Signal level RNA modifications detection in eukaryotic non-coding RNAs

17

EMBL EBI

ISTITUTO ITALIANO DI TECNOLOGIA

Gurdon INSTITUTE

STORM THERAPEUTICS


Adrien Leger @AdrienLeger2


Tommaso Leonardi @tom_leon


Paulo Amaral @ppamaral

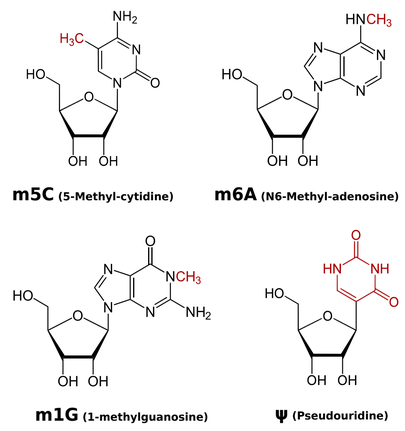
Harvey Che, Luca Pandolfini, Helena Santos-Rosa,
Anton Enright, Tomas Fitzgerald, Ewan Birney, Tony Kouzarides

Eukaryotic transcriptomes contain dozens of covalent RNA post-transcriptional modifications but remain largely uncharted, in particular, in non-coding RNAs (ncRNAs). ONT direct-RNA sequencing signals respond to RNA modifications although the available tools for signal-level detection are still limited in sensitivity and scope.

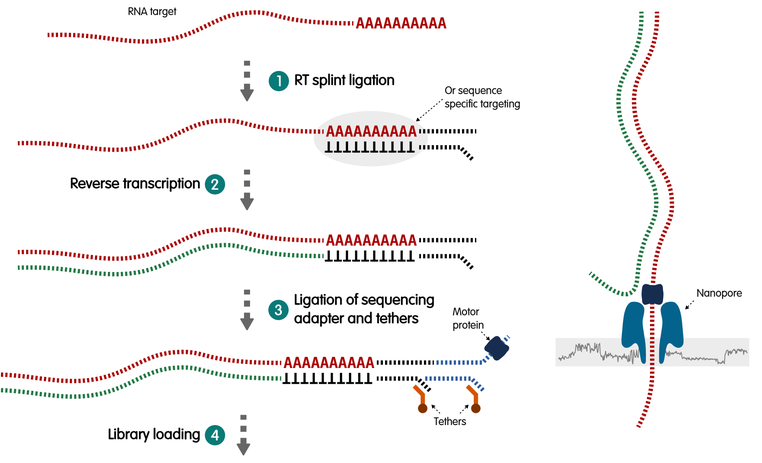
We generated dRNA-Seq datasets from human, mouse and yeast samples, including conditions where several known epitranscriptomic writers were knocked-down or -out. In addition, we in-vitro synthesized control transcripts for selected ncRNAs of interest

To perform comparative analyses of our datasets we developed Nanocompore, a program downstream of Nanopolish that compares samples at the signal level and identifies significantly altered positions corresponding to putative RNA modification sites.

Examples of RNA modifications



Oxford Nanopore direct RNA sequencing



Direct-RNA sequencing datasets

Control condition
(low or no modifications)
↳Replicate 1
↳Replicate 2
↳Replicate ....

Test condition
(wild type)
↳Replicate 1
↳Replicate 2
↳Replicate ...

Nanocompore pipeline nextflow

nanocompore

Automated data pre-processing workflow for multiple samples, including quality control.

Guppy
RNA basecalling

Minimap2
Transcriptome alignment
+
Samtools
reads filtering

Nanopolish
Events alignment
+
NanopolishComp
kmer level collapsing and indexing

Basecalling quality control

pyco QC

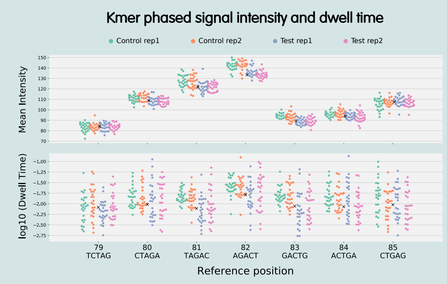
Alignment quality control
(under development)

Compares 2 ONT dRNA sequencing datasets from different experimental conditions expected to have a significant impact on the epitranscriptome. Replicates aware framework with continuous integration testing.

Whitelist
Select reads mapped on transcripts with sufficient coverage

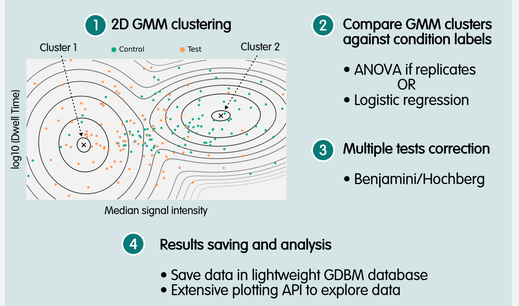
Position-wise data aggregation

Intensity and dwell time values are collected at transcript position level. The replicate structure is conserved



Statistical pairwise comparison

- Mann-Whitney, Kolmogorov-Smirnov or T test
- Gaussian mixture model clustering
- Complex statistical design (multi-factor, batch effects ...)

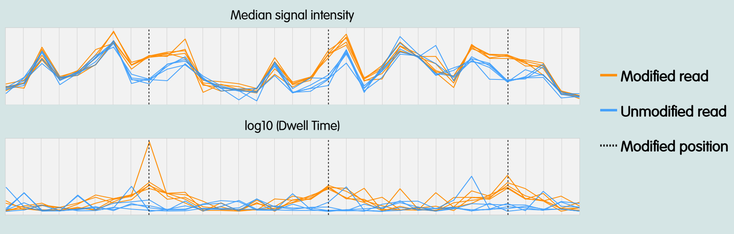


Nanocompore benchmarking using in silico generated datasets

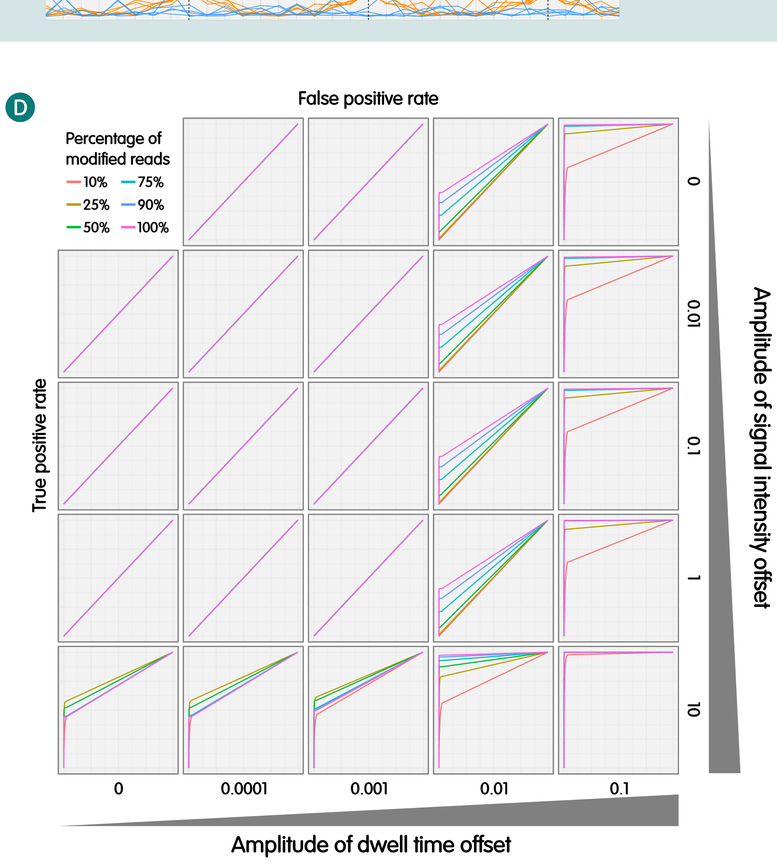
A kmers model
From an in vitro generated non-modified RNA dataset, find the best distribution to fit each kmers for both median intensities (Logistic) and dwell time (Wald)

B Artificial reference
Set of 2000 semi-random references 500 bases long. Covers all 5-mers (median: 970 times per kmer) and 99.67% of the possible 9-mers (median: 4 times per kmer)

C In silico modifications simulation
100 reads per reference using the probability density functions of the kmer models. Simulation of RNA modification effect on signal by offseting the model density function.

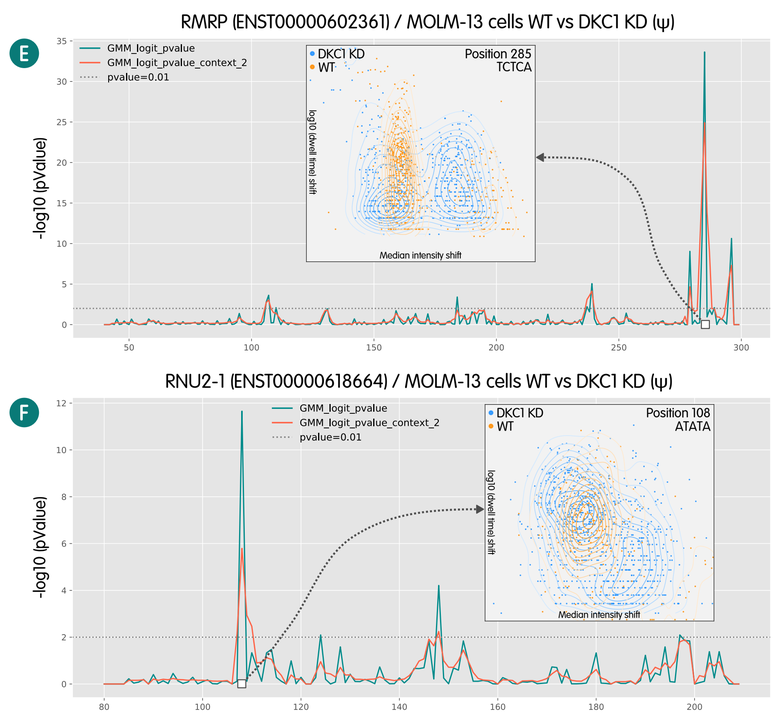


D



D) ROC Curves showing Nanocompore ability to predict the modified sites according to the amplitude of change of signal intensity and dwell time (deviation from unmodified model). For each conditions the percentage of modified reads in the test condition was modulated from 10 to 100%

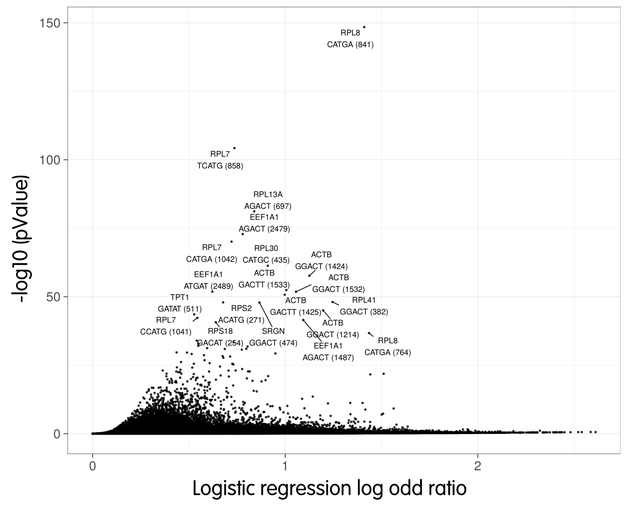
Example of non-coding RNAs targeted sequencing



Distribution of probabilities of the signal to be significantly different between the 2 conditions across all transcript positions with sufficient coverage. All conditions were obtained in MOLM-13 cells in 2 biological replicates, except for the IVT 7SK. **E,F)** Dyskerin gene (ψ writer) knock-down targeting RMRP and RNU2-1 genes. **G)** METTL3 (m6A writer) knock-down targeting RN-7SK gene. **H)** In vitro synthesized 7SK transcript.

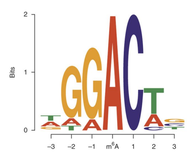
PolyA transcriptome analysis of METTL3-dependent m6A modification sites

I Top candidate kmer sites found by Nanocompore



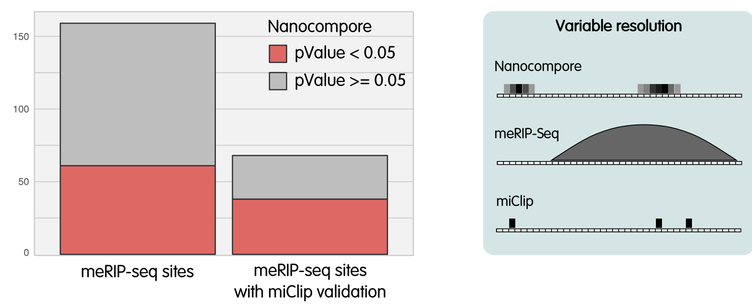
J De novo motif enrichment analysis in top 100 hits

Rank	Motif	P-value	log P-pvalue	% of Targets
1	GGACT	1e-2	-5.971e+00	22.45%
2	ACATG	1e0	-1.172e-01	6.12%



Dominissini et al.
Nat Prot vol 8, (2013)

K Orthogonal METTL3-dependent m6A mapping methods



I) Sharkfin plot of Nanocompore results on the polyA transcriptome of WT MOLM-13 cells vs METTL3 KD. The plot shows the logistic regression p-value plotted against the ratio between WT and KD of the odds of reads belonging to any one of the two GMM clusters. **J)** Results of the Homer motif enrichment analysis on the top 100 most significant kmers (+/- 5nt). Right is the known consensus motif for METTL3. **K)** Number of significant m6A meRIP-Seq peaks (Barbieri et al.) including peaks with at least one significant Nanocompore hit (red). The right bar only shows meRIP-Seq peaks containing at least one significant m6A miCLIP peak (Vu et al). The analysis was limited to transcripts with sufficient coverage (30x) in nanopore sequencing data.

EMBL European Bioinformatics Institute, Wellcome Genome Campus, Hinxton, Cambridge, CB10 1SD, UK / Tel. +44 (0) 1223 494 444 / comms@ebi.ac.uk / www.ebi.ac.uk