LONG-READ SEQUENCING ANALYSIS OF THE LYTICI HUMAN CYTOMEGALOVIRUS TRANSCRIPTOME

Ph.D. Thesis

DR. BALÁZS ZSOLT

SZTE ÁOK Orvosi Biológiai Intézet

SZTE ÁOK Multidiszciplináris Orvostudományok Doktori Iskola

Témavezető: Prof. Dr. Boldogkői Zsolt

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)

Bevezetés

A human citomegalovírus (HCMV, human bétaherpeszvírus 5) egy ubiquiter herpeszvírus ami egészséges felnőtteket tünetmentesen fertőz vagy mononukleózis-szerű tüneteket okoz. A kongenitális HCMV fertőzés abortuszhoz vagy fejlődési rendellenességek kialakulásához vezet (Jin et al., 2014). A központi idegrendszerbeli HCMV fertőzést újabban glioblasztóma-képződéssel hozták összefüggésbe (Dziurzynski et al., 2012). Napjainkig kevéssé ismert a HCMV fertőzés molekuláris biológiája.

Az RNS-ek létfontosságú közvetítői, szabályozói és esetenként végtermékei a génexpressziónak. Ennek folyamán, egy szervezet transzkriptomának a vizsgálata nagyban elmélyítheti ismereteinket az adott szervezetet szabályozó molekuláris mechanizmusokról. Az RNS szekvenálás területén történt rohamszerű fejlődés lehetővé tette a transzkripció genom-szintű vizsgálatát. A hosszú-read szekvenálás a transzkripteket akár teljes hosszukban képes feltérképezni, és így a transzkriptom izoformák szintjén elemezhető (Ozsolak & Milos, 2011).

Bár a hosszú-read szekvenálás egyre fontosabb eszköz a funkcionális genomikai kutatásokban, ez a technológia egyelőre

sokkal kevesebb és sokkal hibásabb readeket produkál, mint a rövid-read szekvenálás (Weirather et al., 2017). Következésképpen a különböző megközelítés szükséges a hosszú-read szekvenálás eredményeinek elemzéséhez.

A humán citomegalovírus transzkriptoma komplex; a policisztonikus, alternatívan splice-olt és transzkripteket különösen nehéz rövid-read szekvenálással elemezni (Ma et al., 2012). A hosszú-read szekvenálás azonban, képes megkülönböztetni a transzkripteket egy ilyen komplex transzkriptomban is.

Célkitűzések

A vírus transzkripciós repertoárjának részletes megismerése érdekében számos harmadik-generációs szekvenáló platformon megszekvenáltuk a HCMV lítikus transzkriptomát. Célul tűztük ki a lítikus fertőzés során keletkező transzkriptek bázispár-pontosságú feltérképezését.

A hosszú-read szekvenálási eredmények elemzéséhez kidolgoztunk egy olyan pipeline-t ami a hosszú-read RNS-szekvenálás adatait feldolgozza és a különöző platformmal kapott eredményeket összehasonlíthatóvá teszi.

Továbbá célunk volt az egyes könyvtárkészítési eljárások jellemzése az alapján, hogy milyen hatékonysággal képesek felismerni a teljes hosszúságú transzkripteket.

Anyagok és módszerek

Két biológiailag független mintát szekvenáltunk. Az egyik mintához 1, 3, 6, 12, 24, 72, 96 és 120 órás HCMV fertőzés után humán tüdő fibroblaszt (MRC-5) sejteket lizáltunk, majd RNS-t izoláltunk és azokat egy mintává összevontuk. A másik minta 24, 72 és 120 órás fertőzés után izolált RNS keverékből készült. Az első mintát cDNS írás után a Pacific Biosciences (PacBio) RSII és Sequel platformjain és az Oxford Nanopore Technologies (ONT) MinION platformján szekvenáltuk meg. A második minta egy részét cap-szelekció és cDNS írás után a MinION platformon cDNS-ként, a minta másik részét ugyanezen a platformon RNS formájában szekvenáltuk meg.

A PacBio readeket a SMRTLink programcsomaggal, az ONT readeket Albacore-ral basecalloltuk. A readeket GMAP-pel és Minimap2-vel mappeltük. Az adatokat saját készítésű scriptekkel elemeztük. Ezek a scriptek a biopython és a pysam modulokon, valamint a bedtools szoftver használatán alapulnak. A statisztikai elemzésre, a transzkriptek jellemzésére és az eredmények összehasonlítására szintén saját készítésű scripteket alkalmaztunk. A vizualizációhoz a ggplot és a Bioconductor csomagokat használtuk. Az illesztések és az annotációk ábrázolásához az IGV és a Geneious

programcsomagokat használtuk (Kearse et al., 2012; Robinson et al., 2011).

Transzkript elemeket, úgymint transzkripciós start helyeket (TSS), intronokat és transzkripciós terminációs helyeket (TES) akkor annotáltunk, ha legalább két szekvenálási könyvtárban detektálhatóak voltak. A transzkripteket az ezen transzkript elemeket tartalmazó readek alapján annotáltuk.

Eredmények

Több, mint 88000 cDNS leolvasást kaptunk a PacBio platformról és több, mint egy milliót a MinION platformról. A direkt RNS szekvenálás 36195 readet eredményezett. Bár a MinION platform jelentősen nagyobb lefedettséget generált, a PacBio readek viszont nagyobb arányban reprezentáltak teljes hosszúságú transzkripteket és pontosabbak voltak. A direkt-RNS szekvenálás eredményeit a cDNS-szekvenálás eredményeinek a validálására használtuk fel.

Létrehoztunk egy pipeline-t a hosszú-read szekvenálás analízisére, amely már mappelt readeket fogad el inputként bármelyik hosszú-read szekvenálási platformról és a readek alapján elkészít egy transzkriptom annotációt. A program a felhasználó beállításainak megfelelően jellemzi és filterezi a transzkript elemeket és transzkripteket.

187 intront sikerült azonosítani legalább két szekvenálási könyvtárban. Ezek egyharmada új splice hely volt. A korábban detektált intronok 90%-át mi is azonosítottuk legalább egy szekvenálási mintában. 233 transzkripciós start helyet (TSS) és 123 transzkripciós terminációs helyet (TES) validált legalább két szekvenálási adatsor. A transzkript elemek egyedi kombinációját tartalmazó readeket transzkript izoformákként

annotáltuk. 440 izoformát sikerült azonosítani az adatok alapján, ezek közül 377 új izoforma. Az új transzkriptek között találhatóak ismert gének TSS-, TES- valamint alternatívan splice-olt izoformái, antiszensz gének és egy intergénikus transzkript a rövid repeat régióban.

A transzkript izoformák közül sok csak néhány nukleotidban tért el egymástól, azonban, érdekes módon, a legtöbb izoforma eltért a bennük foglalt ORF-ek kombinációjában.

A cDNS szekvenálást direkt RNS szekvenálással kiegészítve sikerült elkülönítenünk a technikai műtermékeket a valódi transzkript adatoktól.

Diszkusszió

Eredményeink több, mint megháromszorozták az annotált HCMV transzkriptek számát. A különböző platformok általi validációnak köszönhetően eredményeink megbízhatóak. Hosszú-read szekvenálási adataink sokkal részletesebb képet tudtak mutatni a HCMV transzkriptomáról, amely hasznos mind a virális génexpresszió tanulmányozásához, mind pedig a fertőzés molekuláris mechanizmusának megértéséhez.

A hosszú-read RNS szekvenálási technikák számos új izoformát fedeztek fel minden megvizsgált organizmusban, amelyben eddig alkalmazták őket (Abdel-Ghany et al., 2016; Byrne et al., 2017; Sharon, Tilgner, Grubert, & Snyder, 2013). Az izoformák nagy részének egyelőre nem ismert a biológiai jelentősége. Azonban, az eredményeink azt mutatják, hogy sok izoforma különböző kódoló potenciállal rendelkezik, ami azt jelenti, hogy különböző polipeptideket kódolnak vagy pedig különböző rövid upstream ORF-eket fejeznek ki, amelyek szabályozó szerepet tölthetnek be a transzkript transzlációjában (Barbosa & Romao, 2014).

A hosszú-read szekvenálás rohamos ütemű fejlődése mellett az ezek elemzésére képes bioinformatikai eszközök jelentősége is növekszik. Egy olyan pipeline-t fejlesztettünk ki, amelyik a

különböző hosszú-read szekvenálási platformok adatait gyorsan képes feldolgozni és azok alapján egy olyan transzkriptom annotációt készít, ami egy bioinformatikai ismeretekkel nem rendelkező felhasználó által is kezelhető. Bár a szoftver virális transzkriptomok jellemzésére lett kifejlesztve, elég flexibilis ahhoz, hogy nagyobb eukarióta transzkriptomok elemzésére is alkalmas legyen.

Referenciák

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

Kearse, M., Moir, R., Wilson, A., Stones-Havas, S., Cheung, M., Sturrock, S., … Drummond, A. (2012). Geneious Basic: an integrated and extendable desktop software platform for the organization and

analysis of sequence data. *Bioinformatics (Oxford, England)*, *28*(12), 1647–1649. https://doi.org/10.1093/bioinformatics/bts199

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

Robinson, J. T., Thorvaldsdóttir, H., Winckler, W., Guttman, M., Lander, E. S., Getz, G., & Mesirov, J. P. (2011). Integrative genomics viewer. *Nature Biotechnology*, *29*(1), 24–26. https://doi.org/10.1038/nbt.1754

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

Társszerzői lemondó nyilatkozat

Alulírott Dr Tombácz Dóra kijelentem, hogy Dr Balázs Zsolt PhD értekezésének tézispontjaiban bemutatott - közösen publikált - tudományos eredmények elérésében a pályázónak meghatározó szerepe volt, ezért ezeket a téziseket más a PhD fokozat megszerzését célzó minősítési eljárásban nem használta fel, illetve nem kívánja felhasználni.

2019.01.21. ----------

Dr. Tombácz Dóra

A pályázó tézispontjaiban érintett, közösen publikált közlemények:

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

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