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

Jennifer F. Hu1,10, Daniel Yim2,11, Duanduan Ma3, Sabrina M. Huber2,12, Nick Davis2,13, Jo Marie Bacusmo4, Sidney Vermeulen2, Jieliang Zhou5, Thomas J. Begley6, Michael S. DeMott2,7, Stuart S. Levine2,3,7,8, Valérie de Crécy-Lagard4, Peter C. Dedon2,7,8,9 and Bo Cao2,8,9

Current next-generation RNA-sequencing (RNA-seq) 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 RNA-sequencing (AQRNA-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 microRNA reference library, oligonucleotide standards of varying length, and RNA 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 transfer RNA 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 genomics1,2, current next-generation RNA-seq methods provide precise and accurate analysis of changes in transcript abundance between samples but 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 102-fold variation in efficiency3,4,5,6,7,8,9 causing 105-fold artifacts in sequencing read counts10,11. Highly structured and modified RNA molecules, such as tRNAs, further challenge the quantitative accuracy of RNA-seq12 by causing polymerase fall-off during cDNA synthesis13,14. 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 105-fold range and revealed that the majority of miRNA isomers (‘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 5′ end, with two randomized nucleotides at the 5′ end of linker 1 to maximize T4 ligase efficiency15. 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 intact16. A 50:1 excess of linker 1 resulted in >90% ligation efficiency (Fig. 1b and Supplementary Fig. 1). Ligated RNA can then be treated with AlkB to reduce levels of RT-blocking methyl modifications17. Although not essential for capturing all RNA sequences, it can provide information about the identities of polymerase-blocking modifications. Because the buffer provided with the commercial kit caused RNA degradation (Supplementary Fig. 1k,l), we optimized buffer conditions and AlkB concentration to reduce m7A (90%), m4G (48%) and m5U (96%). Contrary to previous observations18, only 12% of mC was removed.

1Department of Chemistry, Massachusetts Institute of Technology, Cambridge, MA, USA. 2Department of Biological Engineering, Massachusetts Institute of Technology, Cambridge, MA, USA. 3BioMicro Center, Massachusetts Institute of Technology, Cambridge, MA, USA. 4Department of Microbiology & Cell Science, University of Florida, Gainesville, FL, USA. 5KK Research Center, KK Women's and ChildrenBristol Myers Squibb's Hospital, Singapore, Singapore. 6The RNA Institute and Department of Biology, University at Albany, Albany, NY, USA. 7Center for Environmental Health Sciences, Massachusetts Institute of Technology, Cambridge, MA, USA. 8Singapore-MIT Alliance for Research and Technology Antimicrobial Resistance IRG, Singapore, Singapore. 9College of Life Sciences, Qufu Normal University, Qufu, China. 10Present address: Bristol Myers Squibb, Seattle, WA, USA. 11Present address: Theon Therapeutics, Cambridge, MA, USA. 12e-mail: pcedon@mit.edu; caobo@qfnu.edu.cn

NATURE BIOTECHNOLOGY | VOL 39 | AUGUST 2021 | 978–988 | www.nature.com/naturebiotechnology
and m^{2+}G was reduced by only 35%, apparently demethylated to AlkB-resistant m^{2+}G (Fig. 1c). After demethylation, optimized RecJ digestion removed >99% of unligated linker 1 (Fig. 1b), obviating high-performance liquid chromatography (HPLC) purification of the RNA:DNA product.

Reverse transcription is accomplished with a DNA primer complementary to linker 1, and the resulting complementary DNA is ligated to a custom DNA adapter (linker 2) using T4 DNA ligase. Linker 2 possesses a hairpin, a random hexamer 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 and 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 (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 levels of mutation. 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 RNA-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-validation of 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 curating of reference sequences used for alignment of 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 modifications24. For mycobacteria, the 3′ CCA of each tRNA is variably genomically encoded or added post-transcriptionally25.

The resulting nonredundant reference library can be used to align sequencing reads, first separating uniquely aligned reads from those matching multiple sequences. Even with careful library cura-

tion, high sequence similarity or insufficient read length can result in multiply mapped reads. We arbitrarily set a 10-nucleotide (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 (that is, do they significantly alter the final read count for each RNA?) and the cause of multiple mapping (for example, 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 to either the total number of reads in each barcoded sample or an internal RNA standard, to account for sample-to-sample variation in input RNA as well as variable sample pooling before 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 of varying length (25–80 nt) were mixed at varying molar ratios and used as input RNA for library preparation (Supplementary Tables 2 and 3). After sequencing, we found that read counts for each oligonucleotide varied directly and linearly with input RNA abundance (R² = 0.92–1.0; Fig. 2a), with an average sequencing response (slope) of ~300 ± 50 reads fmol⁻¹ 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 miRBase26 (16–28 nt) that possessed all 16 possible dinucleotide combinations at the 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 2-fold of expected abundance (Fig. 2c). The number of jackpot and drop-out 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 reference21,28, established AQRNA-seq as the most quantitatively accurate RNA-seq workflow.

Analysis of sequencing reads for the miRXplore reference set revealed an unexpected correlation between quantitative accuracy and sequence variants induced during library preparation. Here we defined seven classes of sequence variant representing additions and deletions at the ends of the miRNA inserts according to the nomenclature used in the IsomiR Bank29. These sequence variants are not present in the miRXplore set and could have arisen only during library preparation or sequencing. The proportions of sequence variants for the six small RNA-seq methods (green background) and AQRNA-seq (yellow background) are depicted in Supplementary Fig. 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 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 databases29. 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 twofold of that expected (Fig. 2e and Supplementary Fig. 2c). Minor variation in the sequences of inserts does not affect the alignment step during the data mining, 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 analysis of the Escherichia coli K12 tRNA pool on two-dimensional gels with RNA 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 (R² = 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: 45 tRNAs in the Mycobacterium bovis bacille de Calmette et Guérin (BCG) model for the stress-induced, nonreplicative, antibiotic-resistant state of persistence in tuberculosis31–33. Total small RNA was isolated along the time course of BCG persistence caused by nutrient deprivation (Supplementary Fig. 3a,b), with ~1% of 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 expression of 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 axis annotated using the Sprinzl tRNA coordinate system (Fig. 3b)34, with the 3′ end defining the location of linker 1 (Fig. 3b). Alignment plots for the 45 tRNA species in BCG are shown in Supplementary Fig. 4, with stack height directly proportional to the total number of expressed transcripts. The graph (Fig. 3a) can be split into sections: bottom, type 3 or full-length reads spanning the entire tRNA sequence (Fig. 3c); top, reads not reaching the 3′ end of the tRNA (type 1; Fig. 3c) 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. 3c) and represent either tRNA fragments missing a 3′ portion or full-length tRNAs for which cDNA synthesis was prematurely truncated by RT fall-off.

Analysis of three tRNA isoacceptors by RNA blotting (Supplementary Fig. 5) suggests that the majority of 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 aligned to a particular tRNA—changed a large range. In most samples, tRNA Lys-CTT and tRNA IMet-CAT were the most highly expressed tRNAs, together totaling ~20% of the pool during log growth, whereas tRNA Ser-GGA was ~80-fold lower at 0.1–0.3% of
Starvation also induced significant shifts among isoacceptor families (Fig. 4c). For example, Ser, Thr and Leu are specified by six, four and six synonymous codons, respectively, and these codons are read by four, three and five different tRNA species, respectively35. As shown in Fig. 4c, before starvation (S0), tRNAs Ser-CGA, Ser-TGA and Ser-GCT comprised 36, 34 and 25% of the isoacceptor families (Fig. 4c). For example, Ser, Thr and Leu are specified by six, four and six synonymous codons, respectively, and these codons are read by four, three and five different tRNA species, respectively35. As shown in Fig. 4c, before starvation (S0), tRNAs Ser-CGA, Ser-TGA and Ser-GCT comprised 36, 34 and 25% of the isoacceptor pool, respectively. At 4 days of starvation (S4), the

the pool (Fig. 4a and Supplementary Fig. 4). In the transition from rich medium to starvation over 20 days, the abundances of several tRNAs were significantly altered (Fig. 4a,b and Supplementary Fig. 4). For example, tRNA-Ser-CGA and tRNA His-GTG dropped significantly in early starvation and rose again during late starvation and resuscitation. tRNA Leu-CAG and tRNA-Thr-GGT showed the opposite pattern.
abundance of Ser-CGA dropped to 7% while Ser-TGA and -GCT surged 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%, respectively, at S20, 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 responds to persistence-inducing hypoxia by uniquely altering tRNA modifications to cause selective translation of messenger RNAs coding for hypoxia response genes—Dos regulon—that possess codon usage patterns matching hypoxia-altered tRNAs35. To test this mechanism in starvation-induced persistence, we performed quantitative proteomics across the starvation time course, detecting 1,102 proteins common to three separate cultures at all time points (Supplementary Fig. 6). Analysis of codon usage frequencies37 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).
Nearly all forms of RNA contain post-transcriptional modifications, such as RNA editing, ribozyme processing, and tRNA modifications. AQRNA-seq thus provides quantitative information about tRNA levels that can be used to study translation efficiency. These modifications can be quantified using Sanger sequencing or high-throughput sequencing, and these data can be used to investigate the relationship between translation efficiency and amino acid availability.

Fig. 4 | Starvation-induced changes in tRNA abundance correlate with changes in codon-biased translation in M. bovis BCG. a, Plots showing normalized abundance of selected tRNAs across the starvation time course (S0, nutrient-rich medium; S4–20, 4–20 days of starvation; R6, 6 days of resuscitation in nutrient-rich medium). Inset: normalized abundance of tRNA-isoacceptors. Data represent mean ± s.d. for n = 3 experiments. Individual data are omitted for clarity. b, Upper: time courses for changes in abundance of tRNA-isoacceptors with anticodons CGU and GGU. Dot plot shows data for n = 3 experiments with bars for mean ± s.d. Lower: codon usage in mRNAs for the 25 most upregulated proteins at 30 days of starvation. ACG and ACC are cognate codons for tRNA-isoacceptors with anticodons CGU and GGU, respectively. Each circle represents a codon. **P ≤ 0.01, ***P ≤ 0.001, ****P ≤ 0.0001.

c, Plots showing abundance of individual isoacceptors (all reads aligning at the 3' end relative to the total set of tRNAs carrying the same amino acid). Data represent mean ± s.d. for n = 3 experiments. BCG AQRNA-seq data available at BioProject, no. PRJNA579244.

These same genes underutilized the AGC codon read by tRNA-Thr-CTG 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 control44–46. Small RNA regulation of gene expression and toxic–anti-toxin systems41–46. It is difficult to differentiate 5' degradation of full-length tRNA from polymerase fall-off because both generate a fragment that aligns at the 3' end of tRNA (type 2; Fig. 3b). However, 3' degradation and endonuclease cleavage generate fragments with 3' ends positioned inside the reference sequence (type 1; Fig. 3a,c). 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' end relative to the total set of tRNAs carrying the same amino acid.

Quantitative mapping of tRNA modifications and structures. Nearly all forms of RNA contain post-transcriptional modifications, with >150 structures known47. tRNAs are particularly heavily modified at ~10% of the component nucleotides48. In some cases, modifications interfere with RT during RNA-seq library preparation, which allows mapping of modification positions49–51. AQRNA-seq detects RT defects as mutations or read pile-ups at sites along the RNA sequence, as illustrated by small RNA isolated from log-growing E. coli47. As shown in Fig. 5a, several tRNAs had substantial polymerase stops at positions 38 and 48. By overlaying the stop positions on tRNA modification maps47, these two positions were found proximal to known modification sites. One subset of nine tRNAs had 31–83% of mapped reads stop at position 48, which abuts the modification acp3U at position 47 (orange boxes, right, Fig. 5a). With base-pair-blocking m3U previously reported to block RT3, it makes sense that acp3U would also prevent polymerase processes. tRNAs with NNA anticodons had 48–69% of reads end at position 38, which is adjacent to i6A and its hypermodified derivatives (for example, ms2,i). In the anticodon loop (yellow boxes, Fig. 5a, left). While ms2,i induces pausing4, it is plausible that the bulkier i3A 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 between one and two 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 tRNA species were truncated at positions 59–60 (heatmap of RT stops in Fig. 5b). After AlkB treatment, most stops at positions 59–60 had 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 substrate13,17,49, m3A, 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
As reviewed by Motorin and Helm12, this kind of modification mapping (Fig. 5d) for position 34 of tRNA Arg-ACG represents I. Telomerase by expression of a catalytic subunit (hTERT) immortal-engineered tumor-promoting genotypes (Fig. 6a): reactivation of HMEC cell lines represent progressive tumorigenesis conferred by resulting in a more varied distribution of alignment start positions. The heat map shows a significant reduction in the number of aligned reads. Early all the reads begin after position 58, forming a ‘cliff’. After AlkB demethylation, however, the read alignments lengthen and extend past the cliff, (Arg-ACG) in mycobacteria47 tends to pair with C during RT 48,50, 984ision changes in HMEC cells is shown in Supplementary Fig. 8. Quantitative profiling of miRNA dynamics during tumorigen -sion changes in HMEC cells is shown in Supplementary Fig. 8. Fig. 5) points to the potential for quantitative mapping of RNA modifications by AQRNA-seq. Modification mapping is also illustrated by mutations resulting from RT11. For example, wobble inosine (I) in a single tRNA (Arg-ACG) in mycobacteria47 tends to pair with C during RT10,12, 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 Helm22, 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 tumorigenesis51,52. 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 (HRASG12V, ER) and expression of SV40 small T antigen further drive partial transformation and aberrant growth in HMEC2 cells; and additional P53 suppression by short-hairpin RNA knockdown yields HMEC 3 cells fully capable of tumor growth in mice51,52. Based on quantitative PCR with RT, validation of key gene expression changes in HMEC cells is shown in Supplementary Fig. 8.

Before application of AQRNA-seq for quantification of HMEC miRNAs, we modified the data-mining workflow for the complexity of the human genome with numerous repeats and highly similar RNA species, such as tRNA isodecoders, which poses a challenge for uniquely mapping reads. We modified the data-mining 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 nonredundant miRNA sequences for all three cell lines, ranging from 1 to 100,000 normalized read counts (Fig. 6b). Those miRNAs changing in abundance most significantly during tumorigenic transition from HEMC1 to HEMC3 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 Fig. 6d showing that three (15a-5p, 19a-3p, 4,454) significantly increased in the transition from HEMC1, four significantly decreased (24-3p, 4,488, 21-5p, 27a-3p) and seven were unchanged during tumorigenesis. These results are consistent in some cases with literature observations. For example, 15a-5p and 19a-3p were upregulated (>1.8-fold) in patients with triple negative breast cancer13 and 4,454 was upregulated in more aggressive breast cancer types with HER-2 overexpression13,14 as well as in inflammatory breast cancer16. Similarly, miR-27a was downregulated in breast cancer stem cells, with overexpression reducing both the number...
and size of mammospheres and sensitization of breast cancer cells to anticancer drugs by downregulation of genes essential for reactive oxygen species detoxification. 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 in the Validation section above, all RNA-seq methods introduce adventitious sequence variants...
(Supplementary Fig. 2a,b) during library preparation and sequenc-
(ing. However, since adventitious sequence variants should be iden-
tically 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 nonredundant miRNA
sequences present in both the HMEC and miRXplore samples. The
pool of 875 parent miRNAs were filtered to include those with at
least two variants that exceeded ten reads per variant, to ensure that
variants were not the product of sequencing error. Analysis of end
sequences among the subset of 24 miRNAs meeting these condi-
tions revealed a predominance of U additions to both 3’ and 5’ ends
(Supplementary Fig. 8b). To discover variant sequences unlikely to
have been caused by library preparation (that is, biologically rele-
vant isomiRs), we subtracted the number of copies of each of addi-
tional variant in the miRXplore panel from those in HMEC cells for
the 24 miRNAs, selecting those for which HMEC-miRXplore was
>0.1. Graphs depicting the HMEC-miRXplore differential among
the 15 additional variants are shown in Supplementary Fig. 8c,
which reveals a predominance of single 3’ additions expected for
ture true isomiRs15,57.

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 quanti-
tative accuracy and fidelity of next-generation sequencing (NGS)
RNA-seq methods, including biochemical idiosyncrasies of RNA
ligases, RTs and other enzymes and the secondary structures of
adapterand substrates16,58. 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 nonessential 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
adapters to minimize ligation bias and molecular crowding agents
(for example, polyethylene glycol (PEG)) to increase enzyme effi-
ciency, ligation of the 5’ adapter occurs after RT to ensure that
all cDNAs, including truncated species, are captured in the final
library16,59. A nonessential AlkB demethylation step minimizes
premature cDNA truncation and informs about the locations of
RT–blocking or -mutating modifications57. Beyond the linear rela-
tionship between read count and RNA copy number, the RNA-seq
method provides information about modification occupancy, sec-
ondary 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 adapter ends and 20% PEG
to enhance fidelity56. However, the need for miRNA size selec-
tion and ligation of both 3’ and 5’ adapters before 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, Bard-Dascot et al. compared
the ILM, NEB and PEB miRNA kits studied here and two additional
nonligation kits (Clontech SMARTer, Diagenode CATS)16 by
using six miRNA standards, finding that kits using PEG, random-end
adapters and overnight ligation minimized both sequence biases
and interferences caused by 3’-(2′-O-methylation) on certain miR-
NAs60. Wong et al. assessed methods for extraction of miRNA from
plasma, but compared kit performance using only detected diver-
sity as the metric61. Finally, Heinicke et al. compared the TRI, QIA
and LEX kits tested here and the SMARTer and CATS kits tested
by Bard-Dascot et al. using 41 miRNA standards and small RNA
purified from human T cells62. None of these comparisons assessed
quantitative accuracy with the rigor applied by Herbert et al.27 and
in the present study. While the best-performing commercial kits
are to quantify only <50% of miRNAs with twofold accuracy57, we
were able to quantify ~75% of miRNAs within twofold 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 inform about many RNA modifi-
cations critical to RNA function (Figs. 3 and 5).

So how does AQRNA-seq compare to other methods? As
detailed in Supplementary Table 1, there are >40 RNA-seq methods
available for different types of RNA and various purposes. However,
few of these methods (1) optimize ligation and amplification effi-
ciencies, (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 liga-
tion efficiency or evaluated for quantitative bias. The simultane-
ously ligation of 3’ and 5’ linkers before RT introduces significant
loss of sequence information due to polymerase fall-off. Another
example involves the template-switching polymerase methods that
use TGIRT for cDNA production and CircLigase for subsequent
cDNA circularization during library preparation15,59. 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 strand17 while
the efficiency of CircLigase I and II is strongly influenced by the
sequence of cDNA and RNA15,59,60. These ligation biases preclude an
unbiased, quantitative analysis of all RNAs in a sample. A final com-
parison involves the Hydro-tRNA-seq method16 entailing hydro-
lytic fragmentation of tRNA followed by traditional simultaneous
3’ and 5’ linker ligation, RT, PCR and NGS steps. Here, ligation effi-
ciency was not optimized, the RT step loses all sequence information
for polymerase fall-off fragments and hydrolysis of the tRNAs
prevents identification of natural tRNA fragments. Importantly,
Hydro-tRNA-seq was not designed to be quantitative but rather
was intended for tRNA gene annotation for mature and pre-tRNA
sequences15.

AQRNA-seq shares a downside with other RNA-seq technolo-
gies: the introduction of sequence variants during library prepara-
tion and sequencing (Supplementary Fig. 2a). This raises concerns
about the biological relevance of many miRNA isoforms (that is,
isomiRs) noted in the literature16 and online databases18. Given
extensive evidence for the formation of isomiRs by enzymatic
uridylation15,60, 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 rele-
vant isomiRs. It is not clear why the introduction of sequence vari-
ants correlates strongly with the accuracy for quantification of RNA
sequences (Fig. 2a,c,d,f), with AQRNA-seq being the most quanti-
tatively accurate method—at least for miRNAs.

While AQRNA-seq was applied first to tRNA and miRNA anal-
sis, it should be broadly applicable to any form of RNA. Random
priming RNA-seq methods provide relatively accurate quantifica-
tion of mRNAs and long RNAs15, but they do not inform about RNA
fragmentation or modifications. AQRNA-seq is applicable to lon-
ger RNA (for example, mRNA and ribosomal RNA) 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 a length appropriate for library generation for quantifica-
tion 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 content
Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41587-021-00874-y.

Received: 21 September 2019; Accepted: 25 February 2021; Published online: 15 April 2021.

References
1. Wang, Z., Gerstein, M. & Snyder, M. RNA-Seq: a revolutionary tool for transcriptomics. Nat. Rev. Genet. 10, 57–63 (2009).
2. Cech, T. R. & Steitz, J. A. The noncoding RNA revolution-trashing old rules to forge new ones. Cell 157, 77–94 (2014).
3. Hafner, M. et al. RNA-ligase-dependent biases in miRNA representation in deep-sequenced small RNA cDNA libraries. RNA 17, 1697–1712 (2011).
4. Zhang, Z., Lee, J. E., Riemondy, K., Anderson, E. M. & Yi, R. High-efficiency RNA cloning enables accurate quantification of miRNA expression by deep sequencing. Genome Biol. 14, R109 (2013).
5. Fuchs, R. T., Sun, Z., Zhuang, F. & Robb, G. B. Bias in ligation-based small RNA sequencing library construction is determined by adaptor and RNA structure. PLoS ONE 10, e0126049 (2015).
6. Alon, S. et al. Barcoding bias in high-throughput multiplex sequencing of miRNA. Genome Res. 21, 1506–1511 (2011).
7. Zhuang, F., Fuchs, R. T., Sun, Z., Zheng, Y. & Robb, G. B. Structural bias in T4 RNA ligase-mediated 3′-adapter ligation. Nucleic Acids Res. 40, e54 (2012).
8. Pang, Y. L., Levi, R., Levine, S. S. & Dedon, P. C. Diverse cell stresses induce unique patterns of mRNA up- and down-regulation: mRNA-seq for quantifying changes in mRNA copy number. Nucleic Acids Res. 42, e170 (2014).
9. Linsen, S. E. et al. Limitations and possibilities of small RNA digital gene expression profiling. Nat. Methods 6, 474–476 (2009).
10. Machnicka, M. A., Olichowski, A., Grosjean, H. & Bujnicki, J. M. Distribution and frequencies of post-transcriptional modifications in tRNAs. RNA Biol. 11, 1619–1629 (2014).
11. Björg, G. R. et al. Transfer RNA modification. Annu. Rev. Biochem. 56, 263–287 (1987).
12. Motorin, Y. & Helm, M. Methods for RNA modification mapping using deep sequencing: established and new emerging technologies. Genes (Basel) 10, 35 (2019).
13. Motorin, Y., Muller, S., Behm-Ansmant, I. & Branlant, C. Identification of modified residues in RNAs by reverse transcription-based methods. Methods Enzymol. 425, 21–53 (2007).
14. Shigematsu, M. et al. YAMAT-seq: an efficient method for high-throughput sequencing of mature transfer RNAs. Nucleic Acids Res. 45, e70 (2017).
15. Gogakos, T. et al. Characterizing expression and processing of precursor and mature human tRNAs by Hydro-tRNAseq and PAR-CLIP. Cell Rep. 20, 1463–1475 (2017).
16. Kim, H. et al. Bias-minimized quantification of microRNA reveals widespread alternative processing and 3′ end modification. Nucleic Acids Res. 47, 2630–2640 (2019).
17. Dai, Q., Zheng, G., Schwartz, M. H., Clark, W. C. & 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).
18. Zheng, G. et al. Efficient and quantitative high-throughput tRNA sequencing. Nat. Methods 12, 835–837 (2015).
19. Cozen, A. E. et al. ARM-seq: AlkB-facilitated RNA methylation sequencing reveals a complex landscape of modified tRNA fragments. Nat. Methods 12, 879–884 (2015).
20. Xu, H., Yao, J., Wu, D. C. & Lombowitz, A. M. Improved TIGR-seq methods for comprehensive transcriptome profiling with decreased adapter dimer formation and bias correction. Sci. Rep. 9, 7953 (2019).
21. Mohr, S. et al. Thermostable group II intron reverse transcriptase fusion proteins and their use in cDNA synthesis and next-generation RNA sequencing. RNA 19, 958–972 (2013).
22. Jayaprakash, A. D., Jabado, O., Brown, B. D. & Sachidanandam, R. Identification and remediation of biases in the activity of RNA ligases in small-RNA deep sequencing. Nucleic Acids Res. 39, e141 (2011).
55. Wang, X. et al. Differential expression profile analysis of miRNAs with HER-2 overexpression and intervention in breast cancer cells. *Int. J. Clin. Exp. Pathol.* **10**, 5039–5062 (2017).

56. Maltseva, D. V. et al. miRNome of inflammatory breast cancer. *BMC Res. Notes* **7**, 871 (2014).

57. 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).

58. Pirouz, M., Ebrahimi, A. G. & Gregory, R. I. Unraveling 3’-end RNA uridylation at nucleotide resolution. *Methods* **155**, 10–19 (2019).

59. Sorefan, K. et al. Reducing ligation bias of small RNAs in libraries for next generation sequencing. *Silence* **3**, 4 (2012).

60. Dard-Dascot, C. et al. Systematic comparison of small RNA library preparation protocols for next-generation sequencing. *BMC Genomics* **19**, 118 (2018).

61. Wong, R. K. Y., MacMahon, M., Woodside, J. V. & Simpson, D. A. A comparison of RNA extraction and sequencing protocols for detection of small RNAs in plasma. *BMC Genomics* **20**, 446 (2019).

62. Heinricke, F. et al. Systematic assessment of commercially available low-input miRNA library preparation kits. *RNA Biol.* **17**, 75–86 (2020).

63. Chu, Y. et al. Intramolecular circularization increases efficiency of RNA sequencing and enables CLIP-Seq of nuclear RNA from human cells. *Nucleic Acids Res.* **43**, e75 (2015).

64. Lucigen CircLigase II Product Manual, p. 3. https://www.lucigen.com/docs/manuals/MA298E-CircLigase-II-ssDNA-Ligase.pdf (July 2019).

65. Xiong, Y. et al. A comparison of mRNA sequencing with random primed and 3’-directed libraries. *Sci. Rep.* **7**, 14626 (2017).

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

© The Author(s), under exclusive licence to Springer Nature America, Inc. 2021
Methods

Bacterial strains, culturing conditions, growth assays and RNA isolation. All E. coli strains were from the Keio collection [47]. The genotype of each strain was validated before conducting studies. Strains were cultured in 10 ml of lysogeny broth (LB) at 37 °C with constant shaking at 180 rpm. All the cultures reached a final optical density (OD600) of 0.6–0.7. Culture pellets were harvested by centrifugation at 4,000×g for 2 min and immediately used for RNA isolation with the PureLink miRNA isolation kit (Thermo Fisher) following the manufacturer’s protocol. Briefly, cell pellets were resuspended in Trizol reagent (Thermo Fisher) for lysis, followed by treatment with chloroform to separate the aqueous layer containing bulk RNA. The aqueous layer was then subjected to a two-column purification process where genomic DNA, larger RNA fragments (>200 base pairs (bp)) and excess salts were removed. It is important to note that ssRNA cannot be separated from tRNA using this method. Three biological replicates were used in the E. coli RNA study.

Mycobacterium bovis BCG strain Pasteur 1173P2 was grown in roller bottles with 7H9 broth or PBS (with 0.05% v/v tyloxapol, Sigma-Aldrich) at 2 r.p.m. and 37 °C. Exponentially growing cultures with an OD600 of 0.8–1.0 were starved by washing pellicles three times with PBS-tyloxapol. Starvation cultures were inoculated into PBS-tyloxapol at a starting OD600 of 1.0. Samples were retrieved at 4, 10 and 20 days after starvation. At day 20, cultures were resuspended in 7H9 and resuscitated for 6 days before harvesting. At each time point, cultures were plated on 7H10 agar for determination of colony-forming units (CFU). Specific compositions of 7H9 medium, 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) and 0.05% (v/v) glycerol and 0.05% (v/v) Tween 80 (Sigma-Aldrich) as nutrient-replete medium; PBS (137 mM NaCl, 2.7 mM KCl, 10 mM Na2HPO4, 2 mM KH2PO4) was supplemented with 0.5% (v/v) tyloxapol, a nonhydrolysable detergent. 7H10 agar (BD Difco) was supplemented with 0.5% (v/v) glycerol and 10% (v/v) oleic acid albumin dextrose catalase (OADC, BD BHH). For CFU assays, serial dilutions of BCG cultures at various nutrient starvation/resuscitation times and respective CFUs were profiled using the Agilent Bioanalyzer.

RNA isolation and purification from BCG, cells collected at nutrient starvation/resuscitation time points were lysed in the presence of TRI reagent (Sigma-Aldrich) with glass beads in a FastPrep FP120 bead-beater as previously described [48]. Three biological replicates were used in the BCG RNA isolation protocol (Thermo Fisher) for lysis, followed by treatment with chloroform to separate the RNA from the aqueous layer containing bulk RNA. The aqueous layer was then subjected to a two-column purification process where genomic DNA, larger RNA fragments (>200 base pairs (bp)) and excess salts were removed. It is important to note that 5S rRNA cannot be separated from tRNA using this method. Three biological replicates were used in the BCG RNA isolation protocol. The sample was eluted in 20 μl of TRI reagent buffer (137 mM NaCl, 2.7 mM KCl, 0.2% (w/v) glucose, 0.085% (w/v) NaCl, 0.2% v/v glucose, 0.05% v/v sodium acetate, 0.02% v/v sodium deoxycholate, 0.002% w/v sodium azide), 2 μl of 20× RTStar tRF&tiRNA Pretreatment Kit. A 2 μl reaction was then prepared by adding the following: 0.2 μl of 20× RTStar tRF&tiRNA Pretreatment Kit, 2 μl of 5′-deadenylate (NEB, 100 μM –1). The reaction was incubated at ambient temperature. After 2 h, the reaction was stopped by the addition of 1 μl of 20× RTStar tRF&tiRNA Pretreatment Kit (5′-deadenylate (NEB, 100 μM –1)). The mixture was mixed by inversion several times and centrifuged at 16,000×g for 10 min. The top layer was transferred to a new Eppendorf tube. A further 100 μl of 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, no. D4060) according to the manufacturer’s instructions. The sample was eluted in 17 μl of water before proceeding to the next step (linker 1 removal).

Removal of excess DNA linker 1. In this step, the DNA adenylated oligonucleotide adenylate immediately was deadenylated and subsequently digested together with unused linker 1, by exonuclease RecJ. Deadenylation was performed in a 20-μl reaction containing 16 μl of RNA sample from the demethylation step, 2 μl of NEB Buffer 2 (10×) and 2 μl of 5′-deadenylase (NEB, 50 μl –1). After incubation at 30 °C for 1 h, 2 μl of RecJ (NEB, 30 μl –1) was added. The mixture was incubated at 37 °C for 30 min. The reaction was stopped by the addition of a further 2 μl of RecJ digestion for an additional 30 min. The reaction mixture was mixed by heat inactivation 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 was mixed with 1 μl of RT primer (2 pmol μl–1) and 1 μl of deoxynucleotidetriphosphate (10 mM each) and heated at 80 °C for 2 min, followed by cooling on ice. PrimerScSenseTm Buffer (6 μl; Clontech), 1 μl of RNase Inhibitor (NEB) and 1 μl of 5× PrimeScript Reverse Transcriptase buffer (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 the addition of 1 μl of NaOH (5 M) followed by incubation at 90 °C for 3 min. The hydrolysis product was neutralized by the addition of 1 μl of 3 M HCl (Clontech) and the reaction mixture was purified using the Zymo Oligo Clean & Concentrator kit (Zymo Research). The sample was eluted with 15 μl of water before vacuum concentration to 5 μl.

cDNA ligation. Purified cDNA was ligated to linker 2 (Supplementary Table 2) in a 20-μl reaction consisting of 5 μl of cDNA sample, 1 μl of RecJ (50 μl –1), 2 μl of T4 DNA Ligase Buffer (NEB), 1 μl of ATP (10 mM, NEB), 2 μl of 5′-deadenylate (NEB, 50 μl –1) and 9 μl of 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 deadenylated and RecJ was used to digest the deadenylated product. Deadenylation was performed in a 20-μl reaction containing 16 μl of cDNA sample from the cDNA ligation step, 2 μl of NEB Buffer 2 (10×) and 2 μl of 5′-deadenylase (NEB, 50 μl –1). After incubation at 30 °C for 1 h, 2 μl of RecJ (30 μl –1) was added. The reaction was incubated at 37 °C for 30 min. Subsequently, for digestion a further 2 μl of RecJ was added 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 reaction containing 24 μl of cDNA template, 50 μl of 50× KAPA DNA polymerase buffer (2x buffer, Clontech), 2 μl of each PCR primers F and R with unique sequencing barcodes (1 μl each; Supplementary Table 2), 2 μl of seqAMP DNA polymerase (Clontech) and 20 μl of 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 optimized.
purified from an agarose gel using a standard gel purification kit (QiAgel Quick Extraction Kit, Qiagen). Gel-extracted samples were mixed (multiplexing) and submitted for Illumina sequencing. In the studies described, sequencing was performed on the Illumina NEXTSeq sequencing machine (BioMicroCenter, MIT) with custom primers F and R (Supplementary Table 2).

Optimization of AlkB demethylation conditions. Liquid chromatography–tandem mass spectrometry 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⁻¹ 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, 330 °C; flow, 10 l/min; nebulizer, 45 psi; and capillary voltage, 3.5 kV. The first and third quadrupoles (Q1 and Q3, respectively) were fixed to unit resolution and the modifications were quantified by predetermined molecular transitions. The dwell time for each ribonucleoside was 500 ms. The retention time, mass spectrometry analyses.

Custom primers F and R (Supplementary Table 2). The sequences and counts in each sample were further merged and tabulated from linker 1 (Fig. 1a), which were subsequently stripped using fastx_trimmer (fastxtoolkit/0.0.13; http://hannonlab.cshl.edu/fastx_toolkit/). The abundances of unique sequences were calculated in each trimmed sequencing file with fgrep: numberReadsPerFile = $\ltf{fgrep SmiRNA_sequence Strimmed_sequence_file \ wc -l}$. The read counts of 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 give the expected read count, assuming that 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 pair 0 to 80% to (1) stitch read1 and read2 together, (2) cross-validate the sequences of read1 and read2 to eliminate sequencing errors and (3) strip links 1 and 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 the 3′ end. These sequences were then blastn against a reference library using blast (v.2.6.0) with the following parameters: blast -perc_ident 80 -reward 2 -penalty 2 -m 10 -n 300 -evalue 1e-10. Alternatively, readers can use miRNA panels validated in tissue and cell extracts using PCR-based kits (for exact methods and data, see discovery-and-translational-research/pcr-qpcr-dpcr/qpcr-assays-and-instruments/miRNA-qpcr-assy-and-panels/mircury-inna-mirna-mirnome-pcr-panels/clear=true/orderinginformation/).

miRNA mixture. The miRXPlore Universal Reference (Millenyi Biotec) consists of 963 synthetic unmodified, HPLC-purified RNA oligonucleotides. The sample was reconstituted according to the manufacturer’s instructions and aliquoted. Three aliquots were used as separate RNA pool stocks for AQRNA-seq experiments.

Note: The Millenyi 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 (for exact methods and data, see discovery-and-translational-research/pcr-qpcr-dpcr/qpcr-assays-and-instruments/miRNA-qpcr-assy-and-panels/mircury-inna-mirna-mirnome-pcr-panels/clear=true/orderinginformation/).
IncrNA-specific analyses were performed by blasting nonredundant sequences against 107,039 GRCh38/hg38 human IncrNA reference in LNCipedia (https://hg38.cipedia.org). Inserts with ≥20 nt of perfect match to IncrNA reference sequences with an e-value <0.1 were considered as IncrNA sequences for further analyses. Genome-wide analyses were performed by mapping nonredundant sequences against the GRCh38 reference genome using the Burrows–Wheeler 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 gene regions, reads in intergenic regions and reads not mapped to human genome.

**BGC starvation proteomics: isobaric labeling and peptide fractionation.** As the first step in the quantitative analysis of the subtracted BCG proteome, proteins were extracted 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 and S20, respectively), and then resuspended in 7H9 medium for 6 days (S-R6). The extracted proteins were then precipitated, quantified and resuspended in two washes of 8 M urea (S4, S10 and S20, respectively), and Cytoskeleton-digest protein (from 50 µg of total protein) were labeled with TMT 6-plex reagents (Thermo Scientific) according to the manufacturer’s instructions. Aliquots (5 µl) of the labeled peptide mixture were removed from each biological replicate and combined equivolumetrically to reconstitute a single batch, which was analyzed on a Thermo Scientific Easy-LC1200 instrument (Thermo Fisher). The chromatographic runs were performed at pH 3.0–10.0 over 24 wells on an Agilent 3100 OFFGEL fractionator (no. OG24PE00) according to the manufacturer’s protocol. Each of the 24 fractions was collected, dried by vacuum centrifugation, resuspended in 0.1% formic acid in water and analyzed by nano-liquid chromatography–tandem mass spectrometry (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 quadrupole time-of-flight (Q–TOF) LC–MS. The LC system consisted of a capillary pump for sample loading, a nanoflow pump and a thermostatted microcapillary–plate autosampler. The nano-LC pump configuration included a 160-nm emitters column and a 150 mm × 75 µm analytical column (Zorbax 300SB-C18, no. G4340–2601). The following MS-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 100-min linear gradient LC separation was performed at a flow rate of 300 µl/min. The analytical gradient of solvent B was performed at a constant flow rate of 0.3 ml/min 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%; 121–125 min, 75–98%; 125–126 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 °C, drying gas 313 °C, fragmenter 360 V. Capillary voltage was manually adjusted to 1,800–2,150 V to achieve a steady nanospray. Data were acquired from 300–1,700 m/z with an acquisition rate of six spectra s–1 in MS mode, and at 50–1,700 m/z with an acquisition rate of three spectra s–1 in MS/MS mode. A peptide isotope model (charge state 2+) was used to detect a maximum of 20 precursors per cycle at a minimum threshold of 25,000 counts per spectrum at a narrow isolation window (~1.3 m/z). Sloped collision energy (CE) was used to maximize collision-induced dissociation of detected isobarically tagged peptides according to the following rules: charge state 2+, CE slope 4.2, offset 3.5; charge state 3+, CE slope 4.2, offset 4.2. LC–MS data were extracted and evaluated for quality using the minimum free energy (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 the second and third biological replicates. The conditions for maximizing the number of detected proteins 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 was injected in technical duplicate—spectra generated from technical replicate 2 were acquired without use of an exclusion list, whereas those generated from technical replicate 2 were acquired with the exclusion list. Data from MassHunter Qualitative Analysis were 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. Manual analysis of data prefiltered at the 95% confidence interval (minimum of two peptides per protein ID) yielded 1,217 highly quantifiable proteins for the starvation proteomics experiment. Similarly, 965 highly quantifiable proteins were identified at all time points of our published BCG hypoxia isobaric tags for relative and absolute quantitation proteomics studies. The hypoxia proteomics data are available from the CHORUS MS data repository at https://chorusproject.org/, project no. 1107.

**RNA blot analysis of bacterial rRNAs.** Small RNAs were purified from starvation cultures of BCG as described in “Bacterial strains, culturing conditions, growth assays and RNA isolation.” RNA from each time point (~225 ng) was resolved on 10% NovexTBE urea gels (Fisher) along with a New England Biolabs (NEB) low-range, single-stranded RNA ladder (50, 80, 150, 300, 500, 1,000 nt) and NEB microRNA markers (17, 21, 25 nt). The gels were then stained with SYBB Gold before electrotransfer to BrightStar Plus positively charged nylon membranes (Thermo Fisher; Supplementary Fig. 5). Following completion of electrotransfer, membranes were ultraviolet cross-linked and then hybridized with [32P]-labeled oligonucleotide probes specific for the 5′ ends of one of three rRNAs (Asp, Ase and Asn) or 16S rRNA (BCG hypoxia starvation proteomics experiment). Similarly, 965 highly quantifiable proteins were identified at all time points of our published BCG hypoxia isobaric tags for relative and absolute quantitation proteomics studies. The hypoxia proteomics data are available from the CHORUS MS data repository at https://chorusproject.org/, project no. 1107.

**Data availability**

All sequencing and proteomics data that support the findings of this study have been deposited in public databases: RNA-seq studies reported in Figs. 3–5, Supplementary Figs. 2, 4 and 8 and the proteomics studies reported in Fig. 4 and Supplementary Fig. 6 have been deposited in the NCB1 Sequence Read Archive under BioProject ID PRJNA579224; miRNA and standards data shown in Fig. 2 have been deposited in the Gene Expression Omnibus (GEO) under accession no. GSE139936; and data for miRNA studies in HMEC cells shown in Fig. 6 have been deposited in GEO as accession no. GSE159434.

**References**

66. Baba, T. et al. Construction of Escherichia coli K-12 in-frame, single-gene knockout mutants: the Keio collection. Mol. Syst. Biol. 2, 2006.0008 (2006). 67. Hia, F. et 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 Res. 43, e32 (2015).

68. Li, H. et al. The sequence alignment/map format and SAMtools. Bioinformatics 25, 2078–2079 (2009).
Acknowledgements
We thank members of D. Bartel’s laboratory (Whitehead Institute and MIT Department of Biology), especially A. Stefano and D. Bartel, for assistance with RNA blots. We thank P. Ivanov (Harvard Medical School) for sharing synthetic RNA standards. This work was supported by grants from the National Natural Science Foundation of China (no. 32070629 to B.C.), the US National Science Foundation (no. MCB-1412357 to V.C.-L.), the National Institute of Environmental Health Sciences (no. 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 MIT Toxicology Training grant no. T32-ES007820, D.Y. by a postdoctoral fellowship from A*STAR, Singapore and S.M.H. by a postdoctoral fellowship from the Swiss National Science Foundation.

Author contributions
P.C.D., B.C. and 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 RNA 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.

Additional information
Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41587-021-00874-y.
Correspondence and requests for materials should be addressed to P.C.D. or B.C.

Peer review information Nature Biotechnology thanks James Hadfield and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Reprints and permissions information is available at www.nature.com/reprints.
Corresponding author(s): Peter Dedon, Cao Bo
Last updated by author(s): 21 February 2021

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a Confirmed

☐ The exact sample size (n) for each experimental group/condition, given as a discrete number and unit of measurement
☐ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐ The statistical test(s) used AND whether they are one- or two-sided
☐ Only common tests should be described solely by name; describe more complex techniques in the Methods section.
☐ A description of all covariates tested
☐ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐ For null hypothesis testing, the test statistic (e.g. F, t, r) with confidence intervals, effect sizes, degrees of freedom and P value noted Give P values as exact values whenever possible.
☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐ Estimates of effect sizes (e.g. Cohen’s d, Pearson’s r), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection  Sequencing data were collected on an Illumina NextSeq500 system. LC-MS/MS data were collected on an Agilent 6430 triple quadrupole system using Agilent MassHunter software. Bioanalyzer data were collected on a Agilent 2100 Bioanalyzer system with 2100 Expert operating software. Proteomics data were collected on an Orbitrap Q Exactive Mass Spectrometer system.

Data analysis  All custom scripts have been made available at https://github.com/dedonlab/. 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=Blas
tDocs&DOC_TYPE=Download. Peakfit.m version 9.0 available at Tom O’Haver, MATLAB Central File Exchange - https://terpconnect.umd.edu/toh/spectrum/./grep (Linux command) available at https://unix.stackexchange.com/questions/17949/what-is-the-difference-between-grep-egrep-and-grep. fastx_toolkit 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/aqtnaseq for prokaryotic process scripts and https://github.com/dedonlab/general_aqtnaseq for eukaryotic/general pipeline scripts].

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.
Data

Policy information about availability of data. All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

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 2-6, S2, S3, S4, S7, and S8 and the proteomics studies reported in Figures 4 and S6 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 mRNA studies in HMEC cells shown in Figure 6 have been deposited in GEO as accession number GSE159434. All other relevant data are available from the corresponding authors upon request.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

- [x] Life sciences
- [ ] Behavioural & social sciences
- [ ] Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size

A sample size of 50 nanograms of RNA was chosen for AQRNA-seq for three reasons. First, this amount of RNA was found to be sufficient for monitoring the reaction efficiency of each step by Bioanalyzer during AQRNA-seq library preparation. Second, 50 nanograms of RNA showed the highest ligation and linker removal efficiencies during sequencing library preparation. Third, 50 nanograms of small RNA (i.e., ~1-3 pmol) was sufficient for identifying all lRNA isoacceptors in Mycobacterium bovis BCG and for detecting nearly all of the miRNAs in the standard mixture used in the paper.

Data exclusions

No data were excluded from the analyses presented in this paper.

Replication

All studies were carried out with at least three replicates. This replicate number was selected to balance feasibility of experimental scale and cost with robust levels reproducibility.

Randomization

All the samples were allocated randomly, including assignment of sequencing barcode to individual sample, sample pooling, sample injection onto sequencing chip, data collection and analysis.

Blinding

The investigators were blinded to group allocation during data collection and analysis.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

| n/a | Involved in the study |
|-----|----------------------|
| [x] | Antibodies           |
| [x] | Eukaryotic cell lines|
| [x] | Palaeontology and archaeology |
| [x] | Animals and other organisms |
| [x] | Human research participants |
| [x] | Clinical data        |
| [x] | Dual use research of concern |

Methods

| n/a | Involved in the study |
|-----|----------------------|
| [x] | ChIP-seq              |
| [x] | Flow cytometry        |
| [x] | MRI-based neuroimaging|

Eukaryotic cell lines

Policy information about cell lines. Cell line source(s) HMEC 1, HMEC(hTert), HMEC 2, HMEC(hTert- HRas[V12]-ER-EV); and HMEC 3, HMEC(hTert- HRas[V12]:ER-shp53) were developed in the laboratory of Prof. Jit Kong Cheong at the National University of Singapore Department of Biochemistry
| Authentication                                                                 | Expression of oncogenic HRAS and PS3 in the HMEC-derived lines was quantified by PCR and found at expected levels. |
|-----------------------------------------------------------------------------|----------------------------------------------------------------------------------------------------------------|
| Mycoplasma contamination                                                    | Mycoplasma tests were negative                                                                                  |
| Commonly misidentified lines                                               | “HMEC” is a commonly used abbreviation for human mammary epithelial cells.                                      |
| (See IGCLC register)                                                        |                                                                                                                |