DeePlexiCon

Molecular barcoding of native RNAs using nanopore sequencing and deep learning

Martin A Smith[1,2,3,4]#, Tansel Ersavas[1]#, **James M Ferguson**[1]#,
Huanle Liu[1,5], Morgan C Lucas[1,5,6], Oguzhan Begik[1,4,5], Lilly Bojarski[1], Kirston Barton[1,4], Eva Maria Novoa[1,4,5,6]

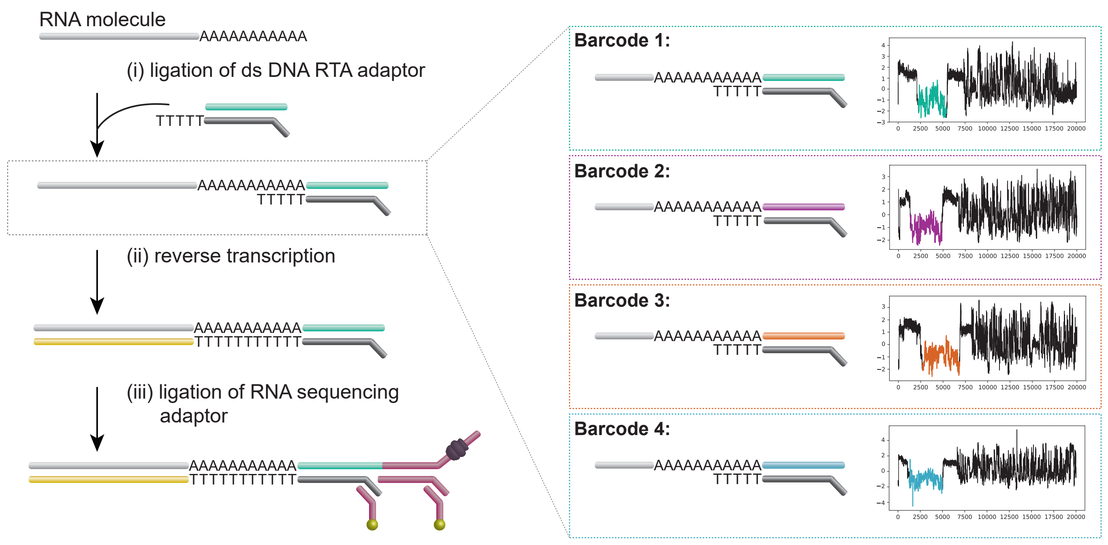
1. Background

Nanopore sequencing has enabled high-throughput sequencing of native RNA molecules without conversion to cDNA. However errors introduced by base calling raw nanopore signal complicates sequence-based analytics including barcode demultiplexing.
Currently there is **no formal barcoding protocol for native RNA sequencing**, limiting the applicability to scenarios where the amount of RNA available is low, such as in the case of patient-derived RNA samples.

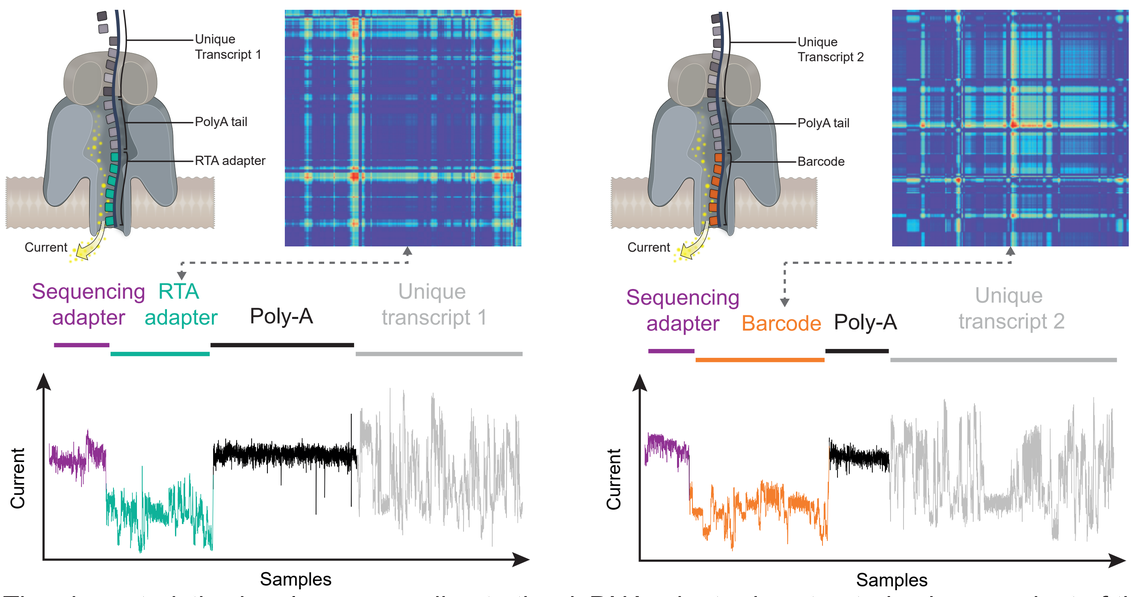
We describe **a novel strategy to barcode and demultiplex direct RNA sequencing**, involving custom DNA oligonucleotides ligated to RNA transcripts during library preparation.
The raw signal associated with the DNA barcode is extracted, transformed into an array of pixels, and demultiplexed using a deep convolutional neural network classifier.
Our method, DeePlexiCon, implements a 20-layer residual neural network model that **can demultiplex 84% of reads with 99% specificity**.

2. Experimental Design

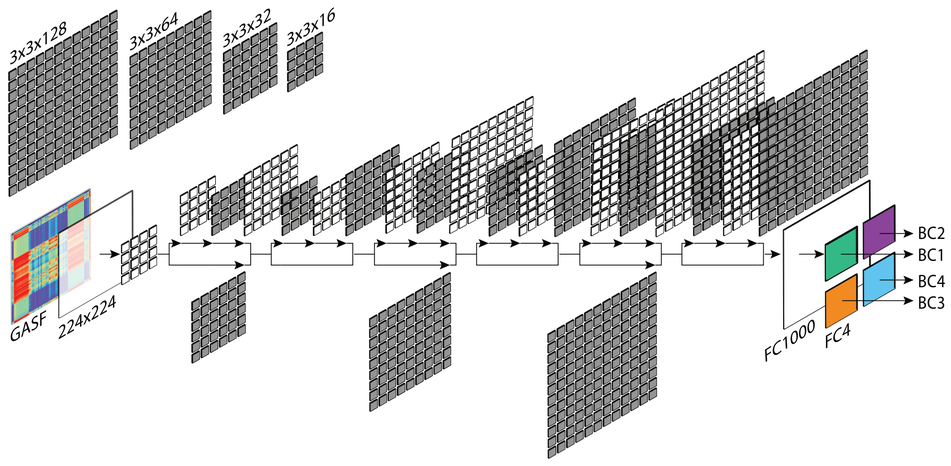
The SQK-RNA002 protocol employs the RTA adapter for splinted ligation to a poly(A) template RNA. We modified the RTA adapter by shuffling its double-stranded segment to generate 3 different dsDNA adapters to serve as molecular barcodes for multiplexing dRNA libraries.



The RTA adapter and the 3 shuffled barcoded adapters were each ligated to a unique in vitro transcribe poly(A) RNA transcript, reverse transcribed and pooled before ligating the RNA sequencing adapter and loading onto a MinION flowcell.



The characteristic signal corresponding to the dsDNA adapter is extracted using a variant of the segmenter algorithm implemented in SquiggleKit[1]. The resulting signal is converted into an image using a Gram matrix and subected to supervised deep learning with the following ResNet-20 deep residual convolutional neural network architecture.



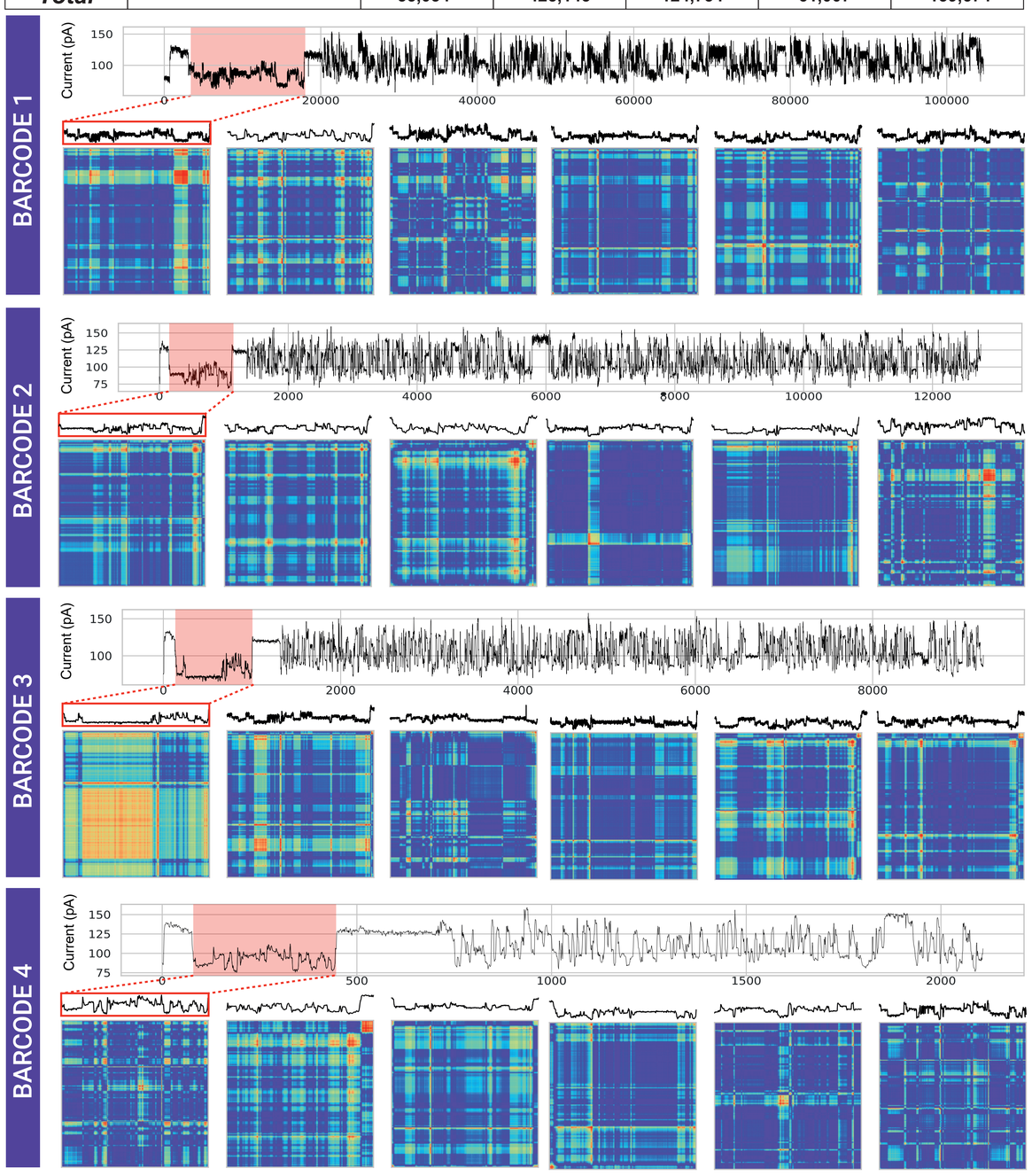
Supervised classification was performed by aligning the basecalled reads to the 4 unique reference sequences, effectively demultiplexing the data into positive controls.Training was performed using 3 replicates (1xRNA001; 2xRNA002) and the data was split as follows: 80% training; 10% testing; 10% validation. Demultiplexing efficiency was further assessed with 2 independent biological replicates: Replicate 1 (SQK-RNA001) and Replicate5 (SQK-RNA002).

Affiliations:
1 Garvan Institute of Medical Research, Sydney, Australia
2 St-Vincent's Clinical School, UNSW Sydney, Australia
3 CHU Sainte-Justine Research Centre, Montreal, Canada
4 Department of Biochemistry and Molecular Medicine, University of Montreal, Canada
5 Centre for Genomic Regulation (CRG), Barcelona 08003, Spain
6 Universitat Pompeu Fabra, Barcelona, Spain
# These authors contributed equally

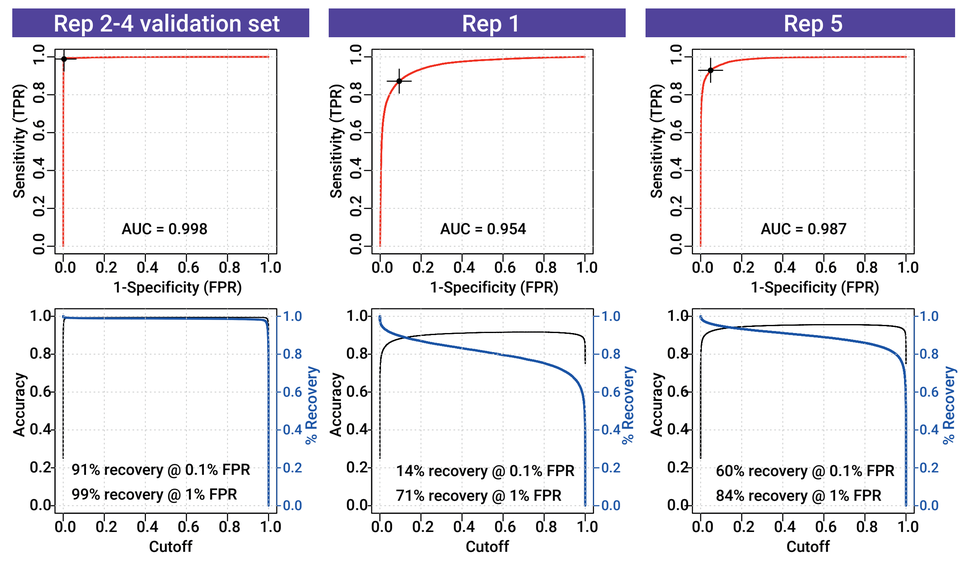
3. Results

40k and 10k empirically demultiplexed reads from each barcode were used for training and testing the model (Rep. 2, 3 & 4).

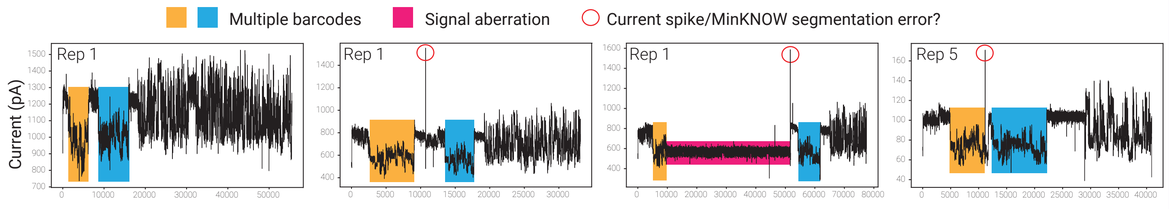
Barcode	Barcode Sequence	Uniquely Mapping Reads				
		Rep. 1	Rep. 2	Rep. 3	Rep. 4	Rep. 5
BC1	GGCTTCTTCTTGCTCTTAGG	17,643	45,489	44,329	15,962	65,461
BC2	GTGATTCTCGTCTTTCTGCG	3,278	151,071	22,331	10,811	16,548
BC3	GTACTTTTCTCTTTGCGCGG	692	55,475	21,192	18,054	35,287
BC4	GGTCTTCGCTCGGTCTTATT	11,421	148,182	36,882	16,180	22,378
Total		**33,034**	**425,145**	**124,734**	**61,007**	**139,674**



Signal segmentation, extraction, and image transformation requires less than 16ms, on average. Deemultiplexing (inference) requires 3ms per read using an NVIDIA Tesla V-100 GPU.



Performance was evaluated using 40k witheld reads from the training set (validation), which suggests overfitting. However, the model performed remarkably well on independent sequencing runs from different chemistries (Rep. 1 & Rep. 5), exposing artifacts in the data (below) and demonstrating the power of deep learning for pattern recognition.



Although using different input (RNA vs DNA), DeePlexiCon performs comparably to Deepbinner[2] (96.5% vs 98.5% precision, respectively), with a similar limitation to both, requiring the model retraining upon the release of new chemistries (RNA003) or barcode mixes.

Contact: Garvan Institute of Medical Research, Sydney, Australia
j.ferguson@garvan.org.au

References:
1. Ferguson JM. et. al (2019) SquiggleKit: a toolkit for manipulating nanopore signal data, *Bioinformatics*
2. Wick RR. et. al. (2018) Deepbinner: Demultiplexing barcoded Oxford Nanopore reads with deep convolutional neural networks. *PLoS Comput Biol*