Detecting Changes in the Epitranscriptome by Re-Interpreting M6aNET

Theodore M. Nelson

Introduction

Recent advances in nanopore technologies have permitted the first ever sequencing of RNA molecules, via a commercially available platform.[1] These sequencers allow for antibody-independent detection of RNA modifications on a transcriptome-wide basis, based on differential current signal within a nanopore, an engineered membrane-like protein.[2] Machine learning classification software, such as *m6anet*, have shown admirable applicability and sensitivity in resolving these modifications, firstly due to their dispensing of a negative control lacking m6a modifications, and secondly due to their performance in comparative studies.[3]

In training *m6anet*, the original source data for direct-RNA peaks were identified with another program, xPORE, by comparing sequenced *METTL3*-KO and WT HEK293T cell lines.[4] Candidates DRACH sites were identified and labeled via m6ACE-Seq data, to generate a training dataset for *m6anet*, which outputs probability estimates for each covered DRACH motif covered within a human sample transcriptome. Within this article, I argue that these probability estimates are unreliable, given that they generalize across different RNA contexts and diverse cell types, which were not originally part of the training data.

VIRMA is a recently characterized component of the m6a methyltransferase complex, along with *METTL3*, regulating m6a deposition in the 3' untranslated region near protein stop codons.[5] *VIRMA*-knockdown has been associated with a decrease in m6a deposition.[6] I analyzed *VIRMA*-knockdown samples from direct-RNA sequencing runs from Oxford Nanopore Technologies 004 platform, comparing them with WT controls and in-vitro transcribed RNA, to propose a novel technique to identify differential methylation in a context-dependent manner.

Instead of assigning methylation status based on *m6anet*, I identified differentially methylated sites by examining the distribution of probabilities within an individual site. By examining these predictions, I can determine whether the model recognizes two distinct probability distributions, irrespective of their assigned mean probabilities.

By determining methylation sets in this way, I aim to propose this methodology as a more sensitive way of identifying *VIRMA*-knockdown affected m6a sites, generalizable to a wider range of biological questions, conditions, and direct-RNA sequencing runs.

Methods

Data. I analyzed six samples from the B lymphoblast cell line MM1S (Table 1) and one in-vitro-transcribed HCV sample, sequenced with the Oxford Nanopore Technologies Direct RNA Sequencing Kit (SQK-RNA004), according to manufacturer's instructions. Sequences were basecalled with Dorado software version 7.2.13+fba8e8925, client-server API version 16.0.0, with base calling model rna_rp4_130bps_hac_prom.cfg. Minimap2 version 2.26 was used to

generate alignments.[7] These were reconnected to the original squiggle alignments, kept in the .blow5 format, with f5c version 1.4.[8] M6anet version 2.1.0 was then applied with the HEK293T_RNA004 model to generate DRACH motif position predictions for human transcripts within the hg38 refMrna.fa.gz file (http://hgdownload.soe.ucsc.edu/goldenPath/hg38/bigZips/).[2]

Table 1: Sample Metadata for Samples

Sample Name	Condition
S007	MM1S-R1 SCRAMBLE
S008	MM1S-R1 gVIRMA 1
S009	MM1S-R1 gVIRMA 3
S010	MM1S-R2 SCRAMBLE
S011	MM1S-R2 gVIRMA 1
S012	MM1S-R2 gVIRMA 3
HCV_IVT_004	In-Vitro Transcribed HCV RNA

Caption: R1 and R2 refer to two biological replicates; SCRAMBLE, gVIRMA 1, and gVIRMA 3 refer to different CRISPR guides with either non-*VIRMA* targeting (SCRAMBLE) or *VIRMA*-specific targeting (gVIRMA 1 and gVIRMA 3). HCV_IVT_004 refers to in-vitro transcribed HCV RNA.

Identifying Epitranscriptome Modifications. The `data.indiv_proba.csv` file output by *m6anet* was transferred to an *R* version 4.3.1 environment, wrapped with *RStudio*. Custom scripts were written to generate the remaining analysis and graphs. Of note, the *diptest* package was utilized to calculate test statistics for the degree of unimodality within the distributions for sites with Hardigan's Dip Test.[9]

Results

Single-Molecule Distribution of M6aNET Predictions Resolves Limited Biological Variability

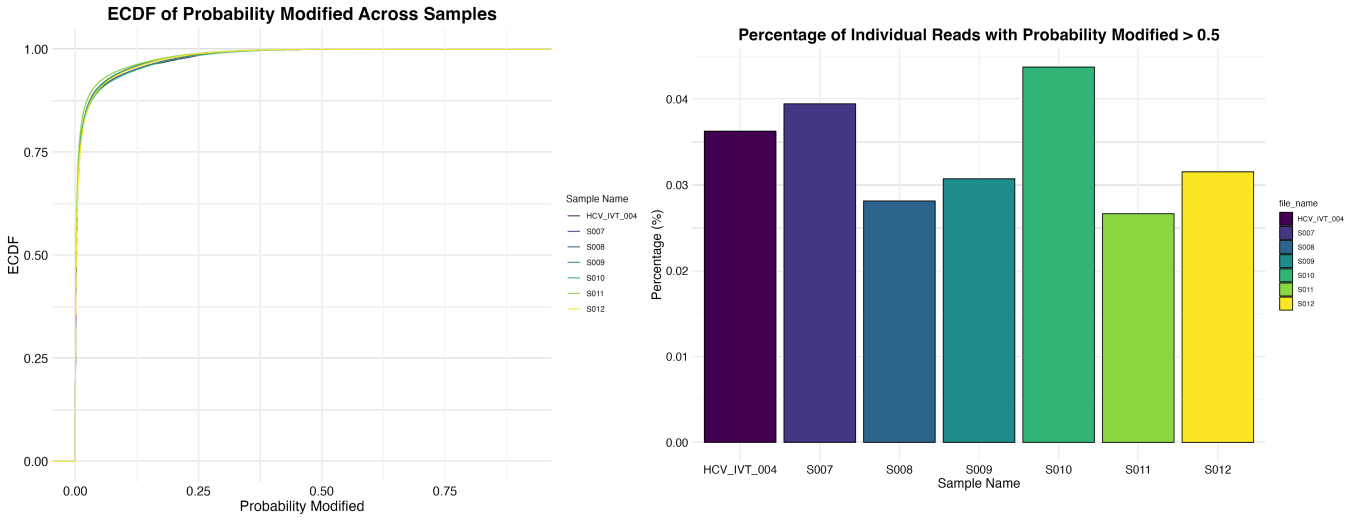
When examining the distribution of probabilities that a particular m6a site on an individual modified, I notice little difference between the *VIRMA*-knockdown (*VIRMA*-KD) conditions [S008, S009, S011, S012] and the wildtype conditions [S007, S010] (Figure 1). Notably, these distributions also did not exhibit any noticeable difference from the distribution of individual site probabilities from an HCV in-vitro transcribed sample, a system which lacks the capacity to form any modifications.

To resolve the capacity of *m6anet* to call high-confidence modifications, I selected an arbitrary cutoff of `.5` to excise the modification calls of highest confidence, encompassing the top `.05` percent of individual molecules. While the wildtype conditions did exhibit a slight `.01` percentage increase when compared to the *VIRMA*-KD, the *VIRMA*-KD also demonstrated a

lower percentage of predicted reads than the HCV in-vitro transcribed sample. Although *VIRMA*-KD is expected to have lower levels of m6a, it should still demonstrate greater methylation probabilities than in-vitro transcribed RNA.

Together, these suggest that the native m6a probabilities output by *m6anet* are not providing meaningful resolution into the epitranscriptomics of these particular sample systems.

Figure 1: Distribution of Site Probabilities at Single-Molecule Resolution, predicted by *m6anet*



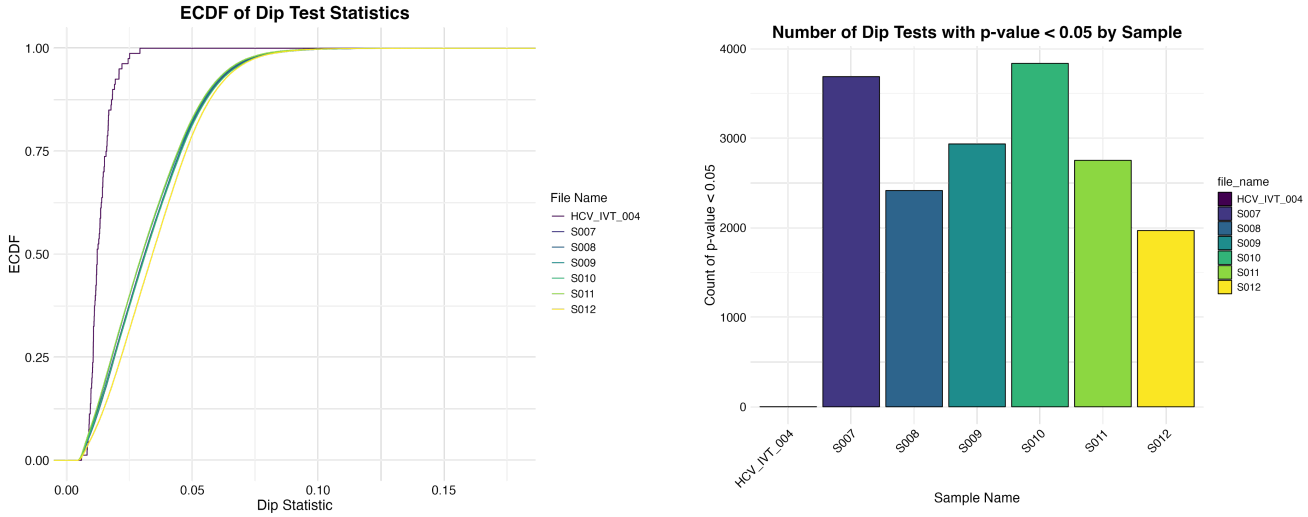
Caption: ECDF of the probability of modification for 1/20 of the individual molecules, randomly extracted from each sample (left), and a bar plot demonstrating the percentage of individual single-molecules with a high probability of modification (right)

Hardigan's Dip Test Resolves Potential Biological Variability in *VIRMA*-KO Dataset

On the assumption that *m6anet* would assign different methylation probabilities to individual molecules in different methylation states, I applied Hardigan's Dip test to test whether the individual distributions would be multi-modal.[9] Hardigan's Dip test calculates a statistic, between zero [uni-modal] and one [multimodal], which quantities the degree to which the cumulative distribution function of a multimodal function needs to be altered in order to become a unimodal distribution. I assume that sites with appreciable methylation will be bimodal, with separate peaks representing unmethylated and methylated states.

After calculating the dip test statistic for all sites within each sample, I noted a difference in the distribution between the biological samples and the in-vitro-transcribed RNA, suggesting the presence of epi-transcriptomic modifications was driving multimodal distributions (Figure 2). I additionally observed many sites with statistically significant multimodal distributions, with the wildtype samples having a greater number of multimodal distributions than *VIRMA*-KD conditions. These trends are affected by but are independent from the amount of coverage within each sample (data not shown).

Figure 2: Distribution of Site Modality Test Statistics, calculated with Hardigan's Dip test

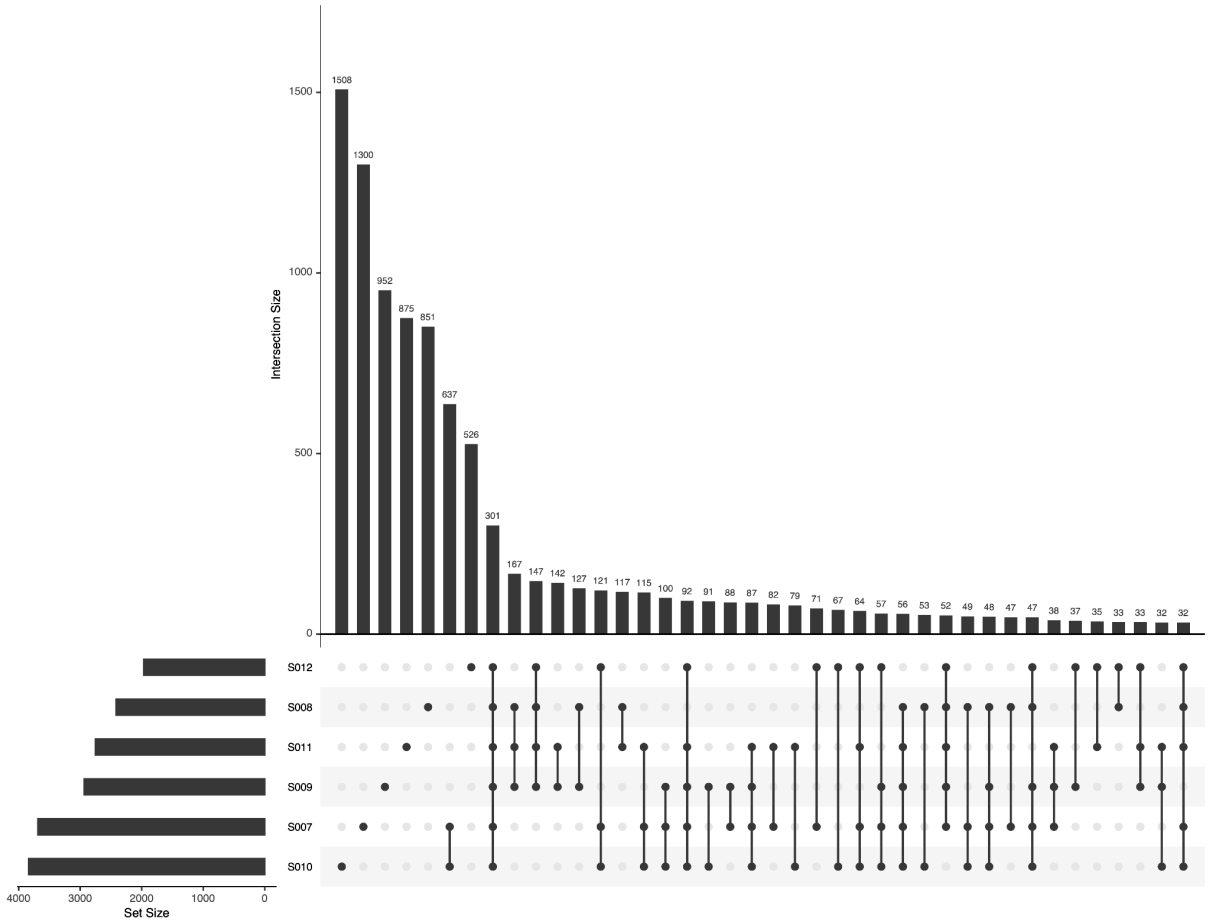


Caption: ECDF of the dip statistic for individual sites (left) and a bar plot demonstrating the number of sites with a statistically significant dip statistic (right)

To understand whether these bi-modal distributions were related to each other in potentially biological relevant conditions, I examined the intersections between the statistically significant dip tests for S007, S008, S009, S010, S011, and S012.

Each of the samples had a large number of sites unique to that specific sample, suggesting potential individualized m6a dynamics. The largest overlap set was between the two wild-type sets, potentially implicating sites of methylation affected by *VIRMA*-KD (Figure 3). The second largest set of overlap consisted of all six samples, implying shared sites of m6a methylation.

Figure 3: UpSet Plot of the Intersection between the Statistically Significant Dip Test Sites



Discussion

The size of the *VIRMA*-KD set (637 genes) compared with the size of the shared gene set (301 genes) reflects the crucial role of *VIRMA* within the m6a-methyltransferase complex. The alignment between biological expectations and the results of Hardigan's Dip test evidences the utility of this method in resolving changes within the epitranscriptome.

The large number of unique statistically significant dip test sites could be caused by underlying biological variation, or by neighboring modifications, inducing current shifts detected by *m6anet* at the DRACH motif. Since this method requires multi-modal distributions, the test will not capture fully biological conditions where every individual RNA molecule is either unmethylated or methylated. However, these states are likely to be rare, given RNA turnover and the dynamic nature of m6a regulation.

Future work will quantify the multimodal peaks in order to allow for quantitative comparisons across samples, allowing for a determination of potential differences within the shared data set between *VIRMA*-KD and wildtype samples.

References

1. Garalde, D. R. *et al.* Highly parallel direct RNA sequencing on an array of nanopores. *Nat. Methods* **15**, 201–206 (2018).
2. Hendra, C. *et al.* Detection of m6A from direct RNA sequencing using a multiple instance learning framework. *Nat. Methods* **19**, 1590–1598 (2022).
3. Zhong, Z.-D. *et al.* Systematic comparison of tools used for m6A mapping from nanopore direct RNA sequencing. *Nat. Commun.* **14**, 1906 (2023).
4. Pratanwanich, P. N. *et al.* Identification of differential RNA modifications from nanopore direct RNA sequencing with xPore. *Nat. Biotechnol.* **39**, 1394–1402 (2021).
5. Yue, Y. *et al.* VIRMA mediates preferential m6A mRNA methylation in 3′UTR and near stop codon and associates with alternative polyadenylation. *Cell Discov.* **4**, 10 (2018).
6. Barros-Silva, D. *et al.* VIRMA-Dependent N6-Methyladenosine Modifications Regulate the Expression of Long Non-Coding RNAs CCAT1 and CCAT2 in Prostate Cancer. *Cancers* **12**, 771 (2020).
7. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).
8. Gamaarachchi, H. *et al.* GPU accelerated adaptive banded event alignment for rapid comparative nanopore signal analysis. *BMC Bioinformatics* **21**, 343 (2020).
9. Hartigan, J. A. & Hartigan, P. M. The Dip Test of Unimodality. *Ann. Stat.* **13**, (1985).