Torchlex: a method for real-time demultiplexing of barcoded ONT reads

David Galevski[5,4], Aleksandar Nikov[5], Anne Kristine Schack[6], Lukasz Krych[6], M. Carmen Garrido Navas[1,2,3], Chris Kyriakidis[7], Zoran Velkoski[7], Gjorgji Madjarov[4,5,7]

1 GENYO Centre for Genomics and Oncological Research: Pfizer, University of Granada, Andalusian Regional Government, Liquid Biopsy and Cancer Interception Group, PTS Granada, Granada, Spain; **2** Genetics Department, Faculty of Sciences, Universidad de Granada, Granada, Spain; **3** CONGEN, Genetic Counselling Services, C/Albahaca 4, Granada, Spain; **4** University Ss Cyril & Methodius, Skopje 1000, N. Macedonia; **5** Netcetera, Zypressenstrasse 71, 8040 Zürich, Switzerland; **6** University of Copenhagen, Rolighedsvej 26, 1958 Frederiksberg, Denmark; **7** gMendel, Fruebjergvej 3, 2100 Copenhagen, Denmark

Introduction

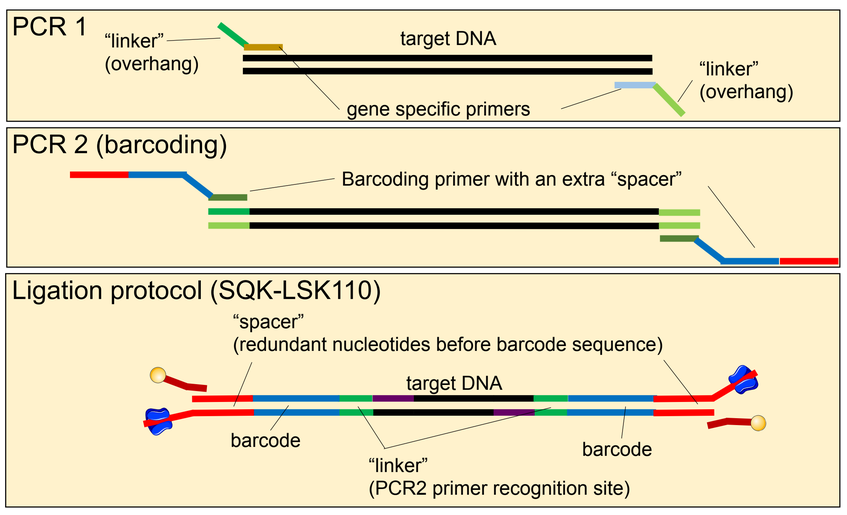
Current state-of-the-art ONT barcode demultiplexing tools (such as guppy) that operate directly on the DNA base-calls are computationally expensive. The **lagging of the demultiplexing algorithms** on the stream of base-called DNA reads that are generated by the ONT device can significantly influence **the real-time monitoring** and deciding capacity about the quality and quantity of the reads per DNA sample.

In order to **reduce the computational complexity** of the demultiplexing process and ensure more real-time data processing, it was our goal to develop demultiplexing system that would keep up with the most efficient basecalling algorithm running on GridION's GPU.

We developed **Torchlex**, a method for **real-time demultiplexing of custom** barcoded reads compatible with Oxford Nanopore Technologies. The proposed method managed to **significantly reduce the computational complexity** of the demultiplexing, while **preserving the quality** of classification compared to the competing methods.

Barcoding system

Multiplexing: We have developed an **optimized PCR based barcoding system** compatible with ONT. The custom designed barcodes (up to 192 combinations) are incorporated via **two step PCR**. PCR1 targets the gene of interest during unsaturated PCR and incorporates **"linker" sequence** that will become a primer binding site for PCR2. The linker sequence is palindromic, hence only one barcode primer is needed in PCR2. Each primer includes **15 bp spacer** separating ONT motor protein adapter from the barcode sequence. The spacer was added to ensure higher tolerance for the low quality at the beginning of the sequence entering the pore and thus higher recovery of barcode sequence. The barcoded DNA constructs are subsequently pooled and motor protein is added using the ligation protocol.



Demultiplexing: The reads obtained from the sequencing need to be demultiplexed and **grouped according to the attached barcode.** Once the reads are grouped per barcode, **further analysis** could be performed **per DNA sample**.

Challenges

The custom designed barcodes are comprised of 3 parts for a total of 57 base pairs, in the following order:

- Spacer (15 base pairs)
- Barcode (25 base pairs)
- Linker (15 base pairs)

They are located at the beginning and at the end of each sequence. A fast and reliable algorithm should be developed to find and identify the correct barcodes in the sequences of length 900-1200. The algorithm must be robust due to the errors created in the sequencing process to avoid the following problems:

1. Finding a different barcode at the start and end

2. Classifying a wrong DNA sequence as part of the barcode

3. Missing part of the barcode

Experimental Setup

Torchlex's computational efficiency and predictive performance is compared with the state-of-the-art demultiplexing method **Guppy** on a next-generation sequencing run using **6 different DNA samples**. The experimental validation was performed on **1 184 898 base-called DNA reads** (sequence length: **900 – 1200**) with a **Phred quality** score higher than **8** as a ground truth.

All the experiments were performed on one referent hardware architecture (**Intel i7 10th generation, 8 cores, 32 GB RAM, no CUDA**) using **thread parallelism of 10**.

Results

In terms of **computational efficiency**, the proposed method demultiplexed the base-called DNA reads by an **order of magnitude faster** than guppy.

	Throughput	Unclassified reads
Torchlex	~1520 reads/s	6.7%
Guppy	~138 reads/s	24%

In terms of **classification performance**, both methods showed **very similar results**.

	Precision	Recall
Torchlex	97.7%	81.4%
Guppy	97.8%	81.3%

Summary

Our solution demonstrated significantly better computational efficiency in comparison to the competing state-of-the-art methods. It exceeds the limits of real-time monitoring and analysis per DNA sample. Our technology was validated on single barcoding, but results can be directly extrapolated to combinatorial barcoding, which can even more reduce the analysis costs per DNA sample.

For more information, contact david@g-mendel.com or visit www.g-mendel.com