA fMet-CAT were the most highly expressed tRNAs, together totaling ~20% of the pool during log growth, whereas tRNA Ser-GGA was ~80-times lower at 0.1–0.3% of the pool (Fig. 4a, Supplementary Fig. 4). In the transition from rich medium to starvation over 20 days, the abundances of several tRNAs are significantly altered (Fig. 4a,b; Supplementary Fig. 4). For example, tRNA Ser-CGA and tRNA His-GTG drop significantly in early starvation and rise again during late starvation and resuscitation. tRNA Leu-CAG and tRNA Thr-GGT show the opposite pattern.

Starvation also induced significant shifts among isoacceptor families (Fig. 4c). For example, Ser, Thr, and Leu are specified by 6, 4, and 6 synonymous codons, respectively, and these codons are read by 4, 3, and 5 different tRNA species, respectively.[35] As shown in Figure 4c, prior to starvation (S0), tRNAs Ser-CGA, Ser-TGA and Ser-GCT comprise 36%, 34%, and 25% of the Ser isoacceptor pool, respectively. At 4 days of starvation (S4), the abundance of Ser-CGA drops to 7% while Ser-TGA and -GCT surge to 47% and 44%, respectively. For Thr, tRNAs Thr-CGT and Thr-GGT represent 51% and 31% during log growth (S0), but flip to 40% and 48% at S20, respectively, before returning to S0 levels during resuscitation. These data illustrate the dynamics of individual tRNAs resulting from stress-induced changes in tRNA gene expression or degradation. However, we are left with the question of how changes in the tRNA pool relate to starvation-induced changes in cell phenotype.

Here we tested the link between starvation-induced tRNA pool changes and shifts in the BCG proteome. We previously discovered that BCG respond to persistence-inducing hypoxia by uniquely altering tRNA modifications to cause selective translation of mRNAs coding for hypoxia response genes – Dos regulon – that possess codon usage patterns matching the hypoxia-altered tRNAs.[36] To test this mechanism in starvation-induced persistence, we performed quantitative proteomics across the starvation time course, detecting 1102 proteins common to three separate cultures at all time points (Supplementary Fig. 6). Analysis of codon usage frequencies[37] in the genes for the 25 most upregulated proteins in late starvation revealed enrichment with the ACC codon read by tRNA Thr-GGT that increased during starvation (Fig. 4b). These same genes underutilized the AGC codon read by tRNA Thr-CGT that was reduced during starvation (Fig. 4c).

AQRNA-seq also captures the dynamics of tRNA fragmentation and degradation, as occurs in tRNA maturation and quality control,[38–40] small RNA regulation of gene expression, and toxin-antitoxin systems.[41–46] It is difficult to differentiate 5’-degradation of full-length tRNA from polymerase fall-off as both generate a fragment that aligns at the 3’ end of tRNA (Type 2, Fig. 3b). However, 3’-degradation and endonucleolytic cleavage generate fragments with 3’ ends positioned inside the reference sequence (Type 1, Fig. 3b). An extreme case is illustrated by tRNA Glu-TTC. While ~80% of reads aligned with the 3’ end in log growth, ~80% of reads at S4 had 3’ ends at position 58 in the TΨC arm (Fig. 3b). These tRNA fragments from tRNA Glu-TTC appeared as a more intense gel band in the S4 sample than the log sample (Supplementary Fig. 7b). The absence of a corresponding number of short (15–20 nt) tRNA fragments representing the 5’-side of the cleaved tRNA (Type 2, Fig. 3e) could result from degradation of the 3’-fragments. AQRNA-seq thus provides quantitative information about tRNA fragmentation in addition to tRNA expression.

Quantitative mapping of tRNA modifications and structures.

Nearly all forms of RNA contain post-transcriptional modifications, with >150 structures known.[47] tRNAs are particularly heavily modified at ~10% of the component nucleotides.[10] In some cases, modifications interfere with RT during RNA-seq library preparation, which allows mapping of modification positions.[17, 48, 49] AQRNA-seq detects RT defects as mutations or read pile ups at sites along the RNA sequence, as illustrated with small

RNA isolated from log-growing *E. coli*.[47] As shown in Figure 5a, several tRNAs had substantial polymerase stops at positions 38 and 48. By overlaying the stop positions on tRNA modification maps,[47] these two positions are found proximal to known modification sites. One subset of 9 tRNAs had 31–83% of mapped reads stop at position 48, which abuts the modification acp^{3}U at position 47 (orange boxes, right, Fig. 5a). With base-pair-blocking m^{3}U previously reported to block RT,[50] it makes sense that acp^{3}U would also prevent polymerase procession. tRNAs with NNA anticodons had 48–69% of reads end at position 38, which is adjacent to i^{6}A and its hypermodified derivatives (e.g., ms^{2}i^{6}A) in the anticodon loop (yellow boxes, left, Fig. 5a). While m^{6}A induces pausing,[50] it is plausible that the bulkier i^{6}A combined with the sharp turn of the anticodon hairpin interferes with RT.

These data corroborate the RT-blocking potential for many modified nucleotides, with read interruption 1–2 nucleotides away from the modification. Further validation comes from a BCG library preparation lacking AlkB treatment. In the absence of AlkB, up to 90% of reads mapping to 23 of 45 of tRNA species were truncated at positions 59–60 (heatmap of RT stops in Fig. 5b) After AlkB treatment, most stops at positions 59–60 have disappeared and reads increased in length with a leftward shift toward the tRNA 5' end (Fig. 5c). The alignment profile of tRNA-Glu-CTC illustrates the AlkB effect: the sharp "cliff" at position 60 in the untreated sample (Fig. 5b) corresponds to a predicted RT-blocking AlkB substrate,[17, 49, 50] m^{1}A, at position 58 (Fig. 5b). After AlkB treatment, the cliff disappears and the reads span the tRNA sequence (Fig. 5c). The presence of 5' cliffs in the alignment plots for nearly all tRNA species in BCG (Supplementary Fig. 4) points to the potential for quantitative mapping of RNA modifications by AQRNA-seq.

Modification mapping is also illustrated by mutations resulting from RT.[50] For example, wobble inosine (I) in a single tRNA (Arg-ACG) in mycobacteria[47] tends to pair with C during RT,[48, 51] which suggests that the near stoichiometric T-to-C sequencing mutation (Fig. 5d) for position 34 of tRNA Arg-ACG represents I. As reviewed by Motorin and Helm,[12] this kind of modification mapping could aid in the discovery of previously unannotated or unlocalized modifications in poorly characterized species.

Quantitative profiling of miRNA dynamics during tumorigenesis.

To demonstrate the utility of AQRNA-seq for human cells, we used it to profile miRNAs in the human mammary epithelial cell (HMEC) model of breast cancer tumorigenesis.[52, 53] The three HMEC cell lines represent progressive tumorigenesis conferred by engineered tumor-promoting genotypes (Fig. 6a): reactivation of telomerase by expression of a catalytic subunit (hTERT) immortalizes HMEC 1 cells; tamoxifen-inducible expression of H-Ras oncoprotein (HRAS^{G12V}-ER) and expression of SV40 small-t antigen further drive partial transformation and aberrant growth in HMEC 2 cells; and additional P53 suppression by shRNA knockdown yields HMEC 3 cells fully capable of tumor growth in mice.[52, 53] RT qPCR-based validation of key gene expression changes in the HMEC cells is shown in Supplementary Figure 8.

Prior to applying AQRNA-seq to quantify HMEC miRNAs, we modified the datamining workflow for the complexity of the human genome with numerous repeats and highly similar RNA species, such as tRNA isodecoders, which poses a challenge to uniquely mapping reads. We modified the datamining pipeline (Fig. 1f) to directly quantify the pair-end assembled inserts based on their sequences, with mapping to reference RNA sequences or the genome serving to annotate the inserts.

When AQRNA-seq was applied to HMEC cells, we observed 875 non-redundant miRNA sequences for all three cell lines, ranging from 1 to 100,000 normalized read counts (Fig. 6b). The miRNAs changing in abundance most significantly during the tumorigenic transition from HEMC1 to HMEC3 were identified by partial least squares regression (PLSR) analysis (Fig. 6c). Here we selected 14 miRNAs that distinguished the three cell lines, with the log plot in Figure 6d showing that three (15a-5p, 19a-3p, 4454) significantly increased in the transition from HMEC1, four significantly decreased (24–3p, 4488, 21–5p, 27a-3p), and seven were unchanged during the tumorigenesis. These results are consistent in some cases with literature observations. For example, 15a-5p and 19a-3p were up-regulated (>1.8-fold) in patients with triple negative breast cancer[54] and 4454 up-regulated in more aggressive breast cancer types with HER-2 overexpression[55, 56] as well as in inflammatory breast cancer.[57] Similarly, miR-27a was down-regulated in breast cancer stem cells, with overexpression reducing both number and size of mammospheres and sensitizing breast cancer cells to anticancer drugs by downregulation of genes essential for ROS detoxification.[58] However, as discussed in Supplementary Information, contradictory behaviors have been observed for these miRNAs in breast and other cancers, which reveals our relatively poor understanding of the role of miRNAs in cell biology and disease.

These observations of canonical miRNAs raise questions about the behavior of isomiRs. As noted earlier, all RNA-seq methods introduce adventitious sequence variants (Supplementary Fig. 2a, b) during library preparation and sequencing. However, since adventitious sequence variants should be identically produced in both HMEC cells and the miRXplore library, we tested the idea of identifying biological variants by comparative analysis of variants associated with the 875 non-redundant miRNA sequences present in both the HMEC and miRXplore samples. The pool of 875 parent miRNAs were filtered to include those with ≥2 variants that exceeded 10 reads per variant to ensure variants were not the product of sequencing error. Analysis of end sequences among the subset of 24 miRNAs meeting these conditions revealed a predominance of U additions to 3’- and 5’-ends (Supplementary Fig. 8b). To discover variant sequences likely not caused by library preparation (i.e., biologically relevant isomiRs), we subtracted the number of copies of each of addition variant in the miRXplore panel from those in the HMEC cells for the 24 miRNAs, selecting those for which HMEC-miRXplore was > 0.1. Graphs depicting the HMEC-miRXplore differential among the 15 addition variants are shown in Supplementary Figure 8c, which reveals a predominance of single 3’-additions expected for true isomiRs.[16, 59]

Discussion

Here we developed, validated, and applied an RNA-seq method that provides precise and accurate absolute quantification – read count directly correlates with molecular copy number – of all small RNA species in a sample. Numerous factors challenge the quantitative accuracy and fidelity of NGS RNA-seq methods, including biochemical idiosyncrasies of RNA ligases, RTs, and other enzymes, and the secondary structures of adaptors and substrates.[16, 60] Following a systematic deconstruction of the RNA-seq NGS library prep workflow, we identified several steps critical to the quantitative precision and accuracy of RNA-seq results: (1) ligation of the 5’ linker after RT, (2) linker structures and biochemical conditions providing >90% ligation efficiency, (3) a non-essential but informative AlkB demethylation step, and (4) a data mining workflow minimizing loss of read information to improve quantitative accuracy. Along with randomized ends on adaptors to minimize ligation bias and molecular crowding agents (e.g., PEG) to increase enzyme efficiency, ligation of the 5’-adapter occurs after RT to ensure that all cDNAs, including truncated species, are captured in the final library.[16, 60] A non-essential AlkB demethylation step minimizes premature cDNA truncation and informs about the locations of RT-blocking or mutating modifications.[17] Beyond the linear relationship between read count and RNA copy number, the RNA-seq method provides information about modification occupancy, secondary structure, and fragmentation.

The optimizations made in AQRNA-seq have been variably used in other RNA-seq methods and subsequently in commercial kits, most notably for miRNA analysis. For example, Kim et al. developed AQ-seq to study isomiRs, with randomized adaptor ends and 20% PEG to enhance fidelity.[16] However, the need for miRNA size-selection and ligation of both 3’- and 5’-adaptors prior to RT limits AQ-seq to miRNA and ignores sequences lost during polymerase fall-off. RNA-seq methods for miRNA quantification have been compared in several publications. For example, Dard-Drescott et al. compared the ILM, NEB, and PEB miRNA kits studied here and two additional non-ligation kits (Clontech SMARTer, Diagenode CATS)[61] using six miRNA standards, finding that kits using PEG, random-end adaptors, and overnight ligations minimized both sequence biases and interferences caused by 3’-(2’-O-methylation) on some miRNAs.[61] Wong et al. assessed methods for extracting miRNA from plasma, but only compared kit performance using detected diversity as the metric.[62] Finally, Heinecke et al. compared TRI, QIA, and LEX kits tested here and the SMARTer and CATS kits tested by Dard-Drescott et al. using 41 miRNA standards and small RNA purified from human T cells.[63] None of these comparisons assessed quantitative accuracy with the rigor applied by Herbert et al.[27] and in the present studies. While the best-performing commercial kits are able to quantify only <50% of the miRNAs with two-fold accuracy,[27] we were able to quantify ~75% of miRNAs within two-fold of expected abundance, with few dropouts and no jackpots. Together with the lack of significant length bias and evidence for a direct correlation between read count and copy number, these studies demonstrate that AQRNA-seq faithfully captures the quantitative landscape of all small RNAs in a sample and also informs about many RNA modifications critical to RNA function (Figs. 3, 5).

So how does AQRNA-seq compare to other methods? As detailed in Supplementary Table 1, there are >40 RNA-seq methods for different types of RNA and different purposes. However, few of these methods (1) optimize ligation and amplification efficiencies, (2) have been validated for quantitative accuracy and lack of bias artifacts, and (3) are broadly applicable to all small RNAs. For example, ARM-seq is a ligation-based RT method that adds an AlkB demethylation step to reduce the impact of methyl modifications. However, in addition to the incomplete removal of AlkB-sensitive modifications (Fig. 1b), the method was not optimized for ligation efficiency or evaluated for quantitative bias. The simultaneous ligation of 3’- and 5’-linkers before RT results in significant loss of sequence information due to polymerase fall-off. Another example involves the template-switching polymerase methods that use Thermostable group II intron RTs (TGIRT) for cDNA production and CircLigase for subsequent cDNA circularization during library preparation.[18, 20] In addition to not being evaluated for RNA capture efficiency by the RNA/DNA duplex or cDNA circularization efficiency, TGIRT is biased by the identity of the overhanging nucleotide in the adapter strand[21] while the efficiency of CircLigase I and II is strongly influenced by the sequence of the cDNA and RNA.[64, 65] These ligation biases preclude an unbiased, quantitative analysis of all RNAs in a sample. A final comparison involves the Hydro-tRNAseq method[15] entailing hydrolytic fragmentation of tRNA followed by traditional simultaneous 3’- and 5’-linker ligation, RT, PCR, and NGS steps. Here, ligation efficiency was not optimized, the RT step loses all sequence information for polymerase fall-off fragments, and the hydrolysis of the tRNAs prevents identification of natural tRNA fragments. Importantly, Hydro-tRNAseq was not designed to be quantitative and instead was intended for tRNA gene annotation for mature and pre-tRNA sequences.[15]

AQRNA-seq shares a downside with other RNA-seq technologies: the introduction of sequence variants during library preparation and sequencing (Supplementary Fig. 2a). This raises concerns about the biological relevance of many miRNA isoforms (i.e., isomiRs) noted in the literature[16] and online databases.[29] Given extensive evidence for the formation of isomiRs by enzymatic uridylation,[16, 59] care must be taken to process RNA-seq data for sequence variants, perhaps by incorporating a set of synthetic RNA standards into each sequencing run along with biological samples, to quantitatively subtract artifacts and enrich for biologically relevant isomiRs. It is not clear why the introduction of sequence variants correlates strongly with the accuracy for quantifying RNA sequences (Figs. 2a, c, d, f), with AQRNA-seq as the most quantitatively accurate method at least for miRNAs.

While AQRNA-seq was applied first to tRNA and miRNA analysis, it should be broadly applicable to any form of RNA. Random priming RNA-seq methods provide relatively accurate quantification of mRNAs and long RNAs,[66] but they do not inform about RNA fragmentation or modifications. AQRNA-seq is applicable to longer RNA (e.g., mRNA, rRNA) as a means to map RNA modifications and cleavage sites (Supplementary Fig. 9). A fragmentation step after ligation of linker 1 reduces longer RNA species to an appropriate length for library generation for quantification of (1) all expressed copies of an RNA (3’-end that maps to the end of the transcribed or mature sequence), (2) polymerase fall-off, and (3) fragmentation sites within RNA molecules (3’-ends mapping within the full-length sequence). Collectively, our results demonstrate that AQRNA-seq is a quantitatively accurate method for sequencing RNAs of all types.

Online Methods

Bacterial strains, culturing conditions, growth assays, and RNA isolation.

All *E. coli* strains were from the Keio collection (GenoBase; http://ecoli.aist-nara.ac.jp/)[67]. The genotype of each strain was validated before conducting studies. Strains were cultured in 10 ml LB broth (Fisher BioReagents) at 37 °C with constant shaking at 180 rpm until the cultures reached a final OD_{600} of 0.6–0.7. Culture pellets were harvested by centrifugation at 4,000×g for 2 min and immediately used for tRNA isolation using the Purelink miRNA isolation kit (Thermofisher) following the manufacturer's protocol. Briefly, cell pellets were resuspended in Trizol Reagent (Thermofisher) for lysis, followed by treatment with chloroform to separate the aqueous layer containing bulk tRNA. The aqueous layer was then subjected to a 2-column purification process where genomic DNA, larger RNA fragments (>200 bp), and excess salts, are removed. It is important to note that 5S rRNA cannot be separated from tRNA using this method. Three biological replicates were used in *E. coli* tRNA study.

Mycobacterium bovis Bacille Calmette-Guérin (BCG) str. Pasteur 1173P2 was grown in roller bottles with 7H9 broth or PBS (with 0.05% v/v tyloxapol, Sigma Aldrich) at 2 rpm and 37 °C. Exponentially growing cultures with an OD_{600} of 0.8–1.0 were starved by washing pellets three times with PBS-tyloxapol. Starvation cultures were inoculated into PBS-tyloxapol at a starting OD_{600} of 1.0. Samples were retrieved at 4, 10, and 20 d after starvation. At day 20, cultures were resuspended in 7H9 and resuscitated for 6 d prior to harvesting. At each time point, cultures were plated on 7H10 agar for CFU determination. Specific compositions of 7H9 media, PBS, and 7H10 agar are as follows. Middlebrook 7H9 (BD Difco) was supplemented with 0.5% (w/v) albumin (Sigma Aldrich), 0.2% (w/v) glucose (Sigma Aldrich), 0.085% (w/v) NaCl (Sigma Aldrich), 0.2% v/v glycerol and 0.05% v/v Tween 80 (Sigma Aldrich) as nutrient replete media; PBS (137 mM NaCl, 2.7 mM KCl, 10 mM Na_2HPO_4, 2 mM KH_2PO_4) was supplemented with 0.05% v/v tyloxapol, a non-hydrolysable detergent. 7H10 agar (BD Difco) was supplemented with 0.5% v/v glycerol and 10% v/v oleic acid albumin dextrose catalase (OADC; BD BBL).

For colony forming unit (CFU) assays, serial dilutions of BCG cultures at various nutrient starvation/resuscitation timepoints were plated on 7H10 agar and incubated at 37 °C for 3–4 weeks. BCG colonies were subsequently counted and respective CFUs determined.

For RNA isolation and purification from BCG, cells collected at nutrient starvation/ resuscitation timepoints were lysed in the presence of TRI reagent (Sigma Aldrich) with glass beads in a FastPrep FP120 bead-beater as previously described[68]. Small RNAs were subsequently isolated using the Purelink miRNA isolation kit (Invitrogen). Small and large RNA fractions were profiled using the Agilent Bioanalyzer.

BCG tRNA fragmentation analysis.

Total RNA was isolated from BCG in log growth in rich medium (log) and on day 4 of starvation in PBS (S4). This RNA (2.9 μg) was mixed with the 50-mer oligo standard (200 pmol; Supplementary Table 2) and an *in vitro*-transcribed human RNA standard (tRNA-Val-AAC; 50 ng). These mixtures were loaded onto 10% NovexTBE urea gels and the resolved gels stained with Sybr Gold dye. The fluorescent image was used to excise bands (red boxes in Supplementary Fig. 7B) for subsequent RNA extraction, with the purified RNA split into two technical replicates for AQRNA-seq analysis.

HMEC derived cell lines, culturing conditions, RNA isolation, and cell line validation.

Immortalized human epithelial cell derived lines (**HMEC 1**, HMEC^{hTert}; **HMEC 2**, HMEC$^{hTert-HRas(V12):ER-EV}$; and **HMEC 3**, HMEC$^{hTert-HRas(V12):ER-shp53}$) were a gift from Jit Kong CHEONG, National University of Singapore Department of Biochemistry. The HMEC-derived lines were maintained in Mammary Epithelial Cell Growth Medium (MEGM) BulletKit (Lonza CC-3150), with media changes every 2 days and passaged before cells reach confluency. For each experiment, HMEC cells were seeded at 80,000 per well in 6-well culture dishes and allowed to attach overnight. The cell lines were subsequently treated with 0.13 mM 4-hydroxytamoxifen (4-OHT; Sigma T5648) for 4 days to activate oncogenic HRAS activity. The experimental culture medium was replaced with fresh MEGM medium supplemented with 0.13mM 4-OHT every 2 days. After the course of 4-OHT, cells were washed in ice-cold PBS once before lysis using TriZol for downstream RNA extraction and small/large RNA separation according to protocols described above. To validate the expression of oncogenic *HRAS* and *P53* in the HMEC-derived lines, total cellular RNA was extracted using TriZol and reverse transcribed with the SuperScript IV Reverse Transcriptase (ThermoFisher 18090010) according to manufacturer’s protocol, followed by quantitative PCR conducted using the Roche LightCycler 480 system and SYBR Green kit (Roche 04707516001). The qRT–PCR primer sequences are as follows: HRAS Fwd: 5’-GACGTGCCTGTTGGACATC, HRAS Rev: 5’-CTTCACCCGTTTGATCTGCTC, P53 Fwd: 5’-GAGGTTGGCTCTGACTGTACC, P53 Rev: 5’-TCCGTCCCAGTAGATTACCAC, GAPDH Fwd: 5’-GGAGCGAGATCCCTCCAAAAT, GAPDH Rev: 5’-GGCTGTTGTCATACTTCTCATGG. Data were initially normalized to the GAPDH signal in each sample and then fold-change data relative to HMEC1 were calculated for HMEC2 and HMEC3.

Optimized AQRNA-seq protocol.

Ligation of input RNA and DNA Linker 1.—Small RNAs (50 ng) were mixed with an 80-mer spike-in RNA oligonucleotide internal standard (DNA and RNA oligonucleotide sequences are detailed in Supplementary Table 2). The mixture was dephosphorylated in a 5 μl reaction containing 0.5 μl of reaction buffer (NEB T4 RNA ligase buffer) and 1 U of shrimp alkaline phosphatase (rSAP, NEB) at 37 °C for 30 min. The reaction was stopped by heat inactivation at 65 °C for 5 min followed by cooling on ice. Linker 1 ligation was performed by adding the following reagents directly to the dephosphorylation product: 1 μl Linker 1 (100 pmol/ul; Supplementary Table 2), 3 μl ATP (10 mM, NEB), 2.5 μl T4 RNA ligase buffer (NEB), 2 μl T4 RNA ligase 1 (30U/ul, NEB), 1.5 μl water and 15 μl PEG8000 (NEB). The ligation mixture was incubated at 25 °C for 2 h and 16 °C overnight. The ligation product was purified using the Zymo Oligo Clean & Concentrator kit (Zymo Research, D4060) according to the manufacturer’s instructions. The sample was eluted in 20 μl water and kept on ice prior to Bioanalyzer analysis (Agilent, small RNA kit) or used directly in the demethylation step.

Demethylation.—The ligated RNA sample was demethylated by AlkB (Arraystar, rtStar tRF&tiRNA Pretreatment Kit). A 2X reaction buffer was freshly prepared before the reaction and consisted of 150 μM 2-ketoglutarate, 4 mM L-ascorbic acid, 150 uM

$(NH_4)_2Fe(SO_4)_2$, 100 µg/mL BSA (NEB, molecular biology grade, 10 mg/mL), and 100 mM HEPES (pH 8.0). The demethylation reaction was performed in a 100 µl volume consisting of 50 µl 2X reaction buffer, 20 µl Linker 1-ligated tRNA sample, 2 µl AlkB demethylase, and 1 µl RNase Inhibitor (NEB, murine, 40,000U/mL). The reaction was incubated at ambient temperature. After 2 h, the reaction was stopped via the addition of 50 µl water and 100 µl phenol:chloroform:isoamyl alcohol 25:24:1, pH 5.2. The mixture was mixed by inverting several times and centrifuged at 16,000×g for 10 min. The top layer was transferred to a new Eppendorf tube. Another 100 µl chloroform was added to the original mixture to remove any remaining phenol. After centrifugation, the top layer was removed and combined with the first extraction. The extracted sample was then purified using the Zymo Oligo Clean & Concentrator kit (Zymo Research, D4060) per the manufacturer's instructions. The sample was eluted in 17 µl water before proceeding to the next step (Linker 1 removal).

Removal of excess DNA Linker 1.—In this step, the DNA adenylated oligonucleotide adenylate intermediate was de-adenylated and subsequently digested, together with unused Linker 1, by exonuclease RecJ. The de-adenylation was performed in a 20 µl reaction containing 16 µl RNA sample from demethylation step, 2 µl NEB Buffer 2 (10X), and 2 µl 5'-deadenylase (NEB, 50U/uL). After incubation at 30 °C for 1 h, 2 µl RecJ (NEB; 30 U/uL) was added. The mixture was incubated at 37 °C for 30 min followed by the addition of another 2 µl RecJ and further digestion for an additional 30 min. The reaction was stopped by heating at 65 °C for 20 min. The reaction mixture was purified using a DyEx spin column (Qiagen).

Reverse transcription.—The RNA sample from the DyEx column purification is mixed with 1 µl RT-primer (2 pmol/µl) and 1 µl dNTPs (10 mM each) and heated at 80 °C for 2 min, followed by cooling on ice. PrimeScript Buffer (6 µl; Clontech), 1 µl RNase Inhibitor (NEB), and 1 µl PrimeScript Reverse Transcriptase (Clontech) were added. The mixture was then incubated at 50 °C for 2 h after which the enzyme was inactivated at 70 °C for 15 min. The RNA strand was hydrolyzed by adding 1 µl NaOH (5 M) followed by incubation at 90 °C for 3 min. The hydrolysis product was neutralized by adding 1 µl HCl (5M) and the reaction was cleaned up using the Zymo Oligo Clean & Concentrator kit (Zymo Research). The sample was eluted with 15 µl of water before vacuum concentration to 5 ul.

cDNA ligation.—The purified cDNA was ligated to Linker 2 (Supplementary Table 2) in a 20 µl reaction consisting of 5 µl cDNA sample, 1 µl Linker 2 (50 pmol/µl), 2 µl T4 DNA Ligase Buffer (NEB), 1 µl ATP (10 mM, NEB), 2 µl T4 DNA ligase (400U/uL, NEB), and 9 µl PEG8000 (NEB). The mixture was mixed and incubated at 16 °C overnight. Ligated product was purified using the Zymo Oligo Clean & Concentrator kit and eluted in 16 µl of water.

Removal of excess Linker 2.—After cDNA ligation, excess Linker 2 was removed in two steps: adenylated linker intermediates were de-adenylated and RecJ was used to digest the de-adenylated product. De-adenylation was performed in a 20 µl reaction containing 16 µl RNA sample from the cDNA ligation step, 2 µl NEB Buffer 2 (10X), and 2 µl

5’-deadenylase (NEB, 50U/ul). After incubation at 30 °C for 1 h, 2 µl RecJ (30 U/ul) was added. The reaction was incubated at 37 °C for 30 min. Subsequently, another 2 µl RecJ was added for further digestion for an additional 30 min. The reaction was stopped through heat inactivation at 65 °C for 20 min.

PCR amplification and Illumina sequencing.—Purified cDNA from the previous step was amplified by PCR in a 100 µl mixture containing 24 µl cDNA template, 50 µl seqAMP DNA polymerase buffer (2X buffer, Clontech), 2 µl each of PCR primer F and R with unique sequencing barcodes (Supplementary Table 2, 1 µM each), 2 µl seqAMP DNA polymerase (Clontech) and 20 µl water. The PCR reaction was performed according to the manufacturer’s instructions with an annealing temperature of 58 °C and 13 reaction cycles. The PCR product was extracted and purified from an agarose gel using a standard gel purification kit (QIAquick Gel Extraction Kit, Qiagen). The gel-extracted samples were mixed together (multiplexing) and submitted for Illumina sequencing. In the studies described, sequencing was performed on the Illumina NEXTseq sequencer (BioMicroCenter, MIT) with custom primers F and R (Supplementary Table 2).

Optimization of AlkB demethylation conditions.

LC-MS/MS analyses.—Ribonucleosides were resolved with a Phenomenex Synergi Fusion reversed-phase column (100 × 2 mm, 2.5 µm particle size, 100 Å pore size) eluted with the following gradient of acetonitrile in 5 mM ammonium acetate (pH 5.3) at a flow rate of 0.35 ml/min and 35 °C: 0–1 min, 0%; 1–10 min, 0–10%, 10–14 min, 10–40%, 14–15 min, 40–80%. The HPLC column was coupled to an Agilent 6430 triple quadrupole mass spectrometer with an electrospray ionization source operated in positive ion mode with the following parameters: gas temperature, 350 °C; gas flow, 10 l/min; nebulizer, 45 psi; and capillary voltage, 3500 V. The first and third quadrupoles (Q1 and Q3) were fixed to unit resolution and the modifications were quantified by pre-determined molecular transitions. The dwell time for each ribonucleoside was 500 ms. The retention time, *m/z* of the transmitted parent ion, *m/z* of the monitored product ion, fragmentor voltage, and collision energy of each modified nucleoside are as follows: m^1A, 3.6 min, *m/z* 282→150, 100 V, 16 V; m^1G, 6.1 min, *m/z* 298→166, 90 V, 10 V; m^1I, 5.9 min, *m/z* 283→151, 80 V, 10 V; $m^{22}G$, 7.8 min, *m/z* 312→180, 100 V, 8 V. Modified ribonucleosides were identified using commercial standards. Three independent tRNA replicates were used to test AlkB demethylation efficiencies in this study.

LC-MS/MS data analysis.—Quantitative comparisons between control and demethylase-treated samples from different buffers were made possible by correcting for variation in tRNA quantities by dividing raw peak are for each ribonucleoside by the ultraviolet absorbance peak areas for the four canonical ribonucleosides. The demethylation efficiencies were calculated by dividing the peak area of the demethylase-treated sample by the peak area of the control sample for each modification.

Optimization of linker ligation conditions and determination of linker removal.

Linker 1 ligation studies.—Tests of Linker 1 ligation efficiencies were performed by combining dephosphorylated tRNA with ATP, T4 RNA ligase buffer, T4 RNA ligase

1, PEG8000, and varying amounts of Linker 1. The reaction was allowed to proceed under conditions consistent with the manufacturer's recommendations and the resulting products were analyzed using the Bioanalyzer small RNA chip. Electropherogram peaks corresponding to ligated and unligated tRNA were fitted and integrated using Peakfit.m (version 9.0; Tom O'Haver, MATLAB Central File Exchange - https://terpconnect.umd.edu/~toh/spectrum/), a Matlab-based peak fitting program that uses an unconstrained non-linear optimization algorithm to decompose a complex peak signal into its fundamental underlying component parts. The fraction of ligated tRNA was calculated by dividing the summed peak area corresponding to the ligated tRNA to the total peak area. Three independent replicates were used for linker 1 ligation efficiency testing.

Linker 2 ligation studies.—The efficiency of the Linker 2 ligation reaction was quantified using single-stranded DNA oligonucleotides with phosphorothioate modifications at the 5' end to simulate the reverse transcription output. Oligonucleotide sequences are listed in Supplementary Table 2. To determine the optimal linker-to-oligonucleotide ratio, we combined the 80-nucleotide oligo with ATP, T4 DNA ligase buffer, T4 DNA ligase, PEG8000, and varying amounts of Linker 2. To determine whether oligonucleotide length altered ligation efficiencies, we used a constant linker:oligonucleotide ratio and varied the length of the oligonucleotide tested. The reactions were allowed to proceed under conditions consistent with the manufacturer's recommendations and the resulting products were analyzed using the Bioanalyzer small RNA chip. Similar to Linker 1 ligation studies, peaks were deconvoluted and integrated using Peakfit.m. The fraction of ligated oligonucleotide was used as a metric for ligation efficiency and used to determine final, optimized ligation conditions. Three independent replicates were used for cDNA ligation efficiency testing.

Linker removal studies.—Linker removal steps were performed on samples from the Linker 1 and 2 ligation studies. The extent of linker removal by deadenylase and RecJ was determined from Bioanalyzer electropherograms using peak fitting and integration. Three independent replicates were used to test linker 1 and 2 removal efficiencies in this study.

Library construction for control AQRNA-seq samples.

Standard RNA oligonucleotides.—Five synthetic RNA oligonucleotides of different lengths and sequences were used to study the effect of oligonucleotide concentration and length on sequencing response. The oligonucleotides, whose sequences are listed in Supplementary Table 2, were diluted from stocks to varying final concentrations and mixed together into standard mix samples A-E according to the scheme presented in Supplementary Table 3. Samples A through E were prepared in triplicate and used as input RNA samples for AQRNA-sequencing.

MicroRNA mixture.—The miRXplore Universal Reference (Miltenyi Biotec) consists of 963 synthetic unmodified, HPLC-purified RNA oligonucleotides. The sample was reconstituted according to manufacturer's instructions and aliquoted. Three aliquots were used as separate RNA inputs for AQRNA-seq library construction.

Note: The Miltenyi miRXplore Universal Reference is apparently no longer commercially available. Interested readers can use other miRNA collections or synthesize their own equimolar panel of miRNAs as performed by Kim et al.[16] Alternatively, readers can use miRNA panels validated in tissue and cell extracts using PCR-based kits (e.g., https://www.qiagen.com/us/products/discovery-and-translational-research/pcr-qpcr-dpcr/qpcr-assays-and-instruments/mirna-qpcr-assay-and-panels/mircury-lna-mirna-mirnome-pcr-panels/?clear=true#orderinginformation).

AQRNA-seq data processing for bacterial tRNA.

Illumina sequencing reads were first assessed using a custom quality control pipeline to ensure sequencing quality. The data were then processed using a workflow that can be customized for any type of transcripts in all eukaryotic and prokaryotic species. For bacterial tRNAs (Fig. 1e), adapter sequences were removed from forward and reverse reads were trimmed using fastxtoolkit (version 0.013). A minimum adapter alignment length of 10 bp was required, and unknown (N) nucleotides were kept. Sequences were blasted against a reference library using blast (version 2.6.0) with the parameters: blast -perc identity 90 -word_size 9 -dust no -soft_masking false. For sequencing libraries prepared from BCG tRNA, a reference library was created based on the 48 entries in the genomic tRNA database for *Myocbacterium bovis* BCG str. Pasteur_1173P2[35] (http://gtrnadb.ucsc.edu/GtRNAdb2/genomes/bacteria/Myco_bovi_BCG_Pasteur_1173P2_BCG_Pasteur_1173P2/Myco_bovi_BCG_Pasteur_1173P2_BCG_Pasteur_1173P2-gene-list.html). Sequences corresponding to duplicate tRNA genes (e.g., tRNA-Ala-TGC-1–2 and tRNA-Ile-GAT-1–2) and tRNA pseudogenes (tRNA-Ser-CGA-2–1) were removed to eliminate redundant entries and reduce the incidence of ambiguous or false positive matches. The terminal (3') CCA sequence was added to tRNA sequences where it is not genomically encoded. The sequence for the 80-nt RNA internal standard was added to the reference library. For the control samples, the sequences of the 5 synthetic RNA oligonucleotides were used to create a reference library, along with the 80-nt RNA internal standard.

For each tRNA and control sample, forward and reverse reads were merged by integrating their start and end positions to generate new start and end positions that reflect their combined coverage. Multiple alignments were reduced by ranking all the alignments for a given read by their e-value and retaining only the alignment with the lowest e-value. Forward and reverse reads were required to match the same target. Paired reads that did not match the same target were stored in a separate file and not analyzed. These manipulations were carried out with the python script cull.py. Uniquely mapped reads were then tabulated and counted. For the microRNA samples, the set of 963 microRNA sequences contained within the miRXplore Universal Reference product was combined with the sequence for the 80-nt RNA internal standard to generate the reference library. In that analysis, the number of exact sequencing reads that matched to the reference microRNA sequence in each trimmed sequencing file was determined with fgrep: numberReadsPerFile=$(fgrep $miRNA_sequence $trimmed_sequencing_file | wc -l). The read counts of the miRNAs were normalized to the summed counts for all detected miRNAs to obtain a "normalized read count". The summed counts of all detected miRNAs were also divided by 963, the total number of detected miRNAs, to obtain the "expected read count" assuming all species were equimolar. The read ratio was calculated by dividing the normalized read count by the expected read count.

AQRNA-seq data processing for human miRNAs.

The data processing workflow for human miRNAs (Fig. 1f) is similar to that noted for bacterial tRNAs. The analyses begin with assembly of read1 and read2 sequences using pear/0.9.10 to (1) stitch read1 and read2 together, (2) cross-validate the sequences of read1 and read2 to eliminate sequencing errors, and (3) strip linker 1 and linker 2 from both 5' and 3' ends of the RNA inserts. The resulting assembly outputs are high quality insert sequences with an additional two random nucleotides (2NN) at 3' end from linker 1 (Fig. 1a), which were subsequently stripped using fastx_trimmer (fastxtoolkit/0.0.13). The abundances of unique sequences were calculated in each sample using fastx_collapser (fastxtoolkit/0.0.13) and merge_count.pl (https://github.com/dedonlab/aqrnaseq). The sequences and counts in each sample were further merged and tabulated using custom scripts (https://github.com/dedonlab/aqrnaseq). for downstream computing and statistics analyses. Unique sequences ≥20 nt were then mapped to tRNA, miRNA, and lncRNA reference sequences using blast/2.6.0 (nucleotide BLAST) and aligned to the human genome using bwa/0.7.12. tRNA-specific analyses were performed by blasting the nonredundant sequences against 432 high-confidence human hg38-tRNA reference sequences in GtRNAdb (http://gtrnadb.ucsc.edu/genomes/eukaryota/Hsapi38/Hsapi38-seq.html). Inserts with at least a 17-nt perfect match to the tRNA references were assigned as tRNA sequences for further analyses. miRNA-specific analyses were performed by blasting nonredundant 20–30 nt sequences against 2656 mature human miRNA reference in mirbase (www.mirbase.org). Inserts with ≥17 nt of perfect match to miRNA references were considered as miRNA sequences for further analyses. lncRNA-specific analyses were performed by blasting the nonredundant sequences against 107039 GRCh38/hg38 human lncRNA reference in LNCipedia (https://hg38.lncipedia.org). Inserts with ≥50 nt of perfect match to lncRNA reference sequences with E-value <0.1 were considered as lnRNA sequences for further analyses. Genome-wide analyses were performed by mapping the nonredundant sequences against the GRCh38 reference genome using the BWA-backtrack algorithm. The sequences were then split into three categories: reads in gene regions, reads in intergenic regions, and reads not mapped to human genome. Reads in gene regions were annotated with the gene name, gene location (allowing an additional 100 nt upstream of the gene start-site and 100 nt downstream of the gene end-site to include some regulatory sequences), gene function, mapping location of the reads, and Compact Idiosyncratic Gapped Alignment Report (CIGAR) string.[69] The sequences mapped to gene desert regions were annotated with mapping location and cigar string. The number of sequences were counted in each step. The sequences, mapping information, and raw counts for each sample were merged and tabulated, resulting in four summary tables specific to tRNA, miRNA, lncRNA, and genome wide analyses.

BCG starvation proteomics: Isobaric labeling and peptide fractionation.

As the first step in the quantitative analysis of the starved BCG proteome, proteins were extracted[26] from biological triplicate cultures of BCG harvested during logarithmic growth in 7H9 and from cultures washed and resuspended in PBS for 4, 10, and 20 days (S4, S10, S20, respectively) and then resuspended in 7H9 medium for 6 days (S-R6). The extracted proteins were then precipitated, quantified, and processed as previously described for BCG hypoxia proteomics[26]. Aliquots of trypsin-digested protein (from 50 µg of total protein) were labelled with TMT 6-plex reagents (Thermo Scientific Tandem Mass Tag Reagents) according to manufacturer's instructions. Aliquots (5 µL) of the labelled peptides mixture were removed from each biological replicate and combined equi-volumetrically to reconstitute a full label set, which was analyzed on a Thermo Scientific EASY-nLC 1200 interfaced to a Thermo Scientific Q Exactive Hybrid Quadrupole-Orbitrap MS. Median total ion intensities for each label were calculated and used to normalize volumetric mixing of the remaining respective labeled samples, so as to avoid signal suppression or bias from any one label. The 6-plex mixture was then desalted with C18 SpinTips (Protea), dried by vacuum centrifugation, and reconstituted in IPD buffer (Agilent) without glycerol. Isoelectric focusing was performed from pH 3 to 10 over 24 wells on an Agilent 3100 OFFGEL fractionator according to the manufacturer's protocol (OG24PE00). Each of the 24 fractions was collected, dried by vacuum centrifuge, resuspended in 0.1% formic acid in water, and analyzed by nano-LC-MS/MS.

BCG starvation proteomics: Nano-LC-MS/MS analysis of the BCG proteome.

The TMT-labeled starvation time course samples were analyzed on an Agilent 1200 nano-LC-Chip/MS interfaced to an Agilent 6550 iFunnel Q-TOF LC/MS. The LC system consisted of a capillary pump for sample loading, a nanoflow pump, and a thermostatted microwell-plate autosampler. The HPLC-Chip configuration consisted of a 160-nL enrichment column and a 150 mm x 75 µm analytical column (G4340–62001 Zorbax 300SB-C18). The following mass-spectrometry grade mobile phases (Burdick & Jackson) were used: 0.1% formic acid in water (solvent A), and 0.1% formic acid in acetonitrile (solvent B). A 130-min linear gradient LC separation was used with 10 min for column wash and equilibration between runs. Samples (1–2 µL injections) were loaded onto the enrichment column at 3% (*v*/*v*) B at flow rates of 3 µL min^{-1}. The analytical gradient of solvent B was performed at a constant flow rate of 0.3 µL min^{-1} using the following solvent transitions on the nanoflow pump: 0–1 min, held at 1% (*v*/*v*); 1–10 min, 1–15%; 10–101 min, 15–35%; 101–121 min, 35–75%; 121–123 min, 75–98%; 123–126 min, held at 98%; 126–127 min, 98–1%; 127–130 min, held at 1%. LC-Q-TOF was operated at high sensitivity (4 GHz) in positive ion mode with the following source conditions: gas temperature 325 ºC, drying gas 13 L min^{-1}, fragmentor 360 V. Capillary voltage was manually adjusted between 1,800 to 2,150 V to achieve a steady nanospray. Data were acquired from 300 to 1,700 *m/z* with an acquisition rate of 6 spectra s^{-1} in MS mode, and from 50 to 1,700 *m/z* with an acquisition rate of 3 spectra s^{-1} in MS/MS mode. A peptide isotope model (charge state 2+) was used to detect a maximum 20 precursors per cycle at a minimum threshold of 25,000 counts/spectra at a narrow isolation window (~1.3 *m/z*). Sloped collision energy (C.E.) was used to maximize collision induced dissociation of detected isobarically tagged peptides according to the following rules: charge state 2+ C.E. slope 4.2, offset 3.5; charge states ≥ 3+ C.E. slope 4.2, offset 4.

LC/MS data was extracted and evaluated for quality using the MFE algorithm in MassHunter Qualitative Analysis software (v B06.00). Test injections (three to four) from each fraction of the first technical replicate were used to optimize injection volumes for second and third biological replicates with the aim to maximize the number of extracted molecules with peptide-like features. For each fraction, the MFE list of molecular ions was exported and used to exclude spectral acquisition of these ions in subsequent technical replicates. Each of the 24 fractions from biological triplicates were injected in technical duplicate — spectra generated from technical replicates 1 were acquired without use of an exclusion list, whereas spectra generated from technical replicate 2 were acquired with the exclusion list. Data from MassHunter Qualitative Analysis was exported to Mass Profiler Professional (v B03.00) for analysis of technical reproducibility. This process was repeated for all three biological replicates. Mass spectra were processed using Spectrum Mill (Agilent, v B06.00) and Scaffold Q+ (v Scaffold_4.8.8), and quantified protein associations were manually analyzed in Excel. Manually analyzing data pre-filtered at a 95% confidence interval (2 peptide minimum per protein ID) yielded 1,217 highly quantifiable proteins for the starvation proteomics experiment. Similarly, 965 highly quantifiable proteins were identified in all time points of our published BCG hypoxia iTRAQ proteomics studies[26]. The hypoxia proteomics data are available from the CHORUS mass spectrometric data repository at https://chorusproject.org/; Project ID 1107.

Northern blot analysis of bacterial tRNAs.

Small RNAs were purified from starvation cultures of BCG as described above. RNA from each timepoint (~225 ng) was resolved on 10% NovexTBE urea gels (Fisher) along with a New England Biolabs (NEB) low-range ssRNA ladder (50, 80, 150, 300, 500, 1000 nt) and NEB microRNA markers (17, 21, 25 nt). The gels were then stained with SYBR Gold prior to electrotransfer to BrightStar Plus positively charged nylon membranes (Thermo Fisher; Supplementary Fig. 5). Upon completion of electrotransfer, membranes were UV crosslinked and then hybridized with [^{32}P]-labeled oligonucleotide probes specific for the 5’ ends of one of three tRNAs (Asp, Asn or Trp). After two washes each in 1X SSC (150 mM NaCl, 15 mM sodium citrate) with 0.1% sodium dodecyl sulfate (SDS) for 30 min at 42 °C, the membrane was analyzed with a Storm Phosphorimager with a 19 h exposure. The membranes were stripped by repeated (4X) washing with heated 0.1X SSC with 0.1% SDS for 10 min and rinsed with distilled water prior to re-hybridizing with probes specific for the 3’ ends of the same tRNAs. Three sets of northern blots were generated by running RNA isolated from BCG starvation cultures (3 experimental replicates of S0, S4, S10, S20, R6 time course) on 3 sets of gels. Each set consisted of 2 gels containing a total of 15 BCG samples (5 timepoints x 3 experimental replicates) along with ladder/control lanes.

Statistics and reproducibility.

Descriptive statistics for all experimental data are defined in the figure legends, with mean and standard deviation or box-and-whisker plots for ≥3 experimental replicates, or individual data points for ≤2 experimental replicates. For inferential statistics in comparisons of tRNA levels in starved BCG in Figure 4, comparisons were made using 2-way ANOVA and P-values calculated using Bonferroni’s multiple comparisons test. GraphPad Prism was used for all graphical plots and statistical analyses.

Life Sciences Reporting Summary.

Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We thank members of Prof. David Bartel’s laboratory (Whitehead Institute and MIT Department of Biology), especially Asia Stefano and David Bartel, for assistance with northern blots. We thank Dr. Pavel Ivanov (Harvard Medical School) for sharing synthetic RNA standards. This work was supported by grants from National Natural Science Foundation of China (32070629 to C.B.), the US National Science Foundation (MCB-1412379 to V.C.L.), the National Institutes of Health (ES002109), and the National Research Foundation of Singapore through the Singapore-MIT Alliance for Research and Technology Antimicrobial Resistance IRG (P.C.D.). J.F.H. was supported by the MIT Toxicology Training Grant T32-ES007020), D.Y. by a postdoctoral fellowship from A*STAR, Singapore, and S.M.H. by a postdoctoral fellowship from the Swiss National Science Foundation.

References

1. Wang Z, Gerstein M & Snyder M RNA-Seq: a revolutionary tool for transcriptomics. Nature reviews. Genetics 10, 57–63 (2009).
2. Cech TR & Steitz JA The noncoding RNA revolution-trashing old rules to forge new ones. Cell 157, 77–94 (2014). [PubMed: 24679528]
3. Hafner Met al.RNA-ligase-dependent biases in miRNA representation in deep-sequenced small RNA cDNA libraries. RNA17, 1697–1712 (2011). [PubMed: 21775473]
4. Zhang Z, Lee JE, Riemondy K, Anderson EM & Yi R High-efficiency RNA cloning enables accurate quantification of miRNA expression by deep sequencing. Genome Biol 14, R109 (2013). [PubMed: 24098942]
5. Fuchs RT, Sun Z, Zhuang F & Robb GB Bias in ligation-based small RNA sequencing library construction is determined by adaptor and RNA structure. PloS one 10, e0126049 (2015).
6. Alon Set al.Barcoding bias in high-throughput multiplex sequencing of miRNA. Genome Res21, 1506–1511 (2011). [PubMed: 21750102]
7. Zhuang F, Fuchs RT, Sun Z, Zheng Y & Robb GB Structural bias in T4 RNA ligase-mediated 3’-adapter ligation. Nucleic Acids Res 40, e54 (2012). [PubMed: 22241775]
8. Pang YL, Abo R, Levine SS & Dedon PC 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). [PubMed: 25348403]
9. Linsen SEet al.Limitations and possibilities of small RNA digital gene expression profiling. Nat Methods6, 474–476 (2009). [PubMed: 19564845]
10. Machnicka MA, Olchowik A, Grosjean H & Bujnicki JM Distribution and frequencies of post-transcriptional modifications in tRNAs. RNA Biol 11, 1619–1629 (2014). [PubMed: 25611331]
11. Björk GRet al.Transfer RNA modification. Annual review of biochemistry56, 263–287 (1987).
12. Motorin Y & Helm M Methods for RNA Modification Mapping Using Deep Sequencing: Established and New Emerging Technologies. Genes (Basel) 10 (2019).
13. Motorin Y, Muller S, Behm‐Ansmant I & Branlant C Identification of Modified Residues in RNAs by Reverse Transcription‐Based Methods. Meth Enz 425, 21–53 (2007).
14. Shigematsu Met al.YAMAT-seq: an efficient method for high-throughput sequencing of mature transfer RNAs. Nucleic Acids Res45, e70 (2017). [PubMed: 28108659]
15. Gogakos Tet al.Characterizing Expression and Processing of Precursor and Mature Human tRNAs by Hydro-tRNAseq and PAR-CLIP. Cell Rep20, 1463–1475 (2017). [PubMed: 28793268]

16. Kim Het al.Bias-minimized quantification of microRNA reveals widespread alternative processing and 3’ end modification. Nucleic Acids Res47, 2630–2640 (2019). [PubMed: 30605524]

17. Dai Q, Zheng G, Schwartz MH, Clark WC & Pan T Selective Enzymatic Demethylation of N(2), N(2) -Dimethylguanosine in RNA and Its Application in High-Throughput tRNA Sequencing. Angew Chem Int Ed Engl 56, 5017–5020 (2017). [PubMed: 28371071]

18. Zheng Get al.Efficient and quantitative high-throughput tRNA sequencing. Nat Methods12, 835–837 (2015). [PubMed: 26214130]

19. Cozen AEet al.ARM-seq: AlkB-facilitated RNA methylation sequencing reveals a complex landscape of modified tRNA fragments. Nat Methods12, 879–884 (2015). [PubMed: 26237225]

20. Xu H, Yao J, Wu DC & Lambowitz AM Improved TGIRT-seq methods for comprehensive transcriptome profiling with decreased adapter dimer formation and bias correction. Sci Rep 9, 7953 (2019). [PubMed: 31138886]

21. Mohr Set al.Thermostable group II intron reverse transcriptase fusion proteins and their use in cDNA synthesis and next-generation RNA sequencing. RNA19, 958–970 (2013). [PubMed: 23697550]

22. Jayaprakash AD, Jabado O, Brown BD & Sachidanandam R Identification and remediation of biases in the activity of RNA ligases in small-RNA deep sequencing. Nucleic Acids Res 39, e141 (2011). [PubMed: 21890899]

23. Lovett ST & Kolodner RD Identification and purification of a single-stranded-DNA-specific exonuclease encoded by the recJ gene of Escherichia coli. Proc Natl Acad Sci U S A 86, 2627–2631 (1989). [PubMed: 2649886]

24. Shepherd J & Ibba M Bacterial transfer RNAs. FEMS Microbiol Rev 39, 280–300 (2015). [PubMed: 25796611]

25. Ardell DH & Hou YM Initiator tRNA genes template the 3’ CCA end at high frequencies in bacteria. BMC Genomics 17, 1003 (2016). [PubMed: 27927177]

26. Kozomara A & Griffiths-Jones S miRBase: annotating high confidence microRNAs using deep sequencing data. Nucleic Acids Res 42, D68–73 (2014). [PubMed: 24275495]

27. Herbert ZTet al.Multisite Evaluation of Next-Generation Methods for Small RNA Quantification. J Biomol Tech31, 47–56 (2020). [PubMed: 31966025]

28. Coenen-Stass AMLet al.Evaluation of methodologies for microRNA biomarker detection by next generation sequencing. RNA Biol15, 1133–1145 (2018). [PubMed: 30223713]

29. Zhang Yet al.IsomiR Bank: a research resource for tracking IsomiRs. Bioinformatics32, 2069–2071 (2016). [PubMed: 27153728]

30. Dong H, Nilsson L & Kurland CG Co-variation of tRNA abundance and codon usage in Escherichia coli at differeng growth rates. Journal of Molecular Biology 260, 649–663 (1996). [PubMed: 8709146]

31. Lewis KPersister cells. Annu Rev Microbiol64, 357–372 (2010). [PubMed: 20528688]

32. Grant SS & Hung DT Persistent bacterial infections, antibiotic tolerance, and the oxidative stress response. Virulence 4, 273–283 (2013). [PubMed: 23563389]

33. Zhang YPersisters, persistent infections and the Yin-Yang model. Emerg Microbes Infect3, e3 (2014). [PubMed: 26038493]

34. Steinberg S, Misch A & Sprinzl M Compilation of tRNA sequences and sequences of tRNA genes. Nucleic Acids Res 21, 3011–3015 (1993). [PubMed: 7687348]

35. Chan PP & Lowe TM GtRNAdb 2.0: an expanded database of transfer RNA genes identified in complete and draft genomes. Nucleic Acids Res 44, D184–189 (2016). [PubMed: 26673694]

36. Chionh YHet al.tRNA-mediated codon-biased translation in mycobacterial hypoxic persistence. Nat Commun7, 13302 (2016).

37. Doyle Fet al.Gene- and genome-based analysis of significant codon patterns in yeast, rat and mice genomes with the CUT Codon UTilization tool. Methods107, 98–109 (2016). [PubMed: 27245397]

38. Phizicky EM & Hopper AK tRNA biology charges to the front. Genes Dev 24, 1832–1860 (2010). [PubMed: 20810645]

39. Chernyakov I, Whipple JM, Kotelawala L, Grayhack EJ & Phizicky EM Degradation of several hypomodified mature tRNA species in Saccharomyces cerevisiae is mediated by Met22 and the 5'−3' exonucleases Rat1 and Xrn1. Genes Dev 22, 1369–1380 (2008). [PubMed: 18443146]

40. Hopper AK, Pai DA & Engelke DR Cellular dynamics of tRNAs and their genes. FEBS Lett 584, 310–317 (2010). [PubMed: 19931532]

41. Anderson P & Ivanov P tRNA fragments in human health and disease. FEBS Lett 588, 4297–4304 (2014). [PubMed: 25220675]

42. Cruz JWet al.Growth-regulating Mycobacterium tuberculosis VapC-mt4 toxin is an isoacceptor-specific tRNase. Nat Commun6, 7480 (2015). [PubMed: 26158745]

43. Schifano JMet al.tRNA is a new target for cleavage by a MazF toxin. Nucleic Acids Res44, 1256–1270 (2016). [PubMed: 26740583]

44. Lyons SM, Fay MM & Ivanov P The role of RNA modifications in the regulation of tRNA cleavage. FEBS Lett 592, 2828–2844 (2018). [PubMed: 30058219]

45. Fu Het al.Stress induces tRNA cleavage by angiogenin in mammalian cells. FEBS Lett583, 437–442 (2009). [PubMed: 19114040]

46. Haiser HJ, Karginov FV, Hannon GJ & Elliot MA Developmentally regulated cleavage of tRNAs in the bacterium Streptomyces coelicolor. Nucleic Acids Res 36, 732–741 (2008). [PubMed: 18084030]

47. Boccaletto Pet al.MODOMICS: a database of RNA modification pathways. 2017 update. Nucleic Acids Res46, D303-D307 (2018).

48. Kietrys A, Velema W & Kool E Fingerprints of modified RNA bases from deep sequencing profiles. J Am Chem Soc 139, 17074–17081 (2017).

49. Hauenschild Ret al.The reverse transcription signature of N-1-methyladenosine in RNA-Seq is sequence dependent. Nucleic Acids Res43, 9950–9964 (2015). [PubMed: 26365242]

50. Motorin Y, Muller S, Behm‐Ansmant I & Branlant C in RNA Modification 21–53 (2007).

51. Levanon EYet al.Systematic identification of abundant A-to-I editing sites in the human transcriptome. Nat Biotechnol22, 1001–1005 (2004). [PubMed: 15258596]

52. Elenbaas Bet al.Human breast cancer cells generated by oncogenic transformation of primary mammary epithelial cells. Genes Dev15, 50–65 (2001). [PubMed: 11156605]

53. Kendall SD, Adam SJ & Counter CM Genetically engineered human cancer models utilizing mammalian transgene expression. Cell Cycle 5, 1074–1079 (2006). [PubMed: 16687931]

54. Qattan Aet al.Robust expression of tumor suppressor miRNA's let-7 and miR-195 detected in plasma of Saudi female breast cancer patients. BMC Cancer17, 799 (2017). [PubMed: 29183284]

55. Xuan P, Li L, Zhang T, Zhang Y & Song Y Prediction of disease-related microRNAs through integrating attributes of microRNA nodes and multiple kinds of connecting edges. Molecules 24, 3099 (2019).

56. Wang Xet al.Differential expression profile analysis of miRNAs with HER-2 overexpression and intervention in breast cancer cells. Int J Clin Exp Pathol10, 5039–5062 (2017).

57. Maltseva DVet al.miRNome of inflammatory breast cancer. BMC Res Notes7, 871 (2014). [PubMed: 25471792]

58. Ueda S, Takanashi M, Sudo K, Kanekura K & Kuroda M miR-27a ameliorates chemoresistance of breast cancer cells by disruption of reactive oxygen species homeostasis and impairment of autophagy. Lab Invest 100, 863–873 (2020). [PubMed: 32066826]

59. Pirouz M, Ebrahimi AG & Gregory RI Unraveling 3'-end RNA uridylation at nucleotide resolution. Methods 155, 10–19 (2019). [PubMed: 30395968]

60. Sorefan Ket al.Reducing ligation bias of small RNAs in libraries for next generation sequencing. Silence3, 4 (2012). [PubMed: 22647250]

61. Dard-Dascot Cet al.Systematic comparison of small RNA library preparation protocols for next-generation sequencing. BMC Genomics19, 118 (2018). [PubMed: 29402217]

62. Wong RKY, MacMahon M, Woodside JV & Simpson DA A comparison of RNA extraction and sequencing protocols for detection of small RNAs in plasma. BMC Genomics 20, 446 (2019). [PubMed: 31159762]

63. Heinicke Fet al.Systematic assessment of commercially available low-input miRNA library preparation kits. RNA Biol17, 75–86 (2020). [PubMed: 31559901]

64. Chu Yet al.Intramolecular circularization increases efficiency of RNA sequencing and enables CLIP-Seq of nuclear RNA from human cells. Nucleic Acids Res43, e75 (2015). [PubMed: 25813040]

65. , Edn. July 2019 (2019).

66. Xiong Yet al.A Comparison of mRNA Sequencing with Random Primed and 3'-Directed Libraries. Sci Rep7, 14626 (2017).

Methods-only References

67. Baba Tet al.Construction of Escherichia coli K-12 in-frame, single-gene knockout mutants: the Keio collection. Mol Syst Biol2, 2006 0008 (2006).

68. Hia Fet al.Mycobacterial RNA isolation optimized for non-coding RNA: high fidelity isolation of 5S rRNA from Mycobacterium bovis BCG reveals novel post-transcriptional processing and a complete spectrum of modified ribonucleosides. Nucleic Acids Res43, e32 (2015). [PubMed: 25539917]

69. Li Het al.The Sequence Alignment/Map format and SAMtools. Bioinformatics25, 2078–2079 (2009). [PubMed: 19505943]

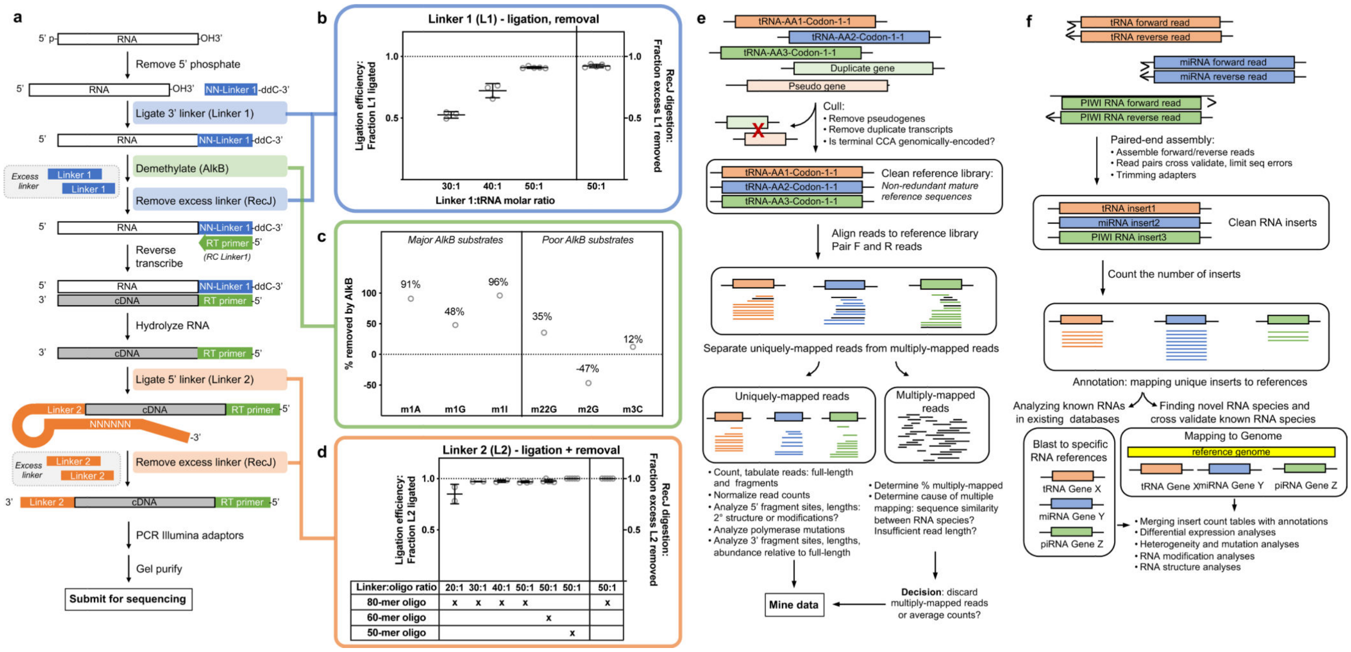
a
5' p-
RNA
OH3'
Remove 5' phosphate
5'
RNA
OH3'
NN-Linker 1-ddC-3'
Ligate 3' linker (Linker 1)
Demethylate (AlkB)
Excess linker
Linker 1
Linker 1
Remove excess linker (RecJ)
RT primer
(RC Linker1)
Reverse transcribe
cDNA
Hydrolyze RNA
Ligate 5' linker (Linker 2)
Linker 2
NNNNNN
-3'
Excess linker
Linker 2
Remove excess linker (RecJ)
PCR Illumina adaptors
Gel purify
Submit for sequencing
b
Linker 1 (L1) - ligation, removal
Ligation efficiency: Fraction L1 ligated
RecJ digestion: Fraction excess L1 removed
1.0
0.5
30:1
40:1
50:1
50:1
Linker 1:tRNA molar ratio
c
Major AlkB substrates
Poor AlkB substrates
% removed by AlkB
100
50
0
-50
91%
48%
96%
35%
-47%
12%
m1A
m1G
m1I
m22G
m2G
m3C
d
Linker 2 (L2) - ligation + removal
Ligation efficiency: Fraction L2 ligated
RecJ digestion: Fraction excess L2 removed
1.0
0.5
Linker:oligo ratio
20:1
30:1
40:1
50:1
50:1
50:1
50:1
80-mer oligo
x x x x
x
60-mer oligo
x
50-mer oligo
x
e
tRNA-AA1-Codon-1-1
tRNA-AA2-Codon-1-1
tRNA-AA3-Codon-1-1
Duplicate gene
Pseudo gene
X
Cull:
• Remove pseudogenes
• Remove duplicate transcripts
• Is terminal CCA genomically-encoded?
Clean reference library:
Non-redundant mature reference sequences
Align reads to reference library
Pair F and R reads
Separate uniquely-mapped reads from multiply-mapped reads
Uniquely-mapped reads
Multiply-mapped reads
• Count, tabulate reads: full-length and fragments
• Normalize read counts
• Analyze 5' fragment sites, lengths: 2° structure or modifications?
• Analyze polymerase mutations
• Analyze 3' fragment sites, lengths, abundance relative to full-length
• Determine % multiply-mapped
• Determine cause of multiple mapping: sequence similarity between RNA species? Insufficient read length?
Mine data
Decision: discard multiply-mapped reads or average counts?
f
tRNA forward read
tRNA reverse read
miRNA forward read
miRNA reverse read
PIWI RNA forward read
PIWI RNA reverse read
Paired-end assembly:
• Assemble forward/reverse reads
• Read pairs cross validate, limit seq errors
• Trimming adapters
tRNA insert1
miRNA insert2
PIWI RNA insert3
Clean RNA inserts
Count the number of inserts
Annotation: mapping unique inserts to references
Analyzing known RNAs in existing databases
Finding novel RNA species and cross validate known RNA species
Blast to specific RNA references
tRNA Gene X
miRNA Gene Y
piRNA Gene Z
Mapping to Genome
reference genome
tRNA Gene X
miRNA Gene Y
piRNA Gene Z
• Merging insert count tables with annotations
• Differential expression analyses
• Heterogeneity and mutation analyses
• RNA modification analyses
• RNA structure analyses

Figure 1. Overview of AQRNA-seq.

(a) Library preparation workflow. **(b-d)** Optimization experiments for the library preparation workflow. **(b)** Linker 1 ligation and removal proceeds with >90% efficiency at a linker:tRNA molar ratio of 50:1. Dot plot shows all data with bars for mean and SD, N=3. **(c)** AlkB demethylation efficiencies for RNA modifications. The data represent percent reduction for a single experiment (N=1). **(d)** Linker 2 ligation is nearly 100% efficient at linker to tRNA molar ratios ≥30:1. A 50:1 ratio is used here, with linker 2 removal nearly ~100% efficient. Dot plot shows all data with bars for mean and SD, N=3. **(e)** Data processing workflow for bacterial tRNAs. Reads are mapped against a non-redundant reference genome. The paired-end protocol of AQRNA-seq yields two FASTQ files per library – one each for forward and reverse reads. After alignment, multiply-and uniquely-mapped reads are separated and mined for abundance and coverage information. (**f**) Data processing workflow for human miRNAs. While bacterial reads are mapped first and then counted according to the mapped RNA species, human inserts are counted first and then blasted to RNA reference sequences or mapped to the entire genome for annotation. Random sequencing errors are corrected and read pairs cross-validated by assembling paired-end forward and reverse reads before counting.

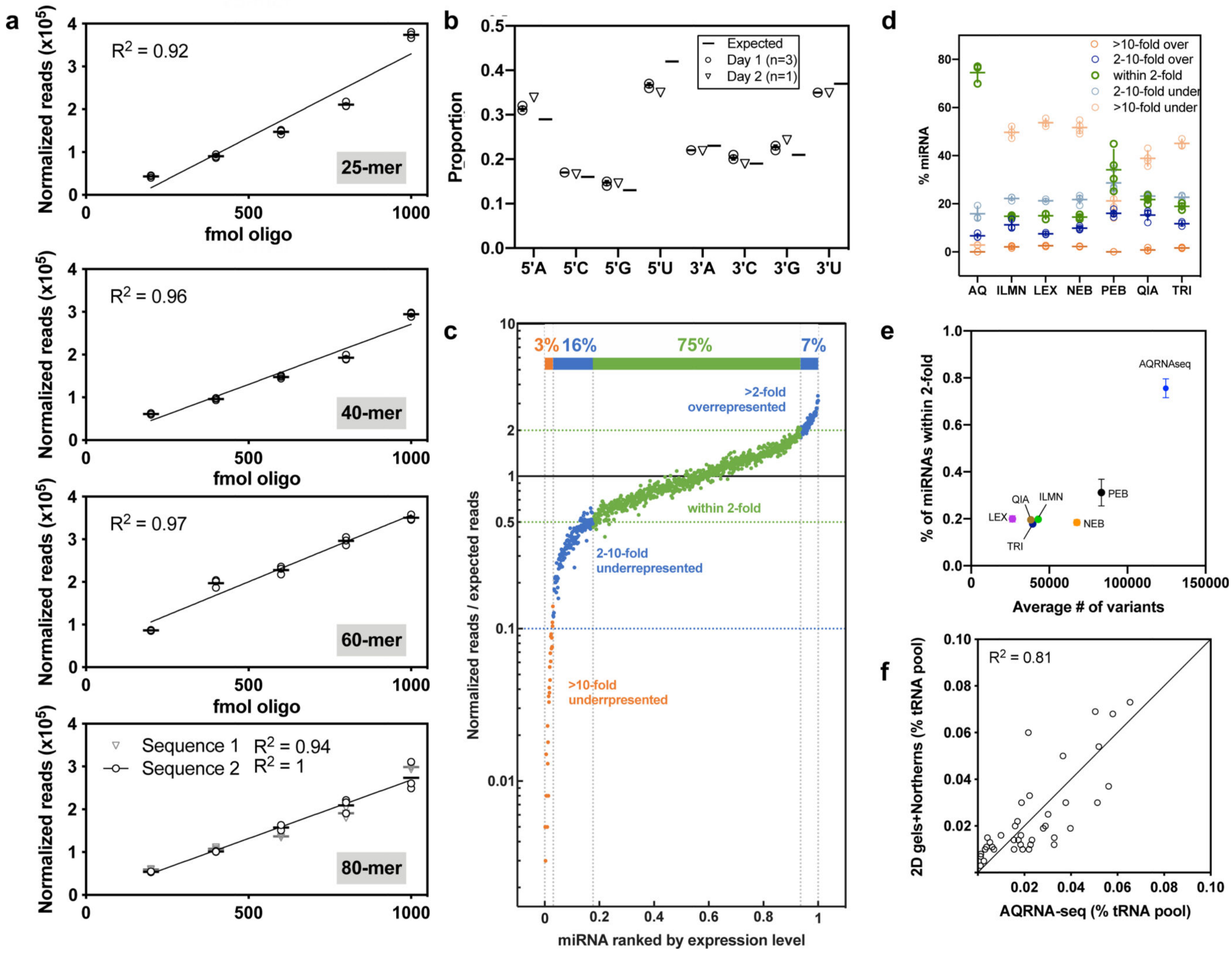

Figure 2. Quantitative validation of AQRNA-seq.

(a) Oligonucleotide spike-ins demonstrate a linear relationship between copy number and read count. Oligonucleotides 25–80 nt long were subjected to AQRNA-seq at different concentrations (GEO accession GSE139936). Dot plots show all data, bar denoting mean, for N=3 experiments. **(b)** Minimal sequence bias in AQRNA-seq analysis of the 963 miRNA Miltenyi miRXplore Universal Reference (GEO accession GSE139936). Among measured miRNAs, the 5’ and 3’ nucleotides were tabulated and their proportions plotted. Dot plot shows data for N=3 experiments on Day 1 and N=1 experiment on Day 2, with a dash denoting expected proportions of A, C, G, and U at each end among all 963 reference miRNAs. **(c)** AQRNA-seq quantitative fidelity was assessed using the miRXplore Reference (GEO accession GSE139936). Sequencing reads for each miRNA were normalized to expected values and sorted into 5 bins as denoted in the graph. The colored bar indicates the percentage of reads within 2- and 10-fold of expected abundance. **(d)** Comparison of the quantitative accuracy of miRNA libraries prepared from the miRXplore Reference using the AQRNA-seq (“AQ”) protocol and the following small RNA or miRNA library kits: Illumina TruSeq (ILM), Lexogen (LEX), NEBNext for Illumina (NEB), Perkin Elmer

NextFlex (PEB), QIAseq miRNA (QIA), and Trilink CleanTag (TRI). Data for the kits was derived from Herbert et al.[26] For each replicate and for each kit, the percentage of total miRNAs found in each bin denoted in panel **c** was calculated. Dot plot shows data for N=3 experiments (ILM, LEX, TRI, AQ) or N=4 (NEB, PEB, QIA), with bars denoting mean ± SD. **(e)** Among AQRNA-seq and the other RNA-seq kits, a positive correlation exists between the average number of sequence variants detected and the percentage of miRNAs quantified within 2-fold of expected value. Sequence variants are defined as additions and subtractions to the insert sequences during library preparation; see Supplementary Figure 2 for the set of sequence variants arising for the kits. Data represent mean ± SD for N=3 experiments (ILM, LEX, TRI, AQ) or N=4 (NEB, PEB, QIA). **(f)** Correlation of tRNA quantification results using AQRNA-seq versus data derived from 2D gel electrophoresis and northern blotting by Dong et al.[27] Data represent individual values derived from Dong et al.[27] plotted relative to mean values from N=3 AQRNA-seq analyses of *E. coli* tRNAs.

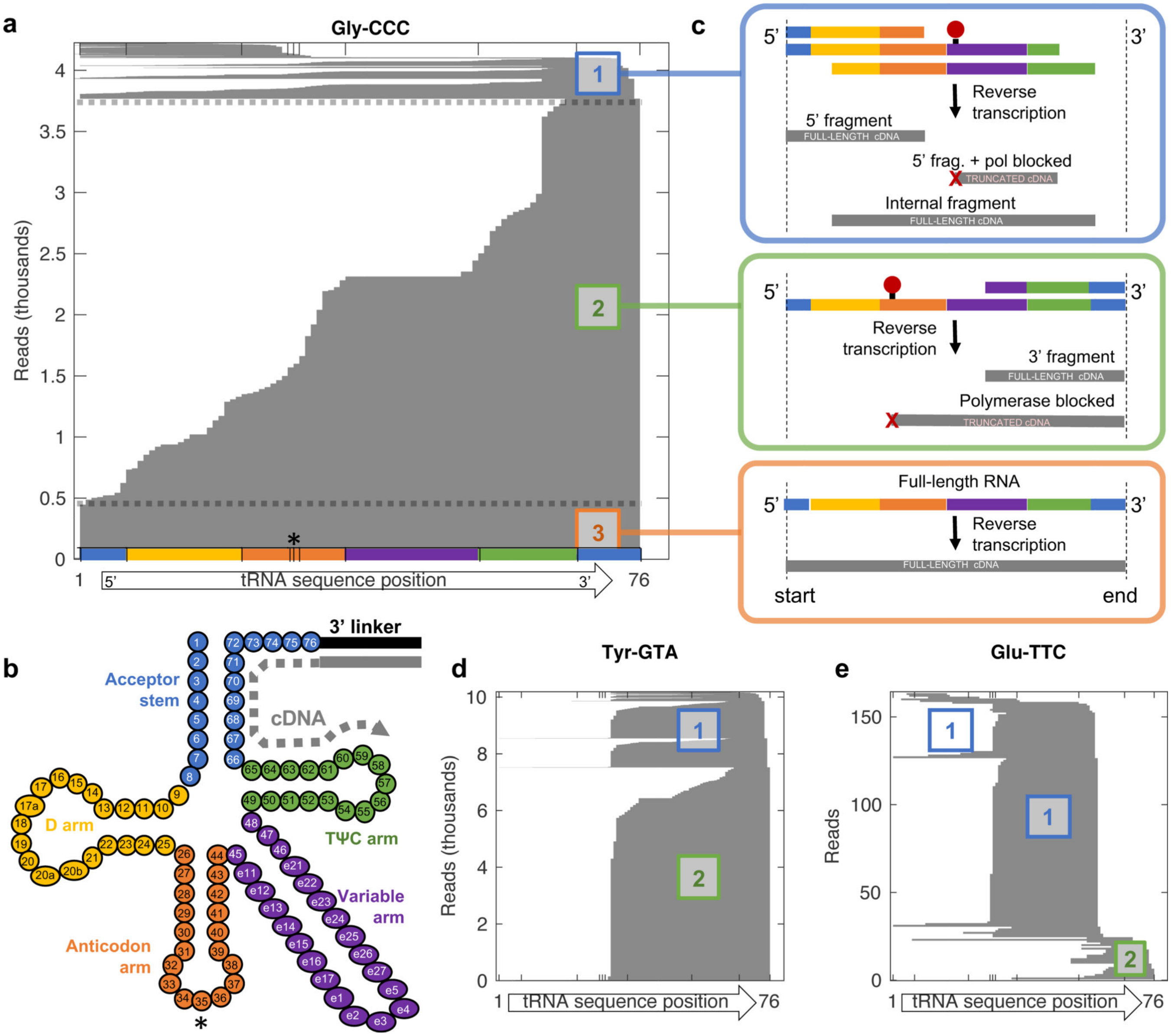

Figure 3. Alignment plots of *M. bovis* BCG tRNAs.

(**a**) Alignment plot showing the start and end position of reads aligning to tRNA Gly-CCC in a stacked horizontal bar graph. The tRNA sequence numbering allows positions 1 and 76 to reflect the 5' and 3' termini, respectively. * Anticodon indicated by three vertical lines. (**b**) Schematic showing linker 1 attachment and reverse transcription along the tRNA sequence (Sprinzl coordinate system).[34] (**c**) Aligned reads fall into three categories with different interpretations – see text. (**d**) Example alignment plot for tRNA Tyr-GTA showing polymerase blockage downstream of the anticodon resulting in a lack of Type 3 (full-length) reads. (**e**) Example of an alignment plot to tRNA Glu-TTC showing polymerase blockage near the anticodon and enrichment of fragments aligning inside the 3'-end, resulting in increased Type 1 reads. BCG AQRNA-seq data available in BioProject #PRJNA579244.

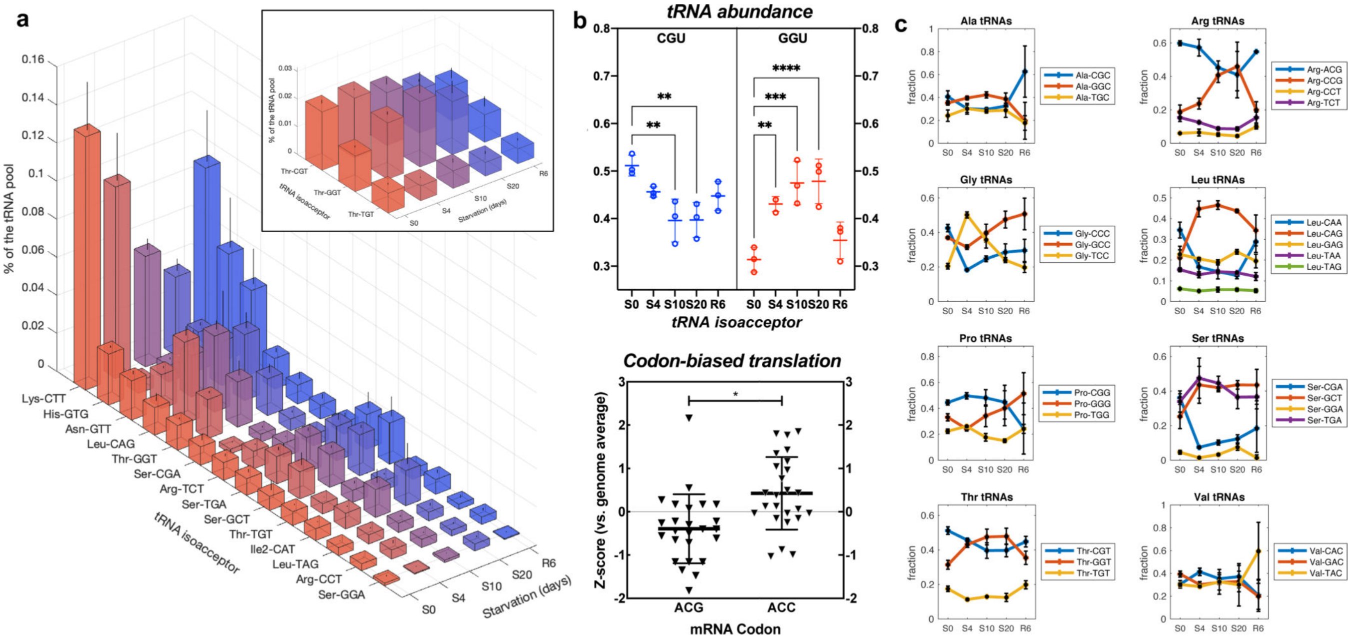

Figure 4. Starvation-induced changes in tRNA abundance correlate with changes in codon-biased translation in *M. bovis* BCG.

(a) Plots show normalized abundance of selected tRNAs across the starvation time course (S0, nutrient-rich medium; S4–20, 4–20 d starvation; R6, 6 d resuscitation in nutrient-rich medium). *Inset*: Normalized abundance of tRNA-Thr isoacceptors. Data represent mean ± SD for N=3 experiments. Individual data omitted for clarity. (**b**) ***Upper***: Time courses for changes in abundance of tRNA-Thr isoacceptors with anticodons CGU and GGU. Dot plot shows data for N=3 experiments with bars for mean ± SD. ***Lower*:** Codon usage in mRNAs for the 25 most upregulated proteins at 30 d starvation. ACG and ACC are cognate codons for tRNA-Thr isoacceptors with anticodons CGU and GGU, respectively, noted in the upper panel. Dot plot with bars for mean ± SD shows Z-scores for codon usage relative to genome averages for the mRNAs for the 25 most upregulated proteins. (**c**) Plots showing abundances of individual isoacceptors (all reads aligning at the 3’-end relative to the total set of tRNAs that carry the same amino acid). Data represent mean ± SD for N=3 experiments. BCG AQRNA-seq data available at BioProject #PRJNA579244.

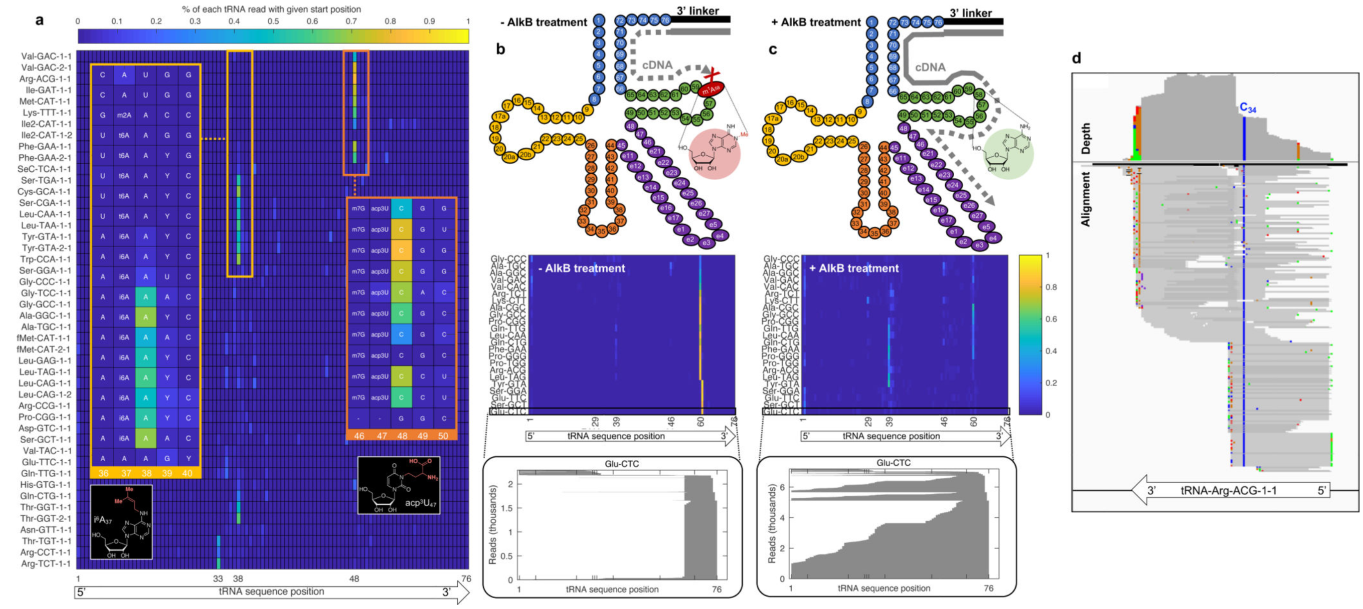

Figure 5.

Application of AQRNA-seq for quantitative mapping of the tRNA epitranscriptome. Many RNA modifications block 3'-to-5' reverse transcriptase-mediated cDNA synthesis or result in mutations in AQRNA-seq (gray line and arrow in panel b). This can be exploited to quantitatively map the modifications. (a) Subsets of E. coli tRNAs exhibit similar reverse transcriptase blockages at positions 38 and 48. The heat map shows the percentage of sequencing reads for which reverse transcription ended at a specific location in the tRNA sequence (columns) for each tRNA isoacceptor (rows). The two positions showing the most significant accumulations of polymerase blockade are noted with a small orange or red box. The RNA sequences surrounding these positions are shown in the larger boxes that magnify the sequence location, with specific modified nucleotides noted based on existing maps of E. coli tRNA modifications. In the orange boxes, the 8 tRNA species showing polymerase blockade at position 38 reveals all possess i6A at position 37. In the red boxes, the 10 tRNAs showing polymerase blockade at position 48 all possess acp3U at position 47. (b) In the absence of AlkB treatment, cDNA synthesis is blocked by m1A at position 58 in nearly half of all BCG tRNAs, which is reflected by the high proportion of aligned reads that do not extend past position 58 in the heatmap of read start positions (light blue to orange line in heat map similar to panel a). This is illustrated for tRNA Glu-CTC in the gray stack plot, which shows that early all the reads begin after position 58, forming a "cliff". (c) After AlkB demethylation, however, the read alignments lengthen and extend past the cliff, resulting in a more varied distribution of alignment start positions. The heat map shows a significant reduction in the number of aligned reads. (d) Many RNA modifications can also be mapped by polymerase-induced mutations in the resulting cDNA. This is illustrated with a striking T-to-C misreading in BCG tRNA Arg-ACG, which is consistent with the presence of inosine at position 34 of the anticodon on nearly all copies of the tRNA. BCG AQRNA-seq data are available in BioProject accession number PRJNA579244.

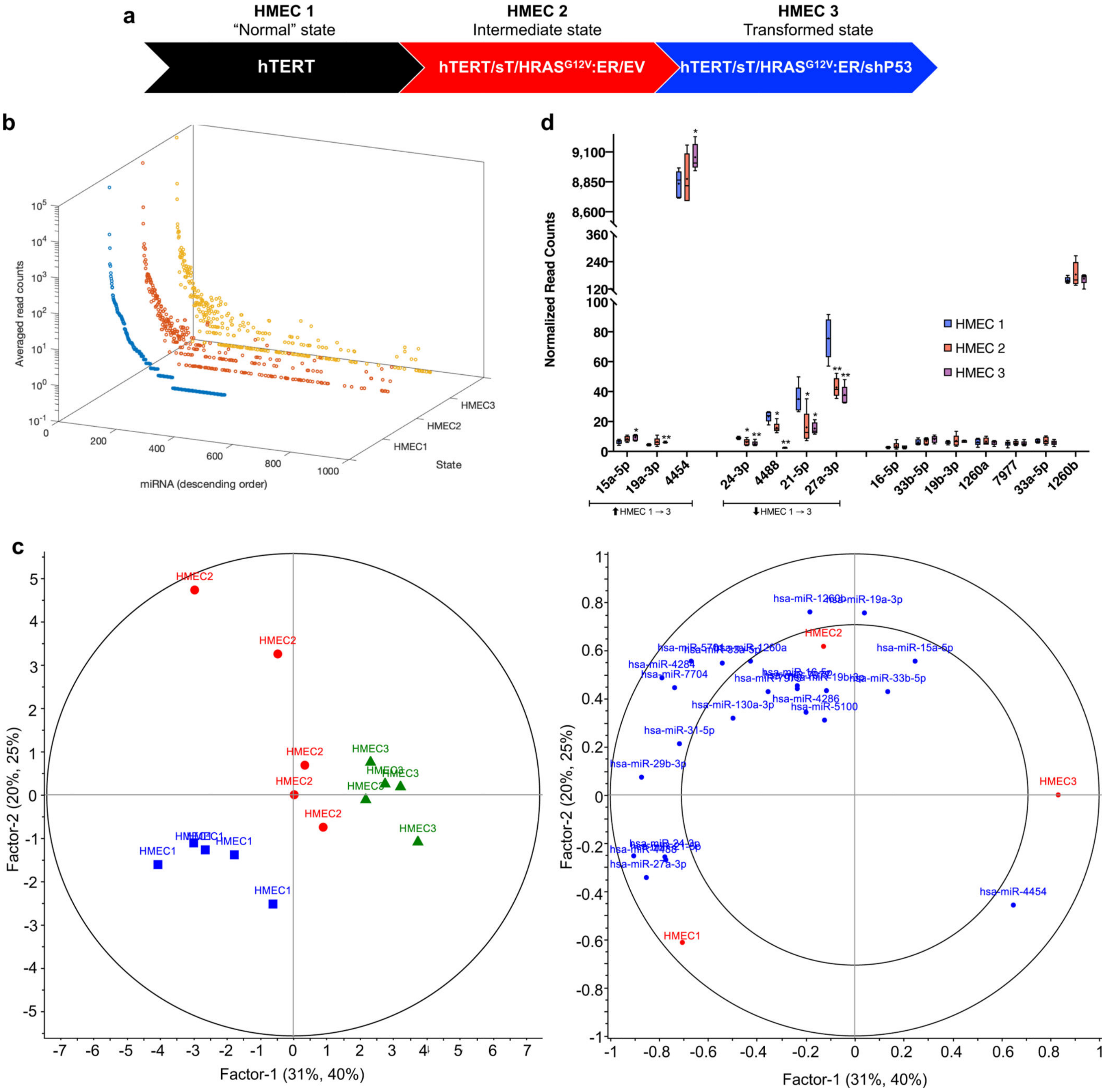

Figure 6. AQRNA-seq analysis of miRNAs in HMEC cancer cells.

(**a**) The HMEC model for progressive tumorigenesis. Primary HMEC cells were immortalized with SV40 large-T antigen and the telomerase catalytic subunit (HMEC1), with subsequent tumorigenic behavior induced by H-Ras oncoprotein (HMEC2) and additional loss of P53 (HMEC3).[52] (**b**) AQRNA-seq reveals a 5 order-of-magnitude range in levels of 875 miRNAs in the HMEC cell lines. Data represent the averaged read count across 5 different cell cultures for each miRNA. Error bars omitted for clarity. The X-axis order of presentation of individual miRNAs is prioritized by decreasing frequency for HMEC1. (**c**) PLSR analysis of the abundance 875 miRNAs associated with the HMEC cells. ***Left***:

scores plot showing strong distinctions among the cell lines. ***Right:*** loadings plot showing the miRNAs most significantly distinguishing the three HMEC cells. (**d**) Changes in the levels of 14 miRNAs strongly associated with the HMEC cell lines from the analysis in panel **c**. miRNAs increasing from HMEC1 to HMEC 3: 15a-5p, 19a-3p, 4454. miRNAs decreasing from HMEC1 to HMEC3: 24–3p, 4488, 21–5p, 27a-3p. miRNAs on the right were unchanged across the HMEC cell lines. Box-and-whisker plot for N=5 experiments: whiskers, maximal and minimal data; box, 25th to 75th percentile; dash, median; and "+", mean.