Ruprecht-Karls-Universität Heidelberg

Faculty of Engineering Sciences

Master Program Molecular Biotechnology

Application of the HELIOS NAD-Seq protocol to human embryonic kidney cells for purification and quantification of non-canonical NAD-capped RNA

Lukas Fesenmeier

Internship protocol

Drug research

01.08.2023 - 10.10.2023

performed at
Jäschke Lab, IPMB

supervision:

Tae Seong

Prof. Jäschke

I herewith affirm that

- I wrote this lab protocol independently under supervision and that I did not use other sources and supporting materials than those indicated
- The adoption of quotation from the literature/internet as well as thoughts from other authors are indicated in the protocol
- I have not submitted this lab protocol for any other examination

I am aware of the fact that a false declaration will have legal consequences.

__________________ __________________

place and date signature

1 Abstract

Nicotinamide adenine dinucleotide (NAD) serves as a pivotal coenzyme in cellular metabolism, ubiquitously present across living cells. Recent advancements in research have unveiled a novel dimension of NAD's involvement in cellular processes, demonstrating its presence as a 5' modification in RNA processing. This non-canonical RNA cap has emerged as a potential key player in mRNA processing, opening avenues for enhanced comprehension of cellular dynamics and promising targets for drug development.

This study delves into the exploration of the molecular functions of NAD in RNA processing by leveraging the innovative HELIOS NAD Seq protocol. The protocol exhibits robust performance in both purification and sequencing reads, as evidenced by its successful application to known RNA fragments or RNA spikes. Specifically, this investigation using human embryonic kidney cells has revealed a substantial enrichment of mitochondrial genes across all treatment conditions. Overrepresentation analysis further underscores the significance of NAD in RNA processing and splicing, highlighting its pronounced involvement in RNA splicing and processing pathways.

While this approach has provided valuable insights, it is essential to acknowledge the challenge posed by the limited number of reads aligning to the reference genome. Only 0.5% of the reads are actually from endogenous origin. Furthermore, the identified NAD-capped genes differ from those discovered through established methods like the NAD Captured Seq protocol (*Winz et al. 2017*). Therefore futher refinement of the methodology might be required to fully unravel the intricacies of NAD's role in cellular dynamics. The findings presented herein could contribute in optimizing the HELIOS protocol, which could be a potential avenues for future drug research targeting NAD-mediated pathways.

Contents

2 Introduction

2.1 RNA capping

RNA capping, traditionally associated with the addition of a 7-methylguanosine (m7G) cap at the 5' end, is a fundamental post-transcriptional modification critical for various aspects of RNA metabolism. This canonical capping process, involving a series of enzymatic reactions, is essential for mRNA stability, processing, and translation initiation. However, recent research has uncovered a whole range of non-canonical RNA capping mechanisms that challenge the conventional paradigm, including non-canonical modifications such as nicotinamide adenine dinucleotide (NAD) and flavin adenine dinucleotide (FAD), both of which are central working aspects of the RNA Modifications group in the Jäschke lab.

2.2 NAD capping and potential role

NAD, a coenzyme involved in cellular redox reactions, has emerged as an unexpected participant in the modification of RNA molecules, adding a new layer of complexity to our understanding of RNA biology. The conventional m7G cap is added through a series of enzymatic reactions involving capping enzymes, such as RNA triphosphatase, guanylyltransferase, and methyltransferase (*Ramanathan, Robb, and Chan 2016*). In contrast, NAD capping appears to be added by the RNA Polymerase II itself during transcription iniation. Because of that it is thought to be found in all organisms including eukaryotes (*Bird et al. 2018*).

The discovery of NAD capping in procaryotes stands against the previous assumption that there is no RNA modification in procaryotes, adding a new level of complexity. Additionally it was shown that NAD capping actually promotes RNA decay in vitro, suggesting a major role in gene regulation (*Jiao et al. 2017*). This observation is in complete contradiction to the previous findings on the m7G cap, which canonically increases RNA stability. It is plausible that NAD capping may have distinct effects on RNA processing and localization, offering a new dimension to the intricate network of post-transcriptional regulation. Understanding this dimension could be crucial for the development of new therapeutics potentially targeting NAD-capped nucleic acids and their associated regulatory function, making them a valuable target for drug discovery research.

2.3 Avaiable protocols

Traditional RNA isolation methods often fall short in preserving the delicate features of non-canonical modifications. However the correct identification of NAD capped RNA and the exploration of its roles in different cellular contexts is crucial for deciphering the functional significance of this unconventional modification. To achieve such identification and isolation of NAD-capped RNA current methods involve a combination of innovative biochemical and molecular biology techniques designed to capture and characterize this unique subset of modified RNA molecules. One of the leading methods for purification and quantification of NAd-capped RNA is currently the NAD Captured Seq protocol developed and published by the Jäschke Lab in 2017. This method relies on the enzymatic transfer of an 'clickable' alkyne moiety to NAD capped RNA. After that fluorescent dyes or a functional biotin group can be added with a catalyzed azide-alkyne cycloaddition (CuAAC). The enzyme found to be capable to do such transfer is the ADP-ribosylcyclase (ADPRC) from *Aplysia california*. After adding a biotin group with CuAAC the RNA fragments are then purified with streptavidin beads. Next a library can be prepared to identify and quantified the RNA with Illumina sequencing (*Winz et al. 2017*).

However the NAD captured Seq protocol requires a considerably high amount of input material from time and cost consuming cell-cultures. Additionally the application of the protocol to several samples is quite time consuming. To overcome such inconveniences Marvin Möhler from Jäschke Lab has developed the HELIOS NAD-Seq protocol. This protocol directly attaches Picolylamin (PA) or Picolylamin biotin (PAB) to the RNA fragments instead of an alkyne moiety. In this protocol barcoded 3' prime adaptors are added before capturing with the streptavidin beads which enables multiplexing of multiple samples reducing the amount of work and time needed for purification. For example in an experiment with four treatment conditions all RNA samples can be processed at the same time instead of starting a new processing round for each treatment conditions, saving about 75% of the workload. The protocol was already successfully applied to *E. coli*.

2.4 Aim and Purpose of the experiment

During the internship, I was responsible for running the experimental workflow in the lab, which took about two weeks. After that, I developed a bioinformatics pipeline to process the reads and read-count tables for the rest of my time. Everything was done under the supervision of Tae Seong, a PhD student in the lab. I would like to express my sincere gratitude to him for allowing me to participate in his work, it was a pleasure.

More specifically he is interested in the effects of several treatment conditions on NAD capping profiles. For that he treated human embryonic kidney cells in four different condition. First, with a reduced form of nicotinamide riboside (NRH), which has been shown to be an effective NAD precursor. Second, with FK866, which is known as a nicotinamide phosphoribosyltransferase inhibitor. Third, with rotenone, which is also known for its inhibitory effect on nicotinamide phosphoribosyltransferase. And fourthly, a wild type without any treatment. The same experiment was already performed with the NAD captured Seq protocol with the same starting material. The goal of applying the HELIOS Seq protocol is to see if purification works successfully, to find NAD capped genes and to compare results with previous analysis to evaluate the protocol.

3 Methods

3.1 ADPRC transglycosylation

Starting material from previous experiments were Trizol extracted RNA from human embryonic kidney cells 293T under 4 different treatment conditions. One group was treated with FK866, one with Rotenone, one with NRH and one without any treatment, just the wild type as control. Triplicates were used for statistical reasons. From each sample, 2 µg RNA was taken twice for the ADPRC reaction, once for the positive reaction and once for the negative control. 1 fmol of internal standards were added to each sample and reaction. Both were mixed in milliQ Water resulting in the 10µL for each sample. The internal standards were known sequences with certain 5'-modifications. More specifically, the internal standard consisted of biotinylated RNA (62bp), NAD-capped RNA (36bp, 60bp, 200bp, 251bp and 501bp), triphosphate-capped RNA (51bp, 102bp, 149bp and 301bp) and m7G-capped RNA (122bp, 250bp and 400bp).

In the final reaction, 15mM 3PAB or 3PA, 38nM ADPRC, 1 µg RNA and 1fmol internal standard are available in 100µL reaction volume, buffered with a 1x ADPRC buffer that can be prepared 5x with 25mM HEPES, 25mM EDTA at pH 7. Each reaction was incubated at 37℃. After 16 minutes, the reaction was quenched with 300 µL PCI solution and 200 µL mQ. PCI extraction was then performed three times by transferring the aqueous phase to a new vial after centrifuging at 13000 rpm for 2 minutes and adding new PCI each time. Two more extractions were performed with Et2O, and the organic phase was removed for further processing. RNA was precipitated by adding NaOAc to a concentration of 0.3M and help of 1 µL RNA-grade glycogen.

3.2 Pre-Bead 3' Adapter Ligation

Each of the triplets and corresponding negative (6 for each condition) received a barcoded 3' adapter enabling later sequencing and demultiplexing. To do so a Ligation Mix was made with 0.6 U/µL RNL1 and 12 U/µL RNL2 (truncated K227Q), 120 µg/mL acetylated BSA, 24% (v/v) DSMO, 14mM DTT, 3 µL RNaseOUT Recombinant RNase Inhibitor in 2.4x T4 RNA Ligase buffer. 8.5µL of this Mix was used for each sample (10µL) to ligate 2.5 µM of each adaptor (1µL) at the 3' end of the RNA of each sample (10µL). Reaction was overnight at 4℃. On the next day the Ligase was heat-inactivated.

3.3 Bead purification and Reverse Transcription

50 µL magnetic hydrophilic streptavidin magnetic beads were prepared 4 times. Since they are magnetic, solutions covering the beads can be easily removed and replaced. The beads were placed in PCR tubes and washed twice with immobilization buffer (50 µL). The beads were then blocked with 100 µg/mL acetylated BSA and washed two more times. After adding 20 µL of immobilization buffer to the ligated RNA samples, they were pooled. Each pool consisted of 3 positive and 3 negative samples with different barcoded 3'-adaptors for later demultiplexing, resulting in 4 pools, one for each condition. The pools were then incubated for 5 minutes with the streptavidin beads for purification. The supernatant was then removed and the beads were washed three times. The first time with 50µl immobilization buffer, the second time with 100µl high salt buffer (1M NaCl, 20mM TrisHCl, pH 7.4), the third time with 100µl low salt buffer (0.1M NaCl, 20mM TrisHCl, pH 7.4).

Each of the pools was reverse transcribed into cDNA while still attached to the beads. For each pool, 20 µL of RT mix was used, consisting of 5 µM RT primer, 500µM dNTP mix, 5mM DTT, 20 U/µL SuperScript IV Reverse Transcriptase and 0.5 RNaseOUT Recombinant RNase Inhibitor. An incubation time of 10 minutes at 55℃ was required for reverse transcription. Then 20µl 2x immobilization buffer was added and incubated at 20℃ for 5 minutes. The pellets were washed three more times, first with 100µL immobilization buffer, then 100µL high salt buffer and finally 100µL low salt buffer.

To release the cDNA from the beads, 50 µL of 0.15 M NaOH was added and incubated at 55 ℃ for 15 minutes. The solution was transferred to a LoBind Eppendorf tube, as was the first wash (with 50 µL mQ) of the beads. Finally, the cDNA was precipitated with 11 µL NaOAc, 1 µL glycogen and 2.5 volumes of absolute EtOH overnight at -20℃.

3.4 5' Adapter Ligation

Before proceeding pellets were pulled down by centrifuging at 14900 g for 60 minutes. Then supernatent was taken off and the pellet washed twice with 70% Ethanol. Remaining Ethanol was removed with the speedvac in 3-5 minutes, so the pellet was resolved in 10 µL mQ. To increase ligation efficiency TdT-tailing was done prior to Ligation with 0.5mM CTP, 0.5 U/µL TdT enzyme in TdT buffer in 10 µL. The 20µL reaction volume was then incubated at 37℃ for 20 minutes and then heat-inactivated at 70℃ for 10 minutes.

To ligate an adaptor a cDNA anchor mix (sense and antisense oligo) was used at 5µM concentration together with 10µM ATP and 1.5 U/µL T4 DNA ligase in T4 DNA ligase buffer, incubated at 4℃ overnight for each sample. Reaction was stopped with 60µL of mQ to reduce viscosity and by incubation at 65℃ for 10 min. cDNA was precipitated with 11 µL NaOAc, 1µL glycogen and 2.5 volumes of EtOH.

3.5 PCR

The resolved cDNA is now amplified by PCR. For this purpose, 0.02U/µL Q5 HotStart High-Fidelity DNA Polymerase 2, 0.2 mM dNTPs, 2µM universal primers, 2µM indexed primers are mixed with the ligated cDNA in Q5 reaction buffer (50µL). An additional index primer is added to each pool. The following PCR protocol is used: Initial denaturation at 98℃ for 40s, then denaturation at 98℃ for 10s, annealing at 69℃ for 20s, extension at 72℃ for 30s, repeating the last 3 steps 8 times. The annealing temperature is then reduced to 66℃ to prevent non-specific binding, and the entire cycle is repeated a further 16 times before ending with the final extension step at 72℃ for 10 minutes. A total of 24 PCR cycles were performed. After the PCR 6µL of NaOAc (3M, pH 5.5) were added to each tube and transferred to lowbind eppendorf vial. Another 2.5 volumes of absolute ethanol were added and the vials stored overnight.

3.6 PAGE purification

cDNA was seperated using PAGE purification. A 10% PAGE Gel was cast. The gel was run in 1x TBE buffer for 1:15h at 20W until the bromphenol blue runned out. For further processing the gel was incubated in TBE with 5 µL SYBR Gold for 5 minutes and analyzed via Typhoon using the SYBR Gold channel with a voltage of 500 Volt. The image was printed and used to cut the gel for everything greater than 150bp. The cut-out bands were crushed with a centrifuge by appling centrifugal force to a vial with a small hole and then the DNA was eluted with 2.4mL NaOAc from the gel overnight at 20℃. DNA was precipitated on the next day. After that it was washed with EtOH twice and dried with a speedvac.

3.7 Sequencing

The concentration of each pool was measured using Qubit and the length distribution of each pool was measured using Bioanalyzer High Sensitivity DNA Analysis Kit from Agilent and a High Sensitivity DNA analysis chip. To estimate the size of the DNA fragments in each pool the median size was taken. With the length and concentration available, the volumes of each pools were calculated to have an equal allocation in the final vial sent for sequencing. A final volume of 27 µL and a concentration of 1.64nM for each pool was chosen.

The DNA was sequenced by the Deep Sequencing Core Facility using the Illumina NextSeq 550. A mid output length of 75 was chosen and reads were sequenced using paired end sequencing. Reads were demultiplexed by the sequencing facility using the different Illumina index-primers. For each Condition four .fastq files were obtained, one for each Lane in the flow chamber of the sequencing machine.

3.8 Hardware and code availability

Computational expensive steps in the analysis were done on a computational cluster. The author acknowledges support by the state of Baden-Württemberg through bwHPC. Software and version management was done with Anaconda. The Software used in this project is available with help of the corresponding .yml file, which can be used to recreate the software environment with anaconda. All scripts, custom software and the environment file can be found on `https://github.com/fesel2/HeliosNAD_cappedRNA_Seq_MasterInternship.git`.

3.9 Processing of the raw reads

Quality of the individual Lanes was assessed with FastQC (*Andrews et al. 2012*). After that lanes were merged. Cutadapt was used for demultiplexing and trimming of the barcoded 3' adaptors on both sides (*M. Martin 2011*). Trimmomatic was used for the 5' adaptors and also the sliding window function to discard low quality reads and reads smaller than 24 nucleotides (*Bolger, Lohse, and Usadel 2014*). Adapter-content was verified using FastQC again. After that paired end reads with a minimum overlap of 15 nucleotides were merged using Flash to unpaired reads (*Magoč and Salzberg 2011*).

3.10 Alignment and feature generation

The reads were aligned to the human genome Release 44 (GRCh38.p14) gathered from the GENCODE project (*F. J. Martin et al. 2023*). Additionally the reads were aligned to the known sequences of the spike-RNA. Both alignment steps were accomplished using HISAT2 (*Kim et al. 2019*) with default settings. Prior to alignment HISAT2 required to built indices for each reference sequence. The logfiles of the alignment to the spikeRNA were fed into a custom python script (available on: `https://github.com/fesel2/HeliosNAD_cappedRNA_Seq_MasterInternship.git`) to generate alignment rates of each spikeRNA. On the other hand the alignment files to the reference genome were further processed using samtools for file conversion and indexing (*Danecek et al. 2021*). With indexed alignment files available, genomics features were extracted using featureCounts resulting in count matrices for downstream processing (*Liao, Smyth, and Shi 2014*).

3.11 Normalization, filtering and annotation

Downstream-processing was done with help of the anndata class (*Virshup et al. 2021*). This python class was originally developed for single cell RNA sequencing analysis with the scanpy library. However it supplies a good foundation to store all kinds of data related to sequencing in one immersive data object. Further because of its unique layer architecture it allows to simultaneously access raw data and normalized data or work on the spikeRNA data within the same object. In fact, the proportion of spikeRNA in relation to the total number of biotinylated spikeRNA was used to standardize each sample to the same number of biotinylated spikeRNA (equation 1).

$$N_{normalized} = (N_{original} + 1) \cdot \frac{B_{total}}{B_{sample}} \tag{1}$$

3.12 Determination of NAD capped RNA

In the original paper enrichment factors were used to determine whether RNA is NAD-capped or not. They took the NAD-spike with the lowest enrichment and decided every feature with higher enrichment was endogenous NAD capped RNA from the cell (*Winz et al. 2017*). However in this analysis it was shown that the NAD capped spikes are the top enriched RNA, so there is no endogenous RNA above that level. Instead statistical tests were used to determine if endogenous RNA is different to a non NAD spike. The non NAD spikeRNA measurements were pooled and considered as a reference to compare each feature's count against it with alternative greater. To test the Null Hypothesis that there is no difference between non NAD spikes and each endogenous gene triplet the non-parametric wilcoxon ranksum test and the independent t-test were applied using the python scipy library (*Virtanen et al. 2020*). Additionally the edgeR pipeline specifically developed for RNA-Sequencing data was applied for comparison. This pipeline relied on the Fisher's exact test (*Robinson, McCarthy, and Smyth 2010*). RNA was determined as endogenous NAD-capped when p-Value was below significance level $\alpha = 0.05$ and logfoldchange $LFC >= 1$ when comparing positive vs negative sample.

3.13 Adding biological meaning

Over representation analysis (ORA) was done with GSEAPY (*Fang, Liu, and Peltz 2023*). The identified NAD capped genes across all conditions and statistical tests were unionised and compared to the genesets from the database GO_Biological_Process_2021. Additionally a closer look was gained by looking at the top genes found in each treatment condition. To identify differences in treatment conditions principle component analysis (PCA) was performed using the scikit-learn library on NAD capped genes which were found in at least two treatment conditions (*Pedregosa et al. 2011*). The PCA was applied on the logfoldchange changes of each feature (positive vs. negative ADPRC reaction).

3.14 Comparing with other Datasets

In previous analysis the same raw material was treated with the original captured sequencing protocol *Winz et al. 2017*. Additionally the analysis pipeline described in this protocol was also applied to an eukaryotic A431 skincancer cell line (only wildtype, without different treatments). The sets of NAD capped RNA were intersected to find similarities and dissimilarities.

4 Results

4.1 Read composition

Biotin RNA is the spike RNA which is found the most in both experimental reactions. Generally there are less reads observed in the negative ADPRC reaction with PA. As expected NAD capped RNA spikes are mostly found in the positive ADPRC reaction when reacting with PAB. However only 0.4% of the total reads are actual DNA fragments aligning to the human reference genome and considered as endogenous RNA (figure 1). From a total of approximately 100 million reads this corresponds to only 450.000 reads. Further endogenous RNA fragments are found to be enriched at least 10 fold in the positive ADPRC sample compared to the negative. This is observed to be consistent across all samples and conditions. Triphosphate capped RNA and m7G capped RNA reads are heavenly depleted in both experimental treatments indicating a successful purification. In both conditions demultiplexed reads are found which do not align to either the spikeRNA or the reference genome. They are labeled as junk and are not considered in further results.

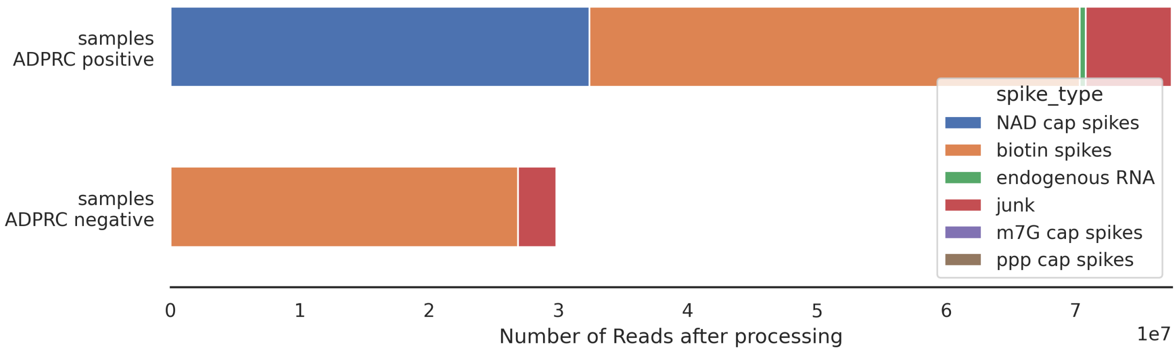


Figure 1: Read Composition: After prepossessing the majority of the reads are referring to the spikeRNAs specifically to the biotinylated RNA-spike. The recovered NAD capped RNA in the positive sample and its abstinence in the negative sample indicates that the experiment worked. Also endogenous RNA is mainly found in the positive reaction. However read count is low.

4.2 Determination of NAD capped endogenous genes

The highest differences between positive and negative sample is found for NAD capped RNA fragments. In conclusion the ratio of positive to negative sample (here referred as enrichmentfactor) indicates potential NAD capping. The enrichment factors found in this experiment vary between 10^{-2} and 10^{4}. Mostly the nonNAD spikes are found in the lower quantile while the NADspikes always are the top values. This can be observed for every sample except Rotenone 1 which is considered as an outlier or a pipetting issue.
About 5000 genetic features passed the filtering condition of having at least 10 raw reads aligning to the reference genome. Their enrichmentfactors are distributed between the NAD and non NAD spikes suggesting being a mixture of canonical and non canonical capped RNA (figure 2).

When combining results of statistical tests 202 genes were found to be significantly enriched in the Control Condition, 1955 in the FK866, 279 in NRH and 385 in the Rotenone condition. Therefore these endogenous genes were labelled as NAD capped. Mostly the statistical tests agree to each other across all conditions. However the Fishers exact test is found to be the most stringent test, identifying the least

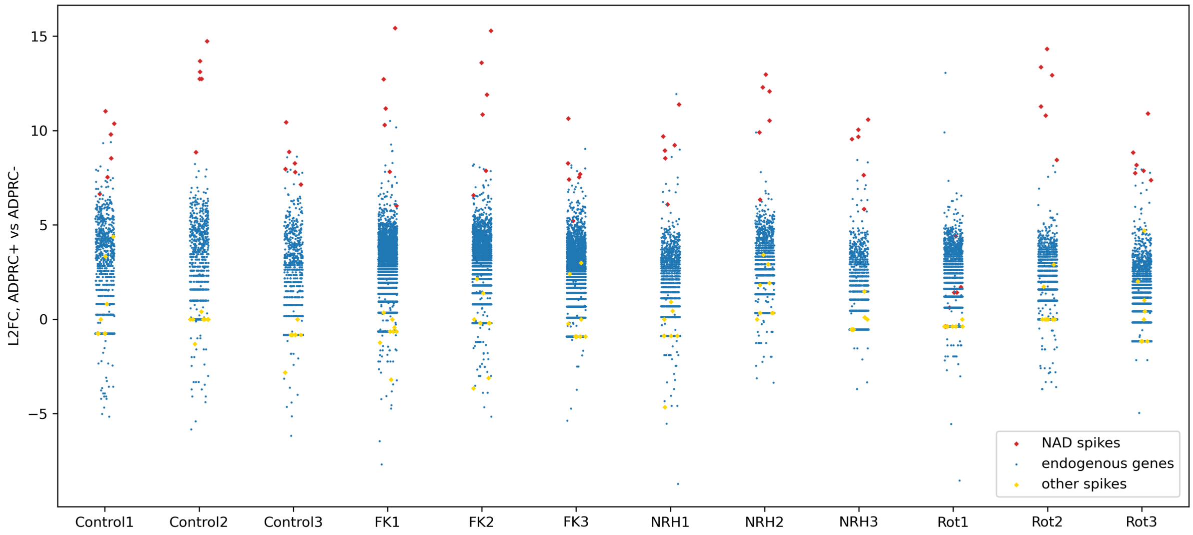


Figure 2: Logfoldchanges for each sample: The highest enrichmentrates and logfoldchanges are found for the NAD spikes for each sample. The non NAD spikes are found to have a low or no enrichment level at all. Endogenous genes are somewhat in between. One sample of Rotenone is considered an outlier because NAD capped spikes are not found in sufficient amount and not to be enriched.

amount of genes to be NAD capped. On the other hand the wilcoxon ranksum test and the independent t-test are less stringent, identifying more genes to be NAD capped (Figure 3).

Closer inspection reveals that all tests perform well at obvious results but lack in determination when one or more samples are zero or below 10 raw counts (figure 4). More specifically genes with high count (> 1000) in positive and low count (< 20) are correctly classified as NAD capped RNA by all tests. Similar observation occurs when having 0 counts for all the negative samples of a feature. However if one or two of the positive samples do not have a corresponded read assigned, statistical test results are not consistent anymore. The t-test favors assigning NAD capping when only one positive sample has sufficient count number (less consistent in regards to outliers at low sample size), referring to *ENSG00000183873* in figure 4. The wilcoxon ranksum test on the other hand is alright with lower numbers but requires more samples, referring to *ENSG00000108854* in figure 4.
In fact the choice of a different test has a significant impact on whether a feature is classified as NAD-capped or not. For simplicity reasons all genes identified by at least one of the three test are considered as NAD capped are considered as NAD capped and sets are combined.

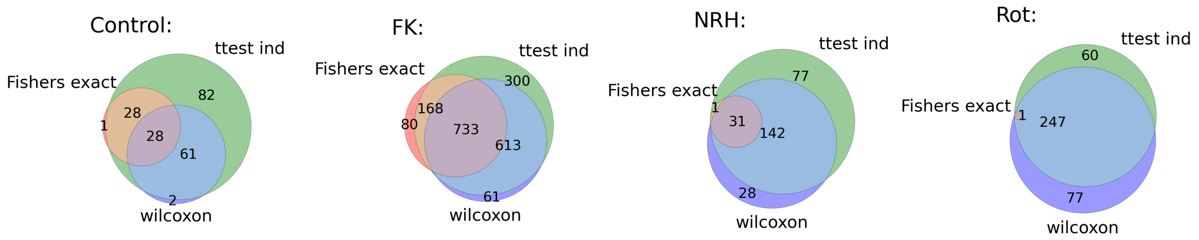
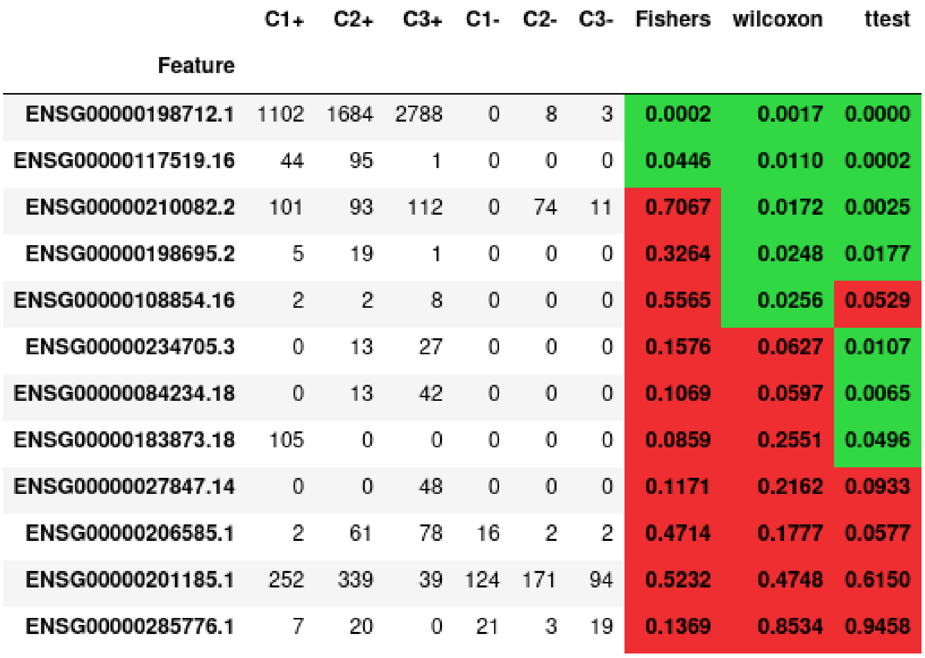


Figure 3: Set Comparison of statistical tests: Generally the NAD capped genes determined by independent t-test, wilcoxon ranksum test and Fishers exact test from edgeR agree. However differences are found. Specifically for Rotenone, where one sample was corrupted with incorrect spike-content the Fishers exact test does not yield determinable genes. The sample FK shows surprisingly a lot of NAD capped genes compared to the other treatment conditions.



	C1+	C2+	C3+	C1-	C2-	C3-	Fishers	wilcoxon	ttest
Feature									
ENSG00000198712.1	1102	1684	2788	0	8	3	0.0002	0.0017	0.0000
ENSG00000117519.16	44	95	1	0	0	0	0.0446	0.0110	0.0002
ENSG00000210082.2	101	93	112	0	74	11	0.7067	0.0172	0.0025
ENSG00000198695.2	5	19	1	0	0	0	0.3264	0.0248	0.0177
ENSG00000108854.16	2	2	8	0	0	0	0.5565	0.0256	0.0529
ENSG00000234705.3	0	13	27	0	0	0	0.1576	0.0627	0.0107
ENSG00000084234.18	0	13	42	0	0	0	0.1069	0.0597	0.0065
ENSG00000183873.18	105	0	0	0	0	0	0.0859	0.2551	0.0496
ENSG00000027847.14	0	0	48	0	0	0	0.1171	0.2162	0.0933
ENSG00000206585.1	2	61	78	16	2	2	0.4714	0.1777	0.0577
ENSG00000201185.1	252	339	39	124	171	94	0.5232	0.4748	0.6150
ENSG00000285776.1	7	20	0	21	3	19	0.1369	0.8534	0.9458

Figure 4: Decision table: Different statistical tests show different behavior. Here, the raw counts of prechosen features from the wildtype are shown. Additionally you can see the corresponding p-values of the tests. Especially if one Feature was not recovered in one sample, results vary. Note that test results were calculated using the normalized counts, not the raw counts which are shown here for simplicity reasons. Independent t-test appears to have the highest sensitivity but struggels with outliers.

4.3 Biological Meaning

NAD capped genes were found to be enriched in pathways regarding mRNA processing, splicing and protein modification. More specifically the top enriched pathways (mRNA processing GO:0006397 and splicing GO:0000377) have an overlap of 79/300 and 69/251. Examples for genes included in these pathways are DBR1, DDX42, CSTF3, MFAP1, SRSF2. However NAD capped RNAs are also found to be associated with non coding RNA, ubiquitination and regulation of the mitotic cell phase transition (figure 5).

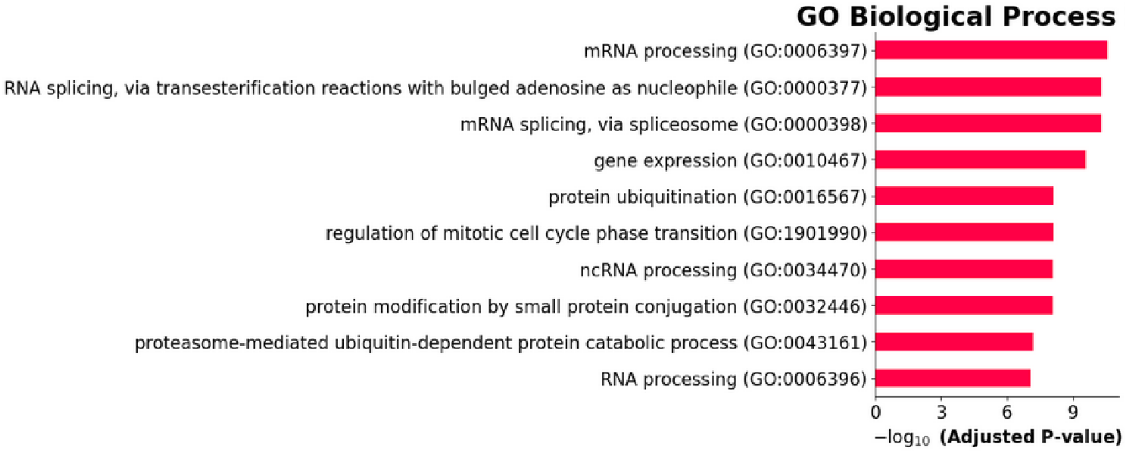


Figure 5: Over represented pathways: Here the top 10 over represented pathways can be seen. Non canonical NAD capping appears to be highly associated with mRNA processing, splicing and protein modification. The pathways are ordered by significance.

It is found that the four conditions have different NAD capping profiles. Principle Component analysis shows that samples with same treatment cluster together. It is possible to separate samples from the FK866 treatment in the first principle component and separate Rotenone treated samples in the third principle component. Both components together are explaining 37.66% of the variance in the data. However no separation is possible between wildtype and NRH treatment within that variance (figure 6).

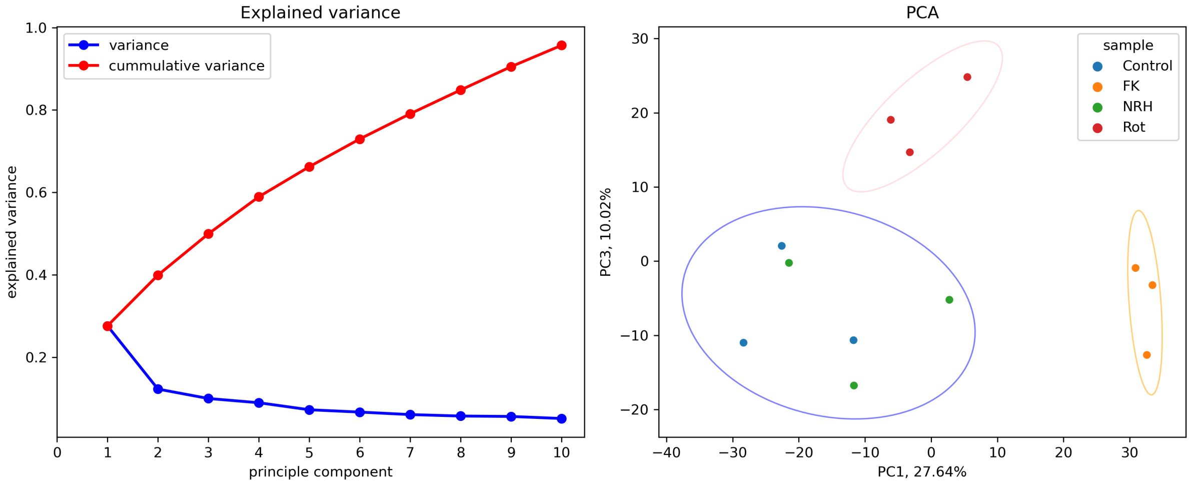


Figure 6: PCA plot: Left: curve of variance and cumulative variance for the first 10 principal components. The first three already explain more than 50%. Separation is revealed of Rotenone and FK866 within the first three components. This suggests there is a different NAD capping profile for these treatment conditions. NRH and control do not show differences here.

It is also notable that the highest significance among the NAD capped genes is found for mitochondrially encoded genes. These include short tRNAs , long non coding RNAs but also mitochondrially encoded proteins like the cytochrome c oxidase which is associated with the oxidative phosphorylation. Often features associated with mitochondrial genes have > 100 reads per positive sample which is considerable high when taking the low read coverage into account.

4.4 Overlap with previous data

When comparing previous data which was gathered by applying the NAD captured Seq protocol *Winz et al. 2017* almost no overlap is observed. For all four conditions completely new NAD capped gene sets are discovered. However the set of genes considered NAD capped in the A341 celline shows overlapping. 42 of the 202 genes are recovered in both celllines. Theses 42 genes are found to be highly significant mitochondrial genes again (figure 7). Note that overlap does not significantly increases when lowering the significance threshold (p-Value or logfldchange)

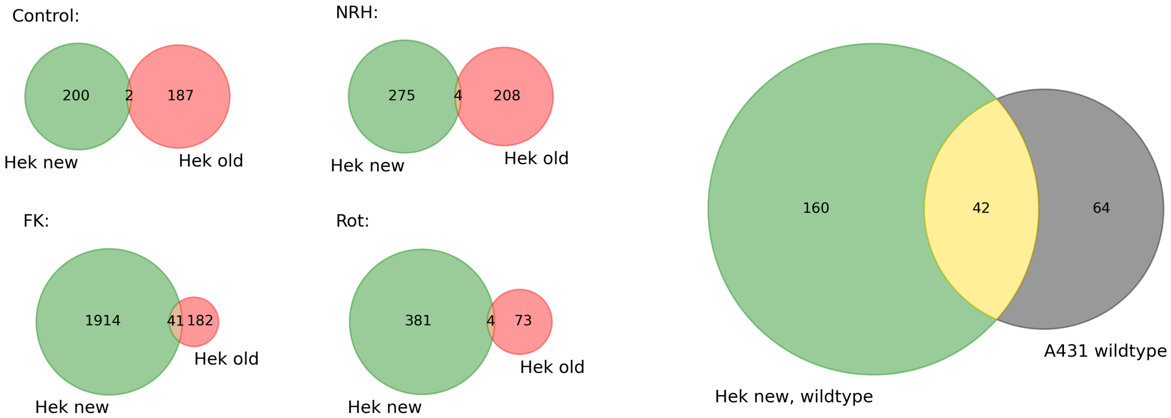


Figure 7: Overlap previous data: The overlap with the NAD Captured Seq protocol is negligible in all four treatment conditions. This is surprising since starting marterial is the same. When compared with cancer cells (the A431 cell line) that have undergone the same procedure and pipeline, a greater overlap is observed.

5 Discussion

5.1 Read composition

The purification of the NAD capped RNA works fine. As expected biotinylated RNA and NAD capped RNA are left for analysis while triphosphate and m7G capped RNA are depleted. Unfortunately the amount of reads aligning to the reference genome is very low, complicating further analysis. Reason for that might be the amount of spike RNA added in the beginning. It was previously shown that the spike RNA concentration used in this experiment works fine for purification and sequencing in *E. coli*. However eucaryotes like human embryonic kidney cells might have other posttranscriptional modifications hindering the experiment, which were not an issue previously.

Another problem might be the bias of Illumina Sequencing towards shorter fragments. Smaller fragments have an advantage during the sequencing process since they might cluster more efficiently in the flow cell and their shorter length can lead to faster sequencing cycles. Spike RNA is considerably small compared to the full RNA fragments of endogenous genes which technically favors them in the sequencing procedure. Potential solutions might be to add a fragmentation step in the protocol to reduce sizes of the endogenous RNA fragments or use a different sequencing technique. Nanopore sequencing for example relies on the entire RNA fragment passing through a pore and not on sequencing by synthesis. Additionally alignment rates might be improved by having a closer look on the alignment algorithm. Here hisat2 was run with default settings. Even though hisat2 on default settings already does exon sensitive local and global alignment some fine tuning might improve alignment rates.

5.2 Determination of NAD capping

The three statistical tests used to determine NAD capping mostly agree. However the Fishers exact test in the R pipeline does not yield much features to be statistical significant. Reason for that might be the generally low read count values or the previously performed biotin normalization which is not a standard processing step in the edgeR pipeline. Generally the independent t-test and the wilcoxon ranksum test identifies more genes to be NAD capped. The unparametric wilcoxon ranksumtest identifies low but consistent counts because it relies on the order of readcounts. However the t-test assumes a t-distribution and might be less suitable when some outliers shift the distributions. In general it seems that the wilcoxon ranksum test performs best because it requires all readcounts of the positive sample to be at least higher than the negative and is not affected by a single high readcount value in one sample.

5.3 Biological meaning

For some reason the most NAD capping was observed in the treatment condition FK866. This is surprising since FK866 is supposed to allosterically inhibit the Nicotinamide Phosphoribosyltransferase resulting in lower NAD concentration throughout the cell. There might be some bias introduced in the library preparation procedure for example all FK-samples were barcoded with the same 3' adaptor. Another indication that something went wrong is the fact that during NAD captured Seq no significant more NAD capped genes were observed in FK866 treatment.

Highest significance in determination of NAD capping in all treatment conditions was found for mitochondrial encoded genes. Previous studies found that mitochondrial RNA Polymerase caps RNA more efficiently than nuclear RNA polymerase, supporting this observation *Bird et al. 2018*. These findings indicate that NAD capping is more likely to occur at higher NAD concentration.

The over representation of NAD capped genes in mRNA processing and splicing pathways suggests that NAD capping plays a role in gene regulation. This could suggest that regulation of splicing and rna tuning proteins is fine tuned with an additional layer of faster decaying NAD capped RNA molecules. For example the cleavage stimulation factor subunit 3 (CSTF3) which is involved in polyadenylation and 3' cleavage of mRNA appears to be NAD capped and found in most of the significant over represented pathways/sets.

Dimension reduction techniques like PCA revealed that an underlying NAD capping structure can be observed across conditions. This suggests that FK866 and Rotenone have a significant impact on NAD capping in general. In further analysis genes with the highest differences could be analyzed to distinguish conditions and potentially identify marker genes for a specific treatment. However it might be recommend to fix the 'bad' read composition first, improving statistical significance and potentially recover more NAD capped genes.

5.4 Overlap with previous data

Unfortunately no or only neglectable intersection is observed when overlapping results from NAD captured Seq and HELIOS NAD Seq. Not even the highly significant mitochondrial genes which are conserved across conditions in HELIOS Seq are found in the NAD captured Seq data. Since both protocols are designed to purify NAD capped RNA fragments in an ideal scenario both protocols should yield similar results when applied to the same starting material. However this is not the case. A potential problem might be the missing Standardization. Data was generated by different persons using different methods. On the other hand there is the A431 skincancer cell data overlapping with the mitochondrial genes and some more of the wildtype indicating there is either biological ground truth found in both cell lines or technical bias of the HELIOS NAD Seq protocol towards those specific genes.

6 References

Andrews, Simon et al. (Jan. 2012). *FastQC*. Place: Babraham, UK Published: Babraham Institute.

Bird, Jeremy G et al. (Dec. 2018). “Highly efficient 5’ capping of mitochondrial RNA with NAD+ and NADH by yeast and human mitochondrial RNA polymerase”. In: *eLife* 7. Ed. by Alan G Hinnebusch and James L Manley. Publisher: eLife Sciences Publications, Ltd, e42179. ISSN: 2050-084X. DOI: 10.7554/eLife.42179. URL: https://doi.org/10.7554/eLife.42179 (visited on 11/30/2023).

Bolger, Anthony M., Marc Lohse, and Bjoern Usadel (Aug. 2014). “Trimmomatic: a flexible trimmer for Illumina sequence data”. In: *Bioinformatics* 30.15, pp. 2114–2120. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btu170. URL: https://doi.org/10.1093/bioinformatics/btu170 (visited on 11/10/2023).

Danecek, Petr et al. (Feb. 2021). “Twelve years of SAMtools and BCFtools”. eng. In: *GigaScience* 10.2, giab008. ISSN: 2047-217X. DOI: 10.1093/gigascience/giab008.

Fang, Zhuoqing, Xinyuan Liu, and Gary Peltz (Jan. 2023). “GSEApy: a comprehensive package for performing gene set enrichment analysis in Python”. In: *Bioinformatics* 39.1, btac757. ISSN: 1367-4811. DOI: 10.1093/bioinformatics/btac757. URL: https://doi.org/10.1093/bioinformatics/btac757 (visited on 11/28/2023).

Jiao, Xinfu et al. (Mar. 2017). “5 End Nicotinamide Adenine Dinucleotide Cap in Human Cells Promotes RNA Decay through DXO-Mediated deNADding”. English. In: *Cell* 168.6. Publisher: Elsevier, 1015–1027.e10. ISSN: 0092-8674, 1097-4172. DOI: 10.1016/j.cell.2017.02.019. URL: https://www.cell.com/cell/abstract/S0092-8674(17)30197-6 (visited on 11/30/2023).

Kim, Daehwan et al. (Aug. 2019). “Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype”. en. In: *Nature Biotechnology* 37.8. Number: 8 Publisher: Nature Publishing Group, pp. 907–915. ISSN: 1546-1696. DOI: 10.1038/s41587-019-0201-4. URL: https://www.nature.com/articles/s41587-019-0201-4 (visited on 11/10/2023).

Liao, Yang, Gordon K. Smyth, and Wei Shi (Apr. 2014). “featureCounts: an efficient general purpose program for assigning sequence reads to genomic features”. In: *Bioinformatics* 30.7, pp. 923–930. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btt656. URL: https://doi.org/10.1093/bioinformatics/btt656 (visited on 11/10/2023).

Magoč, Tanja and Steven L. Salzberg (Nov. 2011). “FLASH: fast length adjustment of short reads to improve genome assemblies”. In: *Bioinformatics* 27.21, pp. 2957–2963. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btr507. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3198573/ (visited on 11/10/2023).

Martin, Fergal J et al. (Jan. 2023). "Ensembl 2023". In: *Nucleic Acids Research* 51.D1, pp. D933–D941. ISSN: 0305-1048. DOI: `10.1093/nar/gkac958`. URL: `https://doi.org/10.1093/nar/gkac958` (visited on 11/10/2023).

Martin, Marcel (May 2011). "Cutadapt removes adapter sequences from high-throughput sequencing reads". en. In: *EMBnet.journal* 17.1. Number: 1, pp. 10–12. ISSN: 2226-6089. DOI: `10.14806/ej.17.1.200`. URL: `https://journal.embnet.org/index.php/embnetjournal/article/view/200` (visited on 11/10/2023).

Pedregosa, Fabian et al. (2011). "Scikit-learn: Machine Learning in Python". In: *Journal of Machine Learning Research* 12.85, pp. 2825–2830. ISSN: 1533-7928. URL: `http://jmlr.org/papers/v12/pedregosa11a.html` (visited on 11/28/2023).

Ramanathan, Anand, G. Brett Robb, and Siu-Hong Chan (Sept. 2016). "mRNA capping: biological functions and applications". In: *Nucleic Acids Research* 44.16, pp. 7511–7526. ISSN: 0305-1048. DOI: `10.1093/nar/gkw551`. URL: `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5027499/` (visited on 11/28/2023).

Robinson, Mark D., Davis J. McCarthy, and Gordon K. Smyth (Jan. 2010). "edgeR: a Bioconductor package for differential expression analysis of digital gene expression data". In: *Bioinformatics* 26.1, pp. 139–140. ISSN: 1367-4803. DOI: `10.1093/bioinformatics/btp616`. URL: `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2796818/` (visited on 11/10/2023).

Virshup, Isaac et al. (Dec. 2021). *anndata: Annotated data*. en. Pages: 2021.12.16.473007 Section: New Results. DOI: `10.1101/2021.12.16.473007`. URL: `https://www.biorxiv.org/content/10.1101/2021.12.16.473007v1` (visited on 11/10/2023).

Virtanen, Pauli et al. (2020). "SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python". In: *Nature Methods* 17, pp. 261–272. DOI: `10.1038/s41592-019-0686-2`.

Winz, Marie-Luise et al. (Jan. 2017). "Capture and sequencing of NAD-capped RNA sequences with NAD captureSeq". en. In: *Nature Protocols* 12.1. Number: 1 Publisher: Nature Publishing Group, pp. 122–149. ISSN: 1750-2799. DOI: `10.1038/nprot.2016.163`. URL: `https://www.nature.com/articles/nprot.2016.163` (visited on 11/10/2023).