Article

Fingerprints of Modified RNA Bases from Deep Sequencing Profiles

Anna M Kietrys, Willem Arend A. Velema, and Eric T. Kool

J. Am. Chem. Soc., **Just Accepted Manuscript** • DOI: 10.1021/jacs.7b07914 • Publication Date (Web): 07 Nov 2017

Fingerprints of Modified RNA Bases from Deep Sequencing Profiles

Anna M. Kietrys, Willem A. Velema, and Eric T. Kool*

Department of Chemistry, Stanford University, Stanford, California 94305

ABSTRACT: Posttranscriptional modifications of RNA bases are found not only in many noncoding RNAs but also have recently been identified in coding (messenger) RNAs as well. They require complex and laborious methods to locate, and many still lack methods for localized detection. Here we test the ability of next-generation sequencing (NGS) to detect and distinguish between ten modified bases in synthetic RNAs. We compare ultra-deep sequencing patterns of modified bases, including miscoding, insertions and deletions (indels), and truncations, to unmodified bases in the same contexts. The data show widely varied responses to modification, ranging from no response, to high levels of mutations, insertions, deletions, and truncations. The patterns are distinct for several of the modifications, and suggest the future use of ultra-deep sequencing as a fingerprinting strategy for locating and identifying modifications in cellular RNAs.

Supporting Information Placeholder

INTRODUCTION

The nucleobases of RNA in the cell are often modified posttranscriptionally to alter their structural and functional properties.[1-3] More than 100 modifications are known and are highly prevalent in noncoding RNAs such as transfer RNAs (tRNA), ribosomal RNAs (rRNA) and small nuclear RNAs (snRNA). Recent studies have reported that some modified nucleotides are also present in long noncoding RNAs (lncRNA) and in messenger RNAs (mRNAs) as well.[4-10] It is becoming evident that at least some of these modifications are dynamic, changing over the lifespan of the RNA to alter their biological activities.[11] The field is rapidly moving as more modifications are found in messenger RNAs, and as the pathways that recognize, add, and remove modifications are identified.[12-14]

To clarify the roles of modified nucleotides in RNA biology, it is necessary to develop methods to locate and identify these modifications both in single specific RNAs as well as in the broader transcriptome. To date, most methods rely on selective chemical properties of modified bases, or on specific antibody recognition.[15-16] Methods vary widely from one modification to another, and can be quite complex. The majority of known modifications (see examples in Fig. 1) are still challenging to locate and identify, and many have not yet been identified, or searched for, in lncRNAs and mRNAs. Thus, new methods that can aid in detection of new species or more easily identify sites and types of modifications would be useful to the field.

Next-generation sequencing (NGS) is rapidly becoming a standard tool for biological and biomedical analysis. RNA sequencing (RNA-seq) is now commonly applied for exploring and characterizing expressed genes.[17-18] While sequencing of short (~50 nt) reads at a depth of 10-100× allows sufficient confidence in homology for alignment and assembly of gene sequences,[19] deeper sequencing that is possible with modern instruments can potentially allow the analysis of smaller differences such as heterozygous mutations or mixed splicing patterns,[20-21] and do so with much greater statistical confidence. Standard NGS relies on a polymerase to accurately copy the nucleic acid target. It has long been recognized that chemical modifications to nucleobases can strongly affect the ability of a polymerase enzyme to copy them; for example, chemical differences that alter shape and H-bonding properties of a nucleobase can affect the efficiency of reading as well as the complementary base that is inserted.[22-23]

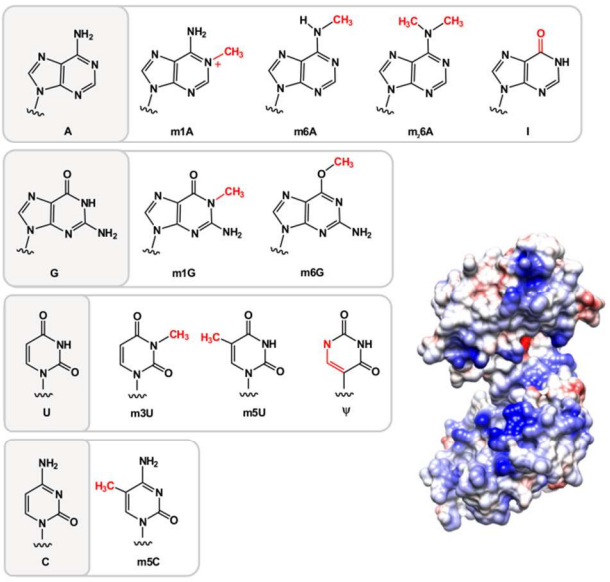

Figure 1. Structures of canonical (shaded) and modified RNA bases in this study (modifications highlighted in red). MMLV reverse transcriptase (PDB: 4mh8), from which Super Script IV was derived, is shown at right.

Here we considered the possibility that posttranscriptionally modified bases in RNA might measurably affect their reading by the reverse transcriptase (RT) enzyme used in next generation sequencing. Biologically relevant modifications, such as methylation of the heterocyclic base or its substituents, might well affect polymerase reading of that base, leading to a pause, an insertion or deletion, or miscoding near the position of modification.[24] For example, we recently described the ability of one reverse transcriptase enzyme to distinguish 6-

methyladenine (m6A) from adenine in RNA by its tendency to pause at the methylated base.[25] Other studies have also noted effects of RNA base modifications on polymerases. For example, hypoxanthine (a product of A-to-I editing) pairs with C, causing miscoding in RNA-seq as a G. Modification of inosine by acrylonitrile results in an RT stop.[26] Similarly, pseudouracil selectively modified with carbodiimide can generate an RT stop just prior to the modification.[7] In a third example, an antibody complex with m1A results in a block to reverse transcriptase.[10] In addition, a sequencing study of m1A noted blocks and mutations associated with this modification.[27] Thus, the early studies show clearly that RNA base modifications can alter polymerase behavior. However, there exists no broader study yet to study and compare multiple modifications, and to test whether such modifications can directly cause distinguishable patterns in deep sequencing data, perhaps even without further chemical modification.

If such altered sequencing patterns were found for modified bases in RNA, it could have two important implications. First, it suggests the use of deep sequencing as a general tool to directly identify new sites and types of base modifications in cellular RNAs. Second, in standard RNA-seq application, specific altered NGS patterns might well be miscalled as mutations in biological samples, when instead the altered patterns are caused by base modifications. Having baseline data for modifications could be useful in clarifying whether mutations are correctly called or are instead better explained by base modifications in the biological sample. In the latter case, this might lead to improved accuracy in clinical analysis of polymorphisms in patient specimens.[28-30]

Here we report experiments aimed at using ultra-deep sequencing information to identify and differentiate modified RNA bases. We used a pool of small synthetic RNA oligonucleotides carrying 10 epitranscriptomic modifications to identify their miscoding fingerprints. Statistical analysis of unique reverse transcriptase miscoding profiles of each modified base allowed us to distinguish them from each other, and from canonical bases as well. We show potential applications of our method in identification and quantification of several modified bases in RNA samples.

RESULTS

Our study is based on the hypothesis that alterations of molecular shapes and structures could result in differences in the deep sequencing data – patterns of mutations, deletions, insertions, and truncations – and thus act as a fingerprint of a given modification. For this initial study, we chose a range of ten base modifications that are known in noncoding RNAs (Fig. 1), [1-10] including methylation of adenine (m6A, $m_2$6A, m1A); methylation of guanine (m6G, m1G); methylation of uracil (m5U, m3U); methylation of cytosine (m5C); as well as hypoxanthine (the base of inosine, I) and pseudouracil (Ψ). Some of the modifications clearly are expected to interfere with Watson-Crick pairing and geometry during reverse transcription (*e.g.* m1A, m1G, $m_2$6A, m3U), while others are expected to be reverse transcribed normally (*e.g.* m5U, m5C, Ψ). To show the maximum signals in pure form, we incorporated single modified bases in synthetic 20 nt RNAs. As controls, we used RNAs with unmodified bases in the same contexts. Sequence contexts were chosen from known sites of these modifications in human RNAs (Table 1).

Table 1. RNA sequences used in RNA-seq library preparation, with biological origins of the sequence context in which that modification is known.

#	Sequence (5'->3')	Biological origin
1	GAG UUC CCC AGU CCU G**A**C UC	snRNA & mRNA
2	GAG UUC CCC AGU CCU G**m6A**C UC	
3	ACG ACG AUU G**U**A CGG CUC CG	mRNA
4	ACG ACG AUU G**Ψ**A CGG CUC CG	
5	AGG CCA UU**A** UCG CGC GAU CG	tRNA
6	AGG CCA UU**I** UCG CGC GAU CG	
7	GCA CGC CCA U**G**U GUA AUC GC	tRNA
8	GCA CGC CCA U**m1G**U GUA AUC GC	
9	CGC AUG U**G**C UUA GCG AUC CG	NA
10	CGC AUG U**m6G**C UUA GCG AUC CG	
11	GCC GUC UUG **A**AA CGC UAG GC	28S rRNA
12	GCC GUC UUG **m1A**AA CGC UAG GC	
13	CCG UAG GUG **A**AC CUC CGG AA	18S rRNA
14	CCG UAG GUG **$m_2$6A**AC CUC CGG AA	
15	AUC GGC UGG G**U**U UAG ACC GC	18S rRNA
16	AUC GGC UGG G**m3U**U UAG ACC GC	
17	GCC GUU AAG GUA GC**C** AAC GC	tRNA & rRNA
18	GCC GUU AAG GUA GC**m5C** AAC GC	
19	GCA AUG CUG G**U**U CGC CAU GC	28S rRNA
20	GCA AUG CUG G**m5U**U CGC CAU GC	

Next, we used these synthetic RNAs to prepare a library for ultra-deep sequencing on a common commercial platform (Illumina MiSeq), which can provide up to 15 million reads. We used a popular commercial kit for library preparation, including 20 indices for the twenty test and control RNAs, and we chose Super Script IV reverse transcriptase (Fig. 1) for conversion of RNAs to DNA amplicons. Polyacrylamide gels were used to check the performance of the reverse transcriptase, which after PCR revealed, not surprisingly, qualitative variations in reading success with different modifications (Fig. S3), indicating pausing and truncations for some of the modifications. All amplicons were combined for the sequencing, which was performed in a single run. Details of the preparation and analysis are given in the SI file.

The results of the sequencing showed strong variations in data patterns for the different modified RNAs. Total numbers of successful reads ranged from 8,158-568,504 reads for unmodified RNAs, while those with modified bases yielded a considerably narrower range of full-length read numbers (range 9,684-288,262). Based on the numbers of reads in each of the twenty data sets, we were able to use multinomial analysis with Gamma-Poisson distribution to generate statistical measures of confidence intervals in variations.[31] An overview of the data (Fig. S4) revealed that virtually all alterations in polymerase sequencing behavior in association with a modified base occurred within 2 nt of the modification, which is consistent with the structural footprint of a reverse transcriptase bound to an RNA-DNA template-primer.[32-33] For in-depth analysis, we focused on positions -2, -1, 0, +1, +2, where "0" is the site of modification (see Fig. 2). We quantified the following frequencies for each modification: A/C/T/G frequency at position 0; frequency and positions of truncations; frequency and position of single-nt deletion; frequency, position, and identity of single-nt insertion. Selected data are shown in Figs. 2, 3, with full data given in Figure S4.

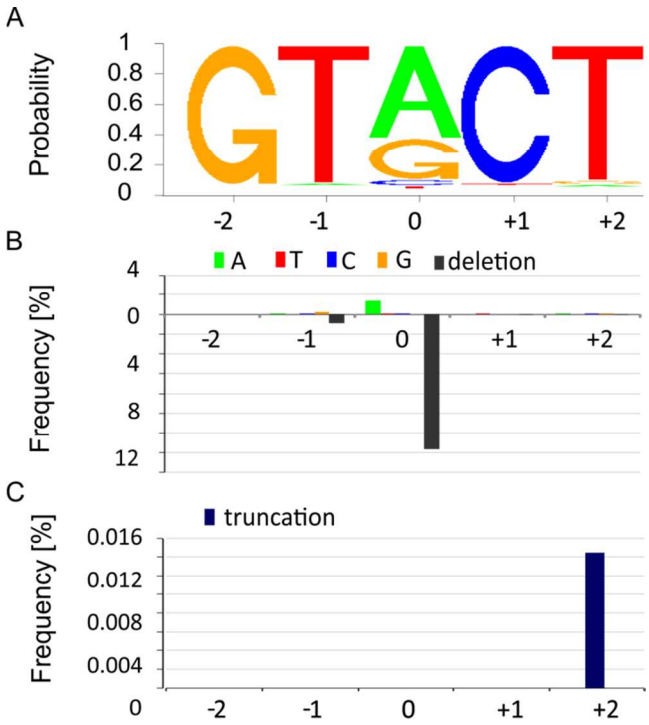

Figure 2. Fingerprint of m6G modification observed in RNA-seq data from a synthetic 20 nt RNA, analyzed from >42,000 reads. Modified position is numbered "0" (see Table 1 for full sequence). (a) Probability of occurrence of each nucleotide replacing m6G shows high miscoding frequency, coding as A>G>C>T. (b) Frequency of insertions of nucleotides (color coded), and single deletions (black) normalized per 10,000 reads. (c) Frequency of truncation (blue) normalized per 10,000 reads.

An overview of the results showed that some modifications yielded strong patterns that were distinct from the parent unmodified base, while a few (as expected) yielded little if any measurable difference (Figs. 3, S4). In the latter class were modifications that were statistically indistinguishable from unmodified bases, at least with this enzyme and protocol. These included m5C, which was not discernible from C (Fig. S4, compare miR9 and miR9m5C), and m5U, which yielded the same pattern as U (Fig. S4, compare miR10 and miR10m5U). Neither of these modifications yielded truncations or insertions or deletions (indels) above the normal background levels seen for the unmodified RNA, and they both coded for their complementary bases at the same fidelity, within error limits, as the unmodified bases. Similarly, pseudouridine (Ψ) showed no distinguishable differences in fidelity, indels or truncations relative to U (Fig. S4, compare miR2 and miR2Ψ). None of these are surprising, since 5-methylation of pyrimidines is known not to be detrimental to polymerase efficiency,[34] and Ψ is also documented to be replicated essentially the same as U.[35] The fourth case that showed little or no measurable difference with respect to modification was m6A, which coded as adenine with the same fidelity as adenine, and had the same frequency of indels and truncations within error limits (Fig. S4, compare miR1 and miR1m6A). An earlier survey of reverse transcriptases showed that most had little hindrance in replicating past m6A,[25] which can pair normally with T if the methyl group rotates away from the Watson-Crick face.[36]

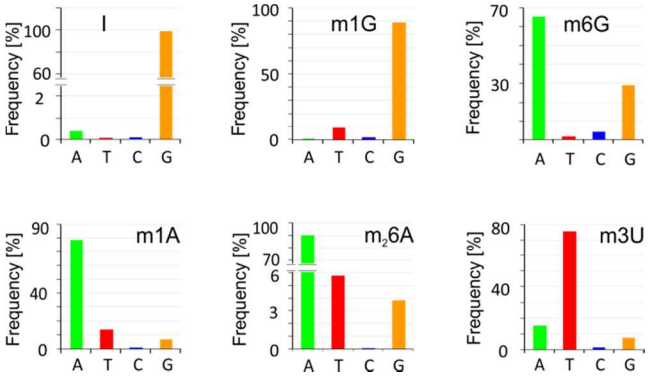

Figure 3. Varied mutation profiles of six modified bases in ultra-deep RNA sequencing. Shown are the coding mutational profiles at position 0, the position of modification. Note that some modifications yield low levels of miscoding (e.g. m1G (8.9% miscoding as T, 1.5% miscoding as C); while others give high degrees of miscoding (e.g. m6G (65.4% miscoding as A)).

In contrast, six of the modified bases studied here revealed clear differences in sequencing behavior relative to the unmodified congeners (Figs. 3, 4, S4, S7, S10). Among modified adenine derivatives, $m_2$6A, m1A and I (which is formed from deamination of A) all gave sequencing profiles quite distinct from adenine. The dimethylated base $m_2$6A gave frequent miscoding as U and G, indicating mispairing with A and C, respectively (Fig. 3). It also showed elevated levels of truncations at positions +1 and +2. This suggests that the polymerase may read it in Hoogsteen pairing mode, with the base flipped to *syn* conformation. The modification m1A also yielded miscoding as U and G, as previously seen by Hauenschild *et al*., [27] but at statistically different frequencies and ratios relative to $m_2$6A (Fig. S4, compare miR6 and miR6m1A). Hypoxanthine was easily distinguished from these, not surprisingly coding essentially as G,[37] with occurrence of deletion at 0 and truncations at -2 position (Fig. S4 miR3 and miR3I). Thus, the deep sequencing analysis can readily distinguish adenine from three modified forms of the base. Although inosine has been well studied by sequencing,[37-38] and one study has been reported for m1A,[27] we are unaware of prior direct deep sequencing data for $m_2$6A in RNA.

Notably, the sequencing data were also able to distinguish the two modifications of guanine studied, m1G and m6G. We found that m1G miscoded as U with a frequency 8.9%, and also yielded relatively frequent truncations at -2 and deletion at 0 position (Figs. 3, S4). The miscoding may be explained by Hoogsteen pairing of m1G with C.[39] In contrast, m6G often miscoded as A (65.4%) (Figs. 2, 3), but yielded deletions at 0 position and statistically few truncations. The known tendency of m6G to pair with U can explain the miscoding seen here.[40] The current data show that the m6G modification in standard low-to-medium depth NGS would simply be assigned as A, yielding an incorrect call. In contrast, our data, obtained at much greater depth, allows us to distinguish this modification from both A and G. Polymerase experiments have previously studied m6G in DNA, and show similar mutation frequencies as seen here in RNA.[41] To our knowledge, no NGS sequencing data have been published before on either of these modifications in RNA.

Lastly, m3U also gave a clear signature of its presence. It yielded frequent miscoding as A (indicating U-T mispairing by the polymerase) and G, and an elevated frequency of

deletion at position +2 (Figs. 3, S4). An early study of a DNA polymerase with this modification in rRNA showed stops prior to the modification but was not analyzed for mispairing.[42] We know of no previous polymerase studies of this modification in RNA, nor any studies of its effect in deep sequencing.

We proceeded to perform a statistical comparison of polymerase responses to evaluate the ability of NGS to distinguish all the modifications together. We used principal component analysis (PCA) to plot the multidimensional data for each modification, and we used numbers of reads to generate 95% confidence ellipsoids about each modification (Figs. S7, S10). Use of the full data set of mutations, truncations and indels over five positions near the modification separated not only most of these modifications from one another, but also separated most of the identical bases from one another in their different contexts (Fig. S7). For example, our input RNAs contained modified adenine in four different sequence contexts, and we observed that there were context-dependent truncations and indels that arise from context alone.[27] While this impressively showed the power of ultra-deep NGS to differentiate bases, it also adds considerable complexity to the analysis of modification, by reflecting the effects of varied context. Thus, we sought to simplify the analysis further by reducing the complexity of the input data.

Varying the amount of input data in our statistical analysis revealed that use only of mutational profiles at position 0 yielded nearly as good separation of modified bases from unmodified ones, but also erased the context-dependent differences of the unmodified bases (compare Figs. S7 and S10). Figure 4 shows a scatter plot of six modifications reduced to the two most dominant dimensional components. The plot shows clearly that six of the modified bases are distinguished from the four unmodified ones in the deep sequencing, with I overlapping with G as expected. The most readily distinguished modification is m6G, which lies far from all unmodified bases. Three others, m3U, m1G and I, also are easily distinguished from their unmodified parents. Other modifications that lie closer to their unmodified congeners are $m_2$6A, m1A and m6G, but all three can still be distinguished clearly at the depth obtained here (~10K or more reads). However, with smaller numbers of reads, as with standard NGS, they would very likely begin to overlap (due to lower confidence) and become indistinguishable. The other four modifications (m5U, m5C, Ψ, m6A) overlap extensively with the unmodified variants (Fig. S10) and thus cannot be distinguished, at least with this enzyme and at this depth of reads.

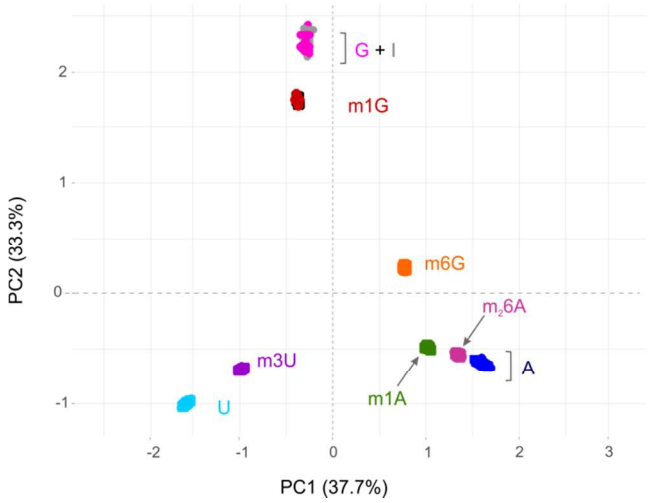

Figure 4. Principal component analysis of miscoding data for six RNA base modifications. Plot is reduced to two most dominant principal components for display. Each cluster represents eleven data points. Confidence ellipses are included for 95% confidence level; their sizes are smaller than the displayed data points.

Thus, our data show clearly that deep sequencing can distinguish several RNA base modifications in a single sequencing run. We also considered whether modifications of these six bases at levels below 100% occupancy would remain distinguishable, and at what point (with reduced fraction) they would overlap with unmodified bases and become undetectable. This could be analyzed by mixing in combinations of sequencing patterns of unmodified and modified bases in the same context. The result is to move the modification pattern progressively closer to that of the natural base (see examples in Fig. 5). Thus, of the six distinguishable modifications, m6G would be the most readily distinguishable at lower occupancy, since it lies furthest from unmodified G and A (Fig. 5). Indeed, it is clearly distinguished even at as low as 10% abundance (Fig. 5c). Conversely, the base $m_2$6A lies closest in the pattern to its unmodified variant, and thus would be less easy to distinguish if, say, only 10% of this position in a given RNA were modified (Fig. 5a). Lying in between is m3U, which (in the simulated mixing data (Fig. 5d) could be detected at ~15% or greater occupancy. Similarly, m1G could also be distinguished from G when it occurs at the level ~15% and higher (Fig. 5b). Overall, the results show that is should be possible to distinguish some of these modifications in RNA from biological samples even at less than 100% frequency, as long as sufficient numbers of reads are available to increase confidence and lower error margins.

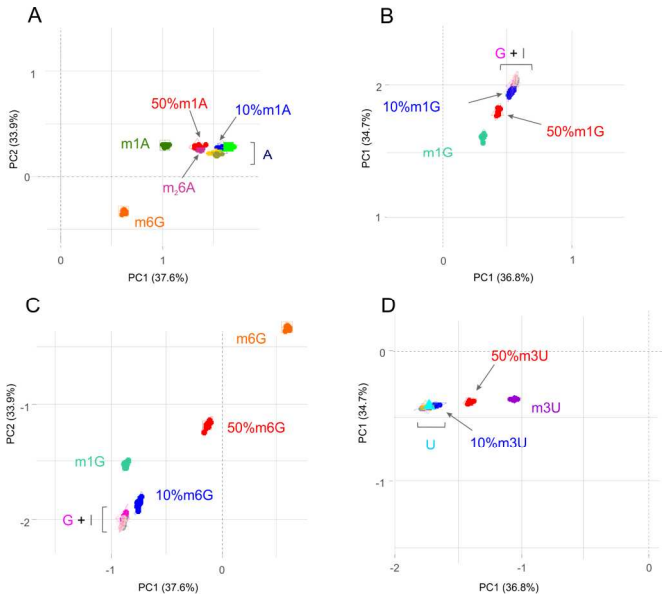

Figure 5. PCA plots showing the ability to distinguish modifications at less than 100% occupancy. Plot uses calculations for 4 variables (M0A, M0T, M0C, M0G); only pertinent local sections of the full PCA plot are shown. (A) Plotting m1A at 50% and 10% occupancy relative to unmodified A and m_26A, shows overlap with m26A and A, respectively; (B) Plotting m1G at 50% and 10% occupancy relative to unmodified G/I, shows the ability to distinguish this modification at above the 10% level; (C) Plotting m6G at 50% and 10% occupancy relative to unmodified G/I and m1G, showing complete resolution even at as low as 10% m1G; (D) Plotting m3U at 50% and 10% occupancy relative to unmodified U, showing discrimination at 50% occupancy but overlap at 10%.

Finally, we employed the data to test whether RNA base modifications might be misassigned as mutations. We used PCA plots of hypothetical mixtures of two unmodified bases (G+A and U+C), to investigate whether modifications might be confused with varied levels of mutations or deamination damage. In Figure 6a we show that as little as 10% of C deamination may be easily mapped in the PCA plot and distinguished from other bases and modifications. Similarly, varied amounts of A to G mutation can also be differentiated from modified residues (Fig. 6b).

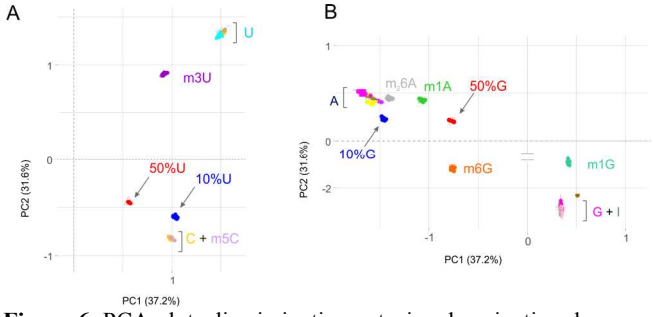

Figure 6. PCA plots discriminating cytosine deamination damage and partial mutations by deep sequencing. Plot uses calculations for 4 variables (M0A, M0T, M0C, M0G); only pertinent local sections of the full PCA plot are shown. (A) Cytosine deamination to U is plotted at 10% and 50% deamination, showing complete discrimination from both C and U. (B) Mixing of A-to-G mutation, showing 10% G and 50% G, with complete discrimination from pure A or G.

DISCUSSION

Our experiments show that it is possible to use high-depth RNA sequencing to directly determine the positions and identities of several different RNA base modifications, based on the varied response of reverse transcriptase to the altered structure and pairing of the modifications. Reverse transcriptase enzymes are known to introduce biases into sequencing experiments by their varied fidelity and processivity.[43-44] In the current strategy, we take advantage of reverse transcription step and use the enzyme's biases in nucleotide incorporation to provide information about noncanonical bases present in the RNA being copied (Fig. 2). We expect that the current data can be useful as a calibration set for employing a specific RT in identifying base modifications in biologically derived RNAs. In our analysis, we chose Super Script IV because of its high fidelity and processivity; it is likely that other reverse transcriptases may well show differential responses and biases to these same modifications (see Fig. S2). Indeed, we expect that use of a different RT with altered base response profiles used parallel to the current one would likely enhance discrimination of modifications. Because the reverse transcription step is the crucial source of the profiles seen here, reaction conditions for that step are important. It is possible that changes in the reaction conditions (e.g. time of reaction, ion concentrations, pH) may substantially change coding patterns.[45] For the current experiments, we chose standardized buffer and conditions optimized by the manufacturer. Using the current protocol, the enzyme enabled us to obtain readily distinguishable coding fingerprints for six distinct base modifications, some of which have not been readily detectable previously.

Amplification by PCR is known to introduce errors in sequencing,[46] and it is a potential source of error and variability in the current approach. For example, a replication error in a PCR template could be amplified alongside a correct template and misinterpreted as a partial mutation. While our data show that base modifications can be distinguished from partial mutations (Fig. 6), it is prudent to minimize this source of uncertainty. To this end, we employed a high-fidelity enzyme, and we limited PCR amplification to 12 cycles or fewer. We note that for small biological samples with limited quantity of cells, it is common to increase numbers of PCR cycles to compensate for a low quantity of RNA. We suggest caution in employing the current methods with large numbers of PCR cycles, and suggest the use of sufficient quantities of input RNA that high amplification can be avoided.

One important source of potential bias in RNA sequencing is differential responses of reverse transcriptase in different sequence contexts. Indeed, our data clearly detect polymerase responses that vary with sequence context (Fig. S7). In particular, we note varied tendencies toward deletions and truncations in different contexts regardless of whether base modifications are present. For example, we have detailed truncation, indel, and miscoding data for canonical adenine in four different contexts, and find frequency of deletions as high as 10% of reads in some contexts, with varied position. Our observations are consistent with previous reports of context-dependent RT responses.[27,47] In the current study, the sequence context of a base changes the frequencies of indels and truncations, but these phenomena are observed both for modified and unmodified bases. Given the high depth and

statistical discriminating power of the current experiments, we were able to discriminate even canonical bases from one another since they arise in different contexts. While this discrimination power is impressive, it could confound the utility in distinguishing modified bases, since every new context could yield a different data fingerprint. Analysis of all possible contexts would be costly in time and resources in actual practice. Thus, we narrowed the data analysis, eliminating variables to find the minimal variables that as much as possible remove context-dependent differences, and focus only on the base modification. We found that limiting the analysis to the miscoding frequencies at the position of modification enables reliable discrimination of modifications, and largely eliminates the influence of sequence context (Fig. 4). In future applications of the method to unknown RNAs, one can in principle simply analyze coding at each position with sufficient sequencing depth to discriminate small miscoding biases that can be assigned to specific base modifications. Spiking in RNAs with known modifications (such as the current ones) may well be helpful for internal calibration in applications with biologically derived specimens.

Our current methods employed ten different common RNA modifications, including all those currently known in messenger RNAs. Since over 100 different modifications are known in cellular RNAs,[4] it is possible that some of the current modification-dependent patterns may overlap with those from other, currently untested, base modifications. Future experiments with more modifications will shed light on this, and offer the prospect of expanding the range of distinguishable modifications. However, in the meantime it seems prudent to limit the current approach to cellular RNAs (such as mRNAs) that contain a smaller diversity of modifications, rather than to those (such as tRNAs) that contain highly diverse modifications. Another possible source of uncertainty when analyzing cellular RNAs by the current approach is incomplete occupancy of the modification at a specific RNA position. Because our data contained unmodified bases as well as modified ones in the same contexts, we are able to calculate the effects of lowered occupancy (see Fig. 5). In most cases, the base modifications were distinguishable at levels as low as 10-15% occupancy. We expect that yet higher levels of sensitivity could be achieved with further increases in depth of sequencing to provide more precision in the data.

Another source of non-homogeneous base coding in biologically derived RNAs is single nucleotide variations that arise from heterozygous alleles and from inhomogeneity in tissues. In the current four-dimensional data, such mixed RNAs fall on an axis between the two unmodified bases (Fig. 6), and are clearly distinguishable at a level at least as small as 10%. Importantly, our data show that such a base mixture does not overlap with the pattern of the modifications tested. For example, we find that partial replacement C by U does not confound the detection of m3U (Fig. 6a), and mixing of canonical A with G does not interfere with detection of m6G, m1G, m6A, or $m_2$6A (Fig. 6b). This is possible because the RT fingerprint of the modified residues characterizes the miscoding profile of all four nucleotides, while partial mutations yield substantial differences in only two nucleotides. This fact should enable the differentiation not only of modified bases in biological samples, but also heterozygous alleles and variations from tissue heterogeneity. Conversely, our data suggest that the assignment of single nucleotide variations in standard-depth RNA seq is likely to be confounded by base modifications, providing a caution for researchers in the field.

Our approach makes use of great sequencing depth to provide the precision to differentiate between canonical and modified bases. While simple sequencing of the transcriptome is possible using only 20× depth of coverage of sequence,[48-49] such a low depth would miss nearly all of the current modifications, and would assign them as the unmodified congener. For some modifications, such a low sequencing coverage would lead to base miscalls. For example, hypoxanthine (I) would be called as G rather than A, and would be missed unless a comparison to the chromosomal DNA sequence was made.[50] Similarly, our data show that m6G could well be miscalled as A at 20× coverage. Higher depths (typically on the order of ca. 100×) are sometimes used in RNA-seq for identification of splicing events.[49,51] Our experiments suggest that of the modifications tested, most would still not be distinguishable at this depth. One exception is m6G, since its miscoding rate is high, but it would likely only be distinguishable at high levels of occupancy. Our data suggest that a depth of 1000× begins to yield sufficient discriminating power to begin to reliably differentiate some of these modifications, and we recommend employing a baseline level of at least 5000×. Fortunately, new ultra-high-throughput instruments can enable such levels of coverage and beyond, given access to sufficient input RNA. According to our data, at depths higher than 8000× we were able to not only identify and differentiate several base modifications, but also distinguish them at relatively low levels of occupancy. It is clear that changes in depth of sequencing will critically influence statistical significance and resolution of the obtained data.

CONCLUSIONS

In summary, our data show that multiple modifications of RNA bases are readily detectable by ultra-deep sequencing patterns in a single sequencing run. This strategy greatly expands the ability to locate and identify several modifications for which current methods are nonexistent or highly laborious. Future experiments are planned to test this approach with biologically derived RNAs. It will also be of interest in the future to test other modifications for their deep sequencing patterns: since over 100 modifications are known, a broader set of data patterns could well be useful in future searches for previously unknown sites of modified bases. In addition, use of varied reverse transcriptase enzymes may well give differences in response patterns, thus possibly providing yet broader discriminating power. New sequencing instruments that are designed to yield great depth of sequencing will also facilitate such studies in the future.

Significantly, our experiments also suggest the possibility that existing biological and clinical sequencing data might be susceptible to error in sequence and mutation calling, due to patterns that were caused not by mutations, but rather by base modification. For example, we have shown that a m6G or m1A posttranscriptional modification could easily be misread as a mutation (A and T respectively). We conclude that researchers and clinicians in the future who make use of low-depth NGS data from RNA should proceed with caution, since base modifications can clearly alter outcomes substantially.

EXPERIMENTAL METHODS

Synthesis of modified RNAs. RNA (Table 1) and DNA (Table S2) oligonucleotides were obtained from the Stanford Protein and Nucleic Acid Facility (PAN), Integrated DNA Technologies (IDT) or GE Dharmacon. Oligoribonucleotide 'miRm6G' was synthesized in the laboratory as described in detail in the SI file. RNA sequences were purified by polyacrylamide gel electrophoresis (PAGE) in denaturing 20% gels and analyzed by MALDI-TOF.

NGS library preparation and sequencing. In the first step, the 3' end adapter was ligated to 200 ng of all RNA sequences. The reaction mixture was incubated for 1 hour at 28˚C, and immediately loaded and separated in a denaturing 15% polyacrylamide gel, next to a 10bp DNA Ladder. Bands of ligated RNA/DNA were cut out and eluted for 2 h at room temperature in 350 µL of Elution buffer (0.3 M NaOAc, pH 5.2, glycogen 0.18 mg/ml). The mixture was decanted, and precipitated overnight. Ligated RNA/DNA was reverse transcribed using Super Script IV and. cDNA was purified in a denaturing 12% polyacrylamide gel, next to a 10bp DNA Ladder. Bands of cDNA were cut out and eluted from gel and precipitated overnight. Next, cDNA was circularized by CircLigase™II ssDNA Ligase. The reaction mixture was incubated for 1 h at 60˚C, then 10 min at 80˚C. Mini Elute columns, and PB Buffer were used to clean up circular cDNA samples as described in the manufacturer's instructions. Eluted circular cDNA template was amplified by PCR (98 ˚C for 2 min; 8-12 times: 98 ˚C for 15 s, 60 ˚C for 30 s, 72 ˚C for 45 s; 72 ˚C for 5 min, 4 ˚C forever) using Phusion High-Fidelity PCR Master Mix according to the NEB's protocol. Reaction products were separated in a 10% polyacrylamide gel to investigate the optimal number of PCR cycles for each sample. Then 17 µL of ccDNA was amplified in 50 µL reaction volume containing: 25 µL of Phusion High-Fidelity PCR Master Mix. The reaction was stopped, and loaded on a native 10% polyacrylamide gel, next to CRL, and HRL DNA Ladders. Bands of dsDNA were cut out, eluted and precipitated overnight. The DNA pellet resuspended in nuclease-free water and its concentration measured by NanoDrop and Qubit Bioanalyzer. Libraries were pooled by mixing 4.5 µg to a final amount of 90 µg dsDNA. The quality and concentration of the sample was determined by High Sensitivity DNA Assay on an Agilent 2100 Bioanalyzer. Sequencing was run using MiSeq Reagent Kits v2 (50cycles, Illumina), MiSeq Instrument with method Single Index by Illumina (1×50 MiSeq with Index).

Sequencing data analysis. RNA-seq reads were filtered according to their indices to 20 groups. Inside each group, only sequences containing the correct sequence of the first 6 nucleotides from a small RNA sequence were taken to further analysis. Inside this group, sequences were separated into 6 bins: i) full length sequences (20 nt length), ii) sequences with one deletion (19 nt), iii) sequences containing 2 deletions (18 nt), iv) truncated sequences (at least 6 nt-long), v) sequences with single insertion (21 nt). We focused on 2 positions forward (+2) and backward (-2) in sequence relative to the modified residue (position 0). Data were normalized to 10,000 reads and plotted as frequencies of indels, truncations and ATCG content. These data values were subject to PCA analysis. Raw sequencing data filtration was performed by Genesis Data Solutions, Omaha, NB.

PCA analysis. The principal components were calculated using standard algorithms from the R software. Varied subsets of the data were used to identify the combinations of variables that would best discriminate between the modifications considered in the experiment. Variables normalized to 10,000 reads were transformed by standardizing the data to a common standard deviation (each variable has a variance of 1). We performed principal components analysis. To investigate the data confidence levels (0.95) Gamma-Poisson distributions for each variable were simulated. Unknown sample simulations were done by the assumption that the sample is a mixture of 10% or 50% modified oligonucleotide with unmodified.

ASSOCIATED CONTENT

Supporting Information

Supporting Information contains detailed experimental methods and additional data and figures. This material is available free of charge via the Internet at http://pubs.acs.org.

AUTHOR INFORMATION

Corresponding Author

kool@stanford.edu

Author Contributions

Notes

The authors declare no competing financial interests.

ACKNOWLEDGMENT

We thank the U.S. National Institutes of Health (GM110050, GM106067, and CA217809) for support. We thank Pedro J. Batista for discussions and suggestions about the experimental design, and Caroline Roost for input in the development of a method for m6G synthesis.

REFERENCES

(1) Frye, M.; Jaffrey, S. R.; Pan, T.; Rechavi, G.; Suzuki, T., *Nature Rev. Genet.* **2016,** *17*, 365-372.
(2) Gilbert, W. V.; Bell, T. A.; Schaening, C., *Science* **2016,** *352*, 1408-12.
(3) Harcourt, E.; Kietrys, A. M.; Kool, E. T., *Nature* **2017,** *541*, 339-346.
(4) Cantara, W. A.; Crain, P. F.; Rozenski, J.; McCloskey, J. A.; Harris, K. A.; Zhang, X.; Vendeix, F. A. P.; Fabris, D.; Agris, P. F., *Nucleic Acids Res.* **2011,** *39*, D195–D201.
(5) Desrosiers, R.; Friderici, K.; Rottman, F., *Proc. Natl. Acad. Sci. USA* **1974,** *71*, 3971-5.
(6) Morse, D. P.; Bass, B. L., *Biochemistry* **1997,** *36*, 8429-8434.
(7) Carlile, T. M., *Nature* **2014,** *515*, 143-146.
(8) Dubin, D. T.; Taylor, R. H., *Nucleic Acids Res.* **1975,** *2*, 1653-1668.
(9) Dominissini, D.; Nachtergaele, S.; Moshitch-Moshkovitz, S.; Peer, E.; Kol, N.; Ben-Haim, M. S.; Dai, Q.; Di Segni, A.; Salmon-Divon, M.; Clark, W. C.; Zheng, G.; Pan, T.; Solomon, O.; Eyal, E.; Hershkovitz, V.; Han, D.; Doré, L. C.; Amariglio, N.; Rechavi, G.; He, C., *Nature* **2016,** *530*, 441-6.
(10) Li, X.; Xiong, X.; Wang, K.; Wang, L.; Shu, X.; Ma, S.; Yi, C., *Nat. Chem. Biol.* **2016,** *12*, 311-6.
(11) Roundtree, I.; Evans, M.; Pan, T.; He, C., *Cell* **2017,** *169*, 1187-1200.
(12) Lewis, C. J. T.; Pan, T.; Kalsotra, A., *Nat. Rev. Mol. Cell Biol.* **2017,** *18*, 202-210.
(13) Zhao, B. S.; Roundtree, I. A.; He, C., *Nat. Rev. Mol. Cell Biol.* **2016,** *18*, 31-42.
(14) Sibbritt, T.; Patel, H. R.; Preiss, T., *Wiley Interdiscip. Rev. RNA* **2013,** *4*, 397-422.
(15) Su, D.; Chan, C. T. Y.; Gu, C.; Lim, K. S.; Chionh, Y. H.; McBee, M. E.; Russell, B. S.; Babu, I. R.; Begley, T. J.; Dedon, P. C., *Nat. Prot.* **2014,** *9*, 828-41.
(16) Helm, M.; Motorin, Y., *Nature Rev. Genet.* **2017,** *18*, 275–291.

(17) Wang, Z.; Gerstein, M.; Snyder, M., *Nature Rev. Genet.* **2009,** *10*, 57-63.
(18) Han, Y.; Gao, S.; Muegge, K.; Zhang, W.; Zhou, B., *Bioinformatics and Biol. Insights* **2015,** *9*, 29-46.
(19) Bussotti, G.; Leonardi, T.; Clark, M. B.; Mercer, T. R.; Crawford, J.; Malquori, L.; Notredame, C.; Dinger, M. E.; Mattick, J. S.; Enright, A. J., *Genome Res.* **2016,** *26*, 705-716.
(20) Merkle, F. T.; Ghosh, S.; Kamitaki, N.; Mitchell, J.; Avior, Y.; Mello, C.; Kashin, S.; Mekhoubad, S.; Ilic, D.; Charlton, M.; Saphier, G.; Handsaker, R. E.; Genovese, G.; Bar, S.; Benvenisty, N.; McCarroll, S. A.; Eggan, K., *Nature* **2017,** *545*, 229-233.
(21) Ameur, A.; Zaghlool, A.; Halvardson, J.; Wetterbom, A.; Gyllensten, U.; Cavelier, L.; Feuk, L., *Nat. Struct. Mol. Biol.* **2011,** *18*, 1435-1440.
(22) Kool, E. T., *Annu. Rev. Biophys. Biomol. Struct.* **2001,** *30*, 1-22.
(23) Ding, Y.; Fleming, A. M.; Burrows, C. J., *J. Am. Chem. Soc.* **2017,** *139*, 2569-2572.
(24) Motorin, Y.; Muller, S.; Behm-Ansmant, I.; Branlant, C., *Methods Enzymol.* **2007,** *425*, 21-53.
(25) Harcourt, E. M.; Ehrenschwender, T.; Batista, P. J.; Chang, H. Y.; Kool, E. T., *J. Am. Chem. Soc.* **2013,** *135*, 19079-19082.
(26) Sakurai, M.; Yano, T.; Kawabata, H.; Ueda, H.; Suzuki, T., *Nat. Chem. Biol.* **2010,** *6*, 733-740.
(27) Hauenschild, R.; Tserovski, L.; Schmid, K.; Thüring, K.; Winz, M. L.; Sharma, S.; Entian, K. D.; Wacheul, L.; Lafontaine, D. L. J.; Anderson, J.; Alfonzo, J.; Hildebrandt, A.; Jäschke, A.; Motorin, Y.; Helm, M., *Nucleic Acids Res.* **2015,** *43*, 9950-9964.
(28) Byron, S. A.; Van Keuren-Jensen, K. R.; Engelthaler, D. M.; Carpten, J. D.; Craig, D. W., *Nat. Rev. Genet.* **2016,** *17*, 257-271.
(29) Chen, J.; Zhou, Q.; Wang, Y.; Ning, K., *Sci. Rep.* **2016,** *6*, 34420.
(30) Serratì, S.; Petriella, D., *OncoTargets and Ther.* **2016,** *9*, 7355-7365.
(31) Argyropoulos, C.; Etheridge, A.; Sakhanenko, N.; Galas, D., *Nucleic Acids Res.* **2017,** 1-22.
(32) Rutvisuttinunt, W.; Meyer, P. R.; Scott, W. A., *PLoS ONE* **2008,** *3*, e3561.
(33) Lapkouski, M.; Tian, L.; Miller, J. T.; Le Grice, S. F. J.; Yang, W., *Nat. Struct. Mol. Biol.* **2013,** *20*, 230-6.
(34) Aschenbrenner, J.; Drum, M.; Topal, H.; Wieland, M.; Marx, A., *Angew. Chem.* **2014,** *53*, 8154-8158.
(35) Bakin, A.; Ofengand, J., *Biochemistry* **1993,** *32*, 9754-9762.
(36) Roost, C.; Lynch, S. R.; Batista, P. J.; Qu, K.; Chang, H. Y.; Kool, E. T., *J. Am. Chem. Soc.* **2015,** *137*, 2107-2115.
(37) Murphy, F. V.; Ramakrishnan, V., *Nat. Struct. Mol. Biol.* **2004,** *11*, 1251-2.
(38) Suzuki, T.; Ueda, H.; Okada, S.; Sakurai, M., *Nat. Prot.* **2015,** *10*, 715-732.
(39) Zhou, H.; Kimsey, I.; Nikolova, E. N.; Sathyamoorthy, B.; Grazioli, G.; McSally, J.; Bai, T.; Wunderlich, C. H.; Kreutz, C.; Andricioaei, I.; Al-Hashimi, H. M., *Nat. Struct. Mol. Biol.* **2016,** *23*, 803-810.
(40) Hudson, B. H.; Zaher, H. S., *RNA* **2015,** *21*, 1648-59.
(41) Delaney, J. C.; Essigmann, J. M., *Chem.and Biol.* **1999,** *6*, 743-753.
(42) Basturea, G. N.; Rudd, K.; Deutscher, M. P., *RNA* **2006,** *12*, 426-34.
(43) Ozsolak, F.; Milos, P. M., *Nat. Rev. Genet.* **2011,** *12*, 87-98.
(44) Archer, N.; Walsh, M. D.; Shahrezaei, V.; Hebenstreit, D., *Cell Syst.* **2016,** *3*, 467-479.
(45) Eckert, K. A.; Kunkel, T. A., *Nucleic Acids Res.* **1990,** *18*, 3739–3744.
(46) Kebschull, J. M.; Zador, A. M, *Nucleic Acids Res.* **2015,** *43*, e143.
(47) Raabe, C. A.; Tang, T. H.; Brosius, J.; Rozhdestvensky, T.S., *Nucleic Acids Res.* **2014,** *42*, 1414–1426.
(48) Liu, Y.; Ferguson, J. F.; Xue, C.; Silverman, I. M.; Gregory, B.; Reilly, M. P.; Li, M., *PLoS ONE* **2013,** *8*, e66883.
(49) Sims, D.; Sudbery, I.; Ilott, N. E.; Heger, A.; Ponting, C. P, *Nat. Rev. Genet.* **2014,** *15*, 121–132.
(50) Suzuki, T.; Ueda, U.; Okada, S.; Sakurai, M., *Nat. Prot.* **2015,** *10*, 715–732.
(51) Kaisers, W.; Schwender, H.; Schaal, H., *Internat. J. Mol. Sci.* **2017,** *18*, E1900.

Graphical abstract