# Comprehensive analysis of nanopore (ONT) direct long-read RNA sequencing data maps *Mettl3*-mediated m6A RNA methylation patterns in colorectal carcinoma cell lines HCT116 and DLD-1

Dr. Yasir Arafat Tamal

Borstel, August 30, 2024

# Results

## 1. Summary of input data

To capture and quantify full-length transcripts and RNA modifications, specifically N6-methyladenosine (m6A) in this study, nanopore direct long-read RNA sequencing was employed using the PromethION device, coupled with flow cell FLO-PRO004RA and the direct RNA sequencing kit SQK-RNA004. This sequencing approach was applied to two colorectal carcinoma cell lines, DLD-1 and HCT116, under two conditions: control and *METTL3* (Methyltransferase 3) knockdown using siRNA.

DLD-1 is a colorectal adenocarcinoma cell line derived from the large intestine of a colon adenocarcinoma patient, while HCT116 is a widely used human colon cancer cell line for studying colon cancer proliferation. *METTL3* encodes the enzyme METTL3, which acts as an m6A methyltransferase. In most cancers, *METTL3* functions as an oncogene by applying m6A modifications to crucial mediators and transcripts, promoting cancer initiation and development. However, *METTL3* can also function as a tumor suppressor, where its m6A mRNA modifications promote tumor suppressor proliferation, migration, and invasion.

Nanopore direct RNA sequencing generates reads as "squiggle" or ionic current signals, stored in pod5 format (Jain et al., 2022). Importantly, since full-length transcripts are sequenced directly, without cDNA conversion or PCR amplification, the base modification information is retained in the sequence data. Transcriptomic profiles were generated for both cell lines under the two conditions, resulting in four samples. The goal of this study is to map the impact of *METTL3* catalytic activity on m6A signatures across these conditions and cell lines.

## 2. From Raw Ionic Signals to Sequences of Canonical and Modified Bases:

As shown in Figure 1, the analysis begins with converting ionic current signals into sequences of four nucleotide bases: A, U (for RNA), G, and C. The *Dorado* basecaller (ONT Dorado, 2024) utilizes a pre-trained model, specific to the analyte type (RNA), flow cell, and sequencing kit, to identify both canonical unmodified bases and modified bases such as N6-methyladenosine (m6A). In addition to basecalling, *Dorado* allows for alignment to a reference transcriptome or genome.

For each sample, the *Dorado* basecaller (version 0.7.0) was employed to convert ionic signals into sequences of both canonical and modified bases, and to align these sequences to the Ensembl reference transcriptome (Ensembl transcriptome GRch38, 2024). As illustrated in Figure 1, the basecaller outputs a binary alignment and map (BAM) file, which serves as input for both differential modification analysis and transcript- and gene-level differential expression analysis.

The objective of this study is to detect changes in m6A modification patterns in the *METTL3*-knockdown cell lines compared to the control group. This objective was addressed through two sets of analyses: 1) analysis of modified bases, and 2) count-based differential transcript- and gene-expression analysis.

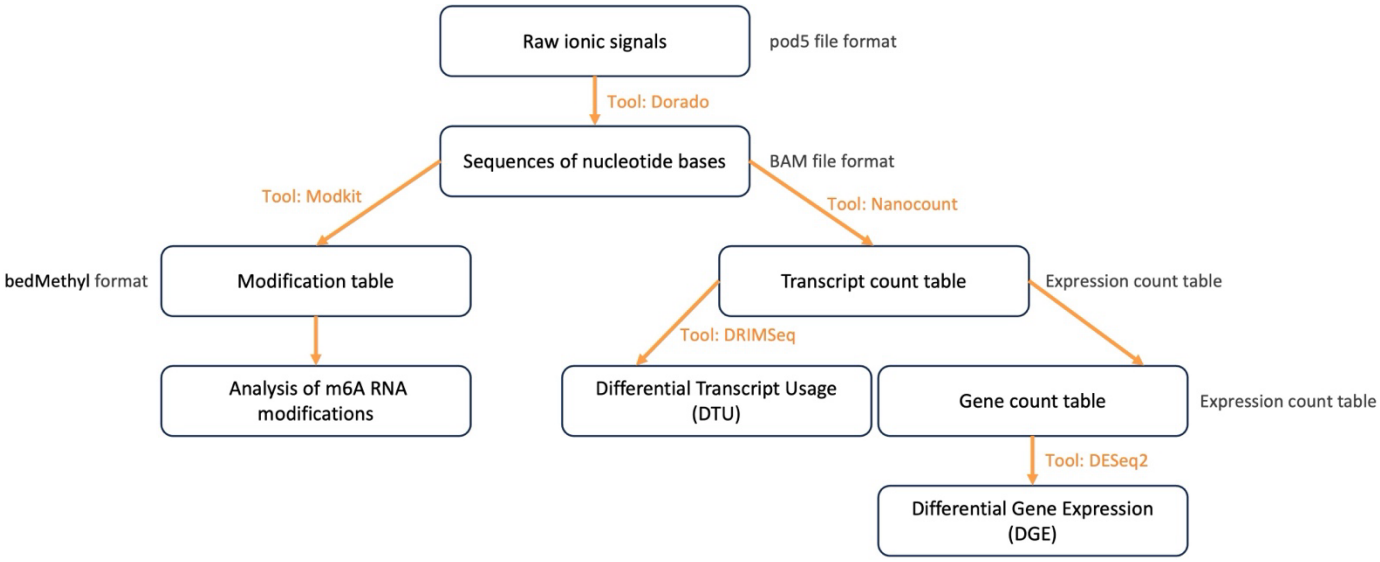


**Figure 1:** Workflow of nanopore direct RNA-sequencing data analysis. The workflow starts with sequencing data (pod5) transformed into nucleotide sequences (BAM). From this point, two complementary analyses are performed: analysis of RNA m6A modifications using modification count data, and differential expression analysis using transcript and gene expression count data.

## 3. Analysis of m6A Modified Bases:

The BAM file generated by the *Dorado* basecaller contains both canonical and m6A-modified base data, which we utilized to quantify base modifications. To facilitate this, we used *Modkit* (ONT Modkit, 2024), a bioinformatics toolset designed for manipulating modified-base data stored in BAM files. The primary function of *Modkit* in this analysis was to create summary counts of modified and unmodified adenosine bases (m6A) in an extended bedMethyl format. This format allows for the tabulation of base modification counts from each sequencing read across all reference genomic positions. Using *Modkit*, we generated m6A modification count summaries for each sample. The subsequent analysis of these counts was performed using the R programming language.

### 3.1 m6A Modification Landscape:

The generated bedMethyl table contains detailed information about the positions of m6A sites (Table 1) at single-base resolution (base A). For each m6A site, the table includes the associated

transcript ID, genomic position and strand, modification rate, number of reads that captured the modified site, and other relevant details. For a more detailed explanation, please refer to Nanoporetech *Modkit* Quick Start (ONT Modkit, 2024).

**Notation:**

- Modification table/bedMethyl table: Generated from *Modkit*.
- Site: The position of an m6A site at single-base resolution.
- Coverage/Score: The total number of reads covering the modified base (sum of canonical and modified bases at a given site).
- Modification rate: The ratio of the number of modified bases to the total bases at a given site.
- Differential modification rate: The difference in modification rates between the two groups.
- cMod table: Combined samples modification table.
- Samples naming, DLD-1 as DLD and HCT116 as HCT.

## 3.2 Preprocessing and Filtering of the Modification Table:

Each sample's modification table undergoes several filtering steps individually. Table 1 shows the initial total number of sites captured in the *Modkit* output modification table and the number of sites retained after each filtering iteration:

1. Step 1: Retain only the sites on the positive strand; reads mapped to the negative strand provide modification information for the "T" base, not the "A" base.
2. Step 2: Retain sites where the proportion of different bases at the site is less than 25%. This step accounts for potential heterozygous variants, where differing bases may represent mutations in one copy of a gene.
3. Step 3: Retain sites where deletions at the site are less than 25%. This step minimizes weak signal or basecalling errors.

As shown in Table 1, the total number of m6A modification sites ranges from ~55 million to ~58.5 million across samples. The varying number of transcripts may reflect inherent biological differences between cell lines and conditions. Despite these differences, for this study, the control conditions of both cell lines were treated as biological replicates of the control group, and the knockdown conditions of both cell lines were treated as biological replicates of the knockdown group. However, comparing conditions within each cell line would likely improve the analysis quality.

| SAMPLES | NO. INITIAL TRANSCRIPTS | NO. INITIAL SITES | NO. SITES AFTER FILTER STEP 1 | NO. SITES AFTER FILTER STEP 2 | NO. SITES AFTER FILTER STEP 3 |
|---|---|---|---|---|---|
| HCT-CONTROL | 86169 | 55929526 | 54186779 | 47265673 | 46111370 |
| HCT-KI | 81631 | 55058384 | 53306268 | 46003944 | 44835710 |
| DLD-CONTROL | 88874 | 56930568 | 54869026 | 47839723 | 46658194 |
| DLD-KI | 89199 | 58689269 | 56505370 | 48474481 | 47233304 |

**Table 1:** Summary of m6A methylation sites and transcripts detected in the sequencing data before and after filtering. The first column shows the sample names. The second column shows the total number of transcripts with detected methylation. The third column shows the total number of methylation sites. Columns four to six display the number of methylated sites retained after each filtering step in order.

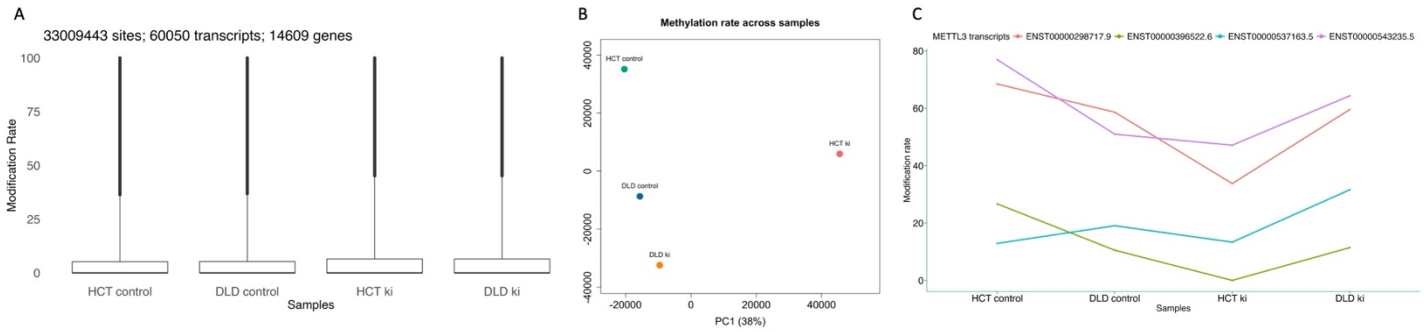


**Figure 2:** m6A modification patterns across cell lines and conditions.
A) Distribution of modification rates at sites (adenosine bases) shared across samples, retained after initial filtering.
B) PCA on methylation rates for all samples.
C) Methylation rates of four sites belonging to four transcripts out of 3228 sites with detected methylation in 10 transcripts of *METTL3*.

After the filtering steps, the total number of sites ranges from ~45 million to ~47 million. Among these, ~39 million sites are shared between conditions in the HCT116 cell line, and ~41 million sites are shared between conditions in the DLD-1 cell line. The cleaned modification tables were then combined to include only sites present across all samples, resulting in a total of ~33 million shared sites and 60,050 transcripts across both cell lines (Figure 2A). Figure 2B shows the PCA projection of the samples. For METTL3, methylation was detected at 3228 sites from its 10 transcripts out of 14. Only four METTL3 transcripts, representing 311 sites, have a coverage of 30 reads and modification rate at least 10 in one or more samples (Figure 2C).

The combined samples modification table (cMod table) still contains a large number of sites, many of which only have the canonical base adenosine (*A*), where the modification ratio is negligible. As the analysis aims to detect differences in modification rates across conditions, sites with no or very low modification rates were removed from the cMod table. The cMod table underwent two additional filtering steps to ensure it contains informative sites with meaningful

modification rates, aiding in highlighting differential modification between conditions. The filtering steps performed on the cMod table are as follows:

4. Step 4: Retain only sites with a coverage of 30 or more in any of the four samples. By setting a coverage filter, we account for both modified and unmodified bases. Changes in modification rate at sites with low coverage add noise to the data and should be filtered out before performing downstream analysis. This step improves computational efficiency and the quality of the analysis.

5. Step 5: Retain only sites where the differential modification rate is at least 10. The differential modification rate is calculated by quantifying the difference between the average modification rate of the control group and the knockdown group. The threshold for the differential modification rate ensures that only sites with a change of 10% or more are considered.

Due to the nature of these filtering steps, they could not be performed on each individual sample's modification table. Step 4 eliminated the highest number of sites, retaining ~15 million sites in the cMod table. The final filtering step, which accounted for differential modification rates between conditions, retained a total of 501,983 sites (1.5% of the total number of shared sites, ~33 million), providing a reduced yet informative representation of the modification sites.

## 3.3 Identification and Characterization of Differential RNA Modifications Across Conditions:

The filtered cMod table contains approximately 0.5 million sites. Our next objective was to identify sites with significant differential modification rates across conditions. To achieve this, we computed the two-proportions Z-test statistics to compare the modification rates in the control and knockdown groups at each site. The analysis implements the same computation described in the *xPore* tool (Pratanwanich et al., 2021). The goal was to statistically determine whether the observed modification rate in the control group differs significantly from that in the knockdown group.

The Z-test provided a test statistic and p-value for each site. We set the significance level at alpha = 0.05, concluding that if the p-value for a given site was less than this threshold (p-value < 0.05), the observed difference in modification rates between the conditions was significant. This analysis identified 58,192 sites with significant differential modification rates between the conditions (Figure 3A). These significant sites were found in 16,393 transcripts representing a total of 9,388 genes.

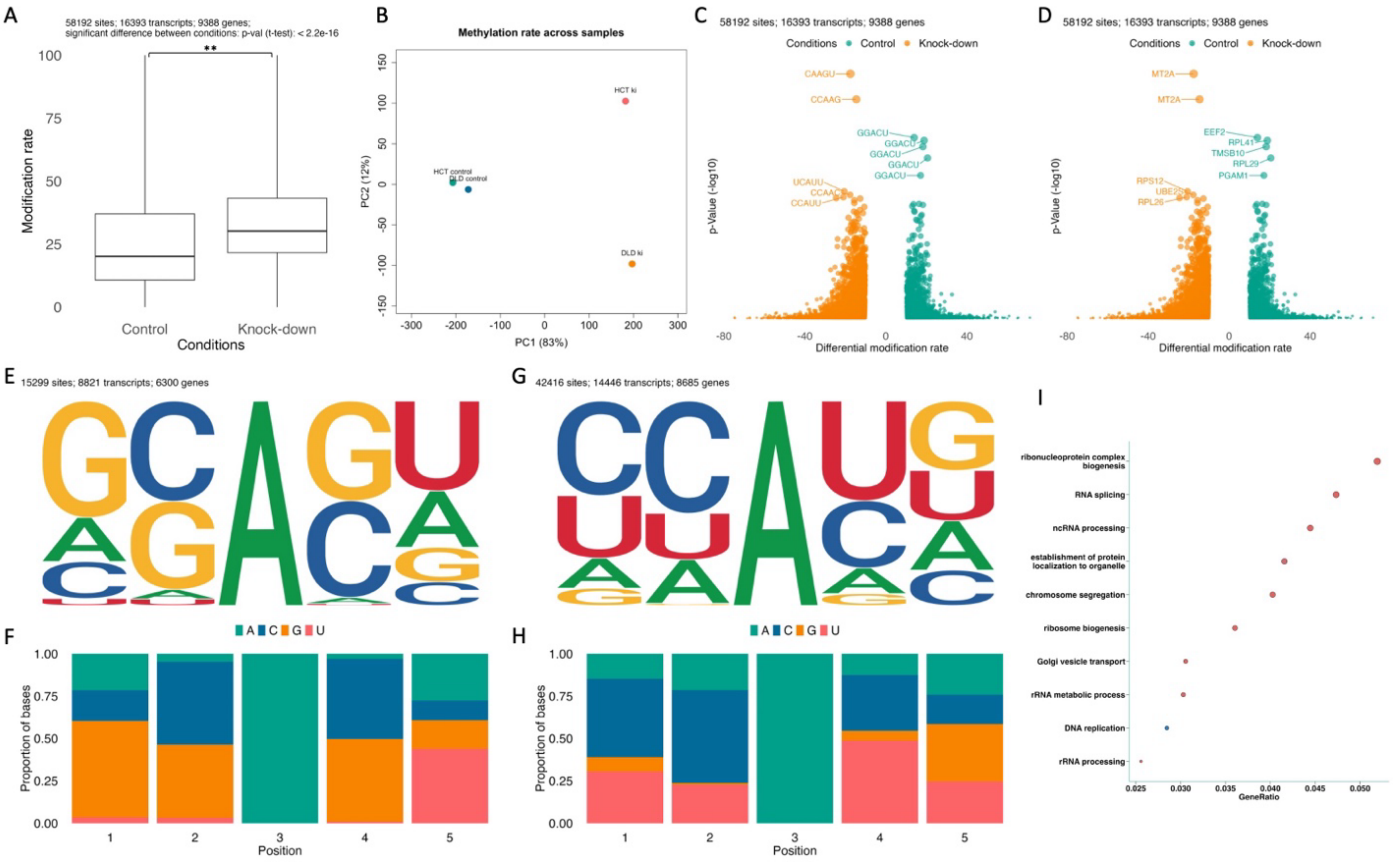


**Figure 3:** Significant differential methylated sites (58,192) and their associated motifs and genes between control and knock-down conditions in HCT116 and DLD-1 cell lines.
A) Distribution of significant differential methylated sites across conditions, retained with a p-value < 0.05 (two proportions z-test). The overall methylation rate between the conditions differing significantly.
B) PCA on significant methylated sites across all samples.
C) Volcano plot of significant methylated sites; dot size indicates p-value, and color represents conditions. Top 10 motifs (5-mers) are shown based on smallest p-values.
D) Volcano plot similar to (C) with gene annotations.
E, F) Consensus motif (E) and base proportions (F) for control condition.
G, H) Consensus motif (E) and base proportions (F) for knock-down condition.
I) Significant GO terms associated with genes linked to significant methylated sites in the knock-down group.

In the control group across both cell lines, 15,311 sites (26.3%) out of the 58,192 had a significantly higher modification rate compared to the knockdown group (Figure 3A). These sites were associated with 8,824 transcripts across 6,301 genes. On the other hand, 73.6% of the significant methylated sites exhibited higher m6A modification signatures due to *METTL3* knockdown. These 42,881 sites were found in 14,486 transcripts across 8,702 genes. I compared the two populations of modification value between the conditions using t-test and found significant difference in the distribution of methylation of the two conditions (Figure 3A). Although, the PCA projection could not capture any clear difference (Figure 3B). The volcano plot in Figure 3C shows all significant methylated sites across conditions and top ten motifs, in this

case 5-mers (2 base – adenosine site + 2 base) associated to the sites. Figure 3D illustrates the same information but top ten genes for the associated significant methylated sites were shown.

The boxplot in Figure 3A illustrates a clear difference between the conditions, underscoring the impact of *METTL3* knockdown on the m6A modification pattern. This raises questions about *METTL3*'s exact role in both cell lines: Does it function as an oncogene, or does it suppress other oncogenes? What gene networks are associated with *METTL3* in these specific cell lines? And how can we narrow down our focus from thousands of genes to a more targeted list? Differential expression analysis can complement this study by identifying overlapping genes between the two sets.

### 3.4 Detection of K-mers (5-mers) and DRACH Motifs:

After identifying the 58,192 significant modification sites (Figure 3A), we focused on identifying DRACH motifs (nucleotide signature: [A/G/U] [A/G] [A] [C] [A/C/U]; D = A, G, or U; R = A or G, H = A, C, or U) at the exact locations of the modified adenosine bases. Previous studies have linked m6A modifications to DRACH motifs, although these motifs are common in mRNAs, and not all are methylated (Martinez De La Cruz et al., 2023). Moreover, recent studies suggest that RNA secondary structure significantly influences m6A modifications, as more disordered or loosely structured regions are more accessible to the *METTL3* methyltransferase complex, and therefore more likely to be methylated (Martinez De La Cruz et al., 2023; Wang et al., 2021).

A primary focus in m6A RNA modification research is specificity. In this study, we already know the exact positions of m6A modifications and have identified a subset of 58,192 sites with significant methylated adenosine base. By locating and characterizing DRACH motifs within these significant methylation sites, we can identify transcripts and genes where both methylation and expression signatures change due to reduced *METTL3* activity. Your insights on the identified genes, transcripts, and DRACH motifs would greatly benefit this section.

We began by identifying all possible 5-mers (*K*-mers; the length of the motif search window was set to 5, matching the length of the DRACH motif) from the significant methylation sites in both conditions. From 57,715 sites (control: 15,299; knockdown: 42,416), 5-mers were extracted, corresponding to 16,362 transcripts (control: 8,821; knockdown: 14,446) and 9,377 genes (control: 6,300; knockdown: 8,685). Figure E and Figure F show the consensus motif sequence and relative proportion of bases at each position in the control condition. On the other hand, Figure G and Figure H show the consensus motif sequence and relative proportion of bases at each position in the knock-down condition. The number of genes and transcripts associated with the 57,715 methylated sites and 5-mers closely matches the input of 58,192 significant

modification sites. To avoid redundancy, further interpretation of these 57,715 5-mers and methylated sites is not included in this report.

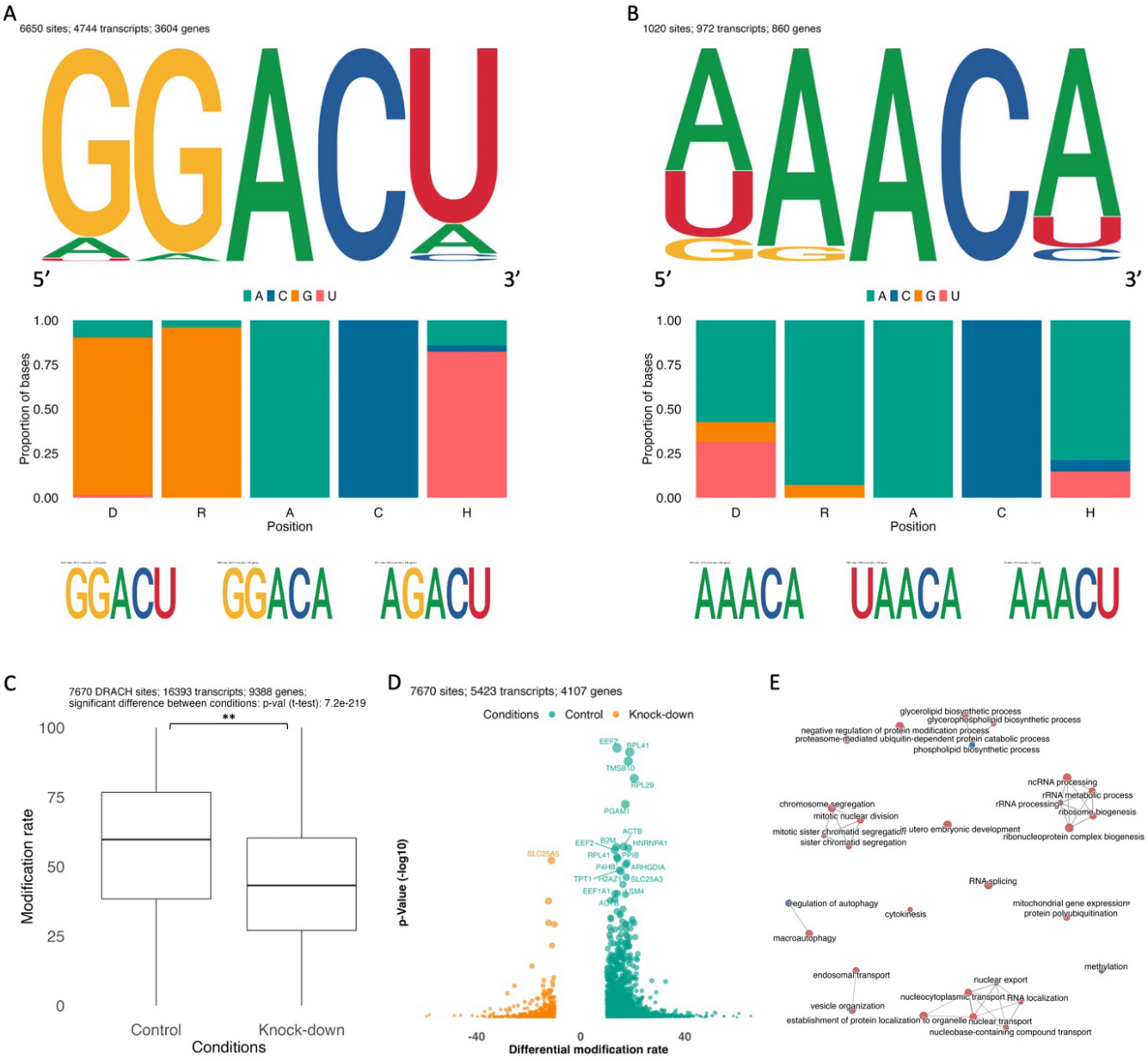


**Figure 4:** Association of DRACH motif with *METTL3*.

A) Consensus DRACH sequence for control group representing 6650 sites (top), and base proportions at each DRACH site (middle). The top three motifs are shown (bottom), with GGACU being the most frequent.

B) Consensus DRACH for knock-down group representing 1020 sites (top), with base proportions at each site (middle); top motifs include AAACA (bottom).

C) Distribution of significantly different methylated DRACH sites (7670) between conditions, with overall methylation rate differing significantly.

D) Volcano plot of significant DRACH methylated sites, with dot size indicating p-value, color for conditions, and top 10 motifs.

E) Significant GO terms for genes (3604) associated with significantly methylated DRACH sites in the control group.

The identified 57,715 motifs (5-mers) represent 246 distinct sequences. Of these, 182 motifs were extracted from methylated sites with higher modification rates in the control group (Figure 3E), while 236 motifs contained sites with higher modification rates in the knockdown group (Figure 3G). Notably, the motif associated with the top significant methylated sites in the control group is a DRACH motif. This raises further questions: What other DRACH motifs are associated with these significant methylated sites? And how are they distributed across the conditions?

## 3.5 Identification of DRACH Motifs Among Significant Methylation Sites:

Among the 57715 significant methylated sites, 7670 are associated with DRACH motif sequences, encompassing a total of 18 distinct DRACH motifs. These 7670 DRACH sites are distributed across 5423 transcripts derived from 4107 genes. This leads to a critical question: What are the consequences of knocking down the methylation writer, *METTL3*?

A key finding is that 86.7% of the 7670 DRACH motif-containing sites exhibit higher m6A methylation levels in the control group cell lines (Figure 4A). Specifically, these 6650 methylated sites correspond to 17 DRACH motifs, with the “GGACU” motif alone accounting for 4523 sites (Figure 4A). Notably, *METTL3* and a subset of METTL family genes share this motif. Among the 14 *METTL3* transcripts, only one protein-coding transcript, ENST00000298717.9, shows a significantly higher methylation rate (17%, p-value = 0.003) at the adenosine base located at position 631 (index starting at 1) on chromosome 14. This DRACH motif spans transcriptomic bases 629 to 633. Figure 4A depicts the consensus DRACH motif, relative position of the bases at the DRACH sites, and top three most frequent DRACH motifs identified in the control condition. These 6650 sites are associated with 4744 transcripts from 3604 genes. Figure 4I shows the significant biological processes in which these genes are involved. Given that m6A methylation often occurs in clusters, these DRACH motifs linked to a unique set of genes could help elucidate *METTL3*'s role in m6A modification.

In contrast, only 1020 sites in the knockdown condition are part of the 18 DRACH motifs, with the most frequent motif, “AAACA,” accounting for 458 sites. This figure is notably smaller compared to the top motif in the control group. These 1020 sites correspond to 972 transcripts from 860 genes (Figure 4B). This smaller gene set might provide insights into the biological or cellular activities that are upregulated due to *METTL3* knockdown. Figure 4B shows the consensus DRACH motif, relative position of the bases at the DRACH sites, and top three most frequent DRACH motifs identified in the knock-down condition. Figure 4D shows the top 20 genes for the sites with significant differential methylation at the DRACH motif.

As previously observed, the overall number of significant methylated sites is much higher in the knockdown group (42416 sites out of 58192) compared to the control group. If *METTL3* acts as a repressor of other genes, its knockdown could induce various cellular activities. Figure 3I shows the biological processes in which the 8685 genes (Figure 3G) in the knockdown condition are involved. However, as can be seen in Figure 4C, the two distribution of methylation rates of the condition, specifically for the 7670 DRACH motif-associated methylation sites, is significantly different. Notably, the methylation rate is higher in the control condition when considering only the DRACH motif sites.

### 3.6 Implications of *METTL3* Knockdown on m6A Methylation:

In conclusion, knocking down *METTL3* leads to significant changes in the methylation patterns observed in the knockdown group cell lines, with a notably higher number of significantly methylated sites in these lines compared to controls. This observation suggests that *METTL3* may have multiple roles, potentially acting as both an activator and a repressor. When examining the relationship between DRACH motifs and m6A methylation, it becomes clear that methylation rates are significantly higher in the control condition than in the knockdown condition. The observed differences in the methylation rate of the *METTL3* transcript, ENST00000298717.9, support the experimental hypothesis, confirming that *METTL3* knockdown drastically reduces the methylation signature at DRACH sites. Furthermore, the overall increase in methylation rates in the knockdown setup, beyond just the DRACH sites, suggests that RNA secondary structure may play an increasingly important role in determining m6A RNA modifications under these conditions.

It is important to note that in this analysis, we treated samples from two cell lines as biological replicates and focused only on sites shared between the conditions across both cell lines. However, I believe that performing cell line-specific modification analyses between conditions could significantly enhance the quality of the results. While focusing on shared sites provides a conservation framework, it can obscure cell line-specific biological insights. For example, *LEF1* (m6A methyltransferase *METTL3* promotes the progression of prostate cancer via m6A-modified *LEF1*) exhibited both methylation and expression in the DLD1 cell line, but these were absent in the HCT cell line. Therefore, a cell line-specific analysis could offer a more robust view of the modification sites between conditions. Your feedback and insights on these results will be invaluable for deriving biologically meaningful interpretations.

# 4 Analysis of Differential Expression (DE):

Recent advances in sequencing technology have made it possible to sequence entire transcripts, significantly enhancing our understanding of biological systems. However, realizing this potential requires bioinformatics tools capable of analyzing such data. The nanopore direct RNA-sequencing technology represents a cutting-edge advancement, but the lack of comprehensive bioinformatics tools for nanopore data analysis poses significant challenges.

In this section, I conducted differential expression (DE) analysis at two levels: differential gene expression (DGE) and differential transcript usage (DTU). These analyses complement the previous analysis of RNA modifications. In the earlier section, we identified sites with significant differential methylation rates, characterized them with DRACH motifs, and associated them with specific transcripts and genes. Now, we explore whether changes in methylation patterns correspond to changes in gene expression. Specifically, how many of the DRACH motif-containing genes exhibit differential expression between conditions, and can we establish a link between differentially expressed transcripts and DRACH methylated sites?

To begin, I conducted an initial comparison of gene expression counts between the control and knockdown conditions in the DLD1 cell line by simply dividing the expression levels. This preliminary analysis identified 868 genes with a log2 fold change of ≥ 0.5, indicating differential expression between conditions. Although this approach provides a basic overview without biological replicates, it helps set an expectation for the number of DEGs. However, such an analysis is exploratory and lacks the statistical rigor necessary to infer differences between groups. Without biological replicates, estimating the biological variability of each gene is impossible. Tools like *edgeR* (Robinson et al., 2009) or *DESeq2* (Love et al., 2014) rely on dispersion estimates from replicates to provide statistically meaningful results, which are not feasible without replication.

## 4.1 Summary of the Count Data:

Referencing Figure 1, the generated BAM file from the *Dorado* basecaller was used to generate transcript expression counts for each sample. *Nanocount* (Gleeson et al., 2022), a bioinformatics tool for counting transcripts from nanopore sequencing, produced a count table with 206723 rows (transcripts) and four columns (samples). The transcript count table was then aggregated into a gene count table with 41116 genes and four samples. This gene count table served as input for DE analysis using *DESeq2*, while the transcript count table was used for DTE analysis with *DRIMSeq* (Nowicka & Robinson, 2016) and *stageR (Van den Berge et al., 2017)*. DE analysis tools generally recommend a minimum of three biological replicates per group for robust analysis.

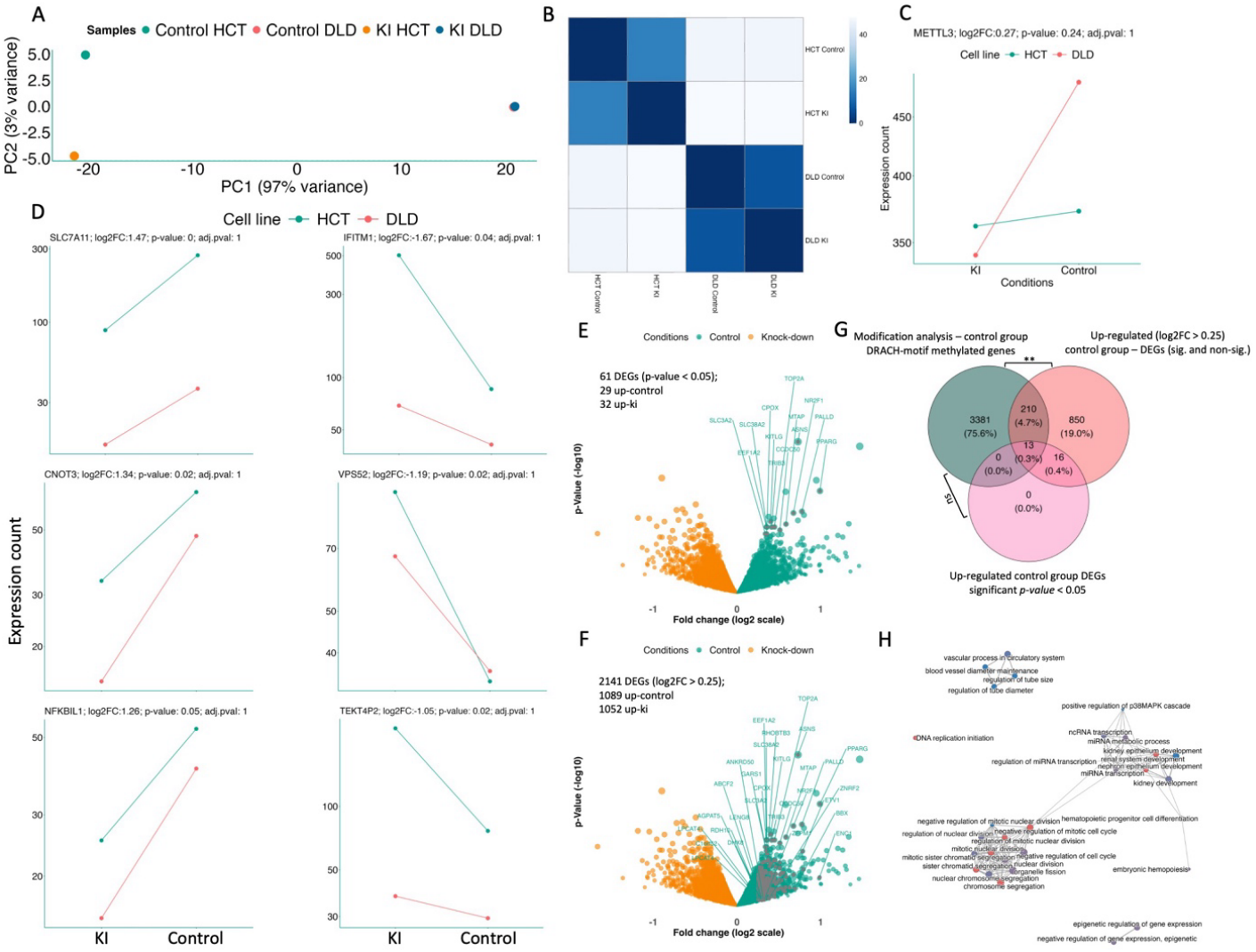


**Figure 5:** DEGs and their enrichment in genes (3604) with significant methylation at DRACH motifs in the control condition.
A) PCA on gene expression data for control and knock-down conditions in HCT116 and DLD-1 cell lines.
B) Similarity matrix of samples; darker indicates higher similarity.
C) Expression of *METTL3* across conditions.
D) Top three significant DEGs upregulated in control (left) and knock-down (right) groups.
E) Volcano plot showing 61 DEGs (p-value < 0.05); dots represent all genes tested for DE, with color for conditions and grey dots for 13 DEGs (upregulated in control) within DRACH methylated genes in the control group.
F) Volcano plot of 2141 DEGs (same analysis with log2FC >= 0.25, no p-value cutoff on DEGs); dots represent all genes tested for DE, color represents conditions, and grey dots are 223 DEGs present in DRACH methylated genes in the control group.
G) Fisher's exact test for enrichment of control group DEGs in DRACH gene set (3604) in the control condition: 13 out of 29 DEGs (p-value < 0.05) upregulated in control group are in DRACH genes, with no significant enrichment (ns); 223 out of 1089 DEGs (log2FC > 0.25) show significant enrichment (**).
H) Significant GO terms for 223 DEGs upregulated in control and present in DRACH gene set (3604).

## 4.1 Differential Gene Expression (DGE) Analysis:

The input gene count table was filtered to retain genes with expression levels of 10 or more in each sample. Due to the absence of biological replicates, I treated the two samples from the control conditions of both cell lines as biological replicates, and similarly, the two knockdown samples as replicates. I then applied the *DESeq2* analysis pipeline to compare the control and knockdown conditions while controlling for cell line-specific effects. The *DESeq2* design formula, "~ cell + condition", was employed to test the effect of *METTL3* knockdown while accounting for the influence of different cell lines.

Figure 5A depicts the PCA projection of the samples, computed on the gene count data. As shown in the PCA plot, samples cluster due to cell line specific biology rather than similarity between the conditions. The similarity matrix (Figure 5B) also agrees with the PCA observation. The DGE analysis was conducted on 11092 genes across four samples, comparing two control samples with two knockdown samples. *DESeq2* returned statistical metrics, including fold change (log2 scale), p-value, and adjusted p-value, for all genes. However, only 61 genes showed significant differential expression based on a p-value cutoff of 0.05 (Figure 3D and 3E). When accounting for false discovery rate (FDR < 0.1), no gene had an adjusted p-value below 0.1. The lack of biological replicates greatly impacts the reliability of these significance values, making it difficult to draw definitive conclusions. Additionally, despite controlling for cell line effects, inherent cell line-specific factors may still influence the results. For example, comparing conditions across cell lines can mask the differential expression signal at the gene level due to differences in transcript expression within the same gene across samples.

In this DE analysis, we focused on the 61 genes that exhibited significant differential expression (p-value < 0.05), with log2-fold changes ranging from 0.31 to 1.668. Among these, 29 genes were upregulated in the control condition, while 32 were upregulated in the knockdown condition. Figure 5D shows expression level of top three significant DEGs across the conditions. Notably, *METTL3* was not identified as a DEG, with a fold change of 0.27 (p-value = 0.24), as shown in Figure 5C. The DE signal for *METTL3* was more pronounced in the DLD1 cell line compared to the HCT cell line, but the overall difference in expression counts masked the differential signal of the gene. This masking effect due to cell-line was observed for many genes.

## 4.2 Enrichment of DEGs with DRACH Motifs:

To assess the reliability of the differentially expressed genes (DEGs) identified in the control group, I compared them to the 3604 genes identified in the control group that had a higher methylation rate at the DRACH motifs (as described in Section 3.5). Of the 29 DEGs, 13 are present in the set

of genes containing methylated DRACH motifs in the control condition (Figure 5E and 5G). This is noteworthy, as this small subset suggests an association between methylation rate and expression level, indicating that changes in methylation could potentially influence gene expression.

To further evaluate this association, I conducted Fisher's exact test for overrepresentation to determine if the DEGs in the control group are enriched in the set of genes with methylated DRACH motifs (Figure 5G). The test yielded a p-value of 0.1, which, while not statistically significant, does hint at a relationship between *METTL3* knockdown and alterations in m6A methylation patterns at DRACH motifs, which could subsequently impact gene expression. Notably, in the knockdown (KI) group, none of the DEGs overlapped with the methylated DRACH-containing gene set. This might suggest that while *METTL3* knockdown alters methylation at DRACH sites, these changes translate to expression differences but may not always —though this requires further validation with a more statistically robust analysis and a larger gene set.

To explore this hypothesis further, I conducted an additional exploratory analysis on the DEGs identified in the control condition, disregarding significance levels and including all p-values (Figure 5F). This analysis identified 1089 DEGs with a fold change (log2-scale) of 0.25 or more. Among these, 223 DEGs, including *METTL3*, were present in the list of genes with methylated DRACH motifs identified in the control condition (Figure 5F and 5G). Fisher's exact test for overrepresentation produced a p-value of 2.148e-15, indicating significant enrichment of the control group DEGs in the list of genes with methylated DRACH motifs (Figure 5G). GO analysis results of the 223 DEGs are presented in Figure 5H. This exploratory analysis underscores the importance of including replicates, demonstrating that the calculation of test statistics heavily relies on the number of replicates.

## 4.3 DGE Analysis on Simulated Replicates:

To determine if the sequencing data and count tables contain sufficient information to support our hypothesis, I conducted an exploratory differential gene expression (DGE) analysis using simulated replicates of the DLD1 cell line.

For the DLD1 cell line, we had two samples representing two conditions: control and *METTL3* knockdown. Genes with detectable expression in both samples were retained. For each sample, I generated two simulated replicates (as biological replicates) by permuting the expression levels of 50% of the genes, introducing a standard deviation of 0.5. This resulted in three replicates (two simulated and one original count) for each condition of the DLD1 cell line (Figure 6A).

The DGE analysis using *DESeq2* was performed on 12,001 genes, comparing the three control samples to the three knockdown samples of the DLD1 cell line. This analysis identified 699 genes with significantly differential expression (log2FC ≥ 0.25; adjusted p-value < 0.05) in the control condition. Of these 373 genes, *METTL3* being one of these significant DEGs (Figure 6B).

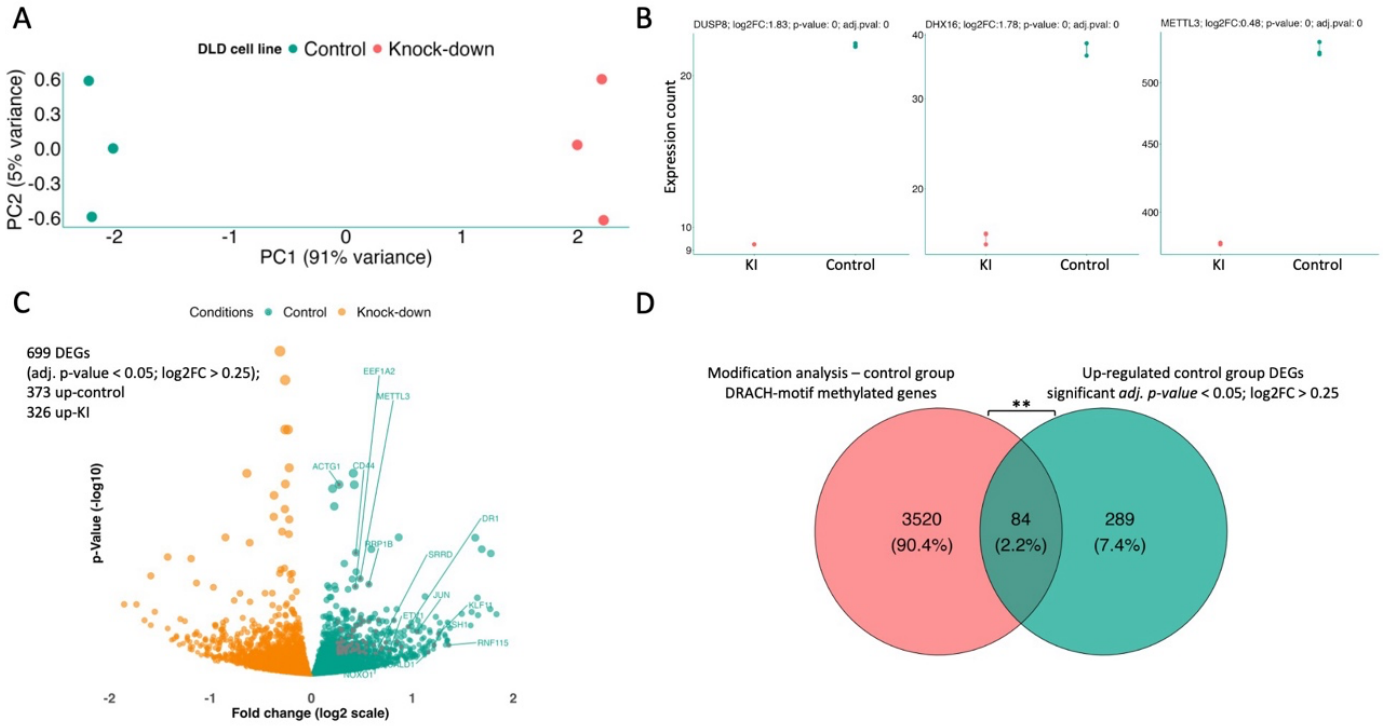


**Figure 6:** Significant DEGs (adjusted p-value < 0.05) from the simulation study of DLD-1 cell line and their enrichment in DRACH genes (3604) in the control condition.
A) PCA on gene expression data for six samples: three control and three knock-down of DLD-1, with two replicates simulated from original counts.
B) Three significant DEGs upregulated in control group.
C) Volcano plot of 699 DEGs (adjusted p-value < 0.05); dots represent all genes tested for DE with size indicating p-value, color for conditions, and grey dots for 84 control-DEGs in DRACH methylated genes (3604) in the control condition.
D) Fisher's exact test for enrichment in DRACH gene set: 84 out of 373 DEGs upregulated in control are in DRACH genes (3604 genes in the control condition), showing significant enrichment.

Next, I compared these 373 DEGs to the 3604 genes (Section 3.5, Figure 3A) with a higher methylation rate at DRACH motifs in the control condition. 84 of the 373 DEGs overlapped with the methylated DRACH-containing gene set identified in the control condition (Figure 6C). Fisher's exact test for enrichment produced a p-value of 0.0004 (Figure 6D), indicating significant enrichment of the control group DEGs in the list of genes with methylated DRACH motifs. This

analysis reinforces the importance of replicate inclusion when calculating test statistics for expression counts.

### 4.5 Differential Transcript Usage (DTU) Analysis:

Detection of differential transcript usage (DTU) from RNA-seq data is a crucial bioinformatic analysis that complements differential gene expression (DGE) analysis. Nanopore direct RNA-sequencing enables the capture of full-length transcripts, and bioinformatics tools like *Nanocount* generate counts for these transcripts.

RNA-seq experiments can be analyzed to detect differences in total gene expression across sample groups, which represents the total expression produced by all isoforms of a gene, as well as differences in transcript isoform usage within a gene. DTU is biologically relevant even when total gene expression does not change significantly, as shifts in the expression of isoforms can have functional consequences. DTU is common when comparing expression across different cell types. For example, recent analyses of the Genotype-Tissue Expression Project (GTEx) dataset revealed that half of all expressed genes contain tissue-specific isoforms (Reyes & Huber, 2018). DTU can lead to functionally different gene products through alternative splicing, changes to the coding sequence, and variations in untranslated regions (UTRs), which may affect RNA binding protein sites. Notably, alternative usage of transcription start and termination sites was found to be more prevalent than alternative splicing in DTU events across tissues in GTEx. DTU patterns have been implicated in various diseases, including cancer, retinal diseases, and neurological disorders (Scotti & Swanson, 2018). Large-scale analyses of cancer transcriptomic data have shown that protein domain losses are a common feature of DTU in cancer, especially domains involved in protein-protein interactions (Climente-González et al., 2017; Vitting-Seerup & Sandelin, 2017).

A typical DTU analysis involves two main questions: (1) Which genes show evidence of DTU? (2) Which transcripts within these genes are involved in the DTU?

In this study, I performed DTU analysis using *DRIMSeq* and *stageR* to identify transcripts with differential usage patterns within a gene across the conditions. Due to the absence of biological replicates for the conditions of the two cell lines, I treated the two samples in the control condition of both cell lines as biological replicates and similarly treated the two knock-down samples as replicates. The transcript count table contains 206723 rows (transcripts) and four columns (samples). I filtered the count table to retain transcripts that met the following criteria: (1) a count of at least 10 in at least two samples, (2) a relative abundance proportion of at least 0.1 in at least two samples, and (3) a total count of at least 10 for the corresponding gene across

all four samples. These filters are important to consider carefully, as they can remove biologically relevant transcripts, especially those with lower expression changes. For instance, a transcript that increases from 0% to 9% of a gene's expression would be excluded by the 10% filter. After filtering, 5978 genes remained in the dataset.

I applied the *DRIMSeq* analysis pipeline to the transcript count table to compare the control and knock-down conditions while controlling for cell line-specific effects. *DRIMSeq* allows for DTU analysis by modeling any fixed-effects experimental design. For this study, I used the formula "~ cell + condition" to test the effect of *METTL3* knock-down while controlling for cell line differences.

*DRIMSeq* generates a p-value per gene to assess whether there is differential transcript usage within the gene and a p-value per transcript to evaluate whether the proportions of a transcript change within its gene. This DTU analysis identified 110 genes showing differential transcript usage and 287 transcripts with changes in their proportions within their genes. These results had p-values of less than 0.05; however, only two genes had adjusted p-values below 0.05. Due to the lack of biological replicates, we cannot rely heavily on adjusted p-values. Figure 7B illustrates the estimated proportions for three significant genes, showing evidence of isoform switching. Figure 7A displays the estimated transcript usage proportions for the *METTL3* gene, which was not identified as significant.

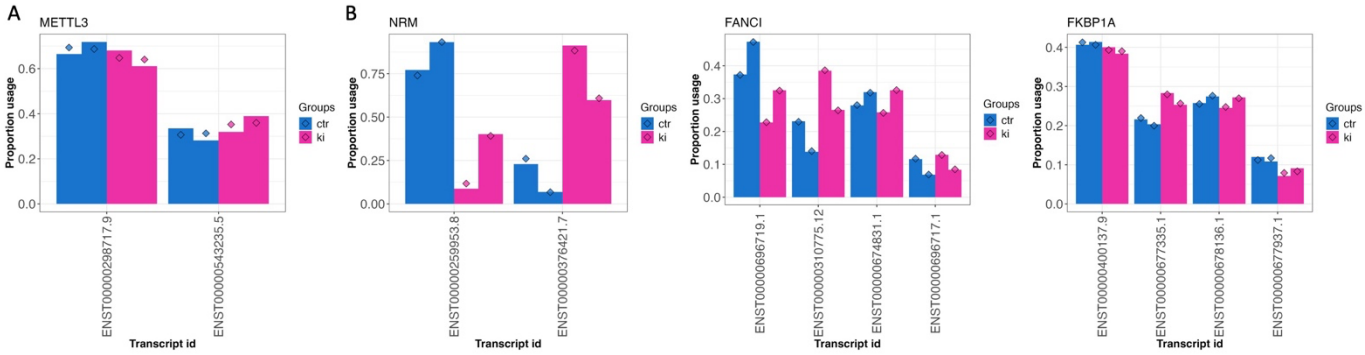


**Figure 7:** Differential transcript usage (DTU) across conditions.
A) Proportion of transcript usage for *METTL3* across control and knock-down conditions.
B) Top three genes showing varying proportions of transcript usage across conditions.

This analysis provides an overview of DTU and its application, but the results could be substantially improved with the inclusion of biological replicates, as discussed in the previous sections (Section 4). While I could have generated simulated replicates and performed the analysis on that data, as done in the DGE analysis, I chose to conclude this analysis with the

understanding that additional biological replicates would enhance the robustness of these findings.

# 5 Conclusion:

These analyses provide a comprehensive overview of how differential gene expression (DGE), differential transcript usage (DTU), and RNA modification analyses complement each other to elucidate the biological effects underlying *METTL3* knock-down. Our primary aim was to quantify and explain the impact of *METTL3* knock-down on m6A methylation signatures. The results indicate that m6A modification patterns on RNA are significantly altered due to *METTL3* knock-down, with a notable increase in the number of methylated sites observed in the knock-down cell lines. However, within DRACH motifs, the methylation rate is markedly higher in the control condition, suggesting that *METTL3* knock-down drastically reduces methylation at these specific sites. This observation may imply that *METTL3* could have multiple modes of action, functioning as both an activator and repressor. Additionally, the overall increase in methylation in the knock-down setup suggests a potential role for RNA secondary structure in determining m6A RNA modifications.

We also identified an association between methylation patterns and differential gene expression (DGE) signatures using *DESeq2* on gene count data (Section 4.2). The initial DGE analysis between conditions (two control and two knock-down samples) revealed 61 significant differentially expressed genes (DEGs), with 29 upregulated in the control condition and 32 in the knock-down condition. Of the 29 control-group DEGs, 13 were found to be present in the set of genes containing methylated DRACH motifs in the control condition. While Fisher's exact test for overrepresentation yielded a p-value of 0.1—likely due to the small number of genes and lack of biological replicates—an expanded analysis that included all DEGs (regardless of p-value) identified 223 DEGs, including *METTL3*, in the DRACH motif-containing gene list. This analysis demonstrated a significant enrichment of control group DEGs within the DRACH motif gene set, suggesting a clear association between *METTL3*-mediated m6A modification and gene expression.

This association was further confirmed through a simulation study of the DLD1 cell line (Section 4.3), where three replicates per condition were compared. In this analysis, 373 significant DEGs (adjusted p-value < 0.05) were identified in the control condition, with 84 of these DEGs present in the methylated DRACH motif gene set. Fisher's exact test for enrichment confirmed significant overrepresentation of these DEGs within the DRACH motif gene set, reinforcing the conclusion

that *METTL3* knock-down significantly alters m6A methylation patterns, which in turn affects gene expression.

In summary, *METTL3* knock-down significantly reduces methylation at DRACH motifs while increasing methylation at non-DRACH sites. These changes in m6A methylation signatures can be captured and further understood through complementary differential gene and transcript expression analyses. The positive correlation between methylation rate and expression pattern in DRACH motif-containing genes highlights the potential of *METTL3* and other genes, identified as common between DEGs and higher methylation signatures, to elucidate the molecular networks through which *METTL3* influences methylation and the underlying biological processes.

Climente-González, H., Porta-Pardo, E., Godzik, A., & Eyras, E. (2017). The Functional Impact of Alternative Splicing in Cancer. *Cell Reports*, *20*(9), 2215–2226. https://doi.org/10.1016/j.celrep.2017.08.012

Ensembl transcriptome GRch38. (2024). *Ensembl transcriptome reference.* ENSEMBL. [(Accessed on 01 June 2024)]. Available Online: Https://Ftp.Ensembl.Org/Pub/Release-112/Fasta/Homo_sapiens/Cdna/.

Gleeson, J., Leger, A., Prawer, Y. D. J., Lane, T. A., Harrison, P. J., Haerty, W., & Clark, M. B. (2022). Accurate expression quantification from nanopore direct RNA sequencing with NanoCount. *Nucleic Acids Research*, *50*(4), E19. https://doi.org/10.1093/nar/gkab1129

Jain, M., Abu-Shumays, R., Olsen, H. E., & Akeson, M. (2022). Advances in nanopore direct RNA sequencing. In *Nature Methods* (Vol. 19, Issue 10, pp. 1160–1164). Nature Research. https://doi.org/10.1038/s41592-022-01633-w

Love, M. I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, *15*(12). https://doi.org/10.1186/s13059-014-0550-8

Martinez De La Cruz, B., Darsinou, M., & Riccio, A. (2023). From form to function: m6A methylation links mRNA structure to metabolism. *Advances in Biological Regulation*, *87*. https://doi.org/10.1016/j.jbior.2022.100926

Nowicka, M., & Robinson, M. D. (2016). DRIMSeq: a Dirichlet-multinomial framework for multivariate count outcomes in genomics. *F1000Research*, *5*, 1356. https://doi.org/10.12688/f1000research.8900.1

ONT Dorado. (2024). *Oxford Nanopore Technologies Dorado.* Oxford Nanopore Technologies Dorado. 2024. [(Accessed on 01 June 2024)]. Available Online: Https://Github.Com/Nanoporetech/Dorado.

ONT Modkit. (2024). *Oxford Nanopore Technologies Modkit*. Oxford Nanopore Technologies Modkit. 2024. [(Accessed on 01 July 2024)].. Available Online: Https://Github.Com/Nanoporetech/Modkit.

Pratanwanich, P. N., Yao, F., Chen, Y., Koh, C. W. Q., Wan, Y. K., Hendra, C., Poon, P., Goh, Y. T., Yap, P. M. L., Chooi, J. Y., Chng, W. J., Ng, S. B., Thiery, A., Goh, W. S. S., & Göke, J. (2021). Identification of differential RNA modifications from nanopore direct RNA sequencing with xPore. *Nature Biotechnology*, *39*(11), 1394–1402. https://doi.org/10.1038/s41587-021-00949-w

Reyes, A., & Huber, W. (2018). Alternative start and termination sites of transcription drive most transcript isoform differences across human tissues. *Nucleic Acids Research*, *46*(2), 582–592. https://doi.org/10.1093/nar/gkx1165

Robinson, M. D., McCarthy, D. J., & Smyth, G. K. (2009). edgeR: A Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, *26*(1), 139–140. https://doi.org/10.1093/bioinformatics/btp616

Van den Berge, K., Soneson, C., Robinson, M. D., & Clement, L. (2017). stageR: A general stage-wise method for controlling the gene-level false discovery rate in differential expression and differential transcript usage. *Genome Biology*, *18*(1). https://doi.org/10.1186/s13059-017-1277-0

Vitting-Seerup, K., & Sandelin, A. (2017). The landscape of isoform switches in human cancers. *Molecular Cancer Research*, *15*(9), 1206–1220. https://doi.org/10.1158/1541-7786.MCR-16-0459

Wang, K., Peng, J., & Yi, C. (2021). The m6A Consensus Motif Provides a Paradigm of Epitranscriptomic Studies. In *Biochemistry* (Vol. 60, Issue 46, pp. 3410–3412). American Chemical Society. https://doi.org/10.1021/acs.biochem.1c00254