Virus-Host Transcriptomics

Caroline C. Friedel
Institut für Informatik, Ludwig-Maximilians-Universität München
caroline.friedel@bio.ifi.lmu.de

Introduction

The application of next generation sequencing technologies (NGS) to sequencing of RNA (RNA-seq) provides novel opportunities for the analysis of transcriptomes beyond the simple quantification of gene expression. In particular, the combination of RNA-seq with powerful techniques for selecting specific types of RNA (e.g. newly transcribed RNA using 4sU-tagging [DRR+08] or actively translated RNA using ribosome profiling [IGNW09]) now allows quantification of real-time changes in RNA synthesis [MLW+12], RNA processing [WBB+12], and translation [IGNW09].

A further interesting application arises from the fact that RNA-seq protocols do not distinguish between RNA from different species. Thus, in case of infections by viruses or bacteria, RNA from the infecting species will automatically be sequenced together with the host RNA. Originally, this application has been proposed in a thought experiment by Westermann et al. [WGV12] and denoted as dual RNA-seq, although it is not limited to just one infecting species and the host. For instance, Castellarin *et al.* [CWF+12] identified a number of microbes in RNA-seq data of colorectal carcinoma and normal tissue samples. To date, dual RNA-seq has been used to annotate and quantify the transcriptome and translatome of several herpesviruses, which are large DNA viruses that replicate in the nucleus. This includes murine and human cytomegalovirus (MCMV and HCMV) [MLW+12, SGWM+12], Kaposi's sarcoma-associated herpesvirus (KSHV) [AWSG+14], and human herpesvirus 1 (HSV-1) [REL+15].

In this presentation, I will provide an overview on methods developed in my group for the analysis of RNA-seq data of infected cells, in particular for the analysis of transcriptional and translational activity, transcription termination and RNA processing during lytic HSV-1 infection [REL+15]. This includes methods for parallel RNA-seq mapping against several read sources [BCZF12, BCZF13, BKC+15] as well as quantification of transcription termination and polyadenylation sites in both host and virus.

Parallel RNA-seq mapping to virus and host

One major challenge in both "standard" and dual RNA-seq is the identification of the transcriptomic origin of sequencing reads (mapping). Accordingly, a number of software programs have been developed for this task, e.g. TopHat [TPS09] or STAR [DDS+13]. However, these approaches to not directly support mapping of reads from multiple species or other read sources (e.g. rRNA sequences, which are not included in the human reference genome). Although additional sequences may be included into the mapping index, this either requires reindexing all reference sequences including the host genome for each new virus investigated or always mapping against all microbe and virus genomes. In addition, non-unique alignments are generally not resolved, which is a problem for rRNA reads which also map to rRNA pseudogenes in the host genome or a meta-transcriptomic screen against all known microbe and virus genomes. To address this problem, we recently extended our context-based RNA-seq mapping approach ContextMap [BCZF12] to allow parallel mapping against different read sources resulting in a unique mapping of each read to only one species/read source [BCZF13].

The parallel mapping approach could be integrated easily into ContextMap as even in the original implementation initial read alignments are clustered into so-called contexts that are treated independently until the last integrating step. Essentially a context represents a set of reads originating from the same

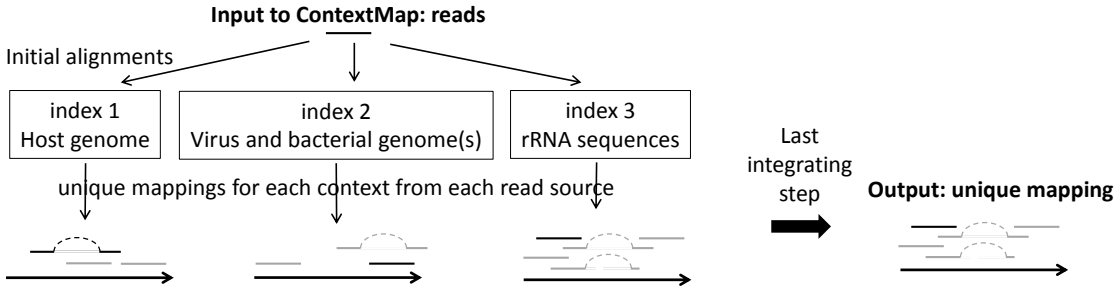

Figure 1: Parallel mapping against host, virus and bacterial genomes as well as rRNA is realized in ContextMap by (1) performing initial alignments against indices for several species/read sources to define contexts, (2) identifying best alignments for each read independently within each context and (3) resolving the resulting multiple alignments in the final integrating step.

stretch of the genome and likely corresponding to transcripts of the same or overlapping genes. Multiple alignments of reads to different contexts are allowed, which are then resolved in the last step. Thus, parallel mapping to multiple species could be included in ContextMap in a straightforward way by aligning against multiple sequence indices in the initial alignment step to recover contexts for different species (Figure 1). This approach is also included in the recent ContextMap 2 release, which allows the use of alternative short read alignment programs and can recover reads containing multiple exon-exon junctions or insertions or deletions [BKC+15].

Wide-spread disruption of host transcription termination in HSV-1

Parallel analysis of host and virus transcription and translation can lead to highly interesting insights not only into the infection process itself but also into important biological processes. This was illustrated by our recent study on HSV-1 lytic infection [REL+15], which established HSV-1 infection as an interesting model system to study transcription termination. HSV-1 is an important human pathogen that causes both common cold sores as well as life-threatening infections and rapidly shuts down host gene expression during lytic infection. In our study, we combined sequencing of 4-thiouridine (4sU)-labeled newly transcribed RNA (4sU-RNA) and ribosome profiling to study both host and virus transcription and translation during the full course of HSV-1 lytic infection. 4sU-labeling was performed in one-hour intervals during the first 8 hours of infection and ribosome profiling was performed at 0, 1, 2, 4, 6 and 8h post infection (p.i.).

Surprisingly, we found that the transcriptional up-regulation of 659 cellular genes was not matched by a respective increase in translational activity. Only 33 (0.34%) of translated genes showed increased translational activity at 8h p.i. When analyzing genes that were transcriptionally induced but not translated, we observed massive transcriptional activity upstream of their 5'-ends at late times of infection originating from neighboring upstream genes (Figure 2). This suggested that the transcription termination and cleavage machinery did no longer recognize or properly function at the termination signals of upstream genes, resulting in transcription into downstream regions by >100,000nt (denoted as 'read-out'). We found that read-out affected the majority of cellular genes and was correlated with a higher prevalence of non-canconical polyadenylation [poly(A)] signals. Although this indicated that non-canonical and likely weaker poly(A) signals were more strongly affected by disrupted transcription termination, the majority of genes with read-out still had the canonical AAUAAA poly(A) signal. Thus, poly(A) signal strength is certainly not the only factor influencing the extent of read-out.

Late in infection, read-out commonly extended over thousands of nucleotides into downstream genes (denoted as 'read-in'). At least 32% of genes showed >15% read-in at 8h p.i. and the extent of read-in depended on the distance to the next upstream gene. For genes with low or no transcription in uninfected

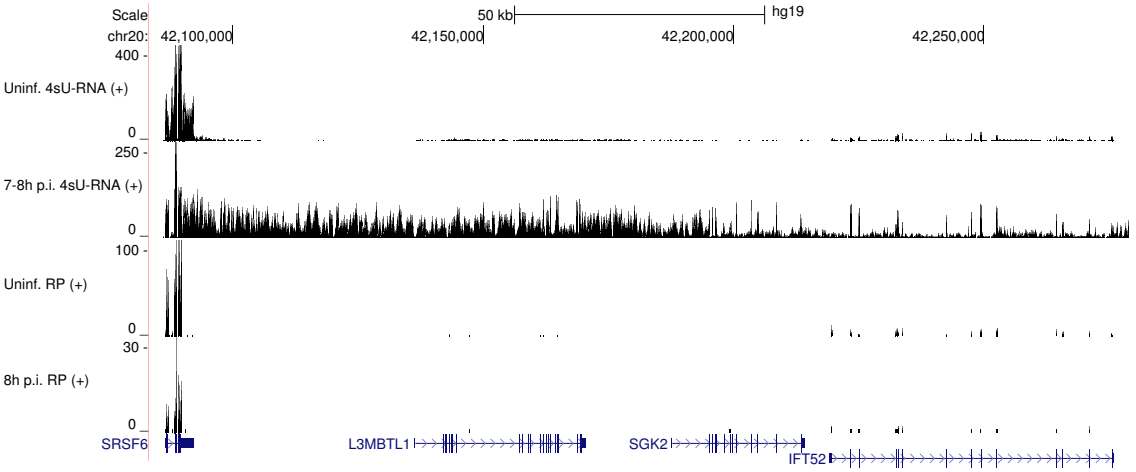

Figure 2: Disruption of transcription termination of the SRSF6 gene and read-in into the downstream SGK2 and IFT52 genes. The top two rows show transcriptional activity (4sU-RNA) and the bottom two rows translational activity (ribosome profiling, RP) in uninfected cells and at 8h p.i., respectively.

cells, read-in often exceeded endogenous transcript level, resulting in seeming 'induction'. This explained the discrepancy between transcriptional and translational induction. Furthermore, HSV-1 infection induced aberrant splicing events, which were enriched among genes with high read-out. Thus, splicing was already affected upstream of poly(A) sites suffering from read-out. Interestingly, 44% of the induced splice junctions were novel and 11% of these represented intergenic splicing between two neighboring genes connected by read-out and subsequent read-in. These intergenic splicing events thus conclusively demonstrated that disruption of transcription termination resulted in large RNA molecules spanning two or more cellular genes.

To investigate whether disruption of transcription termination was specific to the host or also affected HSV-1 genes, we identified reads containing part of a poly(A) tail, i.e. reads for which a partial alignment of the read start to the host or HSV-1 genome was followed by a stretch of A's. As coverage of the poly(A) tails was generally at least two orders of magnitudes lower than of the corresponding transcripts, only few poly(A) reads were recovered for the host genome. Coverage of HSV-1 transcripts, however, was in the order of tens-of-thousands of reads per genome position, allowing us to quantify poly(A) site usage of all but one viral gene (see Figure 3 for the UL39-50 gene segment). Viral poly(A) sites were almost exclusively preceded by an AAUAAA poly(A) signal. To investigate changes in poly(A) site usage in the whole HSV-1 genome throughout infection, we correlated gene expression upstream of each poly(A) site with the number of identified poly(A)-tailed reads. For 80% of poly(A) sites, this correlation was >0.9, which argued against regulated poly(A) site usage in HSV-1 infection and showed that disruption of transcription termination was host-specific.

References

[AWSG+14] Carolina Arias, Ben Weisburd, Noam Stern-Ginossar, Alexandre Mercier, Alexis S. Madrid, Priya Bellare, Meghan Holdorf, Jonathan S. Weissman, and Don Ganem. KSHV 2.0: a comprehensive annotation of the Kaposi's sarcoma-associated herpesvirus genome using next-generation sequencing reveals novel genomic and functional features. *PLoS Pathog*, 10(1):e1003847, Jan 2014.

[BCZF12] Thomas Bonfert, Gergely Csaba, Ralf Zimmer, and Caroline C. Friedel. A context-based approach to identify the most likely mapping for RNA-seq experiments. *BMC Bioinformatics*, 13 Suppl 6:S9, 2012.

[BCZF13] Thomas Bonfert, Gergely Csaba, Ralf Zimmer, and Caroline C. Friedel. Mining RNA-seq data for infections and contaminations. *PLoS One*, 8(9):e73071, 2013.

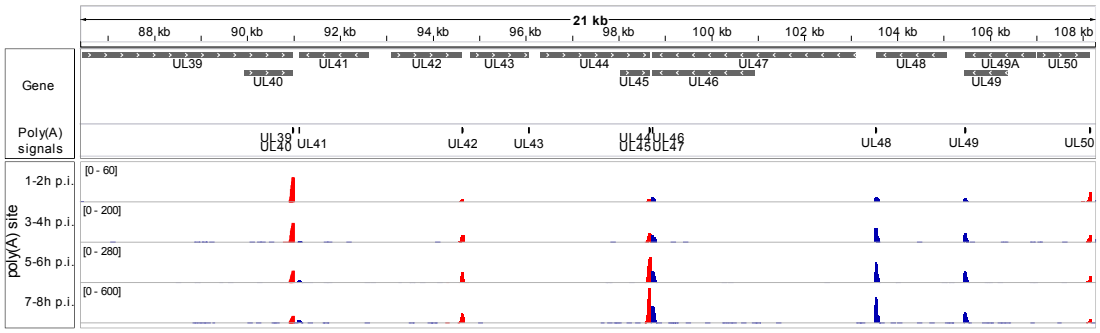

Figure 3: Poly(A) tail read coverage in 4sU-RNA for the UL39-50 gene segment (red = positive strand, blue = negative strand).

[BKC+15] Thomas Bonfert, Evelyn Kirner, Gergely Csaba, Ralf Zimmer, and Caroline C. Friedel. ContextMap 2: fast and accurate context-based RNA-seq mapping. *BMC Bioinformatics*, 16(1):122, 2015.

[CWF+12] Mauro Castellarin, Ren L. Warren, J Douglas Freeman, Lisa Dreolini, Martin Krzywinski, Jaclyn Strauss, Rebecca Barnes, Peter Watson, Emma Allen-Vercoe, Richard A. Moore, and Robert A. Holt. Fusobacterium nucleatum infection is prevalent in human colorectal carcinoma. *Genome Res*, 22(2):299–306, Feb 2012.

[DDS+13] Alexander Dobin, Carrie A. Davis, Felix Schlesinger, Jorg Drenkow, Chris Zaleski, Sonali Jha, Philippe Batut, Mark Chaisson, and Thomas R. Gingeras. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, 29(1):15–21, Jan 2013.

[DRR+08] Lars Dölken, Zsolt Ruzsics, Bernd Rädle, Caroline C. Friedel, Ralf Zimmer, Jörg Mages, Reinhard Hoffmann, Paul Dickinson, Thorsten Forster, Peter Ghazal, and Ulrich H. Koszinowski. High-resolution gene expression profiling for simultaneous kinetic parameter analysis of RNA synthesis and decay. *RNA*, 14(9):1959–1972, Sep 2008.

[IGNW09] Nicholas T. Ingolia, Sina Ghaemmaghami, John R S. Newman, and Jonathan S. Weissman. Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. *Science*, 324(5924):218–223, Apr 2009.

[MLW+12] Lisa Marcinowski, Michael Lidschreiber, Lukas Windhager, Martina Rieder, Jens B. Bosse, Bernd Rädle, Thomas Bonfert, Ildiko Gyry, Miranda de Graaf, Olivia Prazeres da Costa, Philip Rosenstiel, Caroline C. Friedel, Ralf Zimmer, Zsolt Ruzsics, and Lars Dölken. Real-time transcriptional profiling of cellular and viral gene expression during lytic cytomegalovirus infection. *PLoS Pathog*, 8(9):e1002908, Sep 2012.

[REL+15] Andrzej J. Rutkowski, Florian Erhard, Anne L'Hernault, Thomas Bonfert, Markus Schilhabel, Colin Crump, Philip Rosenstiel, Stacey Efstathiou, Ralf Zimmer, Caroline C. Friedel, and Lars Dölken. Widespread disruption of host transcription termination in HSV-1 infection. *Nat Commun*, 6:7126, 2015.

[SGWM+12] Noam Stern-Ginossar, Ben Weisburd, Annette Michalski, Vu Thuy Khanh Le, Marco Y. Hein, Sheng-Xiong Huang, Ming Ma, Ben Shen, Shu-Bing Qian, Hartmut Hengel, Matthias Mann, Nicholas T. Ingolia, and Jonathan S. Weissman. Decoding human cytomegalovirus. *Science*, 338(6110):1088–1093, Nov 2012.

[TPS09] Cole Trapnell, Lior Pachter, and Steven L. Salzberg. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics*, 25(9):1105–1111, May 2009.

[WBB+12] Lukas Windhager, Thomas Bonfert, Kaspar Burger, Zsolt Ruzsics, Stefan Krebs, Stefanie Kaufmann, Georg Malterer, Anne L'Hernault, Markus Schilhabel, Stefan Schreiber, Philip Rosenstiel, Ralf Zimmer, Dirk Eick, Caroline C. Friedel, and Lars Dölken. Ultrashort and progressive 4sU-tagging reveals key characteristics of RNA processing at nucleotide resolution. *Genome Res*, 22(10):2031–2042, Oct 2012.

[WGV12] Alexander J. Westermann, Stanislaw A. Gorski, and Jörg Vogel. Dual RNA-seq of pathogen and host. *Nat Rev Microbiol*, 10(9):618–630, Sep 2012.