LONG-READ SEQUENCING ANALYSIS OF THE LYTICI HUMAN CYTOMEGALOVIRUS TRANSCRIPTOME

Ph.D. Thesis

DR. ZSOLT BALÁZS

Department of Medical Biology

Doctoral School of Multidisciplinary Medicine

Faculty of Medicine

University of Szeged

Advisor: Zsolt Boldogkői habil, PhD, DSc

Szeged
2019

Publications directly related to the subject of the thesis:

I. Balázs, Zsolt; Tombácz, Dóra; Szűcs, Attila; Michael, Snyder; Boldogkői, Zsolt
Dual platform long-read RNA-sequencing dataset of the human cytomegalovirus lytic transcriptome
FRONTIERS IN GENETICS 9 Paper: 10.3389/fgene.2018.00432 (2018)
IF: 4.151

II. Tombácz, Dóra; Balázs, Zsolt; Csabai, Zsolt; Michael, Snyder; Boldogkői, Zsolt
Long-Read Sequencing Revealed an Extensive Transcript Complexity in Herpesviruses
FRONTIERS IN GENETICS 9 Paper: 259 (2018)
IF: 4.151

III. Balázs, Zsolt†; Tombácz, Dóra†; Szűcs, Attila; Michael, Snyder; Boldogkői, Zsolt
Long-read sequencing of the human cytomegalovirus transcriptome with the Pacific Biosciences RSII platform
SCIENTIFIC DATA 4 Paper: 170194 (2017)
IF: 5.305

IV. Balázs, Zsolt; Tombácz, Dóra; Szűcs, Attila; Csabai, Zsolt; Megyeri, Klára; Alexey, N Petrov; Michael, Snyder; Boldogkői, Zsolt
Long-Read Sequencing of Human Cytomegalovirus Transcriptome Reveals RNA Isoforms Carrying Distinct Coding Potentials
SCIENTIFIC REPORTS 7 Paper: 15989, 9 p. (2017)
IF: 4.122

Publications indirectly related to the subject of the thesis:

V. Boldogkői, Zsolt*; Balázs, Zsolt; Moldován, Norbert; Prazsák, István; Tombácz, Dóra
Novel Classes of Replication-associated Transcripts Discovered in Viruses
RNA BIOLOGY 1 Paper 1-10 (2019)
IF: 5.216

VI. Boldogkői, Zsolt*; Szűcs, Attila; Balázs, Zsolt; Sharon, Donald; Snyder, Michael; Tombácz, Dóra*

Transcriptomic study of Herpes simplex virus type-1 using full-length sequencing techniques
SCIENTIFIC DATA 5 Paper: 180266 (2018)
IF: 5.305

VII. Prazsák, István†; Moldován, Norbert†; Balázs, Zsolt; Tombácz, Dóra; Megyeri, Klára; Szűcs, Attila; Csabai, Zsolt; Boldogkői, Zsolt*
Long-read sequencing uncovers a complex transcriptome topology in varicella zoster virus
BMC GENOMICS 19: 1 Paper: 873 (2018)
IF: 3.730

VIII. Moldovan, Norbert; Tombacz, Dora; Szucs, Attila; Csabai, Zsolt; Balazs, Zsolt; Kis, Emese; Molnar, Judit; Boldogkoi, Zsolt*
Third-generation Sequencing Reveals Extensive Polycistronism and Transcriptional Overlapping in a Baculovirus.
SCIENTIFIC REPORTS 8: 1 p. 8604 (2018)
IF: 4.122

IX. Moldován, Norbert; Szűcs, Attila; Tombácz, Dóra; Balázs, Zsolt; Csabai, Zsolt; Michael, Snyder; Boldogkői, Zsolt*
Multi-platform Next-generation Sequencing Identifies Novel RNA Molecules and Transcript Isoforms of the Endogenous Retrovirus Isolated from Cultured Cells
FEMS MICROBIOLOGY LETTERS 365: 5 Paper: fny013, 6 p. (2018)
IF: 1.735

X. Moldován, Norbert; Balázs, Zsolt; Tombácz, Dóra; Csabai, Zsolt; Szűcs, Attila; Michael, Snyder; Boldogkői, Zsolt*
Multi-platform Analysis Reveals a Complex Transcriptome Architecture of a Circovirus
VIRUS RESEARCH 237 pp. 37-46., 10 p. (2017)
IF: 2.484

XI. Tombácz, Dóra; Csabai, Zsolt; Szűcs, Attila; Balázs, Zsolt; Moldován, Norbert; Donald, Sharon; Michael, Snyder; Boldogkői, Zsolt*
Long-Read Isoform Sequencing Reveals a Hidden Complexity of the Transcriptional Landscape of Herpes Simplex Virus Type 1
FRONTIERS IN MICROBIOLOGY 8 Paper: 1079, 16 p. (2017)
IF: 4.019

XII. Tombácz, Dóra†; Balázs, Zsolt†; Csabai, Zsolt; Moldován, Norbert; Szűcs, Attila; Donald, Sharon; Michael, Snyder; Boldogkői, Zsolt*
Characterization of the Dynamic Transcriptome of a Herpesvirus with Long-read Single Molecule Real-Time Sequencing

SCIENTIFIC REPORTS 7 Paper: 73751, 13 p. (2017)
IF: 4.122

XIII. Tombácz, D; Csabai, Z; Oláh, P; Balázs, Z; Likó, I; Zsigmond, L; Sharon, D; Snyder, M; Boldogkői, Z*
Full-Length Isoform Sequencing Reveals Novel Transcripts and Substantial Transcriptional Overlaps in a Herpesvirus
PLOS ONE 11: 9 Paper: e0162868, 29 p. (2016)
IF: 2.806

Publications not related to the subject of the thesis:

XIV. Tombácz, Dóra; Maróti, Zoltán; Kalmár, Tibor; Csabai, Zsolt; Balázs, Zsolt; Takahashi, Shinichi; Palkovits, Miklós; Snyder, Michael; Boldogkői, Zsolt*
High-Coverage Whole-Exome Sequencing Identifies Candidate Genes for Suicide in Victims with Major Depressive Disorder
SCIENTIFIC REPORTS 7 Paper: 7106, 11 p. (2017)
IF: 4.122

XV. Tombácz, Dóra; Moldován, Norbert; Balázs, Zsolt; Csabai, Zsolt; Michael, Snyder; Boldogkői, Zsolt*
Genetic Adaptation of Porcine Circovirus Type 1 to Cultured Porcine Kidney Cells Revealed by Single-Molecule Long-Read Sequencing Technology
GENOME ANNOUNCEMENTS 5: 5 Paper: e01539-16, 2 p. (2017)
IF: 0

Cumulative impact factor: 55.390

(At the time of writing, the 2018 impact factor values were not available, therefore the 2017 values are displayed for all publications newer than 2017)

Introduction

The human cytomegalovirus (HCMV, human betaherpesvirus 5) is a ubiquitous herpesvirus that can cause asymptomatic infections or mononucleosis-like symptoms in healthy adults. Congenital HCMV infection can also lead to abortions or developmental abnormalities (Jin et al., 2014). HCMV infection in the central nervous system has been associated with the development of glioblastomas (Dziurzynski et al., 2012). As of today, little is known about the molecular biology of HCMV infection.

RNAs are essential mediators and regulators and, oftentimes, end products of gene expression. Therefore, the investigation of an organism's transcriptome can provide deeper understanding of the regulatory mechanisms and the molecular functioning of that organism. The recent advances in RNA sequencing have enabled the genome-scale and detailed examination of transcription. Long-read sequencing can even map full-length transcripts, allowing for the isoform-level analysis of transcriptomes (Ozsolak & Milos, 2011).

While long-read sequencing is a powerful tool in functional genomics, the technology is currently producing lower throughput and is more error prone than short-read

sequencing (Weirather et al., 2017). Consequently, different approaches are required for the analysis of long-read sequencing data.

The human cytomegalovirus has a complex transcriptome. Polycistronism and alternative splicing make forming accurate transcript models particularly challenging (Ma et al., 2012). Long-read sequencing, on the other hand, is able to distinguish between isoforms and discern such a complex transcriptome.

Aims

In order to gain a better insight into the transcriptional repertoire of the virus, we have sequenced the lytic HCMV transcriptome on multiple third-generation sequencing platforms. Our main objectives were to determine exon-connectivity, and to annotate the lytic transcriptome of the virus.

In order to utilize the power of long-read sequencing, we have developed a pipeline that is suited for the analysis of long-read RNA sequencing data and is able to compare results obtained from different sequencing platforms.

We also aimed to characterize the performance of each sequencing platform and library preparation method based on their ability to sequence full-length genuine transcripts.

Materials and Methods

Two biologically independent samples were sequenced. One sample contained pooled total RNA from HCMV-infected human lung fibroblast cells (MRC-5) isolated 1, 3, 6, 12, 24, 72, 96 and 120h post infection the other contained pooled total RNA isolated 24, 72 and 120h after the infection. The first sample was subjected to cDNA sequencing on the Pacific Biosciences (PacBio) RSII and Sequel platforms as well as cDNA and dRNA sequencing on the Oxford Nanopore Technologies (ONT) MinION platform. The second sample was used for cap-selected cDNA sequencing on the MinION platform.

The PacBio reads were basecalled by SMRTLink, the ONT reads were basecalled by Albacore. The reads were mapped using GMAP and Minimap2. The data were analysed using a custom pipeline utilizing the biopython and the pysam modules, and the bedtools software. Custom scripts were written to generate read statistics, characterize transcripts and to compare results. Visualisation was carried out in R, using the ggplot and the Bioconductor packages.

Transcritpt features such as transcriptional start sites (TSS), introns and transcriptional end sites (TES) were accepted if they were detected in at least two libraries.

Results

Over 80,000 cDNA reads were obtained from the two PacBio platforms and over 1,000,000 cDNA reads from the MinION platform. The direct RNA sequencing yielded 36,195 reads. Although the MinION platform had a much higher throughput, the majority of the PacBio reads were full-length reads and were less error-prone. The direct RNA sequencing reads were used to validate the cDNA sequencing results.

We have created a pipeline for the analysis of long-read RNA sequencing data which accepts mapped sequencing reads produced by any long-read sequencing platform, and outputs a transcriptome annotation based on the sequenced reads. The toolkit characterizes transcripts and transcript features and lets the user to decide the criteria for filtering them.

187 different introns were detected in at least two separate experiments. A third of these are novel splice junctions. Over 90% of the introns discovered earlier were detected in at least one sequencing library. 233 transcriptional start sites (TSS) and 123 transcriptional end sites (TES) were validated by at least two different sequencing dataset. Reads containing a unique set of features (TSS, splice sites, TES) were considered transcript isoforms. 440 isoforms were detected in our dataset. 377 of

them were novel isoforms. The novel transcripts include TSS-, TES- or alternatively spliced isoforms of known genes, antisense transcripts and a novel intergenic transcript in the short repeat region.

Many of the transcript isoforms only differed from each other in a few nucleotides, however, interestingly, most isoforms differed from each other in the combination of ORFs that they contained.

Discussion

Our results have more than doubled the number of annotated HCMV transcripts. Cross-platform validation gives these novel features high confidence. Using long-read RNA sequencing data we were able to draw a more detailed map of the HCMV transcriptome, which is instrumental both for the analysis of the viral gene expression and for understanding the molecular mechanisms of infection.

Long-read RNA sequencing has discovered countless new isoforms in all the organisms for which it has been used (Abdel-Ghany et al., 2016; Byrne et al., 2017; Sharon, Tilgner, Grubert, & Snyder, 2013). The biological function of most of these isoforms is currently unknown. However, our results show that many of the isoforms have distinct coding potentials, meaning that they code for different peptides of express upstream ORFs which may play a regulatory role during translation(Barbosa & Romao, 2014).

With the headway of long-read sequencing technologies, the importance of bioinformatics tools that can analyse such data is increasing. We developed a pipeline which can rapidly process long-read RNA sequencing data from different platforms and create a transcriptome annotation which can be utilized by user

with no bioinformatics background. Even though the toolkit was developed in order to characterize viral transcriptomes, it is scalable, meaning that is equally utile in the analysis of larger eukaryotic transcriptomes.

References

Abdel-Ghany, S. E., Hamilton, M., Jacobi, J. L., Ngam, P., Devitt, N., Schilkey, F., … Reddy, A. S. N. (2016). A survey of the sorghum transcriptome using single-molecule long reads. *Nature Communications*, *7*, 11706. https://doi.org/10.1038/ncomms11706

Barbosa, C., & Romao, L. (2014). Translation of the human erythropoietin transcript is regulated by an upstream open reading frame in response to hypoxia. *RNA*, *20*(5), 594–608. https://doi.org/10.1261/rna.040915.113

Byrne, A., Beaudin, A. E., Olsen, H. E., Jain, M., Cole, C., Palmer, T., … Vollmers, C. (2017). Nanopore long-read RNAseq reveals widespread transcriptional variation among the surface receptors of individual B cells. *Nature Communications*, *8*, 16027. https://doi.org/10.1038/ncomms16027

Dziurzynski, K., Chang, S. M., Heimberger, A. B., Kalejta, R. F., McGregor Dallas, S. R., Smit, M., … Cobbs, C. S. (2012). Consensus on the role of human cytomegalovirus in glioblastoma. *Neuro-Oncology*, *14*(3), 246–255. https://doi.org/10.1093/neuonc/nor227

Jin, J., Hu, C., Wang, P., Chen, J., Wu, T., Chen, W., … Shen, X. (2014). Latent infection of human cytomegalovirus is associated with the development of gastric cancer. *Oncology Letters*, *8*(2), 898–904. https://doi.org/10.3892/ol.2014.2148

Ma, Y., Wang, N., Li, M., Gao, S., Wang, L., Zheng, B., … Ruan, Q. (2012). Human CMV transcripts: an overview. *Future Microbiology*, *7*(5), 577–593. https://doi.org/10.2217/fmb.12.32

Ozsolak, F., & Milos, P. M. (2011). RNA sequencing: advances, challenges and opportunities. *Nature Reviews. Genetics*, *12*(2), 87–98. https://doi.org/10.1038/nrg2934

Sharon, D., Tilgner, H., Grubert, F., & Snyder, M. (2013). A single-molecule long-read survey of the human transcriptome. *Nature Biotechnology*, *31*(11), 1009–1014. https://doi.org/10.1038/nbt.2705

Weirather, J. L., de Cesare, M., Wang, Y., Piazza, P., Sebastiano, V., Wang, X.-J., … Au, K. F. (2017). Comprehensive comparison of Pacific Biosciences and Oxford Nanopore Technologies and their applications to transcriptome analysis. *F1000Research*, *6*, 100. https://doi.org/10.12688/f1000research.10571.2