HIGH-THROUGHPUT TRANSCRIPTOMIC ANALYSIS OF PSEUDORABIES VIRUS

Ph.D. Thesis Summary

Péter Oláh Msc

Department of Medical Biology
Doctoral School of Interdisciplinary Medicine
Faculty of Medicine
University of Szeged

Supervisor: prof. Zsolt Boldogkői

Szeged
2017

AIMS

1. *De novo* assembly of the PRV strain Ka complete genome using long-read sequencing, in order to provide an accurate reference of the strain used in our studies in place of the previously used composite reference.
2. Generating the first single-base resolution transcriptome map of PRV Ka from a mixed-timepoint infection, using short-read sequencing in order to characterize potential new transcripts and splice isoforms.
3. Independent validation and detailed characterization of novel transcripts in multi-time-point samples.
4. Analysis of potential transcriptional interference events in the viral genome in support of the TIN hypothesis.

Remark: The present thesis focuses primarily on the bioinformatics analysis and sequence categorization aspect of the results, which was the main contribution of the author to the source publications, drawing on the genomic DNA sequencing results from publication III, complete transcriptome results from publication I, and specific results on the CTO non-coding RNA from publication II.

BRIEF SUMMARY

The whole-genome sequence of PRV strain Ka was determined by cutting-edge long-read sequencing in order to facilitate accurate transcriptome mapping and further epigenetic studies of the virus. Illumina short-read, high-throughput sequencing was used for the first time on lytic-infection PRV samples in order to create a single-

base resolution transcriptome map, along with the detailed polyadenylation landscape of the virus by PA-Seq. In accordance with expectations, most of the viral genome was transcribed, with the exception of several small intergenic repetitive sequences, and loci in the large internal and terminal repeats. Among the findings are a novel polyadenylated lncRNA near the OriL origin of replication, and the single-base resolution mapping of 3′ UTRs across the viral genome. A number of genes exhibited alternative polyadenylation sites, while previously described splice sites were confirmed and expanded with a novel alternative splicing event in the key regulator *ep0* gene.

METHODS

Virus, cells and infection

Immortalized Porcine Kidney PK-15 epithelial cells were used for the propagation of strain Kaplan of PRV. PK-15 cells were cultivated in Dulbecco’s modified Eagle medium supplemented with 5% fetal bovine serum (Gibco Invitrogen) with 80 μg gentamycin/ml at 37 °C, 5% CO2 in filter-capped flasks. The following virus stock was prepared for the experiments: semi-confluent PK-15 cells in rapid growth were infected at a multiplicity of infection (MOI) of 0.1 plaque-forming unit (pfu)/cell, and incubation lasted until a complete cytopathic effect was observed. The infected cells were frozen and thawed three times, followed by low-speed centrifugation at 10,000g, 20 min. The supernatant was concentrated and further purified by ultracentrifugation after removal of cell debris, across a 30% sugar cushion at 24,000 rpm for 1h, using a Sorvall AH-628 rotor. The number of cells in a culture flask was 5×10^6. A high MOI (10 pfu/cell) was used for the infection of PK-15 cells in order to generate samples for transcriptome studies. Infected cells were incubated for 1h, followed by removal of the virus suspension and washing with phosphate-buffered saline (PBS). After the addition of new medium to the cells, they were incubated for 1, 2, 4, 6, 8, 10, 12, 14, 16, 18, 20, 22 or 24h p.i. For the control population, mock-infected cells, treated in the same way as the infected cells, were used.

Viral DNA extraction

PK-15 cell monolayers were infected at MOI=10pfu/cell, and cultivated at 37 °C until a cytopathic effect was observed. Culture medium was collected and centrifuged at 4,000 rpm for 10 min using a Sorvall GS-3 rotor. Viral particles were sedimented on a 30% sucrose cushion by ultracentrifugation at 24,000 rpm for 1h using a Sorvall AH-628 rotor. The sedimented virus was resuspended in sodium Tris-EDTA buffer. 100 ug/ml Proteinase-K and 0.5% sodium dodecyl sulfate (SDS) was added, the lysate was incubated at 37 °C for 1h, followed by phenol-chloroform extraction.

TotalRNA extraction

RNA was extracted from samples of each individual time point of infection by using the NucleoSpin RNA II Kit (Macherey-Nagel GmbH and Co. KG). Following centrifugation and cell lysis with buffer containing chaotropic ions, the nucleic acids were purified on silica column. DNA was removed by RNase-free DNase solution (supplied with the NucleoSpin RNA II Kit). Finally, the RNAs were eluted from the column in RNase-free water (supplied with the kit). To eliminate the residual DNA contamination, all RNA samples were treated by an additional digestion with Turbo DNase (Ambion Inc.). The concentrations of the RNA samples were measured by spectrophotometric analysis with a BioPhotometer Plus instrument (Eppendorf) and Qbit fluorometer (Thermo Fisher Scientific). RNA samples were stored at −80°C until further use.

PacBio RS II gDNA prepration and sequencing

SMRTbell template libraries were prepared from DNA using standard protocols for 6-kb and 20-kb library preparation. Sequencing was performed in five single-molecule real-time (SMRT) cells with P5 DNA polymerase and C3 chemistry (P5-C3) yielding a total of 78,111 reads and an extremely high coverage (1,200x) throughout the genome. Sequencing and library preparation were carried out in the Department of Genetics, Stanford University.

Illumina cDNA library preparation and sequencing

Strand-specific total RNA libraries were prepared for paired-end, 2x100bp sequencing by using the Illumina-compatible ScriptSeq v2 RNA-Seq Library Preparation Kit (Epicenter). For polyA-sequencing, a single-end library was constructed through the use of custom anchored adaptor-primer oligonucleotides with an oligo(VN)T-10 primer sequence, in which a two-nucleotide anchor is followed by 10 T nucleotides. Anchored primers compensate for the loss in throughput due to the high fraction of reads containing solely adenine bases, as in the case of conventional oligo(dT) primers. Sequencing was performed on an Illumina HiScanSQ platform at the Genomic Medicine and Bioinformatic Core Facilty of the University of Debrecen.

RT-qPCR analysis of alternative splicing

In order to validate splicing events, two sets of primers were applied, with lengths from 19 to 23 nucleotides, approximately 100 bp upstream and downstream of the given splice site. 5 µl solutions were prepared for reverse transcription reactions, containing 0.02 µg of total RNA, 2 pmol of the gene-specific primer, 0.25 µl of dNTP mix, 1 µl of 5× First-Strand Buffer, 0.25 µl (50 units/µl) of SuperScript III Reverse Transcriptase (Invitrogen) and 1 U of RNAsin (Applied Biosystems Inc.). RT mixes were incubated at 55 °C for 60 min. The reaction was stopped at 70 °C for 15 min. No-RT control reactions (RT reactions without Superscript III enzyme) were run to test the potential viral DNA contamination by conventional PCR. For RT-qPCR reactions, RNA samples with no detectable DNA contamination were used. Real-time quantitative PCR experiments were carried out for each sample in triplicate, on a Rotor-Gene 6000 cycler (Corbett Life Science). RT-qPCR reactions were done in 20-µl mixtures containing 7 µl of ×10 dilution cDNA, 10 µl of ABsolute qPCR SYBR Green Mix (Thermo Fisher Scientific), 1.5 µl of forward and 1.5 µl of reverse primers (10 µM each). The running conditions were as follows: 15 min at 95 °C, 30 cycles of 94 °C for 25 s (denaturation), 60 °C for 25 s (annealing), and 72 °C for 6 s (extension). Products were visualized on 12 % polyacrylamide gel stained with Gel Red dye, gel images were acquired using a ProteinSimple AlphaImager HV gel documentation system.

Data availability

The complete genome sequence of strain Kaplan of pseudorabies virus was assigned DDBJ/EMBL/GenBank accession no. KJ717942. Raw read data from PA-Seq and RNA-Seq experiments are deposited in the European Nucleotide Archive under accession code PRJEB9526.

RESULTS

PRV strain Ka genome map

Although PRV is a widely-studied organism among herpesviruses, and the complete viral genome of strain Kaplan had previously been sequenced, the available draft genomes are mostly poorly annotated, and contain several discrepancies, mainly in low-complexity regions. We carried out DNA sequencing in order to assess intra-strain heterogeneity, and to gather preliminary data for future epigenetic studies. Furthermore, the most well-annotated genome to date (NC006151.1) is a composite of six different strains, and as such could not directly be used in transcriptomic studies.

The especially long read lengths enabled the construction of highly overlapping contigs for assembly. The complete genome consists of 143,423bp, with 73.59% GC-content; sequence identity compared to other PRV strains available in GenBank ranges between 97-99%.

The extent of intra-strain variability was lower than expected, with well-defined variable base positions only outside of protein-coding sequences, and occurring quite infrequently. An intriguing source of heterogeneity was present in a 12bp semi-palindromic repeat in the *ul27* gene, being absent in ~50% of the viral genome copies. Further studies ruled out the possibility of technical error, and showed that the repeat is specific to certain mutant strains of Ka. Protein-coding genes were predicted, and existing, matching annotations lifted by the GATU tool. Manual annotation was used on genomic features such as replication origins and repeat motifs. Annotation of a previously unknown noncoding RNA Close to OriL (CTO), a novel splice site of the ep0 gene, and new isoforms of 11 protein-coding genes were based on our short-read RNA sequencing results. Annotation of PRV miRNAs was created on the basis of previously published precursor miRNAs found in strains NIA-3 and Ea, and as such were included in GenBank annotation for the first time. The rich and up-to-date annotation information was deposited in the corresponding GenBank record, which, through sequence homology, might be useful in annotating and understanding the frequently isolated strains from farms mainly in Asia, but around the world as well.

On a technical level, the study has also been among the first which employed the SMRT technology in sequencing a viral genome, and as such, provided opportunities for methodological improvement, notably in the systematic errors arising in high-GC content organisms, and also effective screening and removal of such errors. With these additions, the long-read sequencing technology has proven efficient and economic for the use in viral gDNA studies, as complete genomes can be assembled from a minimal number of contigs and very little cross-validation using the long reads.

cDNA sequencing

The catalog of PRV transcripts has been updated based on our cDNA sequencing experiments of mixed-timepoint lytic infections on the PK-15 epithelial cell line, providing a detailed view on the transcriptional landscape of the virus, and accurately defining 3’ polyA+ RNA boundaries, which are key to the composition of the characteristic, polycistronic gene arrangement in the Herpesviridae family.

Splice sites

From random-hexamer primed sequencing, data emerged on a splice isoform of the key early transactivator ep0 gene, the functions of which require further investigations, as the disordered nature of the protein prohibits in silico modeling and predictions. Based on amino acid sequence, however, the zinc-finger domain of ep0 is not covered by the spliced intron, indicating retained function of the shorter isoform.

Further splice sites of in the PRV genome have been confirmed and accurately detected in *us1* and *ul15*. The expression of latency-associated transcripts was also detected in our lytic samples at low levels, although the coverage did not enable accurate identification of the LLT splice site. Further, the presence of low-coverage, difficult to validate splice junctions was later confirmed via deep-sequencing isoform analysis reported in publication IV, and led to the identification of several lower-expressed mRNA isoforms and ncRNAs.

Polyadenylation landscape

Through PA-Seq, overall gene expression and transcript boundaries were assessed using highly efficient anchored primers, resulting in the detection of transcripts in the UL7-UL9 cluster, often missed by other PCR-based methods. The detailed organization of PRV gene clusters was also revealed, serving as a foundation for further evaluation of the transcriptional kinetics and regulatory mechanisms that may shed light on transcriptional interference networks. The usage frequency and distribution of both canonical and non-canonical polyadenylation signals was assessed, with sequence motifs corresponding highly to those identified in eukaryotic model organisms, but also distinct features arising from the compact and overlapping 3’ UTR sequences and the polycistronic arrangement.

Novel transcripts

Recent results in herpesvirus research indicated a wide range of non-coding RNAs in clinically important pathogens such as KSHV, HCMV and EBV. The ncRNAs described in lytic infections were of similarly high expression, often accounting to ~50% of the total transcribed RNA quantity. Further, the screening of PRV samples for miRNAs from both latent infection on neuronal ganglia and lytic phase in epithelial cells gave positive results, with pre-miRNAs originating from the LLT intron. As such, the assessment of the totalRNA and polyA+ RNA fractions was both timely and promising, eventually resulting in the characterization of the rather short (268bp), but extremely highly expressed polyA+ lncRNA

CTO-S, near the viral origin of replication. Lacking a direct genomic target in either the host or the viral genome, the initial hypothesis for the function of the gene is the regulation of viral DNA replication, a function which is in agreement with its late expression kinetics. Other novel transcripts include the low-expressed short 3’-antisense RNA SANC, also in the OriL region, and validation of the ORF1.2 hypothetical mRNA, along with several new 3’-UTR variations arising from alternative polyadenylation.

Transcriptional interference

The above findings indicated that key genomic features are present in PRV which may serve as the focus of transcriptional interference studies, including the extensive overlaps between various gene clusters, the location of ncRNAs and short intergenic repetitive sequences, and the newly discovered transcript boundaries within gene clusters. The evaluation of previously studied sites, such as the ul30-ul31 gene pairs added qualitative information to the results of earlier RT-qPCR results.

Publications directly related to the subject of the thesis:

I. **Oláh P**, Tombácz D, Póka N, Csabai Z, Prazsák I, Boldogkői Z. Characterization of pseudorabies virus transcriptome by Illumina sequencing.BMC Microbiology. 2015;15:130. doi:10.1186/s12866-015-0470-0. **IF:2.581 (2015)**

II. Tombácz, D., Csabai, Z., **Oláh, P**., Havelda, Z., Sharon, D., Snyder, M., & Boldogkői, Z. (2015). Characterization of Novel Transcripts in Pseudorabies Virus. Viruses, 7(5), 2727–2744. http://doi.org/10.3390/v7052727 **IF:3.437 (2015)**

III. Tombácz D, Sharon D, **Oláh P**, Csabai Z, Snyder M, Boldogkői Z. Strain Kaplan of Pseudorabies Virus Genome Sequenced by PacBio Single-Molecule Real Time Sequencing Technology. Genome Announcements. 2014;2(4):e00628-14. doi:10.1128/genomeA.00628-14.

Publications indirectly related to the subject of the thesis:

IV. Tombácz D, Csabai Z, **Oláh P**, et al. Full-Length Isoform Sequencing Reveals Novel Transcripts and Substantial Transcriptional Overlaps in a Herpesvirus. PLoS ONE. 2016;11(9):e0162868. doi:10.1371/journal.pone.0162868. **IF: 3.54 (2015)**