doi 10.18699/bgrs2024-9.1-15

Assigning transcriptomic subtypes to CML/CLL samples using nanopore RNA-sequencing and self-organizing maps

Arakelyan A.[1, 2, 3]*, Hakobyan Y.[4], Sirunyan T.[1, 3], Khachatryan G.[1, 3], Ghukasyan L.[1], Minasyan A.[1], Hakobyan S.[2,5], Martirosyan G.[1], Chavushyan A.[1], Binder H.[5,6].

[1] *Bioinformatics Group, Institute of Molecular Biology NAS RA, Yerevan, Armenia*

[2] *Laboratory of Human Genomics, Institute of Molecular Biology NAS RA, Yerevan, Armenia*

[3] *Institute of Biomedicine and Pharmacy, Russian-Armenian University, Yerevan, Armenia*

[4] *Hematology Center after Prof. Yeolyan, Hematology and Transfusion Medicine department of NIH Armenia, Yerevan, Armenia*

[5] *Armenian Bioinformatics Institute, Yerevan, Armenia*

[6] *Interdisciplinary Center for Bioinformatics, Leipzig University, Leipzig, Germany*

* *arakelyan@sci.am*

Key words: nanopore RNA-sequencing; CML; CLL; self-organizing maps; support vector machines; transcriptomic subtypes; RNA sequencing data analysis

Motivation and Aim: Next-generation sequencing technologies have revolutionized the fields of cancer screening, diagnostics, and precision oncology for blood cancers [1]. While Illumina platforms offer robust performance, they require significant initial infrastructure investments, and a high volume of samples for pooling, and may exhibit extended turnaround times when sample numbers are low. Conversely, third-generation (long-read) sequencing from Oxford Nanopore Technologies (ONT) can substantially reduce sequencing costs, making it a valuable tool for cost-effective and accurate cancer testing [2]. However, nanopore sequencing is challenged by lower accuracy and sample throughput compared to Illumina platforms. To enhance its utility, additional computational strategies are necessary. In our study, we integrated publicly available Illumina RNA sequencing data with machine learning methods to classify samples of chronic myeloid leukemia (CML) and chronic lymphocytic leukemia (CLL) into transcriptomic subtypes using nanopore RNA sequencing data.

Methods and Algorithms: Publicly available Illumina RNA sequencing data were obtained from the Gene Expression Omnibus (for CML, accession: GSE100026) and cBioPortal (for CLL, https://www.cbioportal.org/study/summary?id=cll_broad_2015). Gene counts in RPKM format were converted to TPM values. We then utilized the oposSOM R package for portraying high-dimensional data using self-organizing maps (SOM) [3]. This "SOM portrayal" allowed us to create transcriptome maps for CML and CLL, and identify co-expressed gene modules, and transcriptomic subtypes.

Twenty-five patients (six with CML and nineteen with CLL) were recruited from the Hematology Center After Prof. R. Yeolyan. Total RNA was isolated from blood samples using the Quick DNA/RNA Miniprep Plus Kit (Zymo Research, US). Sequencing libraries were prepared using the PCR-cDNA sequencing-barcoding kit (SQK-PCB109, Oxford Nanopore Technologies, UK) according to the manufacturer's instructions, loaded onto an R9.4.1 flow cell, and sequenced using a MinION mk1b instrument. The sequencing run lasted between 24 and 48 hours. Raw sequencing data were basecalled, demultiplexed, and converted to Fastq files using the Guppy tool. Reads were then

aligned to the human reference genome (hg38) using the Minimap2 software. Read counts were estimated using the featureCounts function from the Subread R package, based on GENCODE v44 annotations. Raw reads were transformed to TPM. Subsequently, TPM data from the patients were projected into the SOM space generated with Illumina RNA-seq data using a supervised SOM portrayal (supSOM) approach [4], which integrates a support vector machine regression model with the original SOM algorithm to predict the SOM portrait of a new sample.

The study received approval from the Ethics Committee of the Institute of Molecular Biology NAS RA.

Results: The SOM portrayal of CML enabled the identification of gene expression modules associated with the chronic (CP) and blast crisis (BP) phases. Samples from the CP phase exhibited upregulation of gene sets linked to oxidative phosphorylation, lipid metabolism, and proliferation, along with downregulated immune gene signatures. In the BP phase, there was noticeable upregulation of the RB1 pathway, angiogenesis, and inflammation gene signatures. The projection of ONT RNA sequencing data revealed significant differences from normal peripheral blood mononuclear cells (PBMCs) in the SOM space, with upregulation in gene signatures common to both CP and BP samples. These findings suggest that ONT CML samples are transitioning from the chronic to the blast crisis phase.

In CLL, the SOM space divided samples into eight transcriptomic subtypes (A *, A G *, A K J F * – immune phenotype, G P * – inflammatory, K *, M *, N * – non-immune, P I *, P L * – angioplastic), each characterized by distinct marker gene signatures. The projection of ONT CLL samples onto this SOM space indicated that the samples belonged to the M * and N * transcriptomic subtypes, which are characterized by low expression of ZAP70 and CD38 and associated with a favorable prognosis.

Conclusion: Our findings demonstrate that shallow ONT sequencing, when integrated with publicly available data and machine learning algorithms, can facilitate molecular subtyping CML and CLL. The ability to classify these cancers into subtypes can be beneficial for tailoring treatment strategies and improving patient management.

Funding: This study was funded by a research grant from the Committee of Higher Education and Science of the Ministry of Education and Science of the Republic of Armenia (21AG-1F021, PI: AA).

References

1. Schwartzberg L., Kim E.S., Liu D., Schrag D. Precision Oncology: Who, How, What, When, and When Not? *Am Soc Clin Oncol Educ Book*. 2017;37:160-169. doi 10.1200/EDBK_174176
2. Wang Y., Zhao Y., Bollas A., Wang Y., Au K.F. Nanopore sequencing technology, bioinformatics and applications. *Nat Biotechnol*. 2021;39(11):1348-1365. doi 10.1038/s41587-021-01108-x
3. Löffler-Wirth H., Kalcher M., Binder H. oposSOM: R-package for high-dimensional portraying of genome-wide expression landscapes on bioconductor. *Bioinformatics*. 2015;31(19):3225-3227. doi 10.1093/bioinformatics/btv342
4. Nikoghosyan M., Loeffler-Wirth H., Davidavyan S., Binder H., Arakelyan A. Projection of High-Dimensional Genome-Wide Expression on SOM Transcriptome Landscapes. *BioMedInformatics*. 2022;2:62-76. doi 10.3390/biomedinformatics2010004