

## Report Short Term Mobility

I am currently postdoctoral fellow in Prof. Pesole's lab at the Institute of Biomembranes, Bioenergetics and Molecular Biotechnologies (IBIOM), Bari. A major topic of our research activity is the study of RNA editing in human, that is a biologically relevant co/post-transcriptional process allowing the diversification of transcriptome and proteome. In particular, we are interested in adenosine to inosine (A-to-I) conversions, due to the deamination of adenosines (A). Since inosine (I) is commonly recognized as guanosine (G) by sequencing enzymes, RNA editing substitutions have been detected comparing cDNAs with their corresponding genomic sequences and looking for A-to-G changes. Although million A-to-I events have been uncovered by NGS technologies up to now, the accurate identification of RNA editing sites is yet a challenging task. Indeed, discriminating between G residues that originated from Is and sequencing errors or noise is quite difficult.

Oxford Nanopore platforms offer real-time, scalable, direct DNA and RNA sequencing. This can be performed on the portable MinION device containing hundred nanopores. As the DNA passes the nanopore, a sensor detects changes in ionic current caused by differences in the shifting nucleotide sequences occupying the pore. Nanopore devices have already allowed direct methylation analysis, based on the observation of different current signals at level of methylated/non-methylated cytosines [1]. Directed RNA sequencing through an array of nanopores may allow the detection of epigenetic RNA modifications and, thus, it could be employed to discriminate adenosines from deaminated adenosines (inosines).

During my Short-Term Mobility, I was hosted in the laboratory of Dr Matthew Loose at the School of Life Sciences, University of Nottingham (UK). A collaboration was started with his research group, having a long-standing experience with informatics methodologies to handle Oxford Nanopore data. Dr Loose is also a member of the Nanopore Human Genome Reference Consortium, which has recently sequenced and assembled a reference genome for the human GM12878 (NA12878) Utah/Ceph cell line by using the MinION device [2]. DNA, cDNA and direct RNA sequences, are publically available [3]. I have also been provided with early access to RNA datasets generated on MinION devices by the Consortium, which are still not public domain.

Ionic current changes detected within Nanopore devices are then converted in nucleotide sequences by specific computational tools. Generally, the Oxford Nanopore Technologies basecaller called Albacore is used to identify canonical bases. The identification of non-canonical bases requires *ad hoc* mathematical models and informatics technologies, whose implementation is currently in progress.

Under Dr Loose's supervision, I used the Tombo tool v.1.2 [4] to detect inosine signals within directed RNA sequences from NA12878 cell line in order to evaluate the potentialities of directed RNA sequencing for "direct" RNA editing detection without the need of a comparative analysis with the genomic DNA.

We already know the full set of edited sites and the corresponding editing levels obtained from NA12878 cDNA sequenced by Illumina. From this set, we considered only highly edited sites in order to discriminate between true edited and non-edited adenosines, thus 5984 edited sites were selected considering sites with coverage  $\geq 10$  and editing frequency  $\geq 0.5$ .

A preliminary analysis of the ONT direct RNA sample of NA12878 [5] was performed, including read realignment with GMAP (version 2016-11-07) [6] to hg19 genome assembly, and RNA editing detection with REDIttools (v.1.1) [7], a tool for RNA editing detection in Illumina data developed in Prof. Pesole's lab. Through the comparison with the reference editing dataset, 787 potential editing sites were identified in Nanopore data, 340 of which were filtered by fixing some parameters (3 reads supporting variant alleles, editing frequency  $\geq 0.1$ , coverage  $\geq 10$  within repeated regions).

The Tombo workflow comprises several steps: i) annotation of fast5 files with fastq files, which links basecalls with the raw signal-space data; ii) mapping base calls to a reference (in this case, the human transcriptome GenCode v.27 lift37 [8]) by recalling minimap2 aligner [9], and then assignment of the raw signal to the transcript sequence based on an expected current level model; iii) detection of

modified bases through 3 different methods: *de novo* detection using the Tombo default RNA model, canonical sample comparison (where no model is applied), and alternative model method using reference and alternative models (in this case for unedited/edited RNA, respectively), generated with Tombo model training function; iv) statistics and plotting. Moreover, Tombo allowed model training for both canonical bases only and canonical bases with a single, known, alternative base incorporated randomly instead of one canonical base (in this case, inosine instead of adenosine).

Four different datasets were analyzed using the three methods. Two subsets of reads deriving from the original one [5], showing a variation in the above mentioned 340 and 787 sites, as respectively detected by REDIttools [7], were used as input for *de novo* detection of modified bases. The two edited datasets were also used for the estimation of two alternative reference models via Tombo model training. *De novo* detection was then extended to a more extended fast5 subset including reads covering the reference editing sites in NA12878, used as input dataset also for analysis via comparison with a control sample and through the usage of user-created models. Moreover, an *in vitro* (IVT) RNA sample of NA12878, where no RNA editing occurs, was useful both as control in the detection of modified bases via comparison, and for the generation of a reference model via model training. It was also used as input for *de novo* detection, to estimate Tombo error rate.

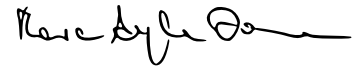
Tombo proved to be a promising tool for the detection of RNA editing sites within direct RNA generated by Nanopore device. The three methods implemented by the tool showed different levels of accuracy. *De novo* detection of modified bases can be quite error prone and may result in a high false positive rate. Indeed, a higher number of modified bases than that expected was found within analyzed datasets, although only 33% of known edited sites within UTR and CDS regions were confirmed (17% ca. when applying a strict filtering). The detection of modified bases through comparison with a control sample (IVT RNA) should be the most accurate, and the most stringent, among the three provided by Tombo, since only 2% of known edited sites were confirmed. The detection of modified bases by using user-created reference and alternative models is still experimental and the usage of synthetic edited samples should generate more reliable results. Currently only 10% of known edited sites were confirmed using this method.

Although further optimization is needed, our preliminary results suggests that directed RNA sequencing may become a reliable alternative methodology for the identification of RNA editing events.

## REFERENCES

1. Simpson JT, Workman RE, Zuzarte PC, David M, Dursi LJ, Timp W. Detecting DNA cytosine methylation using nanopore sequencing. *Nat Methods*. 2017;14:407–10.
2. Nanopore sequencing and assembly of a human genome with ultra-long reads - nbt.4060.pdf [Internet]. [cited 2018 Feb 21]. Available from: <https://www.nature.com/articles/nbt.4060.pdf>
3. Oxford Nanopore Human Reference Datasets [Internet]. Available from: <https://github.com/nanopore-wgs-consortium/NA12878>
4. Stoiber MH, Quick J, Egan R, Lee JE, Celniker SE, Neely R, et al. De novo Identification of DNA Modifications Enabled by Genome-Guided Nanopore Signal Processing. *bioRxiv*. 2017;094672.
5. Direct RNA and cDNA Sequencing of a human transcriptome on Oxford Nanopore MinION and GridION [Internet]. Available from: <https://github.com/nanopore-wgs-consortium/NA12878/blob/master/RNA.md>
6. Wu TD, Watanabe CK. GMAP: a genomic mapping and alignment program for mRNA and EST sequences. *Bioinforma Oxf Engl*. 2005;21:1859–75.
7. Picardi E, Pesole G. REDIttools: high-throughput RNA editing detection made easy. *Bioinforma Oxf Engl*. 2013;29:1813–4.

8. GenCode v.27 lift37 transcriptome [Internet]. Available from: [ftp://ftp.sanger.ac.uk/pub/gencode/Gencode\\_human/release\\_27/GRCh37\\_mapping/gencode.v27lift37.transcripts.fa.gz](ftp://ftp.sanger.ac.uk/pub/gencode/Gencode_human/release_27/GRCh37_mapping/gencode.v27lift37.transcripts.fa.gz)
9. Li H. Minimap2: versatile pairwise alignment for nucleotide sequences. ArXiv170801492 Q-Bio [Internet]. 2017 [cited 2018 Feb 21]; Available from: <http://arxiv.org/abs/1708.01492>

A handwritten signature in black ink, appearing to read 'Karl Lyle' followed by a stylized flourish.