Efficient and quantitative high-throughput transfer RNA sequencing

Guanqun Zheng[#1], **Yidan Qin**[#2], **Wesley C. Clark**[1], **Qing Dai**[3], **Chengqi Yi**[1,6], **Chuan He**[1,3,4,5], **Alan M. Lambowitz**[2,8], and **Tao Pan**[1,4,8]

[1] Department of Biochemistry & Molecular Biology, the University of Chicago, Chicago, IL 60637, USA.

[2] Institute for Cellular and Molecular Biology, the University of Texas at Austin, Austin, TX 78712, USA.

[3] Department of Chemistry, the University of Chicago, Chicago, IL 60637, USA.

[4] Institute of Biophysical Dynamics, the University of Chicago, Chicago, IL 60637, USA.

[5] Howard Hughes Medical Institute, the University of Chicago, Chicago, Illinois 60637, USA.

[#] These authors contributed equally to this work.

Abstract

Despite its biological importance, transfer RNA (tRNA) could not be adequately sequenced by standard methods due to abundant post-transcriptional modifications and stable structure, which interfere with cDNA synthesis. We achieve efficient and quantitative tRNA sequencing using engineered demethylases to remove base methylations and a highly processive thermostable group II intron reverse transcriptase to overcome these obstacles (DM-TGIRT-seq). Our method should be applicable to investigations of tRNA in all organisms.

High-throughput RNA sequencing (RNA-seq) has revolutionized our understanding of gene expression. Widely used RNA-seq methods start with adapter ligation and cDNA synthesis of biological RNA samples followed by PCR amplification to generate sequencing libraries[1]. These standard methods work well for most cellular RNAs, such as mRNA, long

ncRNA, miRNA, or fragments derived from rRNA, snRNA, and snoRNA. tRNA is the only class of cellular RNA for which the standard sequencing methods cannot yet be applied efficiently and quantitatively, although attempts have been made (e.g. ref 2). Significant obstacles for the sequencing of tRNA include the presence of numerous post-transcriptional modifications and its stable and extensive secondary structure, which interfere with cDNA synthesis and adapter ligation. tRNAs are essential for cells and their synthesis is under stringent cellular control. Accumulating evidence shows that tRNA expression and mutations are associated with various diseases such as neurological pathologies and cancer development[3,4]. The lack of efficient and quantitative tRNA-seq methods has hindered biological studies of tRNA.

We applied two strategies to eliminate or substantially reduce the obstacles of tRNA modification and structure for efficient and quantitative tRNA sequencing (**Fig. 1a**). A DNA and RNA repair protein AlkB-derived enzyme mixture was first used to remove methylations at the Watson-Crick face. Three specific modifications, N^1-methyladenosine (m^1A), N^3-methylcytosine (m^3C), and N^1-methylguanosine (m^1G), are abundant in eukaryotic tRNAs and are particularly problematic for reverse transcriptases (RTs) by causing cDNA synthesis to stop or misincorporate. In mammals, known modifications of these types include m^1A present in all tRNAs at position 58, m^3C present in five tRNAs at position 32 and the variable loop and m^1G present in about half of all tRNAs at positions 37 or 9. We applied two recombinant enzymes as a mixture to remove these three methylations in human tRNAs. The first was the wild-type enzyme AlkB (wtAlkB) from *E. coli*, which is known to very efficiently demethylate m^1A and m^3C in single-stranded nucleic acids as its DNA and RNA repair function[5,6]. Wild-type AlkB, however, works very poorly on m^1G modification[7]. Based on its known three-dimensional structure complexed with nucleic acids, we engineered AlkB to generate a specific mutant, D135S that efficiently converted m^1G to G (**Supplementary Figs. 1a,b**). We optimized the reaction conditions and enzyme to RNA ratios using a mixture of the wild-type and the D135S AlkB proteins (**Supplementary Figs. 1c,d**). We were able to remove >80% of total m^1A, m^3C, and ~70% m^1G without a noticeable change in tRNA quality (**Fig. 1b**). The remaining m^1A or m^1G may be more buried in the tRNA tertiary structure (m^1A or m^1G at position 9 of tRNAs) and thus not easily accessible to demethylase treatment without causing tRNA degradation.

We used a thermostable group II intron reverse transcriptase (TGIRT) with high processivity to generate cDNA from highly structured tRNA. The TGIRT reaction does not require adapter ligation; it synthesizes cDNA by template-switching from the adaptor to the 3' end of the target RNA[8-10]. Importantly, demethylase treatment markedly reduced the amount of RT stops at the m^1A58 and m^1G37 positions. At the same time, the amount of longer and full-length cDNAs was substantially increased after demethylase treatment of tRNA (**Fig. 1c**). Similar results regarding demethylase treatment were obtained when the RNA template was either total cellular RNA or gel-purified total tRNA (**Fig. 1c**). These results show that the demethylase treatment is effective in producing longer reads; this property will be crucial for the ability to adequately map mammalian tRNA expression where many tRNAs have just one or a handful of differences in their sequences[11,12].

We performed Illumina sequencing using the libraries generated from tRNAs with and without the demethylase treatment. To facilitate our focus on tRNA at this time, the sequencing reads were mapped only to the genomic tRNA database, which contains 515 predicted tRNA genes distributed over 330 unique sequences and 110 predicted tRNA pseudogenes[12]. In all, 6.6–15.7 million and 2.3–4.7 million reads were mapped to the genomic tRNA database when using gel-purified tRNA or total RNA as template, respectively (**Supplementary Fig. 2a**). These variations in read numbers were derived from sample handling, as similar proportions of reads were obtained for the added internal tRNA standards (**Supplementary Fig. 2b**). For the biological replicate from HEK293T cells, both untreated and treated samples showed similarly high reproducibility with r^2-values from 0.985–0.991 (**Supplementary Figs. 3a-d**). As already indicated by the cDNA bands (**Fig. 1c**), the read length and in particular the proportion of reads corresponding to full-length tRNA is increased substantially upon demethylase treatment, as shown by the position plots along individual tRNA genes (**Figs. 2a,b**). The highest fraction of reads in a LeuAAG tRNA in the untreated samples were RT stops due to the m^1G37 modification, whereas all stops at this position were removed with the corresponding increase of full-length tRNA reads upon demethylase treatment (**Fig. 2a**). Similarly, the substantial RT stop at m^1G9 of a GlnCTG tRNA was markedly reduced upon demethylase treatment, together with a substantially increased generation of full-length reads (**Fig. 2b**). We still detected a strong stop at the m^2_2G26 (N^2,N^2-dimethylguanosine) residue in LeuAAG tRNA that remained unchanged upon demethylase treatment. This result indicates that our demethylase mixture is not effective in removing m^2_2G modifications, which are present in ~20% of tRNAs. Nevertheless, our results indicate that the demethylase treatment is very effective in producing longer reads; this property is crucial for the ability to adequately map the mammalian tRNAome at single base resolution.

We performed additional analysis to further demonstrate the usefulness of our sequencing method. Plotting each tRNA isoacceptor against their genomic gene copy number showed a poor correlation (**Fig. 2c**), consistent with the known tissue specific tRNA expression in humans[13,14]. We compared the fraction read of the tRNA[Arg] isoacceptors to the fluorescence hybridization signals of the Arg-tRNA probes from tRNA microarrays which were obtained through hybridization without the need for cDNA synthesis[13,15]. The sequencing and array results showed the same trend of isoacceptor abundance, thus validating the quantitative nature of tRNA abundance obtained independently through sequencing and hybridization based approaches (**Fig. 2d**). We also compared RT stops and misincorporations at known modification positions with and without demethylase treatment. In the case of m^1A58 in ValCAC and m^3C32 in ThrAGT, the demethylases completely removed these modifications as demonstrated by the reversion to cognate sequence at these positions (**Figs. 2e-f**). In the case of m^1G37 in ProTGG and m^1G9 in GlnCTG, the demethylase treatment removed a large majority of the modification so the mismatch and stops were substantially reduced (**Figs. 2e-f**). Therefore, our DM-TGIRT-seq method can determine differences in the modification dynamics of m^1A, m^1G and m^3C at single base resolution, as well as potentially infer positions of non-demethylated modifications.

We also examined the expression of unique tRNA genes from chromosome 6. Human chromosome 6 contains approximately one third of all tRNA genes, and over 150 genes are clustered within a 2.7 Mbp region near the class I major histocompatibility (MHC) genes[16]. We found that the tRNA expression levels within the tRNA gene cluster were higher compared to the tRNA genes outside of the cluster (**Supplementary Fig. 4**). The expression level of tRNA genes in the cluster was uneven, suggesting that the expression of tRNA genes was not coordinated throughout the entire cluster in HEK293T cells.

In summary, despite its biological importance, tRNA has not been adequately studied at the transcriptome level due to its unique characteristics in modification and structure. Our approach described here makes efficient and quantitative tRNA-seq feasible, and has an additional advantage of being able to study modifications in a high throughput manner. Our method is accessible to any lab and offers broad potentials in basic research and diagnostic applications.

METHODS

Design of the D135S AlkB mutant

The substrates for the wild-type AlkB (wtAlkB), m^1A and m^3C, are positively charged. A close examination of the structures of substrate-bound AlkB[17] revealed that these positively charged substrates are favorably positioned in the active site by interacting with the negatively charged carboxylic group of Asp135 (D135) (**Supplementary Fig. 1a**). We reasoned that single mutants at the D135 position might allow better accommodation of m^1G in the active site (**Supplementary Fig. 1a**), thereby enhancing the demethylation efficiency of m^1G. Truncated but catalytically active AlkB- N11 was used and mutants with single amino acid replacement at D135 were screened. Of several mutations tested, AlkB D135S gave the best demethylation yields towards m^1G (**Supplementary Fig. 1b**). In the AlkB-m^1G structure model constructed by computationally mutating the AlkB structure (**Supplementary Fig. 1a**), the shorter side chain of S135 seems to allow more room to accommodate m^1G while sustaining the crucial hydrogen bond with m^1G, which may explain the improvement in activity.

We also performed pH-activity profiling for the demethylation of m^1A and m^1G in tRNA. Decreasing demethylation activity was observed with increasing pH for both m^1A and m^1G substrates (**Supplementary Fig. 1c**). Therefore, we chose pH 5.0 for detailed kinetic analysis and subsequent experiments.

Cloning, expression and purification of AlkB and mutant

A truncated AlkB with deletion of the amino (N)-terminal 11 amino acids was cloned into a pET30a vector (Novagen) and overexpressed in *E. coli* BL21(DE3)[18]. The proteins were purified following procedures published previously[19]. Briefly, cells were grown at 37 °C in the presence of 50 μM kanamycin until the OD_{600} reached 0.6–0.8. After the addition of IPTG (1 mM) and $FeSO_4$ (5 μM), the cells were incubated for an additional 4 hr at 30 °C. Cells were collected, pelleted and then resuspended in lysis buffer (10 mM Tris pH 7.4, 300 mM NaCl, 5% glycerol, 2 mM $CaCl_2$, 10 mM $MgCl_2$, 10 mM 2-mercaptoethanol). The cells

were lysed by sonication and then centrifuged at 17,418 rcf for 20 min. The soluble proteins were first purified using a Ni-NTA superflow cartridge (Qiagen), and then further purified by ion-exchange (Mono S GL, GE Healthcare) and gel-filtration (Superdex-200, Pharmacia) chromatography. All protein purification steps were performed at 4 °C. The Asp135 to Ser mutation was introduced using the QuikChange Site-Directed Mutagenesis Kit (Agilent). The mutant protein was expressed and purified following the same procedure as that of the wild-type protein.

Mammalian cell culture and RNA preparations

Human embryonic kidney cell line HEK293T (CRL-11268) were obtained from American Type Culture Collection (ATCC) and used without further validation. Cells were cultured in DMEM (Thermo) media supplemented with 10% FBS and 1% $100 \times$ Pen Strep (Gibco). Cells were routinely checked for mycoplasma contamination every 3–6 months, using Universal Mycoplasma Detection Kit (ATCC).

Total RNA was isolated by using a mirVana miRNA Isolation Kit (Life Technologies). Purified total RNA was premixed with the T7 RNA polymerase transcripts of three tRNA standards[13] (0.01 pmol each standard per μg of total RNA) and deacylated by incubating in 0.1 M Tris-HCl, pH 9 at 37°C for 45 min. Although not necessary for studies of mature tRNAs, which all end with 3'CCA, deacylated RNAs with or without demethylation treatment could be treated with T4 polynucleotide kinase (Epicentre) at 37C° for 30 min to further warrant a free 3' hydroxyl group for template-switching.

When necessary, total tRNA was subsequently isolated using a denaturing 10% polyacrylamide gel followed by passive gel elution and ethanol precipitation.

Demethylation reactions

Demethylation activity assay was performed for either gel purified total tRNA or total cellular RNA. For total tRNA, 100 μl of reaction mixture containing 1 μg of tRNA (~40 pmol) was treated with $2\times$ molar ratio of wtAlkB (80 pmol) and $4\times$ molar ratio of D135S mutant (160 pmol). For total cellular RNA, 5 μg of total RNA (estimated to contain ~40 pmol tRNA) was treated with $4\times$ molar ratio of wtAlkB (160 pmol) and $4\times$ molar ratio of D135S (200 pmol). More demethylases were used for total RNA to alleviate potential interference by rRNA and mRNA in the reaction. The reaction buffer contained 300 mM KCl, 2 mM $MgCl_2$, 50 μM of $(NH_4)_2Fe(SO_4)_2 \cdot 6H_2O$, 300 μM 2-ketoglutarate (2-KG), 2 mM L-ascorbic acid, 50 μg/mL BSA, 50 mM MES buffer (pH 5.0). In both cases, the reaction was incubated for 2 hr at room temperature, and quenched by the addition of 5 mM EDTA. After phenol–chloroform extraction, tRNA was recovered by ethanol precipitation.

Thermostable group II intron RT template-switching

Template-switching reactions were performed as described[8-10]. Briefly, we used an initial template-primer substrate consisting of a 41-nt RNA oligonucleotide (5'-AGA UCG GAA GAG CAC ACG UCU AGU UCU ACA GUC CGA CGA UC/3SpC3/-3') that contains Illumina Read1 and Read2 primer-binding sites and a 3' blocking group (three carbon spacer; Integrated DNA technologies, inc.) annealed to a complementary ^{32}P-labeled DNA

primer with a single-nucleotide 3' overhang, T, which facilitates the template switch to full-length tRNAs that mostly contain a 3' CCA end. For TGIRT template-switching reactions, typically 100 ng of demethylated tRNAs or 1 μg of demethylated total RNA were mixed with the initial template-primer substrate (100 nM) and 500 nM TGIRT (GsI-IIC MalE rigid fusion RT[10]) in reaction medium containing 450 mM NaCl, 5 mM MgCl$_2$, 20 mM Tris-HCl, pH 7.5, and 5 mM DTT. The reactions were pre-incubated at room temperature for 30 min, initiated by adding 25 mM dNTPs (an equimolar mix of 25 mM dATP, dCTP, dGTP and dTTP) to a final concentration of 1 mM and incubating at 60°C for 30 min. The reactions were terminated by adding 5 M NaOH to a final concentration of 0.25 M, incubating at 95°C for 3 min, and neutralizing with 5 M HCl. The cDNAs resulting from template-switching were analyzed in a denaturing 6% polyacrylamide gel, electroeluted using a D-tube Dialyzer Maxi with MWCO of 6–8 kDa (EMD Millipore), and ethanol precipitated with 0.3 M sodium acetate, pH 5.2, in the presence of 25 μg of linear acrylamide (Life Technologies) carrier. The purified cDNAs were then circularized with CircLigase II (Epicentre) using the manufacturer's protocol with an extended incubation time of 5 hr at 60°C, extracted with phenol-chloroform-isoamyl alcohol (25:24:1), ethanol precipitated and amplified with Phusion-HF (Thermo Scientific) using Illumina multiplex (5'- AAT GAT ACG GCG ACC ACC GAG ATC TAC ACG TTC AGA GTT CTA CAG TCC GAC GAT C -3') and barcode (5'- CAA GCA GAA GAC GGC ATA CGA GAT BARCODE GTG ACT GGA GTT CAG ACG TGT GCT CTT CCG ATC T -3') primers for 12 cycles of 98°C for 5 sec, 60°C for 10 sec and 72°C for 10 sec. The PCR products were sequenced on an Illumina HiSeq system.

Quantitative analysis of modification levels using LC-MS/MS

Quantitative analysis of modified nucleotides was done as previously described[20]. Briefly, 100 ng of tRNA was digested by nuclease P1 (2 U) in 30 μl of buffer containing 25 mM NaCl, and 2.5 mM ZnCl$_2$ at 37°C for 1 hr, followed by the addition of NH$_4$HCO$_3$ (100 mM) and alkaline phosphatase (0.5 U). After an additional incubation at 37°C for 1 hr, the solution was diluted to 60 μl, and 10 μl of the solution was injected for LCMS/MS. Nucleosides were separated by reverse phase ultra-performance liquid chromatography on a C18 column with online mass spectrometry detection using an Agilent 6410 QQQ triple-quadruple LC mass spectrometer in positive electrospray ionization mode. The nucleosides were quantified using the nucleoside to base ion mass transitions of 282 to 150 (m^{1}A), 268 to 166 (A), 298 to (m^{1}G), 284 to 152 (G), 258 to 126 (m^{3}C) and 244 to 112 (C). Quantification was performed by comparison with the standard curve obtained from pure nucleoside standards running in the same batch of samples. Modification levels were compared by the ratios of methylated base (m^{1}A, m^{1}G, m^{3}C) over regular base (A, G, C).

tRNA microarrays

The tRNA microarray assay consists of four steps starting from purified tRNA or total RNA without the need of cDNA synthesis: (i) deacylation to remove all 3' attached amino acids, (ii) selective fluorophore labeling of tRNA using oligonucleotide ligation with T4 DNA ligase to the 3'CCA of all tRNAs, (iii) hybridization, and (iv) data analysis. The reproducibility of the tRNA microarray method and validation of the results have been extensively described previously[13,15].

Sequencing read mapping

Sequencing reads were aligned using Bowtie to a modified hg19 genomic tRNA database[12]. A single mismatch was allowed in order to identify potential modification misincorporations at a modification site. Briefly, a tRNA library was adapted from the tRNAScan-SE library by appending CCA to tRNAs from the genomic tRNA database (http://gtrnadb.ucsc.edu/ Hsapi19/). Isodecoders with identical scores were consolidated for ease of identity assignment, decreasing the number of reference genes and pseudogenes from 625 to 462. Prior to mapping, reads were processed using Trimmomatic v0.32. Sequences greater than 15 bp were then aligned to the aforementioned culled tRNA library using Bowtie2 with sensitive options. Reads mapping to multiple isodecoders due to length of fragment were discarded. Modification fractions were determined by analyzing at a putative position 'n' the number of correct reads, the number of misincorporations, as well as the number of reads stopped at the n+1 position.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

ACKNOWLEDGEMENTS

REFERENCES

1. Wang Z, Gerstein M, Snyder M. Nat. Rev. Genet. 2009; 10:57–63. [PubMed: 19015660]

2. Pang YL, Abo R, Levine SS, Dedon PC. Nucleic Acids Res. 2014; 42:e170. [PubMed: 25348403]

3. Kirchner S, Ignatova Z. Nat. Rev. Genet. 2015; 16:98–112. [PubMed: 25534324]

4. Abbott JA, Francklyn CS, Robey-Bond SM. Front. Genet. 2014; 5:158. [PubMed: 24917879]

5. Trewick SC, Henshaw TF, Hausinger RP, Lindahl T, Sedgwick B. Nature. 2002; 419:174–8. [PubMed: 12226667]

6. Falnes PO, Johansen RF, Seeberg E. Nature. 2002; 419:178–82. [PubMed: 12226668]

7. Zheng G, Fu Y, He C. Chem. Rev. 2014; 114:4602–20. [PubMed: 24580634]

8. Katibah GE, et al. Proc. Natl. Acad. Sci. U. S. A. 2014; 111:12025–30. [PubMed: 25092312]

9. Shen PS, et al. Science. 2015; 347:75–8. [PubMed: 25554787]

10. Mohr S, et al. RNA. 2013; 19:958–970. [PubMed: 23697550]

11. Goodenbour JM, Pan T. Nucleic Acids Res. 2006; 34:6137–46. [PubMed: 17088292]

12. Chan PP, Lowe TM. Nucleic Acids Res. 2009; 37:D93–7. [PubMed: 18984615]

13. Dittmar KA, Goodenbour JM, Pan T. PLoS Genet. 2006; 2:2107–2115.

14. Gingold H, et al. Cell. 2014; 158:1281–92. [PubMed: 25215487]

15. Pavon-Eternod M, et al. Nucleic Acids Res. 2009; 37:7268–80. [PubMed: 19783824]

16. Horton R, et al. Nat. Rev. Genet. 2004; 5:889–899. [PubMed: 15573121]

17. Holland PJ, Hollis T. PLoS One. 2010; 5:e8680. [PubMed: 20084272]

18. Yu B, et al. Nature. 2006; 439:879–84. [PubMed: 16482161]

19. Mishina Y, Chen LX, He C. J. Am. Chem. Soc. 2004; 126:16930–6. [PubMed: 15612731]

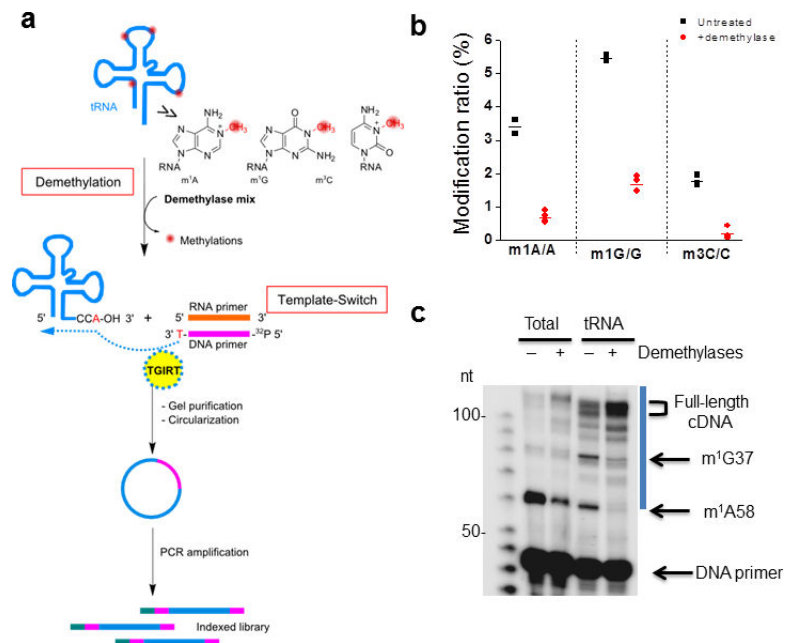20. Zheng G, et al. Mol. Cell. 2013; 49:18–29. [PubMed: 23177736]

Figure 1. Demethylase-thermostable group II intron RT tRNA sequencing (DM-TGIRT-seq)
(**a**) Our method differs from standard RNA-seq in two ways. First, a demethylase mixture was used to remove m^1A, m^1G and m^3C modifications located at the Watson-Crick face. Second, a thermostable group II intron RT (TGIRT) was used that is less sensitive to tRNA structure and adds RNA-seq adaptors by template-switching without RNA ligation. (**b**) Demethylation efficiency for total tRNA as measured by QQQ LC-MS. (**c**) RT reaction for both purified tRNA and total RNA as template with (+) or without (−) demethylase treatment. Blue line shows the gel region excised for library construction.
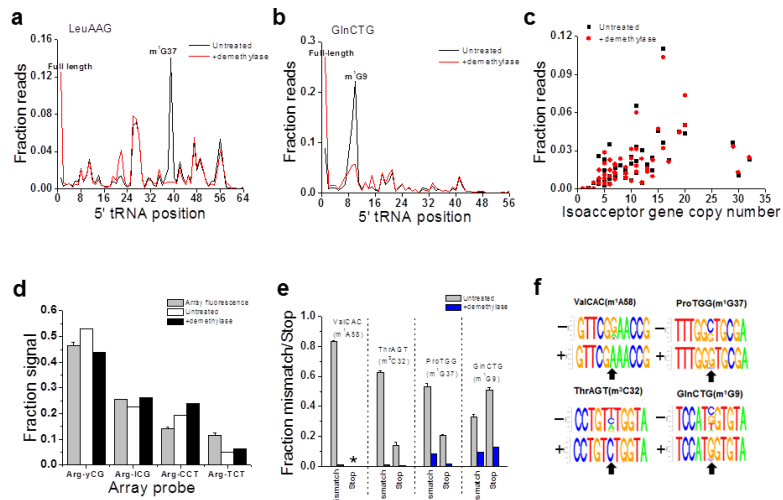
Figure 2. Sequencing results

(**a**) 5' position sequencing plot of a LeuAAG tRNA containing an m^1G37 modification. (**b**) 5' position sequencing plot of a GlnTTG tRNA containing an m^1G9 modification. For panels (a) and (b), m^1A58 stops correspond to very short DNA fragments that are outside the range of our sequenced cDNA, and hence cannot be visualized here. (**c**) As expected, tRNA isoacceptor expression in HEK293T cells does not correlate with the gene copy number, since human tRNA expression is tissue specific. (**d**) Comparison of the array fluorescent signals and sequencing reads for the Arg-tRNA family; error bar, n = 4 ± SD. (**e**) Mismatch and stops with and without demethylase treatment. Shown are modification positions of m^1A58 in ValCAC, m^3C32 in ThrAGT, m^1G37 in ProTGG, and m^1G9 in GlnCTG. Star indicates stops for ValCAC corresponding to very short reads that were not determined in this experiment. n = 4 ± SD. (**f**) Sequence logo at the same modified positions as in (e), centered at the modified residue. Untreated (−) is at the top and the demethylases treated (+) is at the bottom.