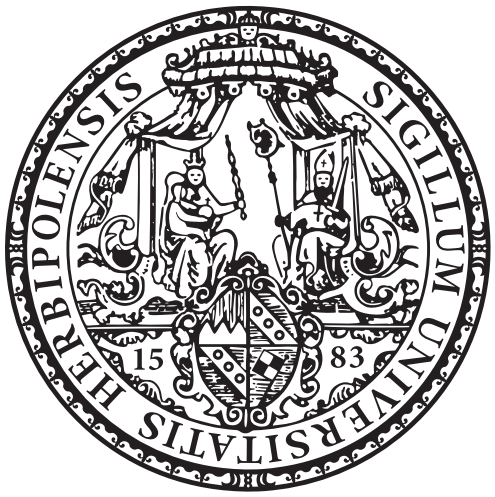


Algorithmic methods for elucidating the transcriptomic landscape of herpesviruses

Algorithmische Methoden zur Aufklärung der transkriptomischen Landschaft von Herpesviren

Doctoral thesis for a doctoral degree

at the Graduate School of Life Sciences

Julian-Maximilians-Universität Würzburg

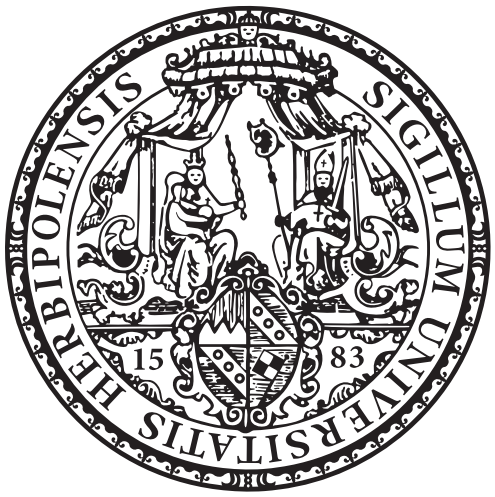
Section: Infection and immunity

submitted by

Christopher Sebastian Jürges

from Stuttgart, Germany

Würzburg 2021



Algorithmic methods for elucidating the transcriptomic landscape of herpesviruses

Algorithmische Methoden zur Aufklärung der transkriptomischen Landschaft von Herpesviren

Doctoral thesis for a doctoral degree

at the Graduate School of Life Sciences

Julian-Maximilians-Universität Würzburg

Section: Infection and immunity

submitted by

Christopher Sebastian Jürges

from Stuttgart, Germany

Würzburg 2021

Submitted on:

.....
Office stamp

Members of the Thesis Committee

Chairperson:	Prof. Dr. Christian Janzen
Primary Supervisor:	Jun-Prof. Dr. Florian Erhard
Supervisor (Second):	Prof. Dr. Lars Dölken
Supervisor (Third):	Prof. Dr. Thomas Dandekar
Supervisor (Fourth):	Prof. Dr. Sibylle Schneider-Schaulies

Date of Public Defence: .....

Date of Receipt of Certificates: .....

Affidavit

I hereby declare that my thesis entitled **"Algorithmic methods for elucidating the transcriptomic landscape of herpesviruses"** is the result of my own work. I did not receive any help or support from commercial consultants. All sources and/or materials applied are listed and specified in the thesis.

Furthermore, I verify that this thesis has not yet been submitted as part of another examination process neither in identical nor similar form.

Würzburg, December 2021

Christopher Sebastian Jürges

Eidesstattliche Erklärung

Hiermit erkläre ich an Eides statt, die Dissertation **"Algorithmische Methoden zur Aufklärung der transkriptomischen Landschaft von Herpesviren"** eigenständig, d.h. insbesondere selbstständig und ohne Hilfe eines kommerziellen Promotionsberaters, angefertigt und keine anderen als die von mir angegebenen Quellen und Hilfsmittel verwendet zu haben.

Ich erkläre außerdem, dass die Dissertation weder in gleicher noch in ähnlicher Form bereits in einem anderen Prüfungsverfahren vorgelegen hat.

Würzburg, Dezember 2021

Christopher Sebastian Jürges

Danksagung

Da ich nun am Ende meines Doktors angelangt bin, möchte ich hier noch einmal die Chance nutzen, um mehreren Personen meinen Dank auszusprechen, die dies alles ermöglicht haben.

Mein erster Dank gilt meinem ersten Supervisor Florian Erhard, der mir es grundlegend überhaupt erst ermöglicht hat meinen Doktor hier starten zu können. Dank ihm durfte ich an sehr interessanten Themen forschen und stand mir zusätzlich auch noch jederzeit bereit wenn ich Fragen hatte. Sein enormes Wissen in der Statistik spornte mich immer an in diesem Gebiet ebenfalls noch besser zu werden! Ebenfalls möchte ich hier auch meinen Dank an meine weiteren Supervisor Lars Dölken, Thomas Dandekar und Sibylle Schneider-Schaulies aussprechen, die mir jederzeit mit hilfreichen Kommentaren zu meinem aktuellen Forschungsthema weiterhelfen konnten.

Ein weiterer Dank gilt meinen Arbeitskollegen der AG Erhard sowie AG Dölken, insbesondere Kevin Berg, Teresa Rummel, Ihsan Muchsin, Thomas Hennig, Manivel Lodha, Lara Djakovic, Adam Whisnant und Andrea Milic, die zu einem sehr angenehmen Arbeitsklima beigetragen haben.

Ein großer Dank gilt meiner Familie und meinen Freunden, die mir immer mit Rat und Tat und aufmunternden Worte zur Seite standen. Insbesondere möchte ich mich hier bei meinem Bruder Tim, sowie meinen Freunden Julian und Mitja bedanken, die es mir zu nahezu jeder Zeit ermöglichten mich bei Videospielen mit ihnen abzulenken, wenn ich dies brauchte.

Mein letzter und größter Dank gilt aber meiner Freundin Steffi. Ohne sie wäre mein Doktor langweilig und trübselig gewesen. Danke, dass du so oft für mich gekocht hast, mir bei meinen biologischen Fragen immer helfen konntest und mich auch in den Zeiten ertragen hast in denen es bei mir mal nicht so gut lief. Auch wenn wir nicht das gemütlichste Zuhause hatten (vor allem im Winter), zu wissen, dass du da bist, hat mir gereicht, um mich jedes mal zu freuen nach Hause zu kommen.

Contents

List of Figures

List of Tables

Summary

Transcription describes the process of converting the information contained in DNA into RNA. Although, tremendous progress has been made in recent decades to uncover this complex mechanism, it is still not fully understood. Given the advances and reduction in cost of high-throughput sequencing experiments, more and more data have been generated to help elucidating this complex process. Importantly, these sequencing experiments produce massive amounts of data that are incomprehensible in their raw form for humans. Further, sequencing techniques are not always 100% accurate and are subject to a certain degree of variability and, in special cases, they might introduce technical artifacts. Thus, computational and statistical methods are indispensable to uncover the information buried in these datasets.

In this thesis, I worked with multiple high throughput datasets from herpes simplex virus 1 (HSV-1) and human cytomegalovirus (HCMV) infections. During the last decade, it has became clear that a gene might not have a single, but multiple sites at which transcription initiates. These multiple transcription start sites (TiSS) demonstrated to have regulatory effects on the gene itself depending on which TiSS is used. Specialized experimental approaches were developed to help identify TiSS (TiSS-profiling). In order to facilitate the identification of all potential TiSS that are used for cell type- and condition-specific transcription, I developed the tool iTiSS. By using a new general enrichment-based approach to predict TiSS, iTiSS proved to be applicable in integrated studies and made it less prone to false positives compared to other TiSS-calling tools. Another improvement in recent years was made in metabolic labeling experiments such as SLAM-seq. Here, they removed the time consuming and laborious step of physically separating new from old RNA in the samples. This was achieved by inducing specific nucleotide conversions in newly synthesized RNA that are later visible in the data. Consequently, the separation of new and old RNA is now done computationally and, hence, tools are needed that accurately quantify these fold-changes. My second tool that I developed, called GRAND-SLAM proved to be capable to accomplish this task and outperform competing programs. As both of my tools, iTiSS and GRAND-SLAM are not specifically tailored to my own goals, but could also facilitate the research of other groups in this field, I made them publicly available on GitHub.

I applied my tools to datasets generated in our lab as well as to publicly available data sets from HSV-1 and HCMV, respectively. For HSV-1, I was able to predict and validate TiSS with nucleotide precision using iTiSS. This has lead to the most comprehensive annotation

for HSV-1 to date, which now serves as the fundamental basis of any future transcriptomic research on HSV-1. By combining both my tools, I was further able to uncover parts of the highly complex gene kinetics in HCMV and to resolve the limitations caused by the densely packed genome of HCMV.

With the ever-increasing advances in sequencing techniques and their decrease in cost, the amounts of data produced will continue to rise massively in the future. Additionally, more and more specialized omics approaches are appearing, calling for new tools to leverage their full information potential. Consequently, it has become apparent that specialized computational tools such as iTiSS and GRAND-SLAM are needed and will become an essential and indispensable part of the analysis.

Zusammenfassung

Transkription beschreibt den Prozess des Umwandelns von DNA-Information in RNA-Information. Obwohl in den letzten Jahrzehnten enorme Fortschritte bei der Aufdeckung dieses komplexen Mechanismus erzielt wurden, ist dieser Prozess bis heute noch nicht vollends verstanden. Mit den Fortschritten und der Kostensenkung bei den Hochdurchsatzexperimenten wurden immer mehr Daten gewonnen, die zur Aufklärung dieses komplexen Prozesses beitragen. Diese Sequenzierungsexperimente erzeugen allerdings riesige Datenmengen, welche in ihrer Rohform für den Menschen unverständlich sind. Darüber hinaus sind Sequenzierungstechniken nicht immer zu 100% genau und unterliegen einer gewissen Variabilität. In besonderen Fällen können sie sogar technische Artefakte enthalten. Daher sind computergestützte und statistische Methoden unerlässlich, um die in diesen Datensätzen verborgenen Informationen aufzudecken.

In dieser Arbeit habe ich mit mehreren Hochdurchsatzdatensätzen von Herpes Simplex Virus 1 (HSV-1) und Humanem Cytomegalovirus (HCMV) gearbeitet. In den letzten Jahrzehnten wurde deutlich, dass ein Gen möglicherweise nicht nur eine einzige, sondern mehrere Transkriptionsstartpunkte (TiSS) besitzt. Diese multiplen TiSS haben nachweislich regulatorische Auswirkungen auf das Gen selbst, je nachdem, welche TiSS verwendet wird. Nachfolgend wurden demnach spezielle experimentelle Ansätze entwickelt, um TiSS zu identifizieren (TiSS-Profiling). Um die Identifizierung aller potenziellen TiSS zu erleichtern, die für die zelltyp- und zustandsspezifische Transkription verwendet werden, habe ich das Programm iTiSS entwickelt. Durch die Verwendung eines neuen, auf allgemeiner Anreicherung basierenden Ansatzes zur Vorhersage von TiSS erwies sich iTiSS in integrierten Studien als anwendbar und war im Vergleich zu anderen TiSS-Erkennungsprogrammen weniger anfällig für falsch positive Ergebnisse. Eine weitere Verbesserung in jüngster Zeit wurde bei metabolischen Markierungsexperimenten wie SLAM-seq erzielt. Hier wurde der zeitaufwändige und mühsame Schritt der physischen Trennung von neuer und alter RNA in den Proben entfernt. Dies wurde erreicht, indem spezifische Nukleotidumwandlungen in neu synthetisierter RNA induziert wurden, die später in den Daten sichtbar sind. Daher wird die Trennung von neuer und alter RNA jetzt per Computer vorgenommen. Dies benötigt daraufhin nun aber neue Programme, welche in der Lage sind diese Werte genau zu quantifizieren. Mein zweites von mir entwickeltes Tool namens GRAND-SLAM hat sich als fähig erwiesen, diese Aufgabe zu erfüllen und übertraf konkurrierende Programme. Da meine beiden Tools, iTiSS und GRAND-SLAM, nicht speziell auf meine eigenen Ziele zugeschnitten sind, sondern auch die Forschung anderer Gruppen in diesem Bereich er-

leichtern könnten, habe ich sie auf GitHub öffentlich zugänglich gemacht.

Ich habe meine Tools auf Datensätze angewandt, die in unserem Labor erzeugt wurden, sowie auf öffentlich verfügbare Datensätze von HSV-1 bzw. HCMV. Für HSV-1 konnte ich mit iTiSS TiSS mit Nukleotidpräzision vorhersagen und validieren. Dies hat zu der bisher umfassendsten Annotation für HSV-1 geführt, die nun als grundlegende Basis für jede zukünftige transkriptomische Forschung zu HSV-1 dient. Durch die Kombination meiner beiden Programme konnte ich außerdem Teile der hochkomplexen Genkinetik von HCMV aufdecken und die durch das dicht gepackte Genom von HCMV verursachten Einschränkungen überwinden.

Mit den zunehmenden Fortschritten bei den Sequenzierungstechniken und den sinkenden Kosten wird die Menge der produzierten Daten in Zukunft weiter massiv ansteigen. Darüber hinaus gibt es immer mehr spezialisierte "Omics"-Ansätze, die neue Programme erfordern, um ihr Informationspotenzial vollständig auszuschöpfen. Folglich ist es offensichtlich geworden, dass spezialisierte Computerprogramme wie iTiSS und GRAND-SLAM benötigt werden und zu einem wesentlichen und unverzichtbaren Teil der Analyse werden.

Chapter 1

Introduction

1.1 Transcriptomics

The Central Dogma of molecular biology states that DNA contains the instructions to create proteins, which in turn are the essential parts driving the cellular machinery. However, the DNA-code needs to be converted into RNA-transcripts first [Crick, 1958]. This fundamental process is called transcription and many advances have been made over the past decades to unravel this complex mechanism, however, even to this day it has not been fully understood.

Transcriptomics describes the set of techniques used to analyze transcription, by capturing the whole or parts of the sample's transcriptome. In 1991, the first attempt on obtaining a portion of the human transcriptome was conducted, and resulted in 609 mRNA sequences being identified [Adams et al., 1991]. Back then, this was revolutionary, however, with the evolution and refinement of new high-throughput sequencing techniques and the accompanying reduction in cost, the volume of sequenced data has increased exponentially ever since. Today, complete transcriptomes of different tissues or disease states are routinely generated with up to single-cell precision [Melé et al., 2015, Kolodziejczyk et al., 2015, Sandberg, 2014].

Consequently, multiple subfields have emerged in biomedical research using transcriptomics to advance new technologies, such as the diagnosis of differences in diseased and normal tissues [Wang et al., 2009], quantifying gene expression changes of pathogen infected cells and host-pathogen immune interactions [Wu et al., 2008], as well as in many other areas [Garg et al., 2016, Govind et al., 2009, Verbruggen et al., 2009, Hobbs et al., 2014, Li et al., 2011].

With the ever increasing amount of data per experiment as well as the total number of conducted sequencing experiments overall, a new challenge emerged. The size of a dataset from a single sequencing experiment easily comprises several gigabytes. To analyze and more importantly interpret those datasets computational solutions are indispensable, which is where the now known field of Bioinformatics steps in. Here, computational techniques

are developed and implemented to extract the information hidden in these datasets and visualize them to guide their interpretation.

In this work, I implemented and applied computational methods to analyze high-throughput RNA-sequencing experiments in the context of herpesviruses. In the first chapter, I provide a brief overview of transcription in general, RNA-sequencing and the approaches applied to analyze them. This chapter ends with an introduction to herpesviruses, their phylogeny, life cycle and role in the clinical context. In the second chapter, I describe the specific experimental and RNA-sequencing techniques used throughout this study and point out their specific advantages and disadvantages. Afterwards, I describe the methods I implemented to analyze these datasets and their results in greater detail, ending with a conclusion of this work. Finally, an outlook is provided.

1.1.1 Transcription

The DNA of an organism contains multiple stretches of information needed to produce proteins, non-coding RNAs or other functionally relevant actors. Such regions are called genes. However, not all of them are active at the same time. Various states of an organism's life-cycle as well as the potential exposure to stress, demands different sets of genes to be active or inactive during those time points. Consequently, regulatory mechanisms need to exist. In cells, this regulation takes place on virtually all levels starting with the chromatin structure, making it physically possible or impossible to access certain DNA regions, respectively [Bell et al., 2011]. This goes on to transcriptional mechanisms controlling if and how much RNA is created [Saxonov et al., 2006, Haberle and Stark, 2018], with finally post-transcriptional or even post-translational mechanisms that control the activity of the produced transcript or protein, respectively [Ogorodnikov et al., 2016, Mann and Jensen, 2003]. All of these regulatory processes are highly interconnected and thus, depend on each other. However, in this thesis, I will mainly cover transcriptional regulation, where information in form of DNA is converted into an RNA-transcript.

Although, the exact details and interplay of the transcriptional machinery have still not been fully elucidated, in the past decades, multiple studies gave insight in the structural organization of factors driving transcription as well as how individual parts are regulated [Cramer, 2019].

The process starts by recruiting a multi-subunit polymerase, which catalysis the synthesis of RNA, named RNA-polymerase. Recruitment of the RNA-polymerase is driven by so called transcription factors (TFs), which, in turn, bind to promoter regions inside the DNA. One of the earliest eukaryotic promoter regions identified is the TATA-box, which was found in 1978 [Lifton et al., 1978] and is located $\approx$ 31 bp upstream of the actual transcription start site (TiSS) [Ohshima et al., 1981]. As the name suggests, it consists of consecutive thymine and adenosine base pairs and is capable of recruiting multiple TFs to initiate transcription [Bell and Tora, 1999]. Interestingly, it is only present for a small fraction of genes [Vo ngoc et al., 2017]. Moreover, it has been shown, that its presence induces a more focused transcription, where the RNA polymerase initiates precisely at a

specific nucleotide position. In contrast, many genomes contain regions with an increase in cytosine and guanine di-nucleotides known as CpG islands. These CpG islands seem to have the opposite effect to TATA-boxes, where the polymerase can initiate at multiple different positions [Carninci et al., 2006]. This behavior will be further analyzed in chapter 3, where I implemented a program to detect such initiation sites in datasets specialized for identifying TiSS. Further, in chapter 6 I show that the human cytomegalovirus (HCMV) evolutionary favored these TATA-boxes in conjunction with the initiator element (Inr), resulting in strong and defined transcription of its genes.

After the initiation process, the RNA-polymerase switches into the elongation phase, transcribing the full transcript of the respective gene. This phase, again, is driven by various TFs as well as the chromatin structure. Most importantly, failing to recruit these additional factors leads to stalling or the termination of the transcription process. Consequently, the initiation of transcription alone is not the driving factor to whether or not productive transcription takes place [Core and Adelman, 2019]. This is an important distinction to make and in chapter 6, I show that this is a process not exclusive to eukaryotic cells but also present in HCMV.

Finally, the whole transcription complex is terminating during the 3'-processing. The exact mechanism of this process varies depending on the organism as well as the specific transcription elongation complex. Consequently, many different models have been published, however, the exact biochemical details are still not fully understood. The general mechanism, however, for the termination of transcription of mRNAs consists of cleavage and polyadenylation [Porrua et al., 2016, Mischo and Proudfoot, 2013, Porrua and Libri, 2015, Beaudoing et al., 2000]).

Due to its highly regulated and regulatory nature, modifications in any of these steps could drastically alter the cell's gene expression profile. The results could be fatal, as for example cancer is one of the many results of misregulated gene expression [Hough et al., 2000]. To detect and analyze such differences for any organism we need to know the location of e.g. its transcripts and open reading frames (ORFs). Consequently, genomes need to be annotated beforehand. If these annotations are missing, consecutive analysis are almost impossible. For large and prominent organisms these annotations already exist (e.g. human, mouse, ...). However, the transcriptomic annotation of herpesviruses is mostly missing completely. In chapter 3, I provided such an annotation for the herpes simplex virus 1 (HSV-1). Here, I used my tool iTiSS [Jürges et al., 2021] to predict TiSS and combine these predictions with those of PolyA-sites. Subsequently, I combined the TiSS and 3'-ends into full length transcripts including all potential alternative splicing events, which was made possible by additional total RNA-seq datasets. In the end, this lead to the most refined annotation of HSV-1 to date.

1.1.2 RNA-sequencing

In order to gain insight into the transcription processes of cells and to obtain (at least part of) its transcriptome, RNA-sequencing is applied. In this process, the RNA of a sample

is captured and their biochemical properties (i.e. the order of nucleotides) are turned into digital data, which can then be analyzed further computationally.

Hereby, the term RNA-sequencing comprises the application of any sequencing approach, which themselves are divided into different generations depending on the time of their emergence [Wang et al., 2009, Chu and Corey, 2012, Garalde et al., 2018]. Second-generation sequencing is the currently most widely used approach and has improved the throughput massively compared to the earlier Sanger sequencing techniques [Sanger et al., 1977, Margulies et al., 2005]. This, in turn, also lowered the cost of conducting such experiments drastically. The improvement was achieved by fragmenting RNA-molecules first and then read all of the fragments in a massively parallelized approach. Reconstruction of the individual sequences is subsequently done computationally. In contrast, in third-generation sequencing the fragmentation part is skipped and the sequences are read in their full length instead. It is currently gaining adoption, however, due to still relatively high error rates as well as lower coverage compared to second-generation, third-generation sequencing techniques are mostly used in a hybrid approach with second generation sequencing for now [Bleidorn, 2016, Koren et al., 2012, Bashir et al., 2012].

Independent of the choice of sequencing technique, in order to conduct a sequencing experiment, first, a library needs to be prepared. This involves the isolation of RNA from the sample and reverse transcribing it into complementary DNA (cDNA) and the addition of adapters needed for their subsequent sequencing. Most importantly, a multitude of different omics approaches exists, where each approach is capable of extracting specific information out of the data that remains otherwise elusive to other techniques [Sharma and Vogel, 2014, Herzog et al., 2017, Policastro et al., 2020]. The precise steps from the treatment of cells over the library preparation up until the final sequencing are archived in sequencing protocols. Depending on the goal of the experiment, one can choose from multiple different sequencing protocols or even implement new ones. Here, cells could be treated differently, or additional steps can be included in the library preparation in order to e.g. only sequence a subset of RNA relevant for the scientific question of the experiment. For example, roughly 95% of RNA inside a cell is ribosomal RNA (rRNA) [Kukurba and Montgomery, 2015]. However, for most studies these RNA-molecules are of no interest. Consequently, in order to prevent wasting 95% of the reads to rRNA, they are depleted during the library preparation. Over time, even more complex and sophisticated library preparations were implemented. For example, in organisms with densely packed genomes and therefore multiple overlapping transcripts, normal RNA-seq fails to capture precise 5'-ends. Therefore, omics approaches such as dRNA-seq [Sharma and Vogel, 2014] were introduced, which remove non-capped RNA-fragments in the library preparation step leaving only fragments aligning to the 5'-ends of their respective RNAs. In turn, this makes it easier to annotate TiSS. In contrast, metabolic labeling approaches such as SLAM-seq [Herzog et al., 2017] introduce nucleotide analogues into the living cells for a few hours. This allows to dissect newly transcribed RNAs from old ones after sequencing. Throughout my thesis, I analyzed sequencing data generated by a multitude of different sequencing protocols, which I introduce in more detail in chapter 2.

No matter what omics approach is chosen, in the end, the sample's RNA is converted into digital data. The most common file-format is the FastQ format, which contains an entry for each sequenced fragment. Each entry consists of an identifier, the read sequence itself as well as a quality string, which depicts for each nucleotide of the sequence the certainty of the base caller.

Most importantly, however, sequencing alone does not help to understand the underlying system. Additional bioinformatic analysis is necessary to extract the information hidden in these datasets. Here, it is very important to consider every aspect of the applied sequencing protocols to the respective samples, as it dictates the consequent analysis steps. Consequently, this process thrives from a good communication between the wet-lab researchers conducting the experiments and the corresponding bioinformatician responsible for the downstream analysis.

1.1.3 Data processing and analysis

The bioinformatic part of hight-throughput sequencing experiments is divided into two parts, the data processing and the actual analysis. The data processing consists of comparatively similar steps independent of the sequencing protocol, whereas the analysis part needs to be individually adapted depending on the scientific question as well as the sequencing protocol used.

The data processing step begins with a quality control of the sequenced reads, where bases with low quality as well as potential adapter sequences are trimmed off. Resulting reads that are too short are discarded. In our case, the threshold was set to a minimum length of 18 bp per read, which makes it statistically almost impossible for random mapping of reads to false locations inside the genome ($4^{18} = 68.7$ billion; the human genome consists of ≈ 3.2 billion bases). This is necessary to ensure that all bases of a read are indeed representations of the respective nucleotides in the original RNA-molecule. Afterwards, the reads are mapped against the reference genomes of potential contaminants. Contaminants such as mycoplasm can induce cellular responses. Consequently, any observed interesting responses of the cell during an infection experiment for example cannot necessarily be traced back to the infection. If no contaminants are detected, the reads are further filtered by mapping and removing those that align to parts of the gnome uninteresting for the conducted study such as ribosomes. This is important, as it has been shown that not excluding rRNA reads could result in falsely annotated proteins [Tripp et al., 2011]. Finally, the remaining reads are mapped against the reference genome(s) of the organism in question (see Fig. 1.1). During my work I used bowtie2 [Langmead and Salzberg, 2012] for very short reads (<35 bp), STAR [Dobin et al., 2013] for medium reads (<200 bp) and GMAP [Wu and Watanabe, 2005] for long read alignments used by third generation sequencing experiments.

Although, for most sequencing experiments these initial processes can be done with default parameters for the respective tools, for certain omics approaches, changes need to be made. Those changes can range from very obvious ones like long read sequencing vs. short read sequencing or paired-end vs. single-end sequencing, where the appropriate files need to be

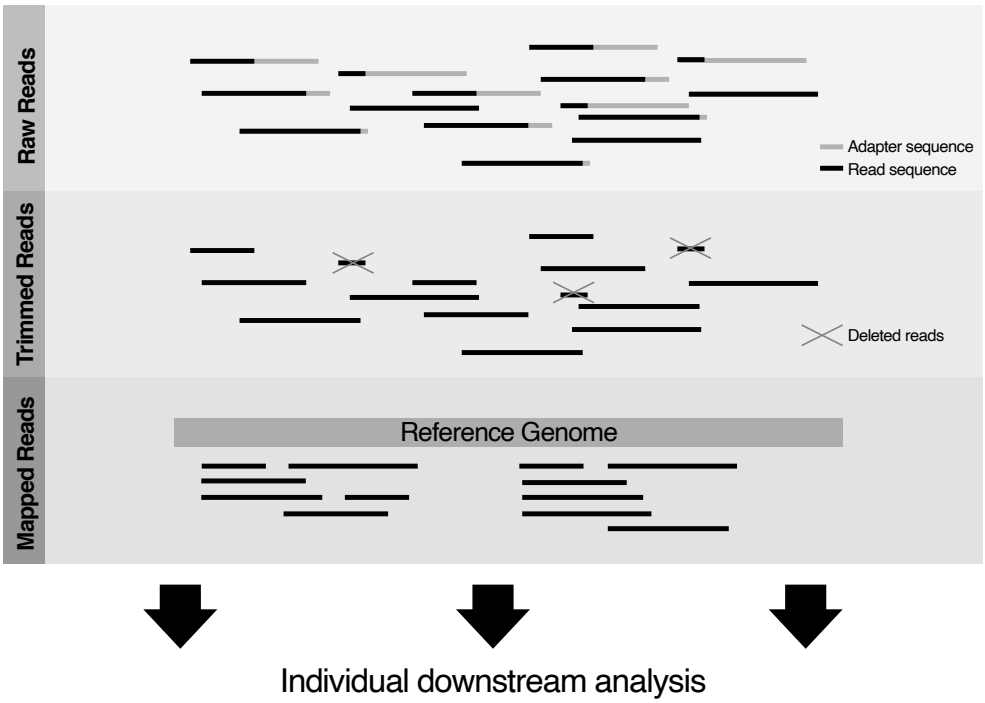


Figure 1.1: Raw read data processing and analysis workflow. The raw data consists of reads potentially containing adapter sequences. Those adapter sequences are trimmed off and remaining reads that are too short are discarded. Afterwards, reads are mapped to the reference genome. This could either be the reference of the investigated organism or of potential contaminants (e.g. Mycoplasm) or undesired parts (e.g. rRNA). If one of the latter is the case, only reads that did not map to the reference genome are mapped in the downstream mapping process. Finally, the mapped reads are used for the downstream analysis, which needs to be highly adapted to the omics approach and scientific question.

provided in the mapping process, up to very small ones like higher error rates induced by e.g. the SLAM-seq [Herzog et al., 2017] protocol. Here, deliberate thymine to cytosine mismatches (T→C) relative to the reference genome are incorporated. Importantly, mapping programs such as STAR [Dobin et al., 2013] use the number of mismatches, amongst other criteria, to filter mapped reads. In such SLAM-seq experiments, however, an overall higher error rate is expected and deliberate. Reads with multiple T→C mismatches contain valuable information and not mapping them would therefore be unfavorable. Consequently, parameters need to be adjusted depending on the chosen omics approach. In chapter 5, I will go into more detail on handling and analyzing SLAM-seq experiments, while introducing our tool GRAND-SLAM, specifically designed to use T→C mismatches to estimate the transcriptional activity of genes during the past hour(s).

After successful data processing, the subsequent specialized analysis starts, which solely depends on the applied omics approach as well as the scientific question. Here, it is important to choose the right tools in order to leverage the full information potential of the datasets

and prevent misinterpretations of the data. For example, the aforementioned SLAM-seq protocol produces data that is very similar to normal RNA-seq data. Consequently, tools designed for such data (e.g. splice-site detection) can be applied and would provide valuable information. However, the information about the current transcriptional activity yielded by the number of T→C mismatches per read would then be completely overlooked. Most notably, often simple approaches such as in this case only counting T→C mismatches still do not leverage the full information potential of these datasets. Therefore, specialized computational analysis methods are required to extract all of the information hidden in these datasets. In chapter 5, I followed this principle and implemented GRAND-SLAM, which uses a statistical model to infer new to total ratios per genes.

Additionally, even though tools specifically designed for a certain type of dataset already exist, it is important to always reevaluate them and make sure they catch every aspect that the particular dataset holds. For example, many TiSS-profiling sequencing protocols exist, generating similar but not identical datasets. Many tools such as CAGEr [Haberle et al., 2015] or CAGEfightR [Thodberg et al., 2019] were implemented to predict TiSS in these datasets. However, they all overlooked the problem that TiSS-profiling datasets can contain a number of reads that do not correspond to actual bona-fide TiSS. This is especially noticeable in datasets, where RNA-fragments downstream of the actual TiSS are not fully depleted and hence, the enrichment of reads at the TiSS is comparatively low [Whisnant et al., 2020]. In chapter 3, I present my tool iTiSS [Jürges et al., 2021], which also predicts TiSS in these types of datasets. However, with the new enrichment-based approach of iTiSS, it was able to differentiate between bona-fide TiSS and artifacts even in depletion-based datasets and was therefore applicable in our integrative HSV-1 annotation project discussed in chapter 4, where those other tools would have failed.

Finally, it is important to mention that each omics approach can only capture specific aspects or parameters of a system. For instance, TiSS-profiling approaches only identify the start sites of transcripts. However, to accurately annotate the whole transcriptome, the 3'-ends as well as all the potential splice-sites are also needed. Consequently, TiSS-profiling data alone cannot be used to annotate genes, but it needs to be combined with 3'-sequencing data as well as normal RNA-seq. Further, all of these approaches only look at transcriptional events, but fail to capture translation. For that, ribosomal profiling (Ribo-seq) is necessary, which in turn requires additional programs analyzing this new type of data. In the end, all of the information provided by the individual programs need to be combined in order to understand the whole system. As individual programs often follow their own file-standards, output files must be merged and pipelines designed to analyze as well as visualize the data. In particular the last part is important to make information accessible for the research community. In chapter 4, we conduct such an integrative analysis of heterogeneous omics data to annotate the HSV-1 transcriptome and translatome. We identified 5'-ends, 3'-ends as well as splice sites from corresponding datasets. Together with the open reading frames (ORFs) identified using Ribo-seq data, we provided the most comprehensive annotation to date for this large DNA virus, as well as a unifying and systematic nomenclature and a viewer to facilitate further research on it. With that, the

highly complex sequencing data and outputs of several prediction programs are visible at once and easy to understand.

1.2 Herpesvirales

The herpesvirales order describes viruses with a linear double stranded DNA packed inside an icosahedral capsid (125nm in diameter) surrounded by tegument proteins, all of which is enveloped inside a protein-lipid membrane. Almost all of herpesviruses target vertebrate species as their host [Domingo et al., 2008]. The major characteristics that they have in common and makes them special is the ability to persist in a lifetime long latent form inside the host, but also being able to switch back into a lytic infection cycle, producing infectious virus particles inside the host's cell destroying it in the process. Its transcription, synthesis and nucleocapsid assembly all occur in the nucleus, but all or at least part of their tegument are received in the cytoplasm and so is the envelope process. Finally, all of the different herpes species use a large array of enzymes in their nucleic acid metabolism, DNA synthesis and processing of proteins. Further, terminal and internal repeat regions as well as the ability to contain sequences that can be lost or duplicated during passage increasing variation of single species, marks an additional interesting but at the same time threatening characteristic. All of this is possible as the earliest form of the herpes virus are dated back as far as 200 million years, giving it enough time to adapt to its hosts .[Fields et al., 2013, McGeoch et al., 1995]

The G+C content of different herpes species varies significantly between 32% and up to 75% [Fields et al., 2013, King et al., 2012]. Further, with a genome size ranging between 125 to 295 kbp they are comparatively large viruses only dwarfed by the more recently identified giant viruses [La et al., 2003]. Consequently, it is no surprise that herpesviruses can contain hundreds of ORFs coding for their respective proteins. In turn, ORFs need to have a transcript they are transcribed from. Of note, prior to my work, the annotations for HSV-1 and HCMV were only based on prediction of ORFs from the genome sequence and lacked a large number of translated small ORFs and the exact positions of most mRNAs. Recent publications showed that the coding potential for Kaposi's sarcoma-associated herpes virus (KSHV) and Epstein-Barr Virus (EBV) is significantly larger than previously thought [Stern-Ginossar et al., 2012, Arias et al., 2014, Bencun et al., 2018, Erhard et al., 2018] with multiple former unknown small ORFs (<100 aa). In chapter 4, we annotate the translatome as well as the transcriptome of HSV-1. Here we showed that a lot of those smaller ORFs are located upstream of bigger ORFs (uORF) on the same transcript. In mammalian cells, uORFs are known to govern the expression of their respective large ORF downstream [Vattem and Wek, 2004]. With herpes viruses relying heavily on their controlled translation of proteins, it is not far fetched to assume that those uORFs have a major impact on transcriptional as well as translational control, too. With our new annotation of HSV-1 we set the foundation to help analyze these phenomenons, as, for the first time, we now provide associations between ORFs and their respective transcripts.

1.2.1 Phylogeny

Based on the NCBI taxonomy database there are a total of 103 virus species identified and categorized under the herpesvirales order (accessed 3.4.20). The order is subdivided into three families, which are the herpesviridae (containing mammal, burd and reptile viruses), the alloherpesviridae (containing bony fish and frog viruses) and malacoherpesviridae (containing only an abalone and ostreid virus). All of which are subdivided into further subfamilies, genera and finally species. In this thesis, we focus only on the herpesviridae family, which is subdivided into three subfamilies consisting of α-herpesvirinae, β-herpesvirinae and γ-herpesvirinae (see Figure 1.2) [Davison, 2010].

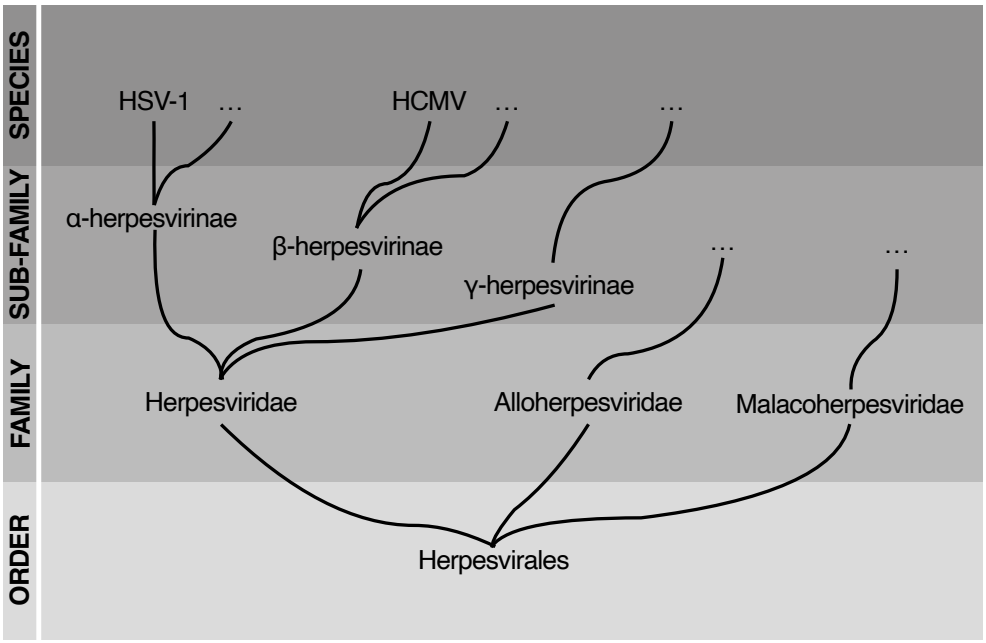


Figure 1.2: The phylogenetic tree of the herpesvirales order up to the species used in this thesis. It roots in the order herpesvirales. The subsequent family contains the herpesviridae, alloherpesviridae and malacoherpesviridae family. The subfamilies of herpesviridae are α-herpesvirinae, β-herpesvirinae and γ-herpesvirinae. The subfamilies of alloherpesviridae and malacoherpesviridae are omitted as they are not touched upon in this thesis. The two species encountered in this thesis are HSV-1, categorized in the α-herpesvirinae subfamily and HCMV, categorized in the β-herpesvirinae subfamily.

The α-herpesvirinae subfamily describes herpesviruses with a variable range of different hosts, a very rapid reproduction cycle, which is further shown in a rapid spread in culture, as well as an efficient destruction of the host's cell and the capability to establish latency in sensory ganglia (although, not exclusively) [Fields et al., 2013]. In chapter 4, I focus on HSV-1, which is part of the α-herpesvirinae subfamily. It infects primarily human cells and has a quick reproduction cycle. Its replication already initiates at 2 hours past infection (h p.i.), first virus particles are released at 4 h p.i. and >80% of translational activity in

the infected cells is viral at 8 h p.i. [Rutkowski et al., 2015].

The β-herpesvirinae subfamily comprises herpesviruses with a very restricted host range and longer reproductive cycles with up to 7 days. Consequently, infection also progresses slowly in culture [Fields et al., 2013]. In chapter 6 we focus on HCMV, which is part of the β-herpesvirinae subfamily. It infects solely human cells and has a reproductive cycle of roughly 3 days.

The third and last subfamily of herpesviridae are the γ-herpesvirinae. Herpesviruses associated with this subfamily are restricted to the family or order of their respective hosts. They replicate in lymphoblastoid cells and can also infect epithelioid and fibroblastic cells in some cases. They can also establish latency, which primarily happens in lymphoid tissue. In general, they are specific to either T- or B-lymphocytes [Fields et al., 2013]. However, during my work I did not encounter any datasets comprising data of γ-herpesvirinae.

1.2.2 Life cycle

The three major components of the herpesvirus life cycle consist of the initial infection, the lytic replication and latency. After initial infection of the cell, the virus has two options by either switching into the lytic infection cycle or persisting in the cell in the latent phase.

During the lytic infection cycle, the virus is replicating itself, suppressing its host immune response and killing it, resulting in the spread of virus particles infecting neighboring cells. Here, various genes are expressed in a highly regulated fashion. In general, genes of herpesviruses are categorized into distinct kinetic classes consisting of immediate early (IE), early (E) and late (L) genes, which are based on viral DNA replication and protein synthesis. However, in chapter 4 and 6, I show that HSV-1 and HCMV employ more granular expression kinetics and, more importantly, that they differ between these two. In HSV-1 genes are generally regulated by starting their expression at different time points. During the course of infection, the expression strength then increases for all of them. In contrast, HCMV genes seem to have a clear timepoint where their expression peaks and subsequently switch into an inactive state. This strict regulation of genes is a necessity for the virus to immediately and continually regulate the immune responses of the host in order to evade it. Further, herpesviruses depend on the cellular mechanisms for their replication process by using the polymerases of the cell as well as its compartmentalization to express different proteins in different places of the cell. The outcome of the lytic cycle is identical for all herpesviruses, where the cell dies and bursts, spreading virus particles to neighboring cells. [Fields et al., 2013]

In the latent phase, the viral genome of herpesviruses of the herpesviridae family generally forms into a closed circular loop. The expression of genes is then reduced to a minimum. However, the virus still retains the capability to switch from its latent phase back into the lytic phase at any point. The exact causes and mechanisms involved in communicating and executing these switches are yet not fully understood [Fields et al., 2013].

1.2.3 Role in the clinical context

Herpesviruses induce a wide variety of different symptoms in human that not only depend on the specific virus genus, but are also based on several other factors such as the amount of particles during infection and the immune responses of the host. However, a common feature of these viruses is their capability to induce latency with spontaneous reactivation during the life cycle of the infected individual as described in the previous section. During this phase, transmission is possible, which results in them being very common and prevalent in the human population. If the immune system of the infected host is not compromised, both HSV and CMV infections usually remain clinically silent [Fields et al., 2013]. However, primary infections of newborns as well as of the immune-compromised can be neurologically devastating or even fatal [Malani, 2010]. Where latent HSV infections are mostly associated with a weak immune system, CMV manages to persist in latent form even in immune-competent individuals [Fields et al., 2013].

Even though effective antiviral therapy is applied, HSV is the causative agent of around 15% of all infants and causes neurological morbidity in more than two thirds of the survivors [Kimberlin et al., 2001]. Unfortunately, those numbers are not declining, but have rather stayed stable in the early 2000s [Morris et al., 2008].

CMV has the highest rate of congenital infections of all known viruses [Fields et al., 2013]. Fortunately, most of them (around 90%) are asymptomatic [Demmler, 1991]. However, congenital CMV infections are still the main cause for subsequent developmental and neurological abnormalities [Williamson et al., 1990, Boppana et al., 1992]. Furthermore, similar to HSV, serious CMV infections in neonatal can cause neurological morbidity and can even be fatal.

This shows that even with all the medical advances in the past years, we are still unable to completely contain CMV and HSV infections. However, with roughly 200 million years to adapt itself to the mammalian species, this is not surprising [McGeoch et al., 1995]. Consequently, to finally defeat the threats originating from herpes infections, fundamental research of it is necessary. I contributed to this: In chapter 4, we extended the annotation of HSV-1 from merely 80 coding sequences to 284 open reading frames as well as adding 201 transcripts, which were completely unknown before. A similar approach was applied to HCMV in chapter 6, where we annotate and categorize its TiSS based on their time of expression, revealing a very complex kinetic cascade of transcription as well as transcriptional regulation.

Chapter 2

Datasets

Every high-throughput sequencing experiment, whether it uses second or third-gen sequencing, produces datasets containing large amounts of reads. In order to retrieve the information buried inside these datasets, the use of existing or the implementation of new computational tools is necessary. However, as mentioned in the previous chapter, the applied analysis steps vary drastically depending on the chosen sequencing technique as well as on the applied omics approach. Consequently, in order to leverage the full information potential buried in each sequencing dataset, knowledge about the effects of different omics approaches on the resulting read-files is required.

During my studies I analyzed a multitude of datasets comprising different sequencing generations with various complex library preparation steps. Each dataset comes with its own advantages and disadvantages, which I will discuss in detail during this chapter.

2.1 Second generation sequencing

Second generation sequencing is still the most widely used sequencing approach. It can be run in two modes, single-end or paired-end sequencing. Where single-end sequencing sequences each cDNA fragment once (from 5' to 3'), paired-end sequencing sequences a single cDNA fragment from both ends. Consequently, for the same cost single-end approaches can sequence twice as many cDNA fragments as paired-end ones (sequencing cost is determined by the number of bases sequenced). However, with paired-end sequencing the additional information about fragment sizes as well as drastically reduced error rates in overlapping regions can, depending on the experimental context, outweigh these disadvantages. This is especially important in metabolic labeling experiments, which introduce intentional mismatches into the reads compared to the reference sequence. Here, the overlapping regions help to differentiate those intentional mismatches from error rates. For example, if the probability of an erroneous sequenced base is at 10^{-4}, the subsequent probability of seeing the same mismatch in both paired-end reads at the same position is at $10^{-4} \cdot 10^{-4} = 10^{-8}$. Therefore, this is highly unlikely to originate from a sequencing error, but rather occurred

due to an incorporated labeled nucleotide. We demonstrate the advantages of this additional information in paired-end sequencing data with our tool GRAND-SLAM introduced in chapter 5 and applied in chapter 6.

In the following sections I will introduce the different omics approaches used throughout my thesis and discuss their information content as well as their caveats. Further, I will link them to the respective chapters they were used in.

2.1.1 Total RNA-seq

The total RNA pool of a cell consists of rRNA, pre-mRNA, mature mRNA and noncoding RNAs (ncRNAs). However, as rRNA makes up roughly 95% of it, even in Total RNA-seq rRNA is depleted [Kukurba and Montgomery, 2015]. Consequently, the term "Total" here refers to the fact that not only mature (polyadenylated) RNAs are sequenced, but also premature ones. After the extraction of rRNA using specialized commercial kits such as RiboMinus (Life Technologies) or RiboZero (Epicentre) RNA is fragmented and converted into cDNA-molecules, which are subsequently sequenced [Kukurba and Montgomery, 2015].

The area of application for Total RNA-seq approaches are versatile. Although, it is not possible to determine nucleotide precise start- and end-sites of RNAs, it is capable of quantifying gene expression as well as identifying splice-junctions and therefore alternative transcript isoforms. In chapter 4, I use Total RNA-seq as an additional factor of validation for our transriptome annotation.

2.1.2 cRNA-seq

The cRNA-seq sequencing protocol was implemented by [Stern-Ginossar et al., 2012]. As it involves a circularization step to make the library, we coined it cRNA-seq, in analogy to dRNA-seq [Whisnant et al., 2020].

The cRNA-seq approach falls in the category of TiSS-profiling techniques and is therefore used to identify TiSS. The library preparation process involves the fragmentation of mRNAs into on average 100bp long fragments using partial hydrolysis in a bicarbonate buffer. As the average fragment size is 100bp, the fragments at the 5'-end and 3'-end are shorter. The subsequent extraction of only fragments with a size between 50-80nt leads to an enrichment of fragments originating from the 5'- and 3'-ends of mRNAs, respectively [Stern-Ginossar et al., 2012]. Of note, in the original paper the theoretical accumulation of reads at the 3'-end was disregarded. Interestingly, in practice, reads only enrich at the 5'-end with no visible enrichment at the 3'-end.

As the fragmentation step is random, many shorter fragments are produced as well throughout the transcript, which are also sequenced. Consequently, after the mapping step, the enrichment of reads at TiSS is relatively low compared to other TiSS-profiling methods, as more reads are also mapping downstream throughout the body of the transcript. However, this expected asymmetry of read coverage at TiSS upstream to downstream can further be used to differentiate between bona-fide TiSS and artefacts [Jürges et al., 2021]. This

information is used in chapter 3, where I present my TiSS-prediction tool iTiSS. We further used cRNA-seq datasets in chapter 4 to annotate the HSV-1 transcriptome.

2.1.3 dRNA-seq

Likewise to the aforementioned cRNA-seq technique, the dRNA-seq approach also belongs into the TiSS-profiling category. However, in contrast to the cRNA-seq technique, the dRNA-seq approach produces reads originating from the respective 5'-ends of mRNAs with only a minor fraction mapping downstream. This is achieved by adding a 5' to 3' exonuclease after the mRNA fragmentation step. This exonuclease then digests all fragments that are not protected by a 5'-cap [Sharma and Vogel, 2014]. The resulting fragments predominantly originate from mRNA 5'-ends. However, I discovered that dRNA-seq and similar approaches result in significant amounts of artifactual read accumulations not corresponding to bona-fide TiSS [Whisnant et al., 2020, Jürges et al., 2021]. As dRNA-seq is missing the additional information of upstream to downstream read coverage asymmetry that cRNA-seq possesses, it is impossible to distinguish artifacts from true TiSS. I therefore concluded that multiple TiSS-profiling methods must be combined in an integrated approach to minimize false positive TiSS. Further, I observed a substantial amount of reads mapping to the 5'-ends of snoRNAs ($> 30\%$), greatly decreasing the overall number of reads for other genes shown in chapter 6.

We used dRNA-seq data in an integrated approach to annotate the HSV-1 transcriptome in chapter 4. Further, it is used during the implementation of iTiSS in chapter 3.

2.1.4 PRO(cap)-seq

PRO-seq and its alteration PROcap-seq enrich reads at the 5'-ends of mRNAs and can therefore be also categorized into the TiSS-profiling category. However, whereas cRNA-seq and dRNA-seq sequence stable mRNAs, PRO-seq sequences nascent mRNAs, i.e. mRNAs that are currently actively transcribed by the polymerases [Mahat et al., 2016]. This is accomplished by first halting the transcription by rapidly isolating the nuclei of the cells and washing away any native nucleotides. Subsequently, isolated nuclei are incubated with biotin-labeled NTPs, allowing the polymerase to continue to transcribe one or at most a few more labeled nucleotides. Finally, by using streptavidin-coated magnetic beads, the labeled nascent RNA is pulled out, fragmented and sequenced. The slightly adjusted PROcap-seq method includes an additional step, which filters only for mRNA fragments that contain a 5'-cap, i.e. originate from the 5'-end of mRNAs [Mahat et al., 2016]. To further increase the amount of fragments originating from 5'-ends the PROcap-seq approach was further modified by adding flavopiridol 1 hour before harvesting the cells. Flavopiridol inhibits release from promoter proximal pausing [Chao and Price, 2001]. Consequently, polymerases already in the elongation phase finish their transcription, whereas newly initiated polymerases are stalled near the 5'-end [Parida et al., 2019]. This leads to an overall very high enrichment of reads at transcription initiation sites while also reducing artefacts.

However, as PRO-seq only captures transcription initiation events, it cannot differentiate between productive and unproductive transcription. Many cellular promoters are bidirectional, where they recruit the transcriptional machinery in both strand directions. However, in those cases only one of the directions usually results in a productive transcription of an mRNA transcript, whereas the transcription process of the polymerases transcribing in the opposite direction is quickly aborted and the resulting small transcripts are rapidly degraded [Mayer et al., 2015]. Therefore, PRO-seq experiments alone are not suitable to identify bona-fide TiSS and need to be used in conjunction with other TiSS-profiling methods to differentiate between productive and unproductive transcription events.

In chapter 6, I use a PROcap-seq dataset to show that unproductive transcription initiation events also occur in HCMV.

2.1.5 4sU-seq

4sU-seq termed datasets originate from experiments that build upon the Total RNA-seq protocol. However, before the cell-harvesting step, the cells are provided with a uridine analogue called 4-thio-uridine (4sU). Interestingly, the 4sU only has minimal adverse effects on the overall transcriptional machinery of the organism [Melvin et al., 1978, Woodford et al., 1988, Ussuf et al., 1995, Kenzelmann et al., 2007]. 4sU is then incorporated into the transcription process, where it replaces roughly 2-6% of the native uridines in newly transcribed RNAs [Herzog et al., 2017, Erhard et al., 2019]. The newly transcribed RNA is subsequently isolated from old RNA by thiol-specific biotinylation followed by affinity purification on streptavidin-coated magnetic beads [Dölken et al., 2008]. Both samples containing the newly transcribed and the pre-existing RNA, respectively, are then sequenced.

Consequently, by using the 4sU-seq approach, it is possible to detect short term kinetic changes even for very short-lived RNAs [Dölken et al., 2008], which is impossible with Total RNA-seq, as it assumes a steady state. However, this comes at the cost of a very laborious and complex preparation phase [Duffy et al., 2015]. Additionally, due to the low incorporation rate of 4sU, this method has a bias towards longer RNAs, as the likelihood of multiple 4sUs being incorporated is higher the longer the sequences are.

2.1.6 SLAM-seq

The SLAM-seq protocol was introduced in 2017 [Herzog et al., 2017]. Similar to the aforementioned 4sU-seq, it uses 4sU to be incorporated into newly synthesized RNA. However, instead of physically separating new from old RNA, a thiol-reactive compound, in this case iodoacetamide (IAA) is added prior to cDNA synthesis. The then alkylated incorporated 4sUs in the RNAs are misread during the reverse transcription step as a cytosine, leading to thymine to cytosine (T→C) mismatches in respect to the reference genome [Herzog et al., 2017]. Consequently, dissecting newly transcribed from old RNA needs to be done computationally after the mapping step. This drastically reduces the laborious complexity of this experiment compared to the 4sU-seq protocol.

However, with a low incorporation rate ranging from 2% up to a maximum of 6% [Herzog et al., 2017, Erhard et al., 2019] of all uridines paired with an average read length of 100bp, a lot of reads originating from newly synthesized RNA do not contain any T→C mismatches. Let the incorporation rate be 4%. Then, on average, we only get $100_{read-length} * 0.25_{thymines} * 0.04_{incorporation-rate} = 1$ T→C mismatches per read. In chapter 5, I introduce our tool GRAND-SLAM, which uses a statistical model to estimate the new to total rate (NTR) for each gene, which proved to be a much more reliable way than simply counting T→C mismatches due to the aforementioned complications.

2.1.7 dSLAM-seq

In chapter 6, I introduce our newly developed dSLAM-seq protocol. It is a combination of dRNA-seq and SLAM-seq. Consequently, it falls into the TiSS-profiling category, by producing read accumulations only at the 5'-end of respective mRNAs (dRNA-seq). However, in addition, those reads contain the characteristic T→C mismatches from the SLAM-seq part, making it possible to dissect newly transcribed from old RNA. This serves as a huge benefit in densely packed genomes such as for HSV-1 or HCMV. Here, it is not uncommon that multiple transcripts use the same polyadenylation site. The only difference between them is the use of varying TiSS. With transcripts mostly overlapping, SLAM-seq, where reads map throughout the whole length of transcripts, is unable to provide enough resolution to accurately calculate NTRs for each of the overlapping transcripts. By only looking at the TiSS, however, this problem is circumvented and accurate NTRs can be calculated for each transcript regardless of the overlap. I demonstrate this in chapter 6, where I apply dSLAM-seq to HCMV in order to accurately detect its TiSS as well as providing a temporal clustering for them.

2.2 Long-read sequencing

Long-read sequencing in comparison to second generation sequencing is capable of sequencing whole RNAs without any prior fragmentation step. Consequently, the read-lengths of the resulting datasets varies and is not fixed to a specific size. With the information of the whole RNA-molecule, long-read sequencing can be used to identify 5'-ends and 3'-ends of RNAs as well as alternative isoforms without the reconstruction of genes from multiple small RNA fragments or special sequencing protocols. However, these datasets are still plagued by relatively high error-rates and low coverage compared to second generation sequencing datasets [Koren et al., 2012, Bashir et al., 2012]. Although, it has to be noted that progress was made in the recent years, where the error-rates have been reduced significantly by sequencing the same RNA-molecule multiple times [Wenger et al., 2019].
Throughout my study we never generated long-read sequencing libraries by ourselves, however, in chapter 4 and 6, I use publicly available long-read sequencing datasets to validate our findings. The two approaches consist of PacBio-sequencing [Eid et al., 2009] and

MinION-sequencing, also known as Oxford Nanopore sequencing [Quick et al., 2014].

2.3 Ribosome profiling (Ribo-seq)

In Ribo-seq experiments, the protective ability of ribosomes to prevent nuclease digestion of the mRNA template currently translated by them is used. Ribosomes are immobilized using cycloheximide and unprotected mRNA-sequences are removed by nuclease digestion. Subsequently, only the mRNA-sequences protected by ribosomes are sequenced, also called the ribosomal footprint [Ingolia et al., 2009]. Consequently, each sequenced RNA-fragment represents the position of a ribosome. Accumulations of reads along a stretch of the reference genome then indicate actively translated ORFs. Recently, the tool PRICE was implemented, which uses an EM algorithm that uses the fact that the ribosomal footprints are of varying length to drastically increase the signal-to-noise ratio especially in regions of overlapping ORFs [Erhard et al., 2018]. I use Ribo-seq data in chapter 4 and 6 to identify large and small ORFs (sORFs) as well as use it as an additional means of validation for our prediction of TiSS.

Chapter 3

Integrative transcription start site identification with iTiSS

Motivation: *In chapter 2 I pointed out the differences in the vast variety of high-throughput sequencing datasets. Even in the same subcategory of transcription start site (TiSS) profiling, datasets obtained from the same biological sample but with different sequencing methods can look completely different. At the time of publication, multiple different tools were already implemented to aid in predicting TiSS. However, they were all tailored for their specific kind of data. While annotating the HSV-1 genome in chapter 4 we learned that relying only on a single type of dataset will produce a significant amount of false positives. Additionally, the available tools all relied on a count based or read-cluster model for detecting accumulations of reads. We realized that for the detection of bona-fide TiSS a local enrichment based method was indispensable. For that reason I implemented iTiSS as the first integrative approach to TiSS-profiling.*

Publication: *This chapter has been published in Bioinformatics [Jürges et al., 2021]. Here, I adapted the layout and incorporated the Supplementary Methods, which were submitted alongside the manuscript into the text.*

Individual author contributions: *See Appendix D*

3.1 Abstract

Many experimental approaches have been developed to identify transcription start sites (TSS) from genomic scale data. However, experiment specific biases lead to large numbers of false positive calls. Here, we present our integrative approach iTiSS, which is an accurate and generic TSS caller for any TSS profiling experiment in eukaryotes, and substantially reduces the number of false positives by a joint analysis of several complementary data sets.

3.2 Introduction

Accurate mapping of potential condition or cell type specific transcription start sites (TSS) with nucleotide precision is fundamental to many studies [Carninci et al., 2006]. Over the last years, multiple sequencing based methods, such as CAGE [Shiraki et al., 2003], dRNA-seq [Sharma et al., 2010], cRNA-seq [Stern-Ginossar et al., 2012], PRO-cap [Kwak et al., 2013] and many more have been developed to identify TSS experimentally. Basic computational tools for calling TSS have been developed for individual approaches [Parida et al., 2019, Georgakilas et al., 2020, Thodberg et al., 2019, Haberle et al., 2015]. However, we found that the integration of multiple TSS experiments is necessary to identify and remove experiment specific bias and to reduce the number of false positive TSS [Whisnant et al., 2020]. So far, no TSS calling method is able to deal with the specifics of individual experiments, and to integrate multiple data sets for accurate TSS mapping.

Here, we present iTiSS (integrative Transcription Start Site caller) to close this gap. We show the method's utility by comparing it to existing TSS callers on multiple data sets from different TSS profiling experiments. The key contributions of iTiSS are the inclusion of local enrichment of reads rather than mere read counts at TSS, and the integration of multiple data sets in its prediction process. Both aspects are currently overlooked by other tools, which can lead to many false positive TSS. We propose that, in general, integrative analyses of complementary experimental approaches are essential to call TSS with high sensitivity and specificity.

3.3 Approach

For iTiSS, we differentiate two categories of TSS profiling approaches: The first, which includes dRNA-seq and PROcap is characterized by extremely high signal (TSS reads) to noise (mRNA internal reads) ratios (see Fig. 3.1A) [Whisnant et al., 2020]. Such data also contain a large number of artefactual peaks, i.e. accumulation of reads that do not correspond to TSS [Whisnant et al., 2020]. The challenge with such data therefore is to discern such artefacts from bona-fide TSS. This is complicated by differences in expression levels among mRNAs spanning multiple orders of magnitude. In the second category, 5'-end reads are merely enriched compared to reads downstream in the gene body, which is e.g. the case for cRNA-seq. Thus, these data sets are noisier, but they have the advantage that the expected asymmetry of read coverage upstream and downstream of a bona-fide TSS can be exploited to exclude artefacts (see Fig. 3.1A).

iTiSS makes use of these features from both categories and can leverage data from multiple datasets for removing false positives. The user can choose which analysis mode (high or low signal to noise) is used for which dataset. Moreover, iTiSS merges TSS into transcription start site regions (TSRs) and is therefore able to identify both focused and dispersed transcription initiation [Haberle and Stark, 2018].

iTiSS is platform independent, and has no other dependencies than a Java (version $\geq$1.8)

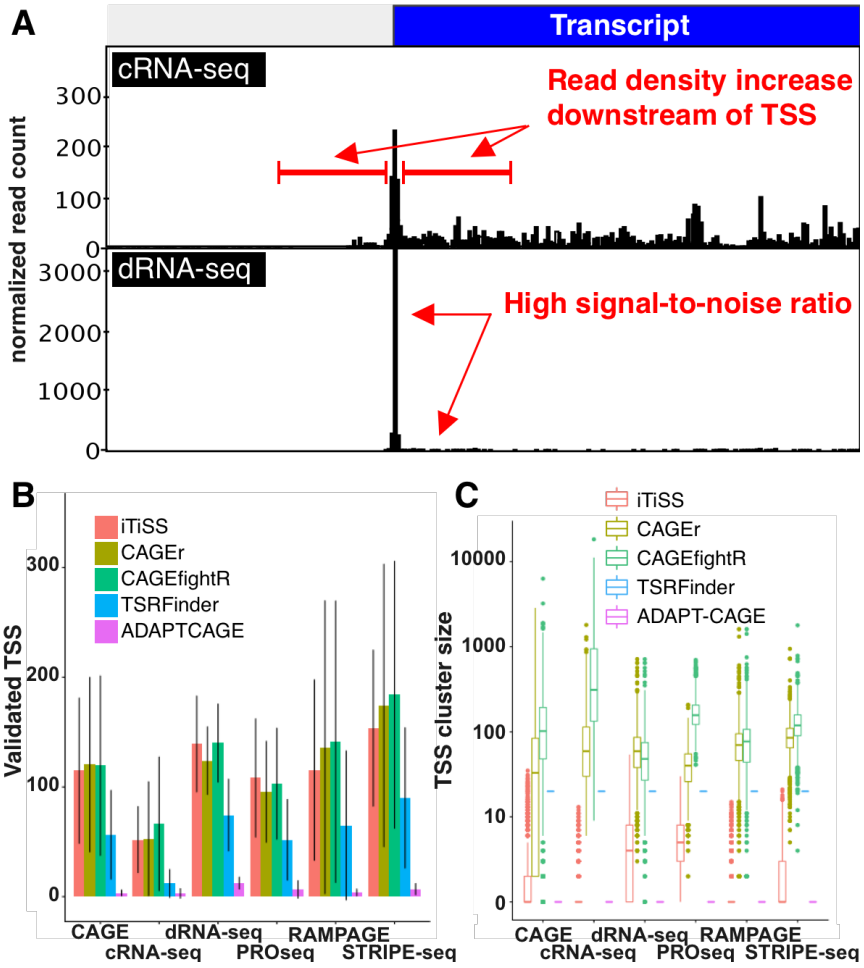


Figure 3.1: **A** Genome browser visualization of the number of reads (5'-ends) for two different TSS profiling methods (Noisy: cRNA-seq; High signal-to-noise: dRNA-seq). **B** Number of TSS from strongly expressed genes (n=1,842) that were identified in each dataset (x axis) and validated in each of the five other datasets by using the same TSS calling approach. The bars depict the mean number of validated TSS over the five other datasets, the error bars the standard deviation. **C** The distribution of the TSS cluster sizes of the 1842 highest scoring TSS per tool and data set. Boxes represent quartiles, whiskers the 1.5 * interquartile range in log10-scale.

runtime environment. In contrast to available tools, it directly processes BAM files and is able to deal with single-end or paired-end data as well as data sequenced in sense or antisense orientation.

Peak detection in high signal-to-noise data

To detect accumulations of reads in high signal-to-noise TSS profiling experiments, iTiSS uses a sliding window-approach. The window w has a user-defined width (default is 100 bp) and is moved in single nucleotide steps along the genome containing the read start counts $c_1, \ldots c\_w$ of each position inside it. For each position the fold change between the number of reads starting at the center position of the window against the mean number of reads starting at all other positions inside the window is computed:

$$F = \frac{c_{\frac{w}{2}} + 1}{\frac{1}{w}\sum_{i=1}^{w} c_i + 1} \tag{3.1}$$

A pseudo-count of 1 is added to prevent division by 0. A TSS is called if the fold change F exceeds a certain threshold. This threshold can be defined by the user, or is set automatically: First, all positions with more reads than the window average are identified (i.e. fold change $\geq$1). The fold changes of these candidates are then sorted in descending order, smoothed using a moving average approach (window size $0.2 \cdot$ #peaks) and log2 transformed. The x-axis and y-axis are subsequently normalized by their highest and lowest value, respectively. The threshold value is set where the curve has a slope of 1. Further noise is removed by filtering values occurring multiple times (default is 100, can be specified by the user), as observing exactly the same values multiple times is mostly due to positions with few reads.

Peak detection in noisy data

To detect accumulations of reads in noisy TSS profiling experiments, iTiSS uses a sliding window approach. For each position p, iTiSS computes the mean values μ_u and μ_d and standard deviations σ_u^2 and σ_d^2 of read start counts in a 100 upstream window u and a 100 downstream window d (windows sizes can be altered by the user), where the two windows u and d contain all read start counts inside their respective window. These are then used for computing the upstream and downstream z scores (z_u,z_d) of the read count c_p at the center position p (see Equation 2 & 3):

$$z_u = \frac{c_p - \mu_u}{\sigma_u^2} \tag{3.2}$$

$$z_d = \frac{c_p - \mu_d}{\sigma_d^2} \tag{3.3}$$

A TSS is called if both z scores exceed a certain threshold. This threshold can be user-defined, or is set by the same procedure as for high signal-to-noise data.

Read density change detection for noisy data

To identify significant changes of read start counts upstream and downstream of a potential TSS, iTiSS first identifies the positions in the upstream window and in the downstream window that are above and below the downstream mean. Enrichment of reads above the mean in the downstream window is then tested using Fisher's exact test. Instead of correcting the p-value for multiple testing, we again use the same thresholding procedure as described above. This is advantageous because we prevent choosing a pre-defined cut-off value and instead use a threshold that is based on the data itself. With the high variability between data sets, this approach yields more consistent results.

Detection of differential TSS usage

Differential TSS usage in multiple conditions was tested by using a likelihood ratio test based on the Dirichlet distribution: For each currently analyzed position p, we identified the vector a composed of read counts c_i starting at p in condition $i \in \{1, \cdots n\}$. In a similar manner, the vector b was computed as the total number of read u_i starting in a window upstream of p (default size 100 bp) in condition $i \in \{1, \cdots n\}$:

$$a = (c_1, \cdots c_n) \tag{3.4}$$

$$b = (c_1, \cdots c_n) \tag{3.5}$$

For non-differentially used TSS, both vectors follow the same multinomial distribution, for differentially used TSS, two multinomials are needed. It is straight forward to compute the maximum likelihood estimates of both models (relative frequencies of $a + b$, or relative frequencies of a and of b) and to compute the likelihood ratio. The p-value is computed by a χ^2 test with $n - 1$ degrees of freedom (when there are n conditions). The p-value threshold is determined by the procedure described above.

Merging of TSS into TSRs

After successful prediction of TSS, iTiSS offers the possibility to merge TSS in close proximity into transcription start site regions (TSRs). Here, the predicted TSS are sorted based on their position first. Then, starting with the lowest TSS position, each TSS and its neighboring TSS are merged into a TSR if their distance is equal or lower than a user defined threshold (default is 10). If the distance from the next following TSS to the TSS lastly added to the current TSR is below the user defined threshold as well, it is added to the same TSR, too. Consequently, if the user provides a very large range-threshold on a dataset with a lot of predicted TSS (e.g. for small densely packed genomes), this could, in theory, lead to all TSS being merged into the same single TSR. Hence, a lower range-threshold ($\geq$50) is recommended.

Merging of TSS can either be done on a per-sample basis or even on multiple samples, combining TSS predicted in e.g. different conditions into a single file.

3.4 Test setup

Datasets used in this study

We compared iTiSS to methods developed for CAGE, dRNA-seq, PROcap-seq, and cRNA-seq using data for human fibroblasts where TSS profiling has been performed using all four experimental approaches, and included additional data from more recent experimental approaches (see 3.1).

Table 3.1: Datasets were chosen based on the cell-type (human foreskin fibroblasts/keratinocytes) to keep the "comparability" high. Only uninfected (mock) samples were chosen. The only exception were the cRNA-seq data. Here, we combined all timepoints to make it comparable with the kinetic analysis, where we used the individual timepoints for iTiSS.

Type	Accession	Notes	Reference study
CAGE	https://fantom.gsc.riken.jp/5/datafiles/latest/basic/human.primary_cell.CAGEScan/	The dataset named Fibroblast Gingival was used	FANTOM5
cRNA-seq	GSE128324	For peak detection, all conditions starting with cRNA-seq were merged. For iTiSS kinetic detection, the 5 conditions of the first replicate was were used	https://www.nature.com/articles/s41467-020-15992-5
dRNA-seq	GSE128324	Only the mock conditions starting with dRNA-seq were used. Replicates were merged together	https://www.nature.com/articles/s41467-020-15992-5
PROcap	GSE113394	Only the PRO-Cap uninfected condition treated with Flavipiridol was used	https://mbio.asm.org/content/10/1/e02047-18#T1
STRIPE-seq	GSE142524	Only the untreated STRIPE-seq conditions for K562 cells were used	https://genome.cshlp.org/content/early/2020/07/02/gr.261545.120
RAMPAGE	https://www.encodeproject.org/experiments/ENCSR115BCB/		ENCODE-project

Testing methodology

A ROC or precision-recall-curve can be used to measure the prediction performance of tools. However, this is not possible here, because there is no gold-standard of TSS. Although, the human genome is very well annotated, we found that a lot of TSS are not very precise. Even though, it is easy to determine the condition and cell type specific active set of genes e.g. using RNA-seq, it is difficult to do this accurately for TSS (in the end this is the point of TSS profiling). Consequently, we had to use a custom performance metric. We took the RNA-seq data set from a time-course experiment of HSV-1 infection of human foreskin fibroblasts [Rutkowski et al., 2015] (GSE59717) and used kallisto [Bray et al., 2016] to calculate TPMs for all mock infected samples. All genes annotated in the ENSEMBL (v90) [Yates et al., 2015] database exceeding a total TPM value (sum over the 6 samples) of 100 were considered highly expressed, which resulted in 1842 genes. Next, we took the 1842 highest scoring TSS, which are inside an annotated gene region from each tool predicted in each of the four data sets. In theory, if one transcript per gene is active, those TSS should all predict the exact same TSS per gene in each data set. This assumption is obviously not 100% true as there exist genes with multiple active transcripts, however, it

is unbiased and does not favor or discriminate against a specific tool. It further mimics a real-world example, where one wants to find the most active TSS for a gene. Additionally, we gain a threshold-independent measurement, as we only look at the best TSS predicted by each tool, which therefore should result in the most favorable outcome for each tool. Subsequently, per tool and data set, we looked at the highest scoring TSS per gene (which should be the TSS of the respective active transcript of the gene) and tested, if this TSS was called in the other data sets for the same gene as well (within a 5 bp window). E.g. for the predictions by iTiSS in the dRNA-seq data set, we count how many of the highest scoring TSS per gene were also predicted in cRNA-seq, CAGE, PROcap, STRIPE-seq and RAMPAGE data set by iTiSS. Consequently, the so determined number of retrieved TSS per data set pair depict how well a certain tool can retrieve a TSS in all types of TSS profiling datasets, which is the essence of a well performing tool in an integrated setting.

TSS cluster size

Earlier studies suggest two types of promoters in context of the number and distribution of TSS in CAGE data. The 'sharp' promoter type is defined by a TSS peak at a single nucleotide position with no or close to no other TSS peak around it. The 'broad' promoter type on the other hand shows multiple TSS peaks distributed along a broad region. This was found to be mostly driven by either the presence of a TATA-box or CpG islands, respectively [Carninci et al., 2006]. Tools like CAGEr, which are specifically designed for CAGE data take that into account by merging TSS that are in close proximity [Haberle et al., 2015] into TSS clusters. However, if data sets have a high number of reads not mapping to bona fide TSS in close proximity, the reported TSS cluster sizes will become unrealistically large. To further emphasize this problem, we tested the reported cluster sizes of the programs once with our full list of TSS and once, where we removed TSS with no TATA-box upstream from it, which should result only in TSS with a 'sharp' promoter type [Carninci et al., 2006].

TSS-prediction tools tested in this study

The six tools were executed in the following way:
TSSPredator (v. 1.07.1beta) is only applicable to dRNA-seq data sets, as it predicts TSS by comparing an XRN-1 treated condition with an untreated one, which are only conducted for dRNA-seq experiments. Consequently, we cannot apply it to all four of our data sets, rendering it uncapable for an integrated approach, hence, its removal from our further analyzis.
TSRFinder was executed with default parameters. Data was provided by running custom scripts that converted our mapped-reads format (CIT) into bed-format usable by TSRFinder. The antisense pair-end sequenced PROcap data was consequently switched to the opposite strand.
ADAPT-CAGE (does not provide version numbers, commit 301fc1e0 was used) was ex-

ecuted with default parameters against the hg38 genome. As ADAPT-CAGE has its reference sequence hard-coded into it, which does not contain contig-annotations found in the ENSEMBL annotation we had to include those in the file "algorithm_init.pm". Further, the file "pol2_features.pl" threw exceptions as it was unable to read the data from the file "extended_tc_rep_flank100.valid.bed". We fixed this by removing three extra characters introduced in the first column. Further, we altered the file "fur.pm" for the PROcap data set to accommodate for its antisense paired-end sequencing method.

CAGEfighR (v. 1.6.0) was executed by first converting the BAM-files into BigWig-files using Samtools and Bedtools. Then, CTSSs were quantified using the quantifyCTSSs-function. Then, TSSs were predicted using the quickTSSs-function. Finally, TPMs were calculated using the calcTPM-function and all TSSs with a TPM smaller than 1 were removed using the subsetBySupport-function.

CAGEr (v. 1.28.0) was executed by first converting the BAM-files into BED-files using custom scripts. This was done to work around the problem, that CAGEr cannot handle anti-sense sequencing files natively. A CAGEexp object was created using the BSgenome.Hsapiens.UCSC.hg38 reference genome. To finally obtain the tag clusters CTSS were obtained using the getCTSS function, normalized using the normalizeTagCount function and clustered using the clusterCTSS function. Finally, the tagClusters function was used to obtain and write out the final TSS.

iTiSS' (v. 1.0) peak detection in high signal-to-noise data module (SPARSE_PEAK-module) was used to predict peaks for the dRNA-seq, PROcap and RAMPAGE. iTiSS already supports antisense sequencing by internally converting reads to the opposite strand. This option was therefore turned on for the PROcap data. For the cRNA-seq data set we used the peak detection in noisy data module (DIRTY_PEAK-module), the density changes detection module (DENSITY-module) and the differentially used TSS detection module (KINETIC-module). As false-positives are already filtered by combining the peak, density and kinetic modules, we lowered the threshold of all algorithms to include more potential TSS. For the peak module we set a z-score threshold of 2.5 and on a separate run a threshold of 6.0 (i.e. weak TSS and strong TSS), for the density module a p-value threshold of 0.001 was set and for the kinetic module a p-value threshold of 0.1 was set. iTiSS' merge module (TiSSMerger) was further used to combined the predictions of all three modules for the cRNA-seq data set. Here, we filtered for those that are either in the group of strong TSS, or are seen either in the weak TSS group and the density module or in the weak TSS group and kinetics module. This way of combining TSS called by the different modules in cRNA-seq is our advised way of dealing with the rather hard to identify TSS in noisy data sets. As all of those modules use different scoring metrices, we chose to always keep the z-score from the peak-calling module to keep things simple. For STRIPE-seq and CAGE only the DIRTY_PEAK-module was applied as they showed overall 'cleaner' peaks compared to cRNA-seq.

Further, for all programs that do not call TSRs we merged TSS predicted in close proximity (5bp) in the same data set onto the one with the highest score. For all programs calling TSRs, we always used the highest scoring TSS inside the TSR.

All programs were run on an Ubuntu Server (v.16.04.6) with two XEON 56 processing cores and 128 Gb of system memory. However, as iTiSS is not optimized for multithreading, yet, only two threads were used simultaneously.

All the custom scripts to extract and convert data, to create the Figures and Tables and to run the different programs as well as the altered ADAPT-CAGE files are available at ZENODO (https://doi.org/10.5281/zenodo.3860525).

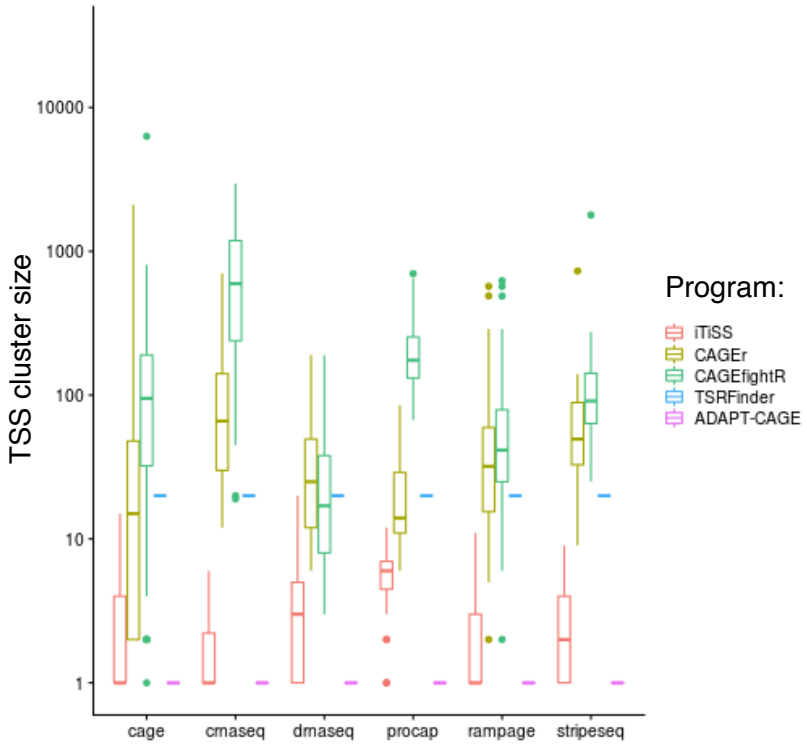


Figure 3.2: The TSS used in Fig. 3.1c were reduced to TSS having a TATA-box (TATAAA) located within 50bp upstream of them. The size of the TSS clusters reported by the individual tools for the respective datasets are then depicted here. Boxes represent quartiles, whiskers the 1.5 * interquartile range in log10 scale.

3.4.1 Results

iTiSS, CAGEr [Haberle et al., 2015] and CAGEfightR [Thodberg et al., 2019] substantially outperformed TSRfinder [Parida et al., 2019] and ADAPT-CAGE [Georgakilas et al., 2020] in terms of this reproducibility meas-ure (Fig. 3.1B and Fig. 3.3). However, CAGEr and CAGEfightR generally reported much larger TSRs than all other tools (Fig. 3.1C). This was also the case for TATA-box promoters (Fig. 3.2), where focused transcription initiation is expected [Carninci et al., 2006]. In extreme examples, CAGEr/CAGEfightR called every exon to be a TSS, with TSRs longer than 1kb (Fig. 3.4). Further, both tools reported substantially more TSS beyond the first exon than iTiSS (Fig. 3.5). Taken

together this indicates that the seemingly high reproducibility of CAGEr and CAGE-fightR is largely due to their tendency to call unrealistically long TSRs. This low specificity of CAGEr and CAGEfightR is prohibitive for their use in integrative TSS calling, as also many false positives might be validated in additional data sets. The very low number of overlapping TSS predicted by ADAPT-CAGE indicates that it is not suitable in an integrative approach. TSRFinder calls a TSS if the total number of reads in a moving window exceeds a certain threshold. Such an approach works reasonably well for high signal to noise data, however, this also leads to calling a lot of false positives in noisy data with dozens to hundreds of TSS called per gene similar to CAGEr and CAGEfightR.

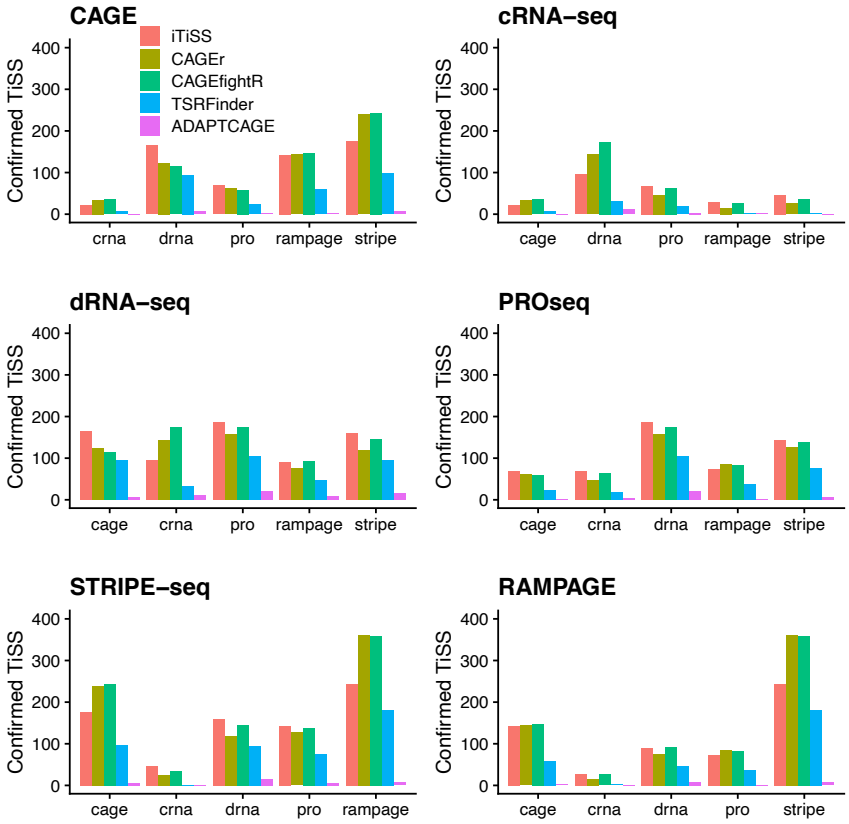


Figure 3.3: For each tool and dataset the 1842 highest scoring TSS that are located inside known genes are taken. Then, for each dataset the number of overlapping TSS with all the other datasets are depicted.

To highlight the advantages of an integrated approach, we used iTiSS to search for TSS identified by all four data sets (see section 3.3). This identified 34 yet unannotated TSS confirmed by all data sets (see Tab A.1 and Fig. 3.6). Many of those were either in close proximity to an annotated TSS (suggesting that the current annotation of these TSS is inaccurate), or in unannotated locations (suggesting novel transcripts).

Figure 3.4: Visualization of the predicted TSS clusters by iTiSS, CAGEr and CAGEfightR for the gene GAPDH.

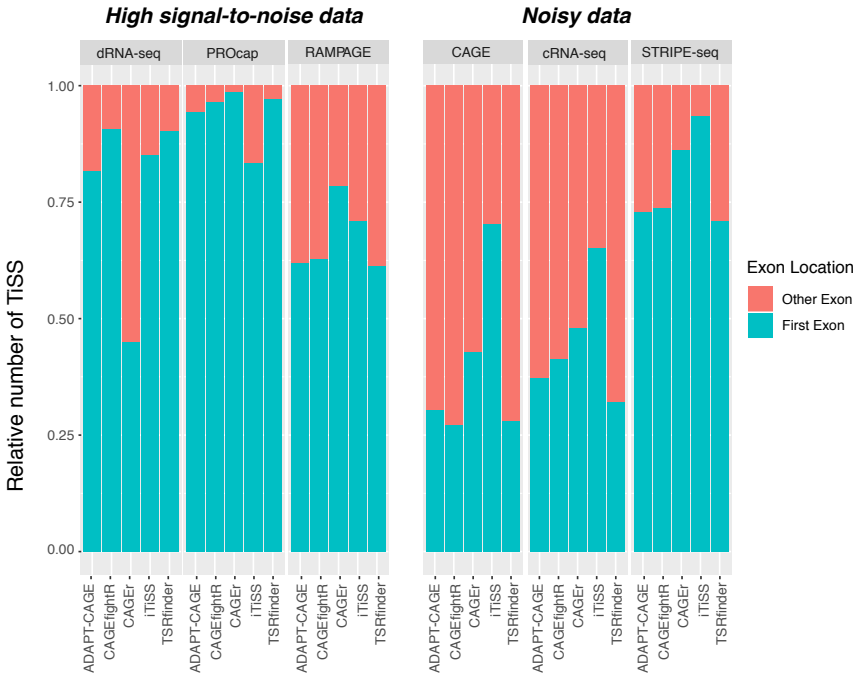


Figure 3.5: Depiction of the relative number of TSS located either in the first exon of spliced genes or in any of the later ones. As noisy datasets do not fully eliminate reads mapping throughout the genes body, spliced genes produce read accumulations at each of their exons. If a tool does not take the individual enrichment of singular positions compared to their surroundings into account, each of the read accumulations at the mentioned exon will be called as a false positive TSS. This in turn will lead to supposedly bona fide TSS validated by multiple data sets, which are in fact false. This effect can be seen here for TSRfinder, ADAPT-CAGE, CAGEr and CAGEfightR.

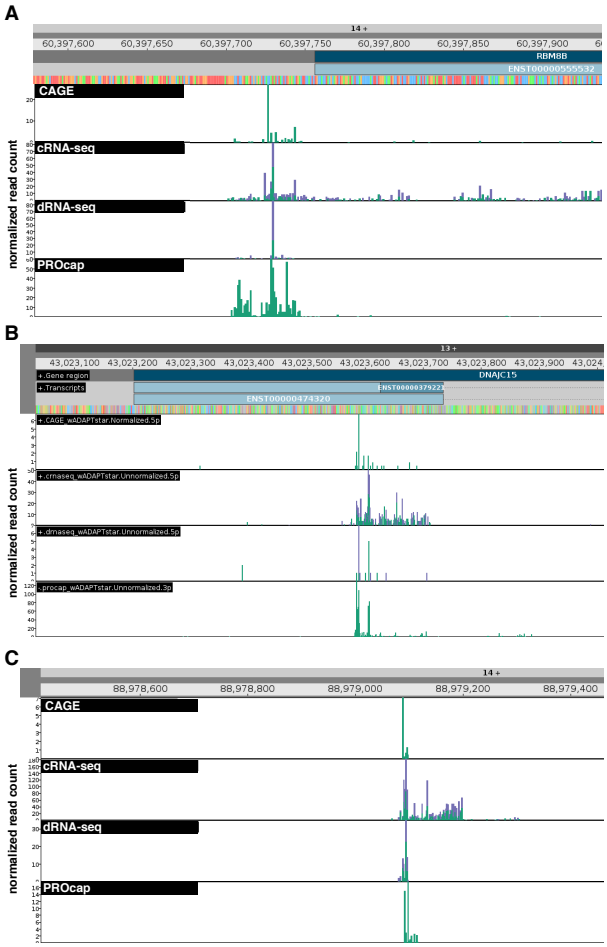


Figure 3.6: Three of the 34 TSS identified by iTiSS and confirmed by all four datasets, which are not annotated in the ENSEMBL (v90) database. **A** Chromosome 14, plus strand at around position 60397750 is a TSS, which suggests that the current annotation of the transcript following downstream is inaccurate. **B** Chromosome 13, plus strand at around position 43023600 is a TSS suggesting either a new transcript isoform or inaccurately annotated TSS of the transcripts starting upstream. **C** Chromosome 14, plus Strand at around position 88979100 is a TSS with no annotated gene around it. Of note, also on the negative strand is no annotated transcripts, suggesting a novel transcript.

3.5 Conclusion

We previously observed that all individual TSS profiling approaches produce a large number of false positives in the genome of herpes simplex virus 1 [Whisnant et al., 2020]. This can predominantly be attributed to artefacts that cannot be distinguished from bona-fide TSS in a single data set. Here, our evaluations demonstrate that this is also the case for much larger genomes such as the human genome. To remove false positive TSS, it is therefore essential to utilize integrative analysis of multiple approaches as facilitated by iTiSS.

Chapter 4

Integrative functional genomics decodes herpes simplex virus 1

Motivation: *The HSV-1 reference genome had 80 annotated open reading frames (ORFs) at the time of writing. An annotation of its transcriptome was missing completely. Although, the importance of the genomic origins of proteins is unquestionable, without knowledge about the transcripts they are translated from, the whole picture of HSV-1's transcription and translation machinery stays hidden. Especially, as recent studies showed that different herpesviruses code for multiple small ORFs with yet unknown functions. We hypothesized that those smaller ORFs are located on the same transcripts as the larger ORFs and serve as regulators of such. With a huge amount of various sequencing data available comprising TiSS-profiling, RNA-seq, mass spectrometry and Ribosome-profiling data sets, we wanted to answer these questions and in the process of it provide a complete annotation of HSV-1's transcriptome as well as its translatome. Here, I implemented the first version of iTiSS mentioned in chapter 3 and applied it to the TiSS-profiling data sets.*

Publication: *This manuscript was published in Nature Communications [Whisnant et al., 2020] as a joint-first authorship. Here, I adapted the layout and made minor corrections to the text. Further, I incorporated important parts of the Supplementary Material into the text.*

Individual author contributions: *See Appendix D*

4.1 Abstract

The predicted 80 open reading frames (ORFs) of herpes simplex virus 1 (HSV-1) have been intensively studied for decades. Here, we unravel the complete viral transcriptome and translatome during lytic infection with base-pair resolution by computational integration of multi-omics data. We identified a total of 201 transcripts and 284 ORFs including all known and 46 novel large ORFs. This includes a so far unknown ORF in the locus deleted in the FDA-approved oncolytic virus Imlygic. Multiple transcript isoforms expressed

from individual gene loci explain translation of the vast majority of ORFs as well as N-terminal extensions (NTEs) and truncations. We show that NTEs with non-canonical start codons govern the subcellular protein localization and packaging of key viral regulators and structural proteins. We extend the current nomenclature to include all viral gene products and provide a genome browser that visualizes all the obtained data from whole genome to single nucleotide resolution.

4.2 Introduction

Herpes simplex virus 1 (HSV-1) is the causative agent of the common cold sores but also responsible for severe, life-threatening diseases including generalized skin infections, pneumonia, hepatitis and encephalitis [Fields et al., 2013]. The HSV-1 genome is about 152kb in size and known to encode at least 80 open reading frames (ORFs), many of which have been extensively studied. Large-scale RNA-seq and ribosome profiling recently revealed that the coding capacity of three other herpesviruses, namely human cytomegalovirus (HCMV), Kaposi's sarcoma-associated herpesvirus (KSHV) and Epstein-Barr Virus (EBV) is significantly larger than previously thought [Stern-Ginossar et al., 2012, Arias et al., 2014, Bencun et al., 2018, Erhard et al., 2018]. For HCMV and KSHV, in particular, hundreds of new viral gene products were identified. These result from extensively regulated usage of alternative transcription and translation start sites throughout lytic infection. Moreover, these viruses were found to encode hundreds of short ORFs (sORFs) of unknown function. Similar to their cellular counterparts, these may either regulate translation of viral gene products or encode for functional viral polypeptides [Hinnebusch et al., 2016, Starck et al., 2016, Young and Wek, 2016, Cabrera-Quio et al., 2016, Chu et al., 2015]. To date, the majority of novel viral gene products have not been experimentally validated. Furthermore, the lack of a complete annotation and a revised nomenclature severely hampers functional studies.

Here, we employed a broad spectrum of unbiased functional genomics approaches and reanalyzed recently published data to comprehensively characterize HSV-1 gene products (Fig. 4.1). Our analysis of the viral transcriptome included: previously published time-course experiments of (i) total RNA-seq and 4sU-seq data [Rutkowski et al., 2015], (ii) new transcription start site (TiSS) profiling using two complementary approaches (cRNA-seq [Stern-Ginossar et al., 2012] and dRNA-seq [Sharma and Vogel, 2014]), (iii) incorporation of viral transcripts identified by two other groups using PacBio [Tombacz et al., 2017] and MinION [Depledge et al., 2019] platforms, and (iv) RNA localization by RNA-seq of subcellular fractions of both wild-type HSV-1 [Hennig et al., 2018] and the deletion mutant of the key viral RNA export factor ICP27. Analysis of the viral translatome included (i) standard ribosome profiling [Rutkowski et al., 2015] as well as so far unpublished translation start site (TaSS) profiling using (ii) Harringtonine and (iii) Lactimidomycin. Novel viral ORFs were validated using whole-cell quantitative proteomics and reverse genetics. To make the annotation and all the obtained data readily accessible to the research community, we provide an in-house genome browser software tailored the visualization of

HSV-1 and our collection of data (available at http://software.erhard-lab.de) as well as all data files to browse our annotation and data with any available genome browser at Zenodo https://zenodo.org/record/3465873). Thereby, viral gene expression and all data can be visually examined from whole genome to single-nucleotide resolution.

In total, we expanded the number of known of HSV-1 genomic elements to 201 viral transcripts encoding a total of 284 ORFs; including N-terminal peptide extensions and truncations of several classically described viral proteins as well as previously un-annotated protein-coding sequences in the loci of genes for major regulatory proteins ICP0 and ICP34.5.

4.3 Results

4.3.1 Characterization of the HSV-1 transcriptome

To identify the full complement of viral transcripts, we performed TiSS profiling employing a modified RNA sequencing protocol that is based on circularization of RNA fragments (here termed cRNA-seq) [Stern-Ginossar et al., 2012]. It enables quantification of RNA levels as well as identification of 5' transcript ends by generating a strong enrichment (≈18-fold) of reads that start at the 5' RNA ends. With cRNA-seq, we identified 266 potential TiSS that explained the expression of many previously annotated viral coding sequences (CDS). To comprehensively identify and validate putative novel TiSS, we applied a second 5'-end sequencing approach termed differential RNA-seq (dRNA-seq) [Sharma and Vogel, 2014], which provides a much stronger (≈300-fold) enrichment of TiSS at increased sensitivity (446 potential TiSS). It is based on selective cloning and sequencing of the 5'-ends of cap-protected RNA molecules that are resistant to the 5'-3'-exonuclease Xrn1. The two approaches provided highly consistent data at single nucleotide resolution (Fig. 4.3a). Furthermore, we reassessed viral transcripts called by two other groups exclusively based on third generation sequencing techniques (MinION [Depledge et al., 2019] and PacBio [Tombacz et al., 2017] platforms). This confirmed many of our TiSS (Fig. 4.2 A,B). The 80 viral transcripts (corresponding to a total of 89 TiSS, some of which were only separated by 5 nt), which were recently identified using MinION data, generally lacked 7-18 nucleotides (nt) at the 5' end due to technical limitations of the MinION direct RNA sequencing method (Fig. 4.2B) [Moldován et al., 2018]. After correcting this bias using our data, MinION-derived TiSS were highly consistent with our cRNA-seq and dRNA-seq data (Fig. 4.3b). Only 11 of the 89 TiSS (12%) identified by Depledge et al. could not be confirmed. We thus did not adopt them into our final genome annotation. Nevertheless, our genome browser encodes a separate track that visualizes all MinION and PacBio transcripts. Around half of all the TiSS that were previously identified using PacBio sequencing [Tombacz et al., 2017] matched to our data with single nucleotide resolution. The remaining TiSS (108 of 201; 54%) could neither be confirmed by cRNA-seq nor dRNA-seq (Fig. 4.3c). Most of them were only called from very few reads and presumably represent cleavage products of larger viral RNAs. In total, 102 TiSS were identified by at least two of the four approaches.

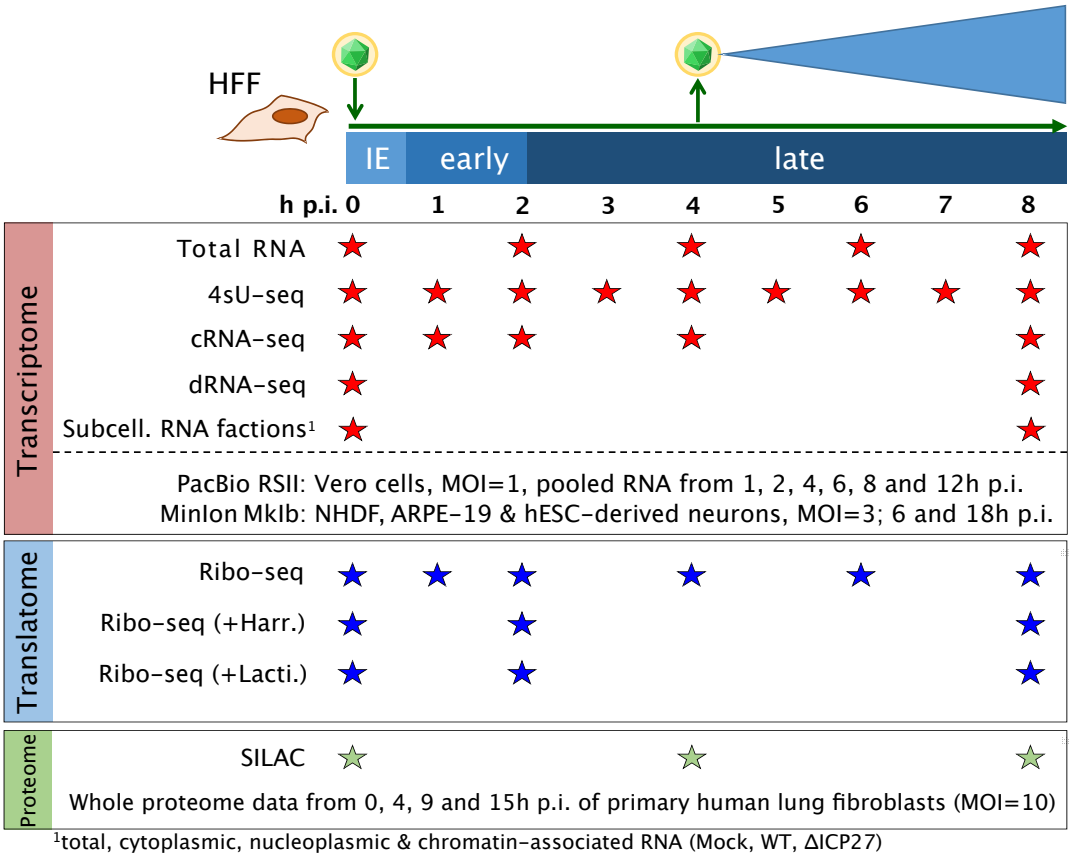


Figure 4.1: Viral gene expression was analyzed in primary human fibroblasts (HFF). The total RNA-seq, 4sU-seq and ribosome profiling data were recently published [Rutkowski et al., 2015]. To comprehensively identify transcription start site (TiSS), we performed cRNA-seq [Stern-Ginossar et al., 2012] and dRNA-seq [Sharma and Vogel, 2014] as well as RNA-seq on subcellular RNA fractions from mock, wild-type and ΔICP27 infected cells. For all of these, two biological replicates were performed. Furthermore, we incorporated the recently published transcripts originating from PacBio [Tombacz et al., 2017] and MinION [Depledge et al., 2019] sequencing data. Translation start site (TaSS) profiling was performed by ribosome profiling following treatment of cells for 30 min with either Harringtonine or Lactimidomycin [Stern-Ginossar et al., 2012]. Proteome analysis included two whole proteome data sets using SILAC and label-free mass spectrometry. The available time points and conditions are indicated by stars.

This demonstrates that complementary experiments are essential to exclude false positives and that none of the approaches by themselves is sufficient to reliably identify all viral TiSS.

To comprehensively assess the viral TiSS candidates that were only identified by a single approach, we developed a computational pipeline termed iTiSS (integrative Transcriptional Start Site caller). It screens potential TiSS for clustered accumulation of read 5'-ends in dRNA-seq (i) and cRNA-seq (ii) data. It evaluates our cRNA-seq data for an increase in upstream to downstream read coverage at potential TiSS (iii), and temporal changes in the potential TiSS read clusters in the cRNAs-seq time-course data (iv). It accounts for TiSS already identified by MinION (v) and PacBio (vi) sequencing. In addition, we also analyzed our 4sU-seq time-course data to both score potential TiSS that explained temporal changes in expression levels throughout infection (vii), and an increase in upstream to downstream read coverage (viii). Finally, we also scored TiSS that explained translation of viral ORFs, for which no other transcript had otherwise been identified (ix). For more details on each criterion see section 4.5.9. Thus, 9 criteria were utilized to confirm a potential TiSS. All identified TiSS were manually assessed and curated. In total, this resulted in 189 bona-fide viral TiSS, of which 161 (85%) were called by at least 2 criteria (Fig. 4.3d). Three of the five transcripts (LAT [Stevens et al., 1987], AL-RNA [Perng et al., 2002] and US5.1 [Jovasevic and Roizman, 2010]), which we could not confirm by any method, had previously been convincingly validated by other groups and were thus included. The other two were included after careful manual inspection (see section 4.5.10). The complete set of HSV-1 transcripts with their respective scores is provided in Tab. B.1.

TATA-boxes are a key element of eukaryotic promoters located 25 to 30 bp upstream of the TiSS [Smale and Kadonaga, 2003]. They are also prevalent for herpesvirus genes [Sandri-Goldin, 2007]. The presence of a TATA-box or TATA-box-like motif upstream of the viral TiSS strongly correlated with the expression levels of the respective transcripts. For weakly transcribed viral RNAs, the respective motifs were rarely observed ($p<10\text{-}6$, Fisher's exact test). In mammalian cells, the TiSS is marked by the initiator element (Inr), the core of which is a pyrimidine-purine (PyPu) dinucleotide [Carninci et al., 2006]. Interestingly, PyPu was also prevalent for the viral TiSS independent of expression levels (Fig. 4.3e). This provides strong evidence for the TiSS of even the most weakly expressed viral transcripts.

We next looked at splicing within the HSV-1 transcriptome based on our total RNA-seq and 4sU-seq data [Rutkowski et al., 2015]. We first identified all unique reads that spanned putative exon-exon junctions by at least 10 nt. This confirmed all 8 well-described splicing events and identified an additional tandem acceptor site ("NAGNAG") [Hiller et al., 2004] for both the third exon of the ICP0 gene (RL2) and the UL36.6 gene. Recently, Tombácz et al. proposed 11 novel splicing events based on PacBio sequencing data [Tombacz et al., 2017]. Our data confirmed all of these splicing events. However, only 4 of them occurred at relevant levels compared to the overall transcript levels (Tab. B.2). Two of these explained translation of novel small ORFs (UL40.5 iORF and UL40.7 dORF). Finally, we identified 44 novel putative splicing event sites based on our Illumina data (Fig. 4.4 and Tab. B.2).

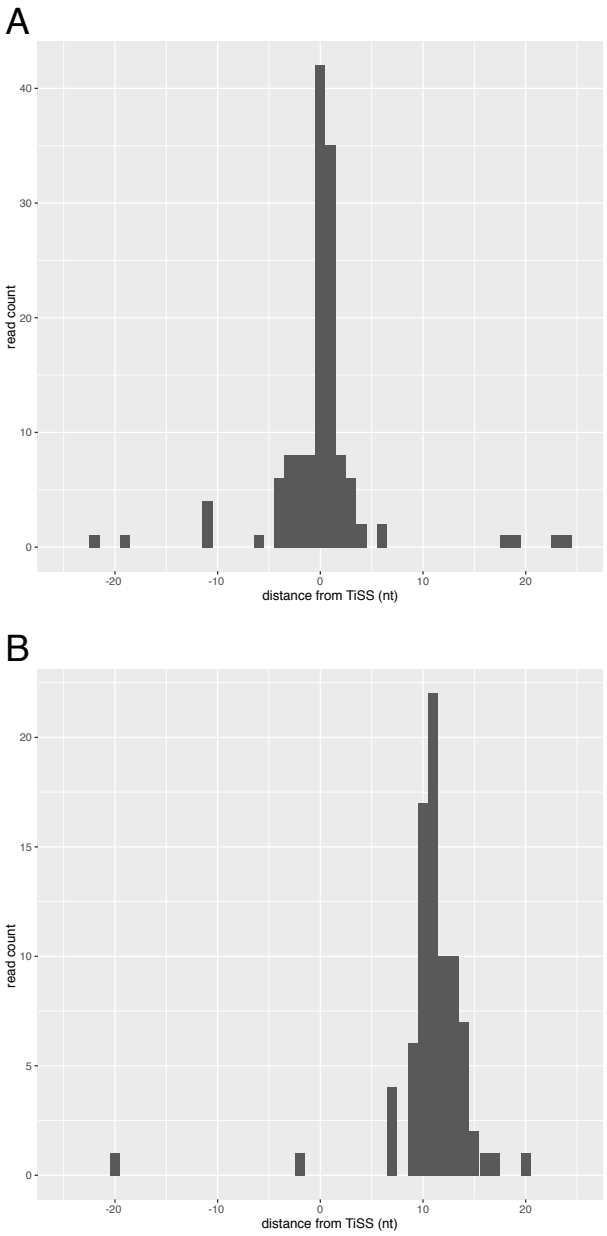


Figure 4.2: **(A)** Distance of transcription start sites (TiSS) identified by PacBio [Tombacz et al., 2017] to the TiSS positions obtained by both cRNA-seq and dRNA-seq is shown in relation to read counts. This confirmed TiSS identified by cRNA-seq and dRNA-seq at single nucleotide resolution. **(B)** Same as for **(A)** but for TiSS obtained from MinION sequencing data [Depledge et al., 2019]. The 89 TiSS called by MinION generally lacked 7-18 nucleotides (nt) at the 5' end for technical limitations of the MinION direct RNA sequencing method. After correcting for this, the manually curated MinION data confirmed many of the TiSS we identified.

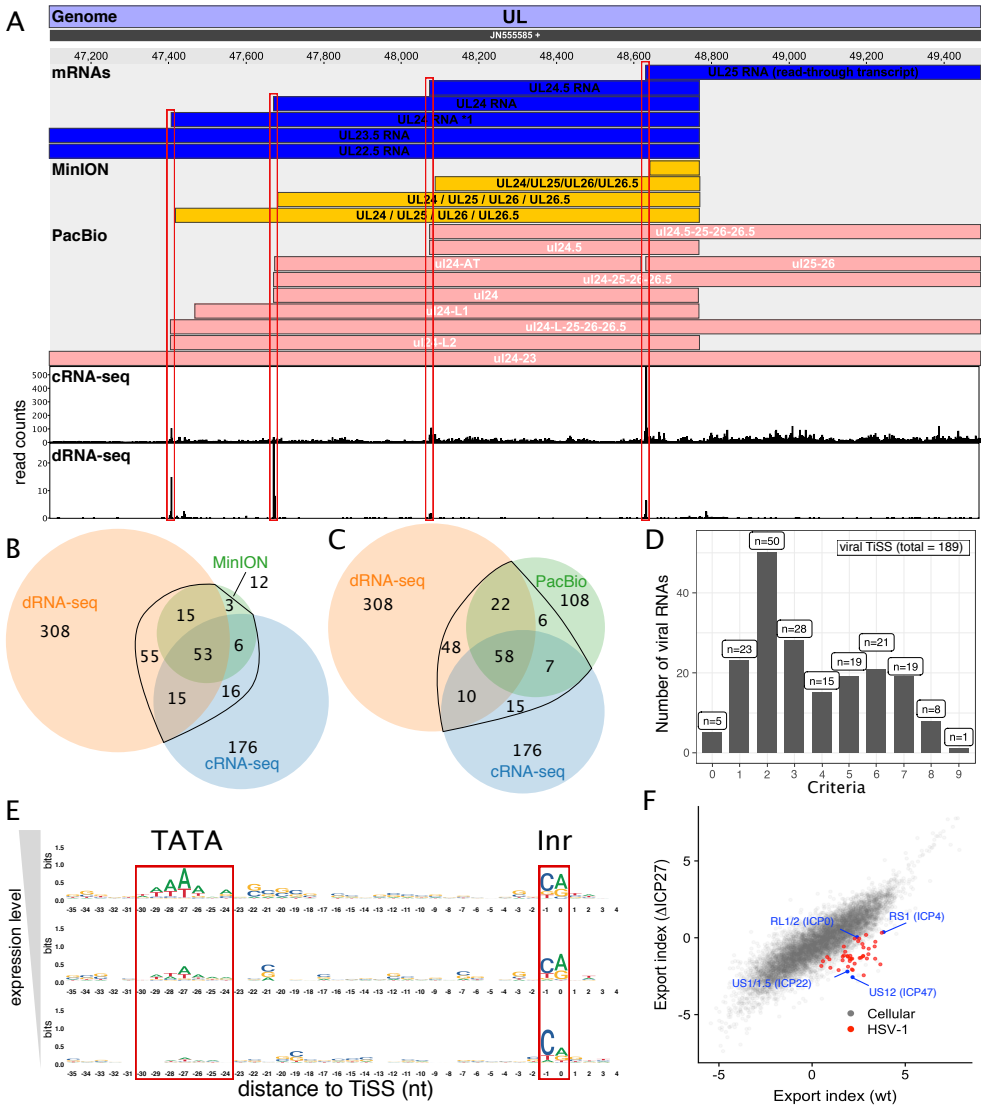


Figure 4.3: **(A)** Screenshot of our HSV-1 viewer displaying the annotated mRNAs of MinION, PacBio and coverage of read 5'-ends for cRNA-seq and dRNA-seq of the UL22.5-UL25 gene locus. Transcripts in our reference annotation are indicated in blue. **(B,C)** Venn diagram depicting the number of Transcription start sites (TiSS) that were identified by cRNA-seq, dRNA-seq and MinION [Depledge et al., 2019] **(B)** or PacBio [Tombacz et al., 2017] **(C)** sequencing, respectively. TiSS included into the final annotation are indicated by the black circle. **(D)** Histogram depicting the number of TiSS criteria that were fulfilled by the individual viral transcripts. **(E)** Sequence logos upstream of viral TiSS with viral TiSS grouped into 3 equally sized bins according to their transcription rates (top: highest; bottom: lowest). The TATA-box and initiator element (Inr) are shown. **(F)** Log fold-change between cytoplasmic and chromatine associated FPKM-normalized read counts (export index) of cellular (grey) and viral (red and blue) gene clusters compared between wild-type HSV-1 (wt) and a null mutant of the viral RNA export factor ICP27 (ΔICP27). Viral immediate early genes are indicated in blue.

However, all of these showed substantially lower read coverage than the surrounding exons, indicating that they only represented rare events at best. Therefore, we decided not to include these low abundance splicing events into our final reference annotation.

A recent paper by Tang et al. [Tang et al., 2019] proposed 71 novel HSV-1 splicing events. We also observed 15 of these in our Illumina data. Interestingly, about half (28 of 71) of the splicing events reported by Tang et al were exclusively observed upon infection with an ICP27-null mutant. Of note, none of our 44 putative splicing events were found to be more abundant upon infection with an ICP27-null mutant (subcellular RNA factions from δICP27 infected cells). We conclude that they do not reflect aberrant splicing events that originate from infected cells which express insufficient levels of ICP27. In total, we thus identified 189 viral TiSS that give rise to at least 201 transcripts and transcript isoforms.

4.3.2 RNA 3'-end processing and export of viral transcripts

Previous studies reported regulated usage of the 47 viral poly(A) sites during productive infection, which appeared to be mediated or at least influenced by the viral ICP27 protein [McLauchlan et al., 1989, McGregor et al., 1996, Hann et al., 1998, Rajcani et al., 2004, McLauchlan et al., 1992, Tang et al., 2016]. We recently reported that lytic HSV-1 infection results in a widespread but nevertheless selective disruption of transcription termination of host genes [Rutkowski et al., 2015]. In contrast to the extensive read-through transcription at host poly(A) sites that we observed by 4-8 h p.i., viral gene expression remained mostly unaffected. Recently published third-generation sequencing data proposed numerous very large viral transcripts spanning multiple viral genes [Depledge et al., 2019]. To address the nature of these transcripts and their role in translation, we performed RNA-seq on subcellular RNA fractions (total RNA, cytoplasmic RNA, nucleoplasmic RNA and chromatin-associated RNA) using both wild-type HSV-1 and a null mutant of the viral RNA export factor ICP27 (δICP27). The data from wild-type HSV-1 and mock infected cells were recently published [Hennig et al., 2018]. The δICP27 infection had been performed in the same experiment. Consistent with the well-characterized role of ICP27 in viral mRNA export [Sandri-Goldin, 2011], all viral transcripts were more efficiently ($\approx$11-fold) exported to the cytoplasm in wild-type than in δICP27 HSV-1 infection (Fig. 4.3f). Interestingly, this even included the spliced immediate early (IE) genes ICP0 ($\approx$5-fold), ICP22 ($\approx$17-fold) and ICP47 ($\approx$27-fold) as well as the unspliced (IE) ICP4 gene ($\approx$11-fold). In chromatin-associated, nuclear and total cellular RNA, considerable numbers of reads were observed within the first 500 nt downstream of viral poly(A) sites (PAS). However, in the cytoplasmic RNA fraction of infected cells, read levels dropped substantially immediately downstream of the PAS (Fig. 4.6a). This indicates that reads mapping downstream of the PAS reflect mRNA precursors, which remain nuclear and, thus, do not contribute to the translated viral transcriptome. However, for some viral genes, e.g. UL30, UL38 and UL43, considerable numbers of reads that mapped downstream of the respective PAS were present in the cytoplasmic RNA fraction. For the UL30 PAS, this became substantially more prominent late in infection (8 h p.i., Fig. 4.6b). Furthermore,

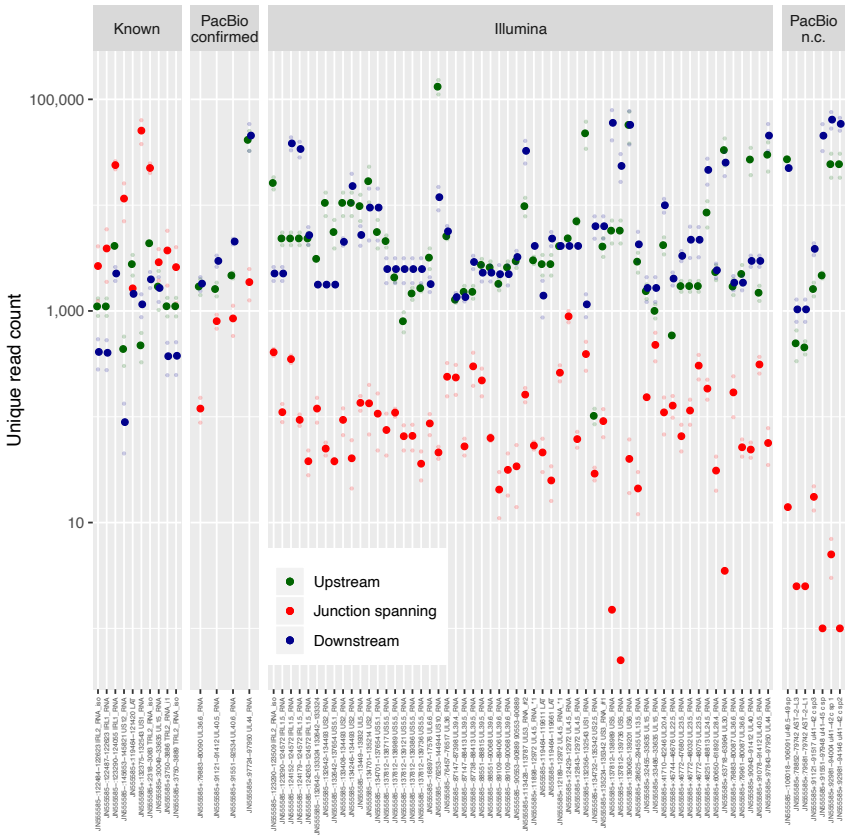


Figure 4.4: The mapped reads of the 4sU-seq and total RNA-seq experiment were used to examine splicing in the HSV-1 genome. Shown are the locations of known and putative splicing events in the HSV-1 transcriptome as well as one representative transcript for each of them. The putative splicing events are separated into three groups. The ones that were called by PacBio, which we could confirm to be present at reasonable expression levels (PacBio confirmed), the ones that were called by PacBio but did not occur at reasonable levels in our data (PacBio non-confirmed; PacBio n.c.) and the ones that were only identified in our data, but did not occur at reasonable levels (Illumina). For each splice junction the number of unique reads spanning it (red) is depicted as well as the non-spanning reads upstream (green) and downstream (blue). Besides the 8 known splicing events and a NAGNAG event in the ICP0 mRNA, Illumina sequencing confirmed 4 splicing events identified by PacBio [Tombacz et al., 2017]. Furthermore, screening our RNA-seq and 4sU-seq data for splicing events with at least 10 nt exon-spanning uniquely mapping reads identified 58 putative additional splicing events. Reads with mismatches around the splice-site were removed to assure the NAGANG event is not due to bad mapping. However, exon-spanning reads were >10-fold less prevalent than reads mapping to the flanking regions. We therefore, decided not to include them into our reference annotation. Nevertheless, they may explain some of the orphan ORFs. Our data confirmed all 11 splicing events. However, only four of them occurred at relevant levels and were included into our reference annotation. The small and bright dots represent the read counts of two biological samples (n=2). Their mean is indicated as larger dot.

transcription of UL25, which initiates 107 nt upstream of the UL24 PAS, efficiently bypassed the UL24 PAS already from 2 h p.i. on (Fig. 4.5). The same was observed for UL24.5 which represents an N-terminal truncated isoform of UL24. These data confirm previous findings on differential polyadenylation of selected viral genes during productive infection [McLauchlan et al., 1989, McGregor et al., 1996, Hann et al., 1998, Rajcani et al., 2004, McLauchlan et al., 1992, Tang et al., 2016].

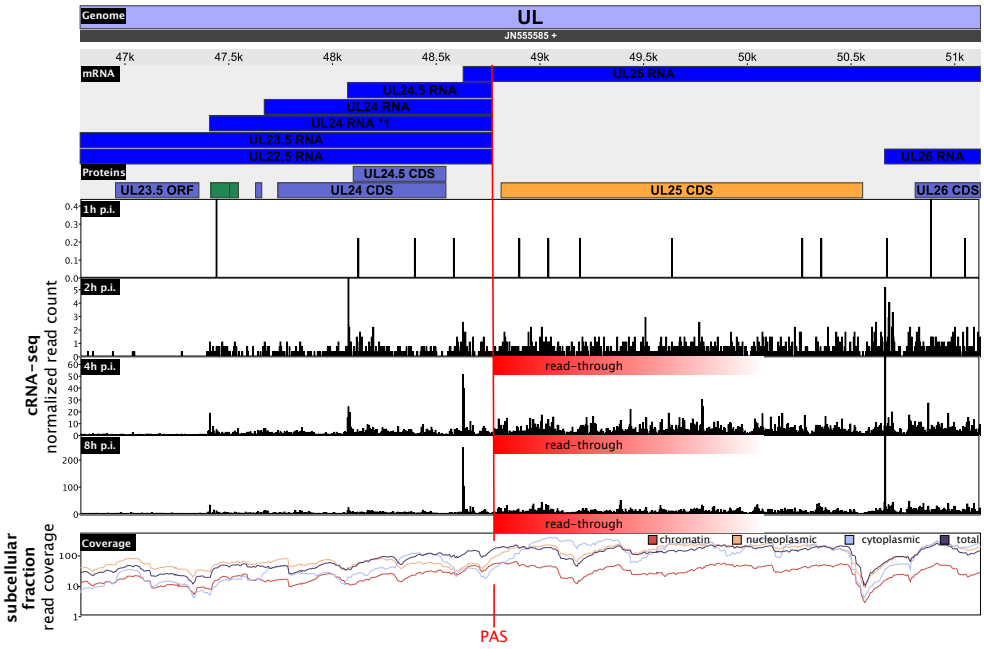


Figure 4.5: Screenshot of our HSV-1 viewer depicting the annotated transcripts (mRNA) and proteins (Proteins) as well as the number of 5' ends of reads for each position of the cRNA-seq dataset. Additionally, the read coverage (Coverage) for chromatin associated, nucleoplasmic, cytoplasmic and total RNA is shown. Cytoplasmic read numbers were notably higher upstream than downstream of the PAS when compared to chromatin-associated dRNA. The read-through (marked in red) of the UL24 polyadenylation site (PAS) that results from transcription of UL25 can be seen as early as 2 h p.i. This confirms previous findings on differential polyadenylation of selected viral genes during productive infection.

4.3.3 HSV-1 expresses hundreds of so far unknown ORFs and sORFs

To comprehensively identify the viral translatome, we performed time-course analysis of ribosome profiling as well as translation start site (TaSS) profiling (see Fig. 4.1 and Methods section). The obtained data confirmed the expression of all 80 previously annotated ORFs (CDS) and detected 46 additional large ORFs and 134 small ORFs (3-99 aa). We also identified 7 N-terminal truncations (NTTs) and 17 N-terminal extensions (NTEs) of CDS.

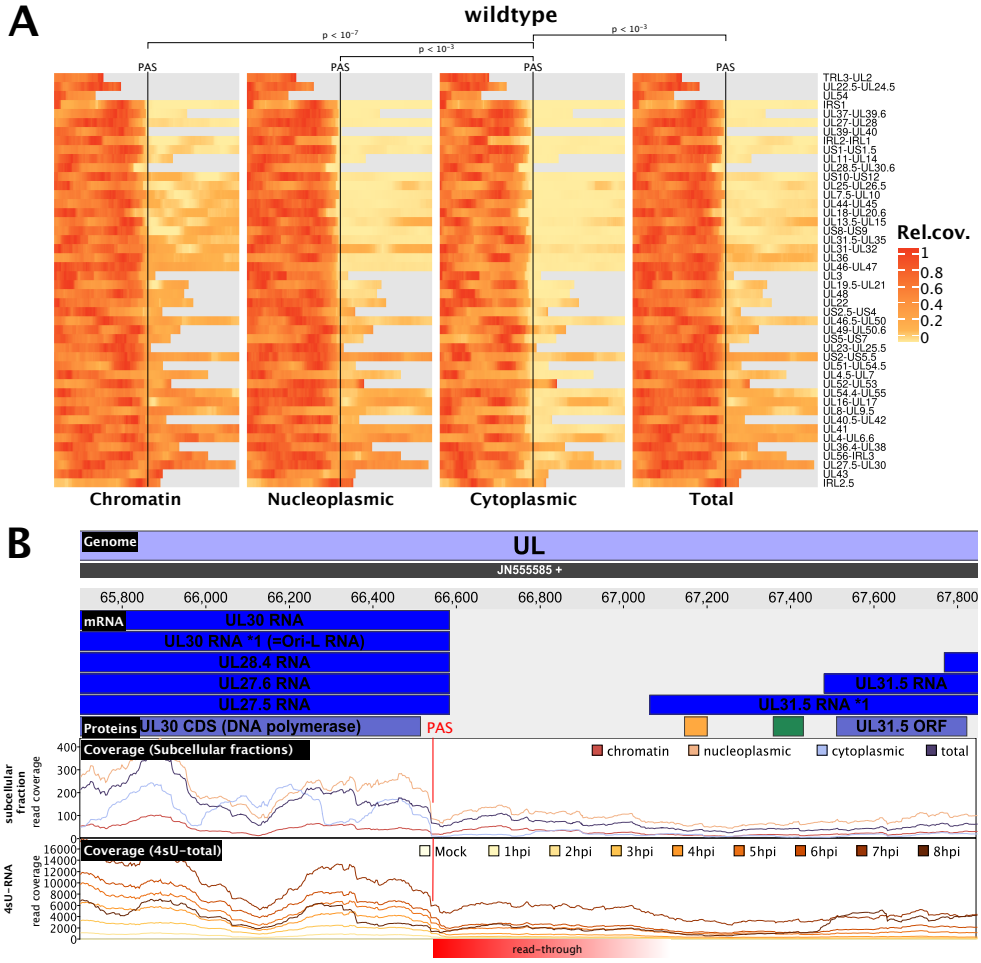


Figure 4.6: **(A)** Read levels 500 bp upstream (left) and downstream (right) of the Poly(A)-site (PAS) of viral genes for wild-type HSV-1. Grey bars indicate overlapping parts with other genes, for which reads could not be uniquely assigned. In the cytoplasmic RNA fraction, read levels dropped substantially immediately downstream of the PAS. p-values were calculated using a one-sided paired t-test over the mean fold-change of read levels 500 bp before against after the PAS (Cytoplasmic to Nucleoplasmic: p-value$=8.157 \cdot 10^{-4}$, Cytoplasmic to chromatin-associated: p-value$=6.956 \cdot 10^{-8}$, Cytoplasmic to total p-value$=4.06 \cdot 10^{-4}$). **(B)** Screenshot of our HSV-1 viewer depicting poly(A) read-through at the UL30 PAS in cRNA-seq data at 2, 4 and 8 hours post infection (hpi) of replicate 1. The annotated transcripts (mRNA), proteins (Proteins) and read coverage (Coverage) for chromatin-associated, nucleoplasmic, cytoplasmic and total reads are shown for the positive strand only. Read-through transcription is schematically indicated in red. Downstream of the UL30 PAS, chromatin-associated and nucleoplasmic reads show substantial read-through, whereas cytoplasmic reads drop down to only a fraction of what they were before (blue arrow)

In total, our data provides evidence for the translation of 284 viral ORFs (Supplementary Data B.3 and B.4). Translation predominantly initiated from AUG start codons (79%). However, non-canonical initiation events also substantially contributed to the HSV-1 translatome with CUG, GUG, ACG and AUC together initiating translation of about 15% and 20% of all large and small viral ORFs, respectively (Fig. 4.7a,b).

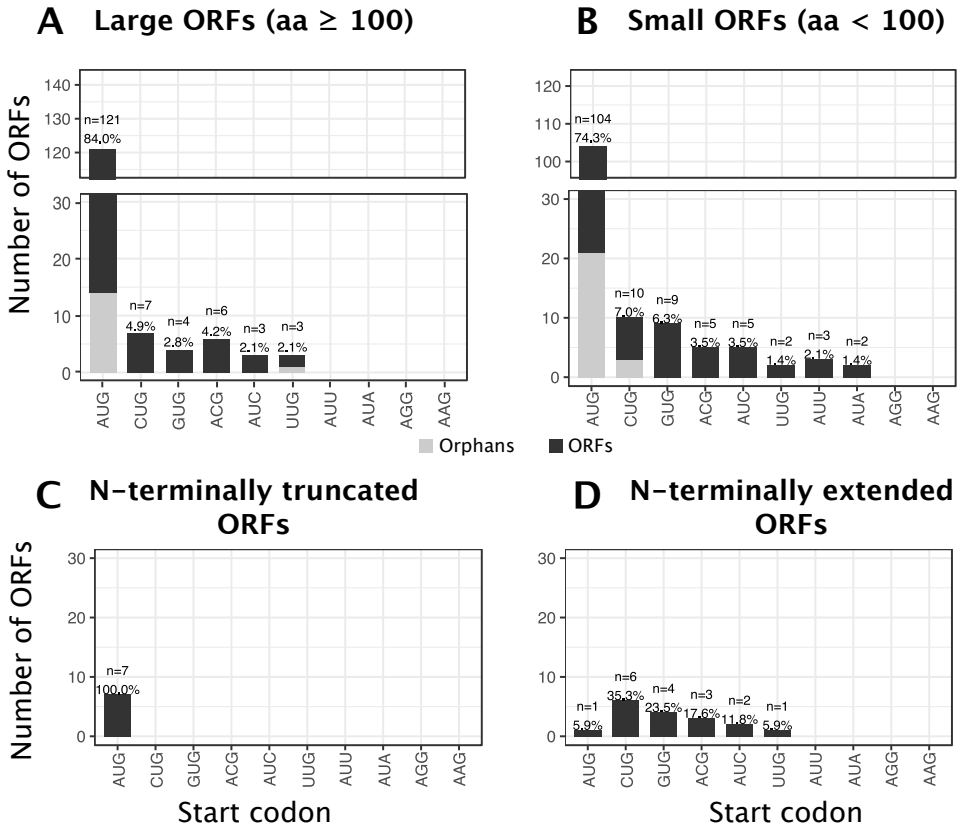


Figure 4.7: Distribution and frequency of possible start codons used by HSV-1 open reading frames (ORFs) **(A)**, short ORFs (sORFs) **(B)**, N-terminal truncated ORFs (NTTs) **(C)** and N-terminal extended ORFs (NTEs) **(D)**. Orphan ORFs are depicted in light grey. Six of the previously identified CDS (UL11, UL49.5, US5, US9, US12 and RL2 iso1) are <100 aa and were thus included in **(B)**.

We observed seven NTTs originating from downstream translation initiation events of previously described viral coding sequences (Tab. 4.1). All of these initiated with AUG start codons (Fig. 4.7C). Alternative TiSS downstream of the main TiSS explained translation of 6 of 7 NTTs (Fig. 4.8). Only for US3.5, we could not identify a corresponding transcript. It thus remains unclear whether US3.5 is translated from an independent transcript or due to leaky scanning. Six of these NTTs (UL8.5, UL12.5, UL24.5, UL26.5, US1.5 and US3.5) had already been reported [Baradaran et al., 1994, Draper et al., 1986, Dridi et al., 2018, Liu and Roizman, 1991, Ogle and Roizman, 1999, Poon et al., 2006]. Only the NTT of the major DNA-binding protein pUL29 (ICP8; comprising aa 516 to 1212) had so far

not been described. Interestingly, this NTT initiates at an AUG start codon immediately downstream of the metal-(Zn)-binding loop (residues 499-512) [Gupte et al., 1991, Mapelli et al., 2005]. While the ribosome profiling data showed a strong peak at the respective AUG start codon, which was further enriched by LTM treatment (Fig 4.8), we were unable to validate the truncated UL29.5 protein using a C-terminally 3X-Flag-tagged mutant virus. Further experiments are thus required to clarify the existence and stability of UL29.5 as well as its corresponding transcript.

Interestingly, 16 of the 80 viral reference ORFs (20%) showed in-frame NTEs (Tab. 4.2) with translational activity exceeding 10% of the main downstream ORF. The majority of NTEs (16 of 17, including 2 NTEs in UL50) initiated translation from non-AUG start codons (Fig. 4.7d). This included key viral proteins like the major immediate early protein ICP27 (UL54), the major capsid protein (VP5, UL19) and the well-studied viral kinase US3. For 5 viral genes, we generated mutant viruses by introducing a 3X-FLAG-tag either into the NTE or downstream of the canonical AUG start codon. This confirmed the expression of 6 NTEs including the 2 UL50 NTEs (Fig. 4.9a-e). Interestingly, the introduction of a 3X-FLAG-tag into the N-terminal extension of both ICP27 and VP5 resulted in dead viruses, which could only be reconstituted upon transfection of complementing cells. For UL54, expression of the NTE was already observable when the 3X-FLAG-tag was introduced downstream of the canonical AUG-start codon (Fig. 4.9e). For UL19 (VP5) major capsid protein (MCP), the 3X-FLAG-tagged NTE appeared to even be dominant negative. Virus reconstitution in non-complementing cells resulted in a partial deletion of the NTE within two passages. This indicates that the 3X-FLAG-tagged NTE-MCP is assembled into virus particles but renders them dysfunctional due to the N-terminally inserted 3X-FLAG-tag.

To test the impact of the respective NTEs on protein localization, we performed immunofluorescence microscopy of both the NTE- and AUG-tagged viruses. While subcellular localization of the NTEs of UL54 and US5 were indistinguishable from their canonical counterparts (Fig. 4.10), the NTEs of US3 and UL50 notably altered subcellular localization (Fig. 4.9F,G). While canonical US3 was predominantly nuclear, the NTE-US3 localized to the cytoplasm. The US3 NTE contains a leucine-rich stretch indicating a putative nuclear export signal. Pseudorabies virus (PRV), a porcine alphaherpesvirus, expresses two isoforms of US3, both of which initiate from AUG start codons on separate transcripts (Fig. 4.11). The longer isoform encodes a mitochondrial localization signal resulting in the cytoplasmic localization and a failure of the respective protein to be incorporated into the tegument [Calton et al., 2004]. The DNA sequence of the US3 NTE is conserved in HSV-2 and its role as a nuclear export signal fits data demonstrating that HSV-2 US3 fails to accumulate in the cytoplasm when nuclear export is inhibited [Finnen et al., 2010]. Similar to US3, localization of NTE-UL50 also shifted to the cytoplasm (Fig. 4.9G). UL50 dUTPase activity in PRV-infected cells was reported to be nuclear [Ns and Mettenleiter, 1996], while it appears to be predominantly cytoplasmic with HSV-2 [Wohlrab et al., 1982] and nearly equally distributed in HSV-1. We conclude that NTEs initiating from non-AUG start codons are common in alphaherpesvirus proteomes. They allow the expression

Table 4.1: List of truncated ORFs including information about their location, name, and start and stop codons used.

Name	Transcript	Strand	Length (aa)	Start-codon	Stop-codon	TaSS	Stop
UL12.5 CDS (Exonuclease activity; truncated isoform of UL12)	UL12.5_RNA	-	500	AUG	UGA	26509	25007
UL24.5 CDS (truncated isoform of UL24)	UL24.5_RNA	+	148	AUG	UGA	48100	48546
UL26.5 CDS (truncated isoform of UL26)	UL26.5_RNA	+	329	AUG	UGA	51727	52716
UL29.5 CDS (truncated isoform of UL29)	UL29.5_RNA	-	681	AUG	UGA	60508	58463
UL8.5 CDS (ATP binding; truncated isoform of UL9; Bahradaran et al, 1994)	UL8.5_RNA	-	487	AUG	UAA	22167	20704
US1.5 CDS (truncated isoform of US1)	US1.5_RNA	+	250	AUG	UGA	133156	133908
US3.5 CDS (truncated isoform of US3)	US3_RNA	+	405	AUG	UGA	135452	136669

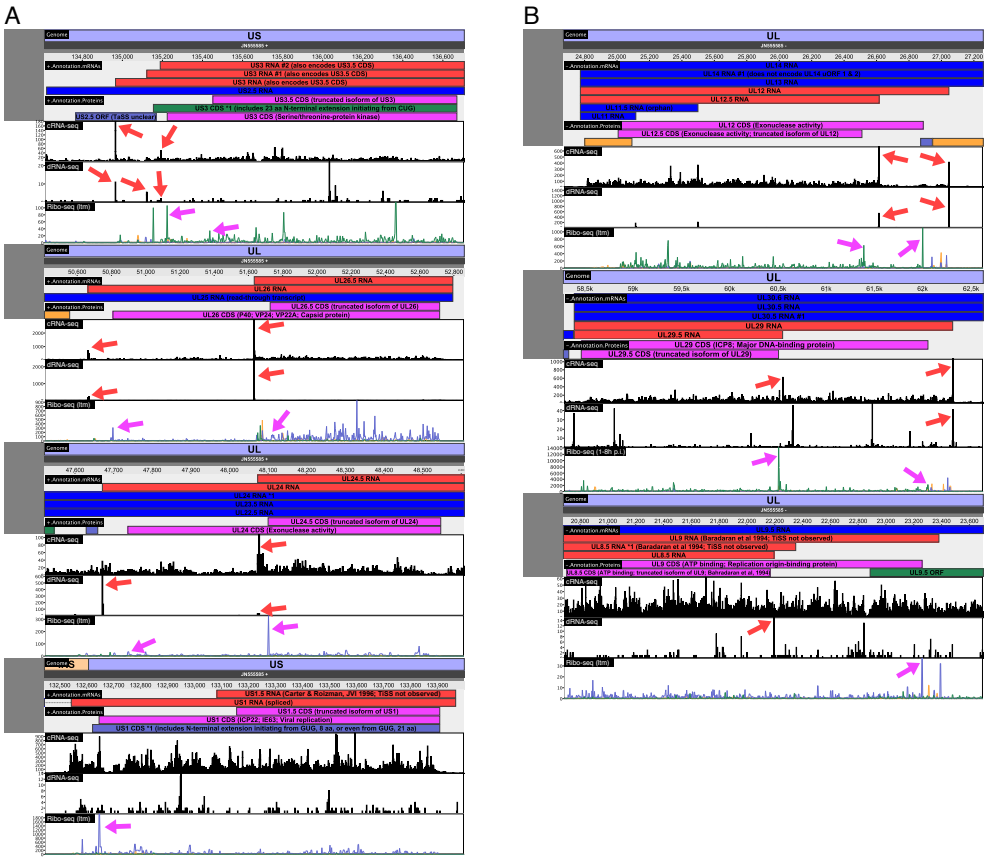


Figure 4.8: Screenshots of our HSV-1 viewer depicting the seven viral open reading frames (ORFs) with N-terminal extended ORFs (NTEs). Alternative transcription start sites (TiSS) downstream of the main TiSS explained translation of 6 of 7 N-terminal truncated ORFs (NTTs). Only for US3.5, we could not identify a corresponding transcript. The NTTs as well the original ORFs are highlighted in pink. The corresponding transcripts for both are highlighted in red. Red arrows in the cRNA-seq and dRNA-seq track point at the specific TiSS validation, if present. Pink arrows point at the translation start site (TaSS) validation if present. The three colored graphs depict the read counts for the three possible ORFs (yellow=1, blue=2, green=3) **(A)** NTTs encoded in the sense strand. **(B)** NTTs encoded in the antisense strand. For UL29/UL29.5 the combined ribosome profiling data for 1 h p.i. up to 8 h p.i. is shown instead of the Lactimidomycin data. This was a pure esthetical decision as the TaSS peak there was too prominent and therefore translation throughout the ORF could not be seen anymore.

Table 4.2: List of ORFs with N-terminal extensions (NTEs) initiating from non-canonical start codons.

Name	Transcript	Strand	Length (aa)	Start-codon	Stop-codon	TiSS	Stop
UL19 CDS *1 (includes 11 aa N-terminal extension initiating from AUC)	UL19_RNA	-	1385	AUC	UAA	40561	36404
UL21 CDS *1 (includes 65 aa N-terminal extension initiating from CUG)	UL21_RNA	+	600	CUG	UAA	41879	43681
UL23 CDS *1 (includes 15 aa N-terminal extension initiating from GUG)	UL23_RNA	-	391	GUG	UGA	47847	46672
UL27 CDS *1 (includes 43 aa N-terminal extension initiating from ACG)	UL27_RNA	-	947	ACG	UGA	55923	53080
UL33 CDS *1 (includes 21 aa N-terminal extension initiating from ACG)	UL33_RNA	+	151	CUG	UGA	69097	69552
UL36 CDS *1 (includes 10 aa N-terminal extension initiating from AUG)	UL36_RNA	-	3139	AUG	UAG	80467	71048
UL39 CDS *1 (includes 38 aa N-terminal extension initiating from UUG)	UL39_RNA	+	1175	UUG	UGA	86328	89855
UL5 CDS *1 (includes 9 aa N-terminal extension initiating from GUG)	UL5_RNA	-	891	GUG	UAA	15158	12483
UL50 CDS *1 (includes 10 aa N-terminal extension intitating from CUG)	UL50_RNA	+	381	CUG	UAG	106980	108125
UL50 CDS *2 (includes 23 aa N-terminal extension intitating from ACG)	UL50_RNA	+	394	ACG	UAG	106941	108125
UL54 CDS *1 (includes 38 aa N-terminal extension initiating from ACG)	UL54_RNA	+	550	ACG	UAG	113620	115272
UL7 CDS *1 (includes 15 aa N-terminal extension initiating from CUG)	UL7_RNA	+	311	CUG	UGA	17090	18025
US1 CDS *1 (includes N-terminal extension initiating from GUG, 8 aa, or even from GUG, 21 aa)	US1_RNA	+	428	GUG	UGA	132622	133908
US11 CDS *1 (includes 4aa N-terminal extension initiating with GUG)	US11_RNA	-	165	GUG	UAG	145262	144765
US3 CDS *1 (includes 23 aa N-terminal extension initiating from CUG)	US3_RNA	+	504	CUG	UGA	135155	136669
US5 CDS *1 (includes 16 aa N-terminal extension initiating from CUG)	US5_RNA	+	108	CUG	UAA	137685	138011
US8 CDS *1 (includes 7aa N-terminal extension initiating from AUC)	US8_RNA	+	557	AUC	UAA	141225	142898

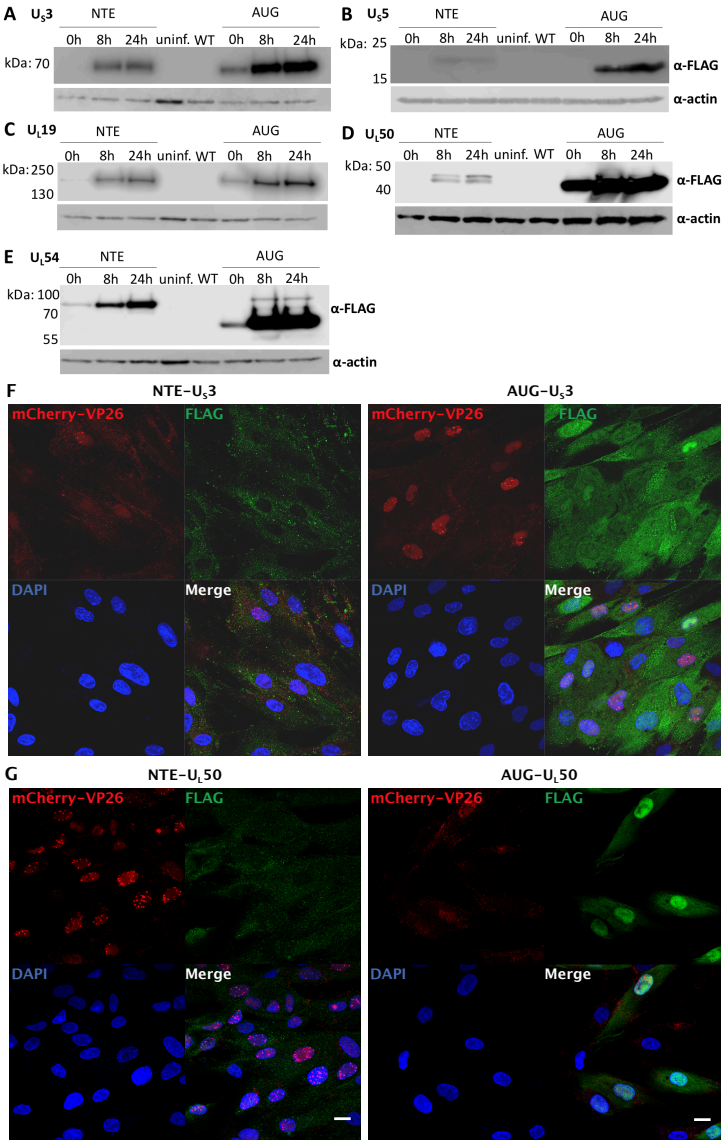


Figure 4.9: Tagged viruses were generated by inserting a 3X-FLAG-tag either upstream of the canonical start codon into the N-terminal extension (NTE) or downstream of it (AUG). Western blots of 3X-FLAG-tagged N-terminal extensions following infection of human foreskin fibroblasts with the indicated viruses are shown. Expression at the given hours (h) post-infection are compared to uninfected (uninf.) and the parental (WT) virus, both at 24 h p.i. for the HSV-1 genes **(A)** US3, **(B)** US5, **(C)** UL19, **(D)** UL50, and **(E)** UL54. Expression of the NTE of UL54 (ICP27) was already visible when the 3X-FLAG-tag was inserted downstream of the canonical AUG. **(F)** Immunofluorescence of human foreskin fibroblasts infected with VP26-mCherry HSV-1 containing 3X-FLAG-tags inserted upstream of the canonical start codon into the N-terminal extension (NTE) or downstream of it (AUG) for US3 and **(G)** UL50. Cell nuclei were stained using DAPI. Scale bars depict 20 microns. Protein localization of both NTEs shifts to the cytoplasm.

of alternative protein isoforms with different subcellular localization and regulatory motifs.

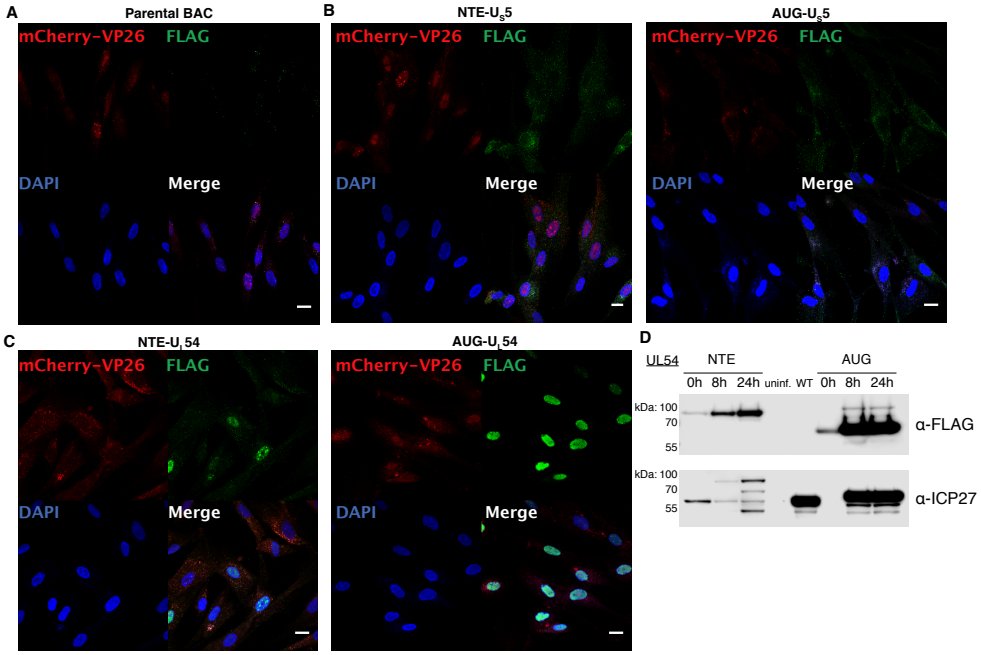


Figure 4.10: Immunofluorescence (IF) of human foreskin fibroblasts infected with parental BAC-derived VP26-mCherry HSV-1 **(A)** or viruses containing 3X-FLAG-tags inserted upstream of the canonical start codon into the N-terminal extension (NTE) or downstream of it (AUG) for US5 **(B)** and UL54 **(C)**. Cell nuclei where stained using DAPI. Scale bars depict 20 microns. No differences in the subcellular localization between the NTE and the canonical protein were observed. **(D)** Western blot for UL54 showing the signal for FLAG as shown in Figure 5 and total UL54/ICP27 levels in the same samples.

In 2015, the first oncolytic virus (talimogene laherparepvec (Imlygic)) was approved for therapy of melanoma [Pol et al., 2016]. This modified herpes simplex virus 1 lacks two viral genes (ICP34.5 and ICP47) and expresses GM-CSF to recruit and stimulate antigen-presenting cells. Within the ICP34.5 (RL1) locus, we found a novel 93 aa ORF, which we termed RL1A (Fig. 4.12a). It initiates from an AUG start codon 46 nt upstream of the AUG start codon of RL1 and is translated from the same transcript at >4-fold higher levels. Imlygic thus lacks a third viral protein, namely RL1A.

The RL2 locus encodes the major viral immediate early protein ICP0. Here, we identified an additional spliced ORF (termed RL2A) of 181 aa that initiates from an ACG start codon 116 nt upstream of the ICP0 TaSS (Fig. 4.12b). Expression of RL2A was confirmed by generating a mutant virus (3X-Flag-RL2A) with a 3X-FLAG-tag inserted 12 nt downstream of the ACG start codon (Fig. 4.12c). Interestingly, RL2A expression by the mutant virus could only briefly be detected immediately upon virus reconstitution and was readily lost upon serial passaging. This indicates that insertion of the 3X-FLAG-tag into the RL2A repeat region severely impaired viral fitness resulting in DNA

Virus	Accession number	Amino acid sequence
HSV-1	JN555585	L/MPLLKTPGPVVRGARWLALTVRRM
HSV-2	NC_001798	L/MPLLKTPGPVARGARWLARATRQM
BHV-1	NC_001847	IAGVDRVRLGVRLPFLPQARSRDTTRRSWAPM
FeHV-1	NC_013590	RFLLFRKCIRLANMDRFPRVGLSCCIPTSKGDIDTGDNYKLQSTM
MaHV-1	KT594769	TCLSFQITGSLCM
PRV	NC_006151	MLAMWRWVTKRSRLRRGHAHLGGNKGVRGICSLYLAGLSRGLSRVHAQRSHAATM

Figure 4.11: Primary peptide sequences for validated (HSV-1, PRV) and predicted (HSV-2, BHV-1, FeHV-1 and MaHV-1) US3 NTEs are depicted from the start codons (canonical or non-canonical) to the annotated US3 start codon ("M" in bold). Hydrophobic resides are indicated in red. Putative nuclear export signals matching the motif [LIVFM]-X2,3-[LIVFM]-X2,3-[LIVFM]-X-[LIVFM] are highlighted in yellow. The mitochondrial localization signal predicted for PRV [Calton et al., 2004] is underlined.

recombination with the other wild-type repeat (data not shown). To address this issue, we generated a second mutant virus (3X-FLAG-RL2A-ΔRL) by subsequently deleting the wild-type RL2A and part of RL2 from the second repeat to prevent recombination and removal of the inserted 3X-Flag-tag upon virus reconstitution and passaging. This resulted in stable expression of 3X-Flag-tagged RL2A of the expected size (21.8 kD; Fig. 4.12c). Interestingly, however, ICP0 expression of this mutant was almost completely abolished. We subsequently noted that the 3X-FLAG-tag contains an out-of-frame AUG start codon (GATTACAAGG**ATG**ACGACGATAA) in every of the three FLAG-tag repeats. Translation initiation at the respective start codons and ribosomes bypassing the ICP0 TaSS explains the observed near-complete loss of ICP0 expression and thus the rapid recombination of our primary 3X-FLAG-RL2A mutant upon serial passaging. Furthermore, this may also explain some of the attenuation, which we observed for the mutant viruses with 3X-FLAG-tagged NTEs, namely for ICP27 and VP5. Accordingly, protein levels of the canonical ICP27 protein were dramatically reduced for the 3X-FLAG-tagged NTE-ICP27 virus (data not shown). These observations highlight the need to carefully consider ectopic translation start site usages when manipulating herpesvirus genomes.

Transcription of all viral genes continuously increases throughout lytic infection with the exception of the transcript encoding ORF-O and ORF-P. These two partially overlapping ORFs are expressed antisense to the ICP34.5 (RL1) gene [Randall et al., 1997]. Consistent with the previous report, the respective transcript was already well detectable in 4sU-seq data at 1 h p.i. but transcriptional activity declined rapidly afterwards (Fig. 4.13A). Nevertheless, translation of the respective transcripts remained detectable until late times of infection. Interestingly, the absence of a canonical start codon resulted in the hypothesis that ORF-O initiates from the same AUG start codon as ORF-P but then diverges within the first 35 codons due to a ribosomal frame shift. We did not observe any evidence for frame-shifts in the HSV-1 translatome. While translation of ORF-O was rather weak, our data indicate that it rather initiates from a non-canonical ACG start codon 76 nt upstream of the ORF-P (Fig. 4.13B).

Finally, we also aimed to validate novel HSV-1 ORFs by whole proteome mass spectro-

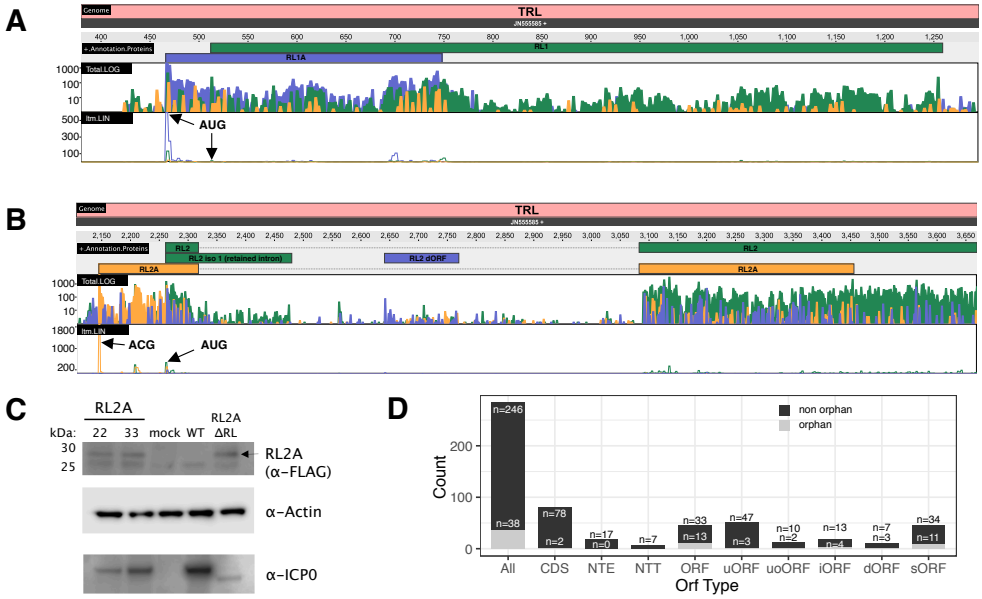


Figure 4.12: Ribosome profiling data visualizing expression of the two viral open reading frames (ORFs), RL1A and RL2A (the three colors depicting the read count for each of the three possible frames, yellow=1, blue=2, green=3), expressed from the **(A)** RL1 and **(B)** RL2 locus of the terminal repeats (TRL). Both standard ribosome profiling data in log scale (Total.log) as well as translation start site profiling data obtained using Lactimidomycin (linear scale, ltm.LIN) are shown. The three possible reading frames are colored in yellow (frame 1), blue (frame 2) and green (frame 3). Both ORFs are well expressed and validated by the strong peak of their respective Translation start site (TaSS) in the ltm-track (black arrows). While RL2A initiates from a non-canonical ACG start codon, the 93 aa RL1A protein initiates from an AUG start codon and was previously missed due to its length of <100 aa. **(C)** Validation of RL2A expression by Western blot. Primary human fibroblasts were infected with two viral clones (22, 33) with one RL segment expressing a 3X-FLAG-tagged RL2A (RL2A), mock, wild-type HSV-1 (WT; for 24 h) or 3X-FLAG-RL2A-ΔRL (=RL2A-ΔRL) for 24 h. Interestingly, insertion of the 3X-FLAG-tag resulted in a loss of ICP0 expression presumably due to the introduction of three out-of-frame AUG start codons (within each FLAG-tag) upstream of the ICP0 TaSS. This was most pronounced when the second repeat was deleted (RL2A-ΔRL). Actin served as house-keeping control. A representative experiment of two independent experiments is shown. **(D)** Distributions of all identified types of ORFs (CDS=coding sequence, NTE=N-terminal extended ORFs, NTT=N-terminal truncated ORFs, uORF=upstream ORF, uoORF=upstream overlapping ORF, iORF=internal ORF, dORF=downstream ORF, sORF=short ORF) of HSV-1 classified by ORFs and orphan ORFs. Source data are provided within the Source Data file.

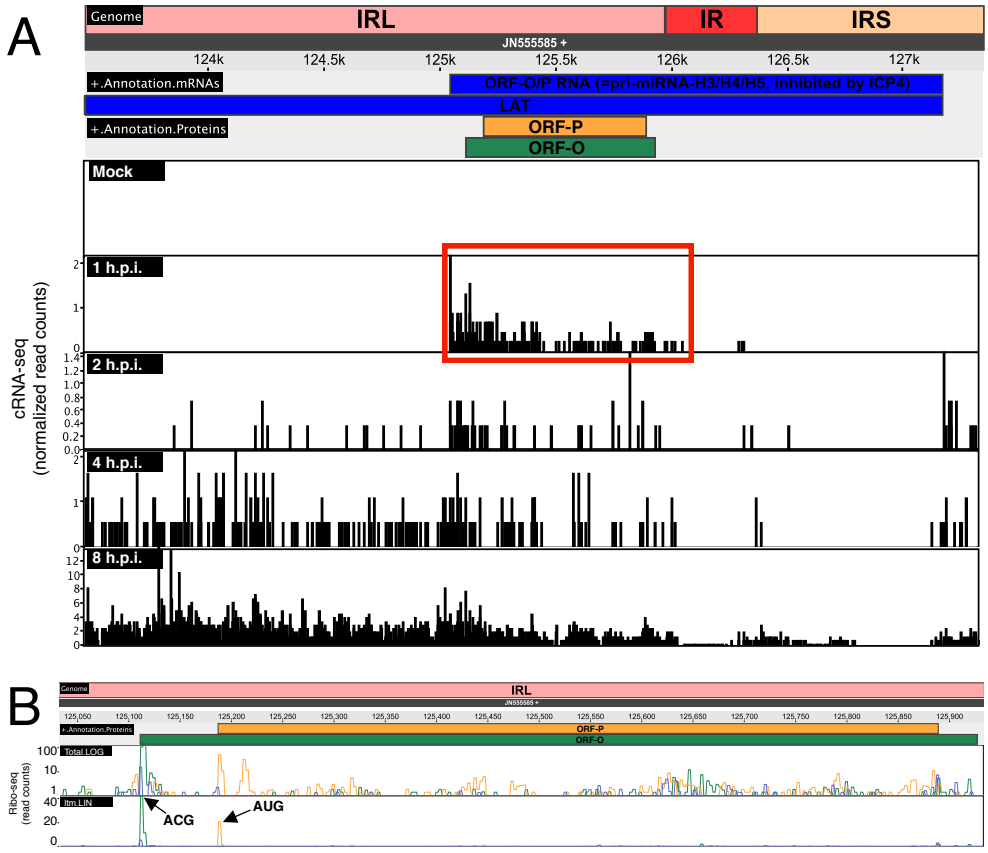


Figure 4.13: **(A)** Expression kinetics of the ORF-O/ORF-P mRNA depicted by cRNA-seq. While the mature transcript is well expressed at 1 h p.i., transcriptional activity rapidly declines thereafter and is obscured by transcription from upstream genomic regions later on in infection. **(B)** Ribosome profiling (Ribo-seq) data for ORF-O and ORF-P. Combined data from all time points analyzed by standard ribosome profiling (Total log) is shown to account for the overall low translation rates. While translation of ORF-P is well represented, translation of ORF-O is less prominent. While we cannot fully exclude the previously proposed frameshift within ORF-O, a strong translation start site peak obtained by Lactimidomycin treatment 76 nt upstream of the AUG start codon of ORF-P is consistent with ORF-O initiating from an ACG start codon upstream of ORF-P. Colors in the Ribo-seq data depict the three possible open reading rames (yellow=1, blue=2, green=3).

metry. We obtained triple-SILAC whole proteome data from HSV-1 infected HFF at 0, 4 and 8 h p.i (n=3). Furthermore, we performed whole proteome mass spectrometry from primary lung fibroblasts infected with HSV-1 for 0, 4, 9 and 15 h. In total, this confirmed only 11 (6%) of the 186 novel ORFs and sORFs (Tab. B.5; excluding NTEs and NTTs). This rather small fraction is consistent with previous work on novel HCMV ORFs [Stern-Ginossar et al., 2012] and presumably reflects that the majority of viral sORFs are inherently unstable and rapidly degraded upon translation similar to their cellular counterparts. Nevertheless, they may play an important role in regulating translation of the viral ORFs encoded downstream of them. Furthermore, the novel large viral ORFs were expressed at substantially lower (10x) levels than the previously identified viral protein coding sequences (Fig. 4.14).

4.3.4 Development of a new integrative nomenclature of HSV-1 gene products

The large number of viral gene products required the extension of the current nomenclature. We first compiled viral gene units comprising transcript isoforms, ORFs and regulatory entities, e.g. uORFs and uoORFs (Tab. B.3). A detailed description of the applied rules is provided in section 4.5.11. In brief, we fully maintained the current nomenclature for all ORFs in the reference annotation [Fields et al., 2013] and attributed each ORF to the most highly expressed transcript in its vicinity. The large number of novel viral gene products required the extension of the current nomenclature. We annotated the novel viral gene products and compiled viral gene units comprising transcript isoforms, ORFs and regulatory entities, e.g. uORFs and uoORFs (Tab. B.3). We fully maintained the current nomenclature for all ORFs in the reference annotation [Fields et al., 2013] and expanded upon it. Alternative transcript isoforms initiating within less than 500 nucleotides were labeled with "*" (extended) or "#" (truncated), e.g. UL13 mRNA *1. Finally, alternatively spliced transcripts were labeled with "iso 1" and "iso 2". Short ORFs (<100 aa), were named upstream ORF (uORF), upstream overlapping ORF (uoORF), internal ORF (iORF) and downstream ORF (dORF) in relation to the next neighboring large ORF. Any ORF for which no transcript could be identified to be responsible for its translation was labeled as "orphan". An overview of the status of the various kinds of ORFs that we identified is shown in Fig. 4.12D. In accordance, any transcript, which was not found to encode an ORF or sORF within its first 500 nt was also labeled as "orphan". Interestingly, we identified 41 "orphan" transcripts (Tab. B.6), which showed predominantly nuclear localization indicating that they may represent novel viral nuclear long non-coding RNAs (lncRNAs). However, all of them were expressed at rather low levels. Accordingly, we were unable to validate five of them by Northern blots despite extensive efforts. We conclude that HSV-1 does not express any highly transcribed viral non-coding RNAs during lytic infection. We uploaded the fully-reannotated HSV-1 genome information to the NCBI GenBank Third Party Annotation database (accession number BK012101).

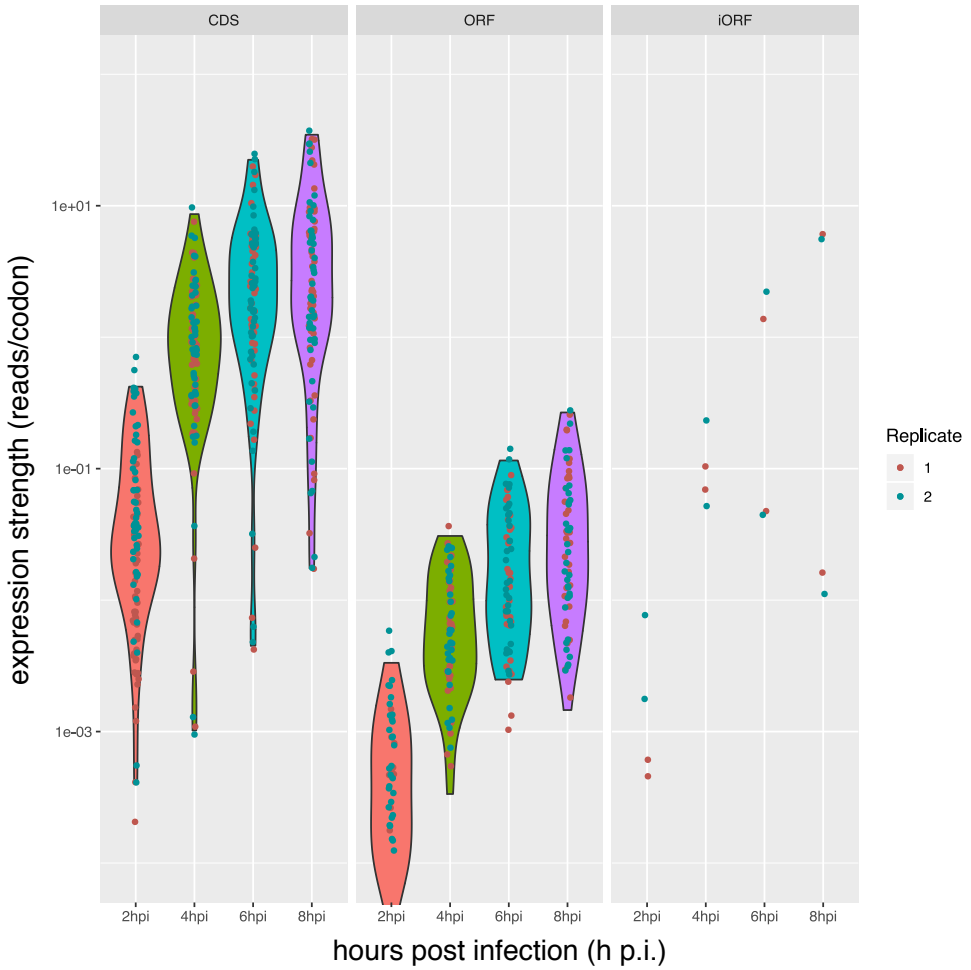


Figure 4.14: Expressions strengths of all open reading frames (ORFs) with an amino acid length $\geq$100 over the course of the infection classified by their respective ORF type. This includes all known large ORFs (CDS) and 41 ORFs and 2 internal ORFS (iORFs). Overlapping ORFs translated in the same frame were excluded. All identified ORFs show lower expressions than the previously identified ones (CDS). Most of the large ORFs are expressed at relatively low levels compared to the known large viral ORFs.

4.4 Discussion

In recent years, major advances in high-throughput experimental methodology have revealed that herpesvirus gene expression is surprisingly complex. While a number of studies in the last few years described hundreds of novel viral transcripts and ORFs, a systematic analysis, validation and integration into gene modules, which attribute individual ORFs and sORFs to specific transcripts they are expressed from, was not attempted. Moreover, the lack of a standardized nomenclature has hampered functional studies on these new viral gene products. Based on a wide spectrum of new and published functional genomics data, we here provide a state-of-the-art, fully revised annotation of the HSV-1 genome.

Calling novel gene products based on big data poses the risk of false positives. As virus replication already initiates at 2 h p.i., first virus particles are released at 4 h p.i. and >80% of translational activity in the infected cells is viral at 8 h p.i. [Rutkowski et al., 2015], we restricted our analysis to the first 8 h of infection to reduce the risk of detecting aberrant gene expression in cells with extensive cytopathic disruption of the transcriptional machinery. Integrative analysis of transcription start site (TiSS) data obtained by both second (cRNA-seq, dRNA-seq) and third (PacBio, MinION) generation sequencing approaches highlighted the necessity to validate viral TiSS by multiple means to exclude such experimental artifacts generated by the individual approaches. Similarly, the different transcription profiling approaches identified numerous putative novel splicing events. However, the vast majority of these only occurred at very low levels. We restricted our analysis to the first 8 h of lytic HSV-1 infection and only included splicing events observed by at least two approaches. While MinION sequencing recently identified novel intergenic splice sites resulting in novel fusion proteins (e.g. between ICP0 and glycoprotein L) [Depledge et al., 2019], the respective transcripts only arise very late in infection and their functional relevance remains unclear. We thus did not include them into our annotation. We conclude that splicing in the HSV-1 transcriptome was already well described by the previous reference annotation but rare splicing events may explain some of our "orphan" viral ORFs and sORFs.

In eukaryotic cells, RNA polymerase II (Pol II) may continue transcribing for thousands of nucleotides downstream of the PAS until transcription is terminated and Pol II is released from the chromatin [Proudfoot, 2016]. With viral gene expression rapidly increasing throughout lytic infection, mRNA precursors with unprocessed 3'-ends that still extend beyond the canonical polyadenylation site are likely to be prevalent in the infected cells. Thus, unprocessed viral pre-mRNAs can easily be misinterpreted as mature viral transcripts. This presumably explains previous reports of near-complete transcription of herpesviral genomes during productive infections [Stern-Ginossar et al., 2012, Gatherer et al., 2011, Marcinowski et al., 2012]. Analysis of cytoplasmic rather than total RNA provides a more accurate picture of the mature viral transcriptome. Consistent with previous reports, we confirmed that the major viral RNA export factor ICP27 was required for efficient export of all viral transcripts [Sandri-Goldin, 2011]. Interestingly, this also included all immediate early genes and spliced viral transcripts.

Ribosome profiling identified 134 novel sORFs expressed during lytic HSV-1 infection. The majority of these represent so called upstream open reading frames (uORFs). Interestingly, a relatively large fraction of transcript isoforms (20%) encode their own uORF, which preferentially (54%) initiate from AUG start codons. Cellular uORFs constitute an important regulatory network governing gene expression at the level of translation by affecting translation initiation of the downstream ORF [Johnstone et al., 2016]. Reliable annotation of both viral transcripts and their respective uORFs will now enable functional studies on these cryptic viral gene products. sORF-encoded polypeptides are usually highly unstable and thus remain undetectable by whole proteome mass spectrometry (WP-MS). Accordingly, we were only able to confirm about 5.5% of our novel ORFs by WP-MS. Interestingly, we could recently show that peptides derived from cellular sORFs are nevertheless efficiently incorporated into and presented by MHC-I molecules on the cell surface despite remaining virtually undetectable by whole proteome mass spectrometry [Erhard et al., 2018]. sORF-derived peptides thus may constitute a new viral class of antigens that are efficiently presented by MHC-I but, due to their instability and extremely low abundance within the cell's proteome, represent poor substrates for cross-presentation and CD4-CD8 augmentation. Further studies are necessary to assess the role of HSV-1 sORFs in the regulation of viral protein expression, antiviral T cell control and evasion thereof.

Based on our revised annotation of 201 viral transcripts and 284 ORFs, we extended the existing nomenclature to include these novel viral gene products. This did not involve any renaming of previously described viral gene products. Our nomenclature thereby explains gene expression of the majority of viral ORFs in the context of different transcript isoforms, uORF and uoORFs. This will facilitate functional studies on the novel viral gene products as well as their transcriptional and translational regulation.

4.5 Methods

4.5.1 Cell culture, viruses and infections

Human foreskin fibroblasts (HFF, #86031405, purchased from ECACC), 293T, Vero 2-2 (Smith, Hardwicke, & Sandri-Goldin, 1992) BHK-21 and BHK-21 dox-UL19 (described below) cell lines were cultured in flasks containing Dulbecco's Modified Eagle Medium (DMEM), high glucose, pyruvate (ThermoFisher #41966052) supplemented with 1x MEM Non-Essential Amino Acids (ThermoFisher #11140050), 1mM additional sodium pyruvate (ThermoFisher #11360070), 10% (v/v) Fetal Bovine Serum (FBS, Biochrom #S 0115), 200IU/mL, penicillin (pen) and 200µg/mL streptomycin (strep). All cells were incubated at 37°C in a 5% (v/v) CO2-enriched incubator.

HFF were utilized from passage 11 to 17 for all high-throughput experiments. Virus stocks were produced on baby hamster kidney (BHK) cells except for the viruses described below. Stocks of the ICP27 null mutant (strain KOS) [Smith et al., 1992] were produced on complementing Vero 2-2 cells [Sekulovich et al., 1988]. HFF were infected for 15 min at 37°C about 24 h after the last split using a multiplicity of infection (MOI) of 10. Subsequently,

the inoculum was removed and fresh media was applied to the cells.

To reconstitute the 3X-FLAG-tagged UL19 NTE, BHK-21 dox-UL19 cells with doxycycline-inducible expression of UL19 were generated by cloning the HSV-1 Syn17+ UL19 coding sequence using primers described in Tab. B.7 into the SalI and NheI sites of pTH3, a derivative of pCW57.1 with a custom multiple cloning site in lieu of the gateway cloning site and the addition of the TRE tight promoter from pTRE-Tight. Lentiviral vectors were generated by cotransfection of this construct with psPAX2 and pCMV-VSV-G into 293T cells. Lentivirus-containing supernatants were sterile-filtered with Minisart® NML 0.45µm cellulose acetate filters (Sartorius #17598) and added to BHK-21 cells. Polyclonal populations were selected 48 h post-transduction and maintained in 1µg/mL puromycin.

4.5.2 Viral mutagenesis and reconstitution

All viral mutants were generated via en passant mutagenesis [Tischer et al., 2010] using Escherichia coli strain GS1783 with the bacterial artificial chromosome (BAC) HSV1(17+)-LoxCheVP26 [Sandbaumhüter et al., 2013] expressing a fusion protein of mCherry on the N-terminus of the UL35 gene product (VP26). Full primer and construct sequences can be found in the Tab. B.7. BAC DNA was purified using the NucleoBond BAC 100 kit (Macherey-Nagel #740579) and transfected for virus reconstitution into BHK-21 cells with Lipofectamine 3000 (ThermoFisher #L3000-075). HSV-1 expressing the 3X-FLAG-tagged N-terminal extension of UL54 virus were reconstituted and titrated on Vero 2-2 cells [Sekulovich et al., 1988]. The virus expressing the tagged N-terminal extension of UL19 was generated in BHK-21 dox-UL19 cells. BHK-21 dox-UL19 cells were plated the day before in media containing 1µg/mL doxycycline (Sigma #D3072), which was maintained throughout virus generation. 3X-FLAG-tagged RL2A BAC-derived viruses were constructed by insertion of the tag with a kanamycin cassette into one genomic repeat followed by replacement of the second repeat (region upstream of RL1 through the second exon of RL2) with the ampicillin resistance gene from pcDNA3. The kanamycin cassette was removed thereafter by traceless mutagenesis.

Virus produced by transfected cells was expanded on minimally five T175 flasks of the corresponding cell type. Virus-containing supernatants were harvested upon >90% cytopathic effect and centrifugation at 8,000 RCF at 4°C for 10 min to pellet cells. Cell pellets were snap-frozen in liquid nitrogen and thawed at 37°C three times to free cell-associated virus. Cellular debris was pelleted at 10,000 RCF, 4°C for 10 min and supernatant combined with the supernatant in the previous step. Virions were pelleted by centrifugation at 19,000 RCF for two hours at 4°C, resuspended in phosphate-buffed saline (PBS), and pelleted once more over a 20% (w/v) sucrose cushion in PBS 16,000 RPM for two hours at 4°C in a SW 28 swinging-bucket rotor (Beckman). Virus pellets were resuspended in PBS, snap-frozen in liquid nitrogen, stored at -80°C, and titrated by plaque assay. Infections were carried out in serum-free DMEM containing penicillin and streptomycin for 1 h at 37°C. The time at which inoculum was replaced with growth media was marked as the 0 h timepoint.

4.5.3 Western blot

Samples were harvested at the indicated timepoints by removal of growth media and direct lysis in 2x Laemmli buffer containing 5% (v/v) β-mercaptoethanol. Samples were sonicated and heated for 5 min at 95°C before loading onto a Novex™ WedgeWell™ 4-20% Tris-Glycine Gel (ThermoFisher #XP04200BOX). Proteins were transferred to polyvinylidene difluoride (PVDF) membranes, blocked for 1 h at room temperature in 1xPBST containing 5% (w/v) milk (Carl Roth T145.3), and probed using α-FLAG M2 (Sigma #F1804) overnight at 4°C at a 1:1000 at dilution and α-mouse IgG (whole molecule)-peroxidase (Sigma #9044) for 1 h. β-actin was probed using α-β-actin C4 antibody (Santa Cruz #sc-47778) at a 1:1000 dilution for 1 h, followed by IRDye® 800CW goat α-mouse IgG (LI-COR #926-32210) at 1:5000 or α-mouse IgG (whole molecule)-peroxidase (Sigma #9044) for 1 h. ICP0 was probed using α-ICP0 clone 5H7 (Santa Cruz #sc-56985) at a 1:1000 dilution for 1 h, followed by IRDye® 680RD goat α-mouse IgG (LI-COR #926-68070) at 1:5000 or α-mouse IgG (whole molecule)-peroxidase (Sigma #9044) for 1 h. Samples were washed with 1xPBST and blocked before addition of each antibody in the milk/PBST buffer. Blots were visualized with a LI-COR Odyssey® FC Imaging System.

4.5.4 Immunofluorescence

105 HFF cells were plated on glass coverslips in 12-well dishes 24h prior to infection. At 8 h post infection cells were fixed in 4% formaldehyde in PBS for 1 h at room temperature, washed three times in PBS and stored at 4°C overnight in PBS. Cells were incubated in permeabilization buffer (10% FBS, 0.25M glycine, 0.2% Triton X-100, 1xPBS) for 1 h at room temperature before incubating them in blocking buffer (10% FBS, 0.25M glycine, 1xPBS) for 1 h at room temperature. Anti-FLAG antibody (GenScript #A00187) was incubated in 10% FBS and 1xPBS for 1 h at 37°C at a concentration of 1μg/mL. The secondary anti-mouse IgG, Alexa Fluor 488 (ThermoFisher #A11017) was incubated in 10% FBS in 1xPBS for 1 h at room temperature with 0.5 μg/mL 4',6-diamidino-2-phenylindole (DAPI). All steps were followed by three 5-minute washes in PBS except for after the primary antibody, which was washed with 1xPBS and 0.05% Tween-20. Coverslips were washed in water before mounting them in medium containing Mowiol 4-88 and 2.5% (w/v) 1,4-diazabicyclo[2.2.2]octane (DABCO).

4.5.5 Transcription start site (TiSS) profiling

Total cellular RNA was isolated using Trizol (Invitrogen) following the manufacturer's instructions. RNA was resuspended in water and stored at -80 °C until use. TiSS profiling dataset using cRNA-seq utilizes a similar approach as employed for decoding HCMV [Stern-Ginossar et al., 2012]. Following rRNA depletion and extensive chemical RNA fragmentation, 50-80 nt RNA fragments are recovered by gel extraction. Library preparation is performed by 3'-adaptor ligation and circularization. This inherently enriches for transcript 5'-ends by 20- to 30-fold. Of note, our cRNA-seq library preparation protocol introduces

a 2 + 3 nt unique molecular identifier (UMI), which facilitates the subsequent removal of PCR duplicates from sequencing libraries.

TiSS profiling dataset using dRNA-seq was prepared according to the published protocol [Sharma and Vogel, 2014] with some modifications by the Core Unit Systems Medicine (Würzburg). In brief, for each sample 3 μg of DNase-digested RNA was treated with T4 Polynucleotide Kinase (NEB) for 1 h at 37 °C. RNA was purified with Oligo Clean & Concentrator columns (Zymo) and each sample was split into an Xrn1 (+Xrn1) and a mock (-Xrn1) sample. The samples were treated with 1.5 U Xrn1 (NEB; +Xrn1) or water (-Xrn1) for 1 h at 37 °C. Digest efficiency was checked on a 2100 Bioanalyzer (Agilent) and 5' caps were removed by incubation with 20 U of RppH (NEB) for 1 h at 37 °C. Afterwards, RNA was purified and eluted in 7 μL and 6 μL were used as input material for the NEBNext® Multiplex Small RNA Library Prep Set for Illumina®. Library preparation was performed according to the manufacturer's instruction with the following modifications: 3' adapter, SR RT primer and 5' adapter were diluted 1:2, 13 cycles of PCR were performed with 30 sec of elongation time, and no size selection was performed at the end of library preparation. Concentrations of libraries were determined using the Qubit 3.0 (Thermo Scientific) and their fragment sizes were determined using the Bioanalyzer. Libraries were pooled equimolar. Sequencing of 75 bp single-end reads was performed on a NextSeq 500 (Illumina) at the Cambridge Genomic Services (cRNA-seq) and the Core Unit Systems Medicine in Würzburg (dRNA-seq). To validate TiSS identified by cRNA-seq, dRNA-seq, PacBio or MinION (no reads were reanalyzed, only the called transcripts were used for PacBio and MinION), total RNA-seq and 4sU-seq data that were previously published [Rutkowski et al., 2015] were reanalyzed (see below).

4.5.6 RNA-seq of subcellular RNA fractions

Subcellular RNA fractions (cytoplasmic, nucleoplasmic and chromatin-associated RNA) were prepared by combining two previously published protocols [Rosner et al., 2013, Pandya-Jones and Black, 2009]. Data from uninfected and wild-type HSV-1 infected cells were published recently [Hennig et al., 2018]. Infection with the ICP27-null mutant was performed in the same experiment. As for wild-type HSV-1 infection, total cellular RNA was isolated using Trizol at 8 h p.i. Fractionation efficiencies were confirmed on the RNA-seq data by comparing expression values of known nuclear and cytoplasmic RNAs as well as intron contributions (see Fig. 4.15) [Hennig et al., 2018]. Sequencing libraries were prepared using the TruSeq Stranded Total RNA kit (Illumina) following rRNA depletion using Ribo-zero. Sequencing of 75 bp paired-end reads was performed on a NextSeq 500 (Illumina) at the Cambridge Genomic Services and the Core Unit Systems Medicine (Würzburg).

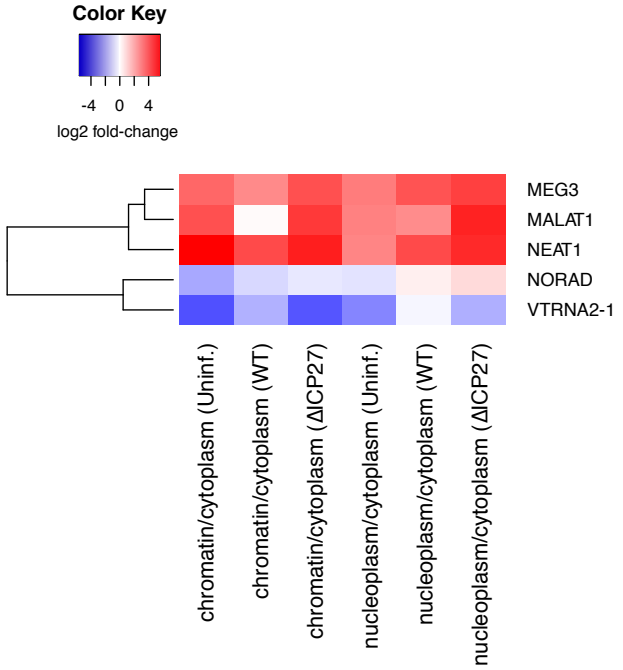


Figure 4.15: Enrichment or depletion levels of known nuclear or cytoplasmic RNAs, respectively in chromatine associated or nucleoplasm RNA versus cytoplasm RNA.

4.5.7 Ribosome profiling

The ribosome profiling time-course data (lysis in presence of cycloheximide) have already been published [Rutkowski et al., 2015]. Additionally, so far unpublished data we generated include translation start site (TaSS) profiling performed by culturing cells in medium containing either Harringtonine (2 µg/ml) or Lactimidomycin (50 µM) for 30 min prior to harvest. Harringtonine samples were obtained for 2 h and 8 h p.i., Lactimidomycin was employed for mock, 4 and 8 h p.i. Two replicates of each condition were analyzed. All libraries were sequenced on a HiSeq 2000 at the Beijing Genomics Institute in Hong Kong.

4.5.8 Proteomic analysis

WI-38 lung fibroblasts grown in SILAC medium were infected at MOI 10 and harvested after 4, 9, and 15 h. After cell lysis, protein concentration was determined by Bradford assay and 200 μg of each sample were mixed with equal amount protein extracted from untreated cells grown in SILAC light medium. Primary human foreskin fibroblasts were grown for five passages in DMEM lacking lysine and arginine (Thermo Scientific) supplemented with 10% dialysed FCS (Gibco), 100 units/mL penicillin and 0.1 mg/mL streptomycin, 280 mg/L proline (Sigma) and light (K0, R0; Sigma), medium (K4, R6; Cambridge Isotope Laboratories) or heavy (K8, R10; Cambridge Isotope Laboratories) 13 C/15N-containing lysine (K) and arginine (R) at 50 mg/L. Pre-labeled cells were infected with HSV-1 at a

Table 4.3: Description of read yield, total and uniquely mapped reads for all the sequencing datasets used.

	million reads per dataset				
	cRNA-seq	dRNA-seq	4sU-seq	total RNA-seq	RNA from subcellular fractions
total sequenced reads	264	353	744	345	391.4
total mapped reads	31.6	227.9	259.1	270.3	332
total mapped reads (human)	25.6	227.2	198	260.1	261.6
total mapped reads (HSV1)	6	0.7	61.1	10.2	70.4
uniquely mapped reads	22.6	140	224.6	218.7	317.4
uniquely mapped reads (human)	16.6	139.3	165.5	209.1	253.2
uniquely mapped reads (HSV1)	6	0.7	59.1	9.6	64.2

multiplicity of infection (MOI) of 10 for 4 or 8 h, and uninfected cells were included as a control. The experiment was conducted in triplicate (biological replicates), with a 3-way SILAC label swap.

4.5.9 Data analysis, statistics and reproducibility

Western blotting and immunofluorescence images are representative of at least two independent biological replicates. Random and sample barcodes in cRNA-seq and ribosome profiling data were analyzed by trimming the sample and UMI barcodes and 3' adapters from the reads using our in-house computational genomics framework gedi (available at https://github.com/erhard-lab/gedi). Barcodes introduced by the reverse transcription primers included three random bases (UMI part 1) followed by four bases of sample-specific barcode followed by two random bases (UMI part 2). Reads were mapped using bowtie 1.0 against the human genome (hg19), the human transcriptome (Ensembl 75) and HSV-1 (JN555585). The HSV-1 genome consists of two components (L and S) that are both flanked by long repeats. To mitigate the effect of multi-mapping reads, we masked the terminal repeats by NNN. The three mappings were merged and only the alignments for a read with minimal number of mismatches were retained. Reads were assigned to their specific samples based on the sample barcode. Barcodes not matching any sample-specific sequence were removed. PCR duplicates of reads mapped to the same genomic location were identified by counting UMIs. If two observed UMI differed by only a single base, one likely is due to a sequencing error. Thus, we discarded one of the two in such cases. If the reads at this location mapped to k locations (i.e., multi-mapping reads for $k > 1$), a fractional UMI count of $1/k$ was used (see Table 4.3 for read yields) [Rutkowski et al., 2015]. Finally, all read mappings in the repeats were copied into the previously masked regions.

dRNA-seq, 4sU-seq, total RNA-seq and RNA-seq data of subcellular fractions were processed similar to cRNA-seq and ribosome profiling data with the exception of STAR (v.2.5.3a) being used to map the reads and PCR duplicates were not collapsed as no random barcodes were used (see Table 4.3 for read yields).

All libraries were prepared with strand-sensitive protocols. Consequently, all reads were

mapped only to their respective strand. Further, reads were weighted by the number of different locations they map to (i.e., if a read mapped to three locations, its weight was 1/3).

Our dRNA-seq and cRNA-seq TiSS profiling data were analyzed with our TiSS analysis pipeline iTiSS (integrative Transcriptional Start Site caller). It screens potential TiSS for clustered accumulation of read 5'-ends in dRNA-seq (i) and cRNA-seq (ii) data. It evaluates our cRNA-seq data for an increase in upstream to downstream read coverage at potential TiSS (iii), and temporal changes in the potential TiSS read clusters in the cRNAs-seq time-course data (iv). It accounts for TiSS already identified by MinION (v) and PacBio (vi) sequencing. Here, for the MinION data we used the already predicted transcripts by Depledge et al [Depledge et al., 2019]. However, we observed that their transcripts usually started around 15 bp downstream of the potential TiSS that we called by cRNA-seq, dRNA-seq as well as the transcripts called by PacBio from Tombacz et al. This is a common observation for MinION datasets caused by 5' degradation. Therefore, we used their transcriptional start sites as a criterion for a potential TiSS found by (i), (ii), (iii), (iv) or (vi), if Depledge et al called a transcript starting in window of up to 20 bp downstream. Similar to the MinION dataset, called transcripts were already published for the PacBio data set. We used these as an additional criterion for potential TiSS found by (i), (ii), (iii), (iv) or (v) if a transcriptional start site was called by Tombacz et al in their PacBio dataset in a window of +/- 5 bp. Further, we analyzed our 4sU-seq time-course data to both score potential TiSS that explained temporal changes in expression levels throughout infection (vii). Similar to (iv), a region downstream of a potential TiSS that has different expression behavior over time as compared to the region upstream is an indicator of a bona fide TiSS. The expression behavior can be analyzed using our previously published 4sU-seq data. First, the number of reads with a 5' end in between two potential TiSS was determined for each of the 4sU-seq samples. For two subsequent potential TiSS of the same gene locus, distinct kinetic behavior was then tested using a likelihood ratio test: Two linear models were constructed with the log2 ratio of the two corresponding read counts as dependent variable, and either no independent variable (offset only; model 1), or the time after infection of the corresponding samples as independent variable (model 2). Statistical significance was then determined using the χ^2 distribution based on the likelihood ratio of these two nested models. If the log2 ratio changes during infection, the model 2 better fits the data than the model 1. This criterion is fulfilled if the (Benjamini-Hochberg adjusted) p value was below 1%. Further, we used the increase in downstream to upstream read coverage as an additional marker for bona fide TiSS (viii). Similar to (iii), an increase of the 4sU-seq read coverage in any sample downstream of a potential TiSS compared to upstream of it is an indicator of a bona fide TiSS. First, the effective length of each range in between two subsequent potential TiSS of the same gene locus was determined based on the fragment length distribution from the experiment (paired-end) and actual length of the range. The effective length of the range between TiSS a and b is a measure of the expected number of RNA fragments that can be sequenced and can originate from a transcript starting at a but not from a transcript starting at b. If the transcription

termination site (TTS) is far away with respect to the fragment length distribution, the effective length equals the actual length. Closer TTS restrict the number of possible RNA fragments and the effective length is reduced accordingly. The effective length and the read count from (vii) were then used to compute a sample-specific coverage for each pair of subsequent potential TiSS. This criterion was fulfilled if the TiSS was associated with at least a 2-fold increase in coverage for at least 4 different samples. Finally, we also scored TiSS that explained translation of viral ORFs, for which no other transcript had otherwise been identified (ix). Every ORF needs a transcript from which it is translated. Commonly, translation initiates with the first 250 bp of an mRNA. Thus, this criterion is fulfilled if a yet unexplained ORF initiated within the next 250 bp downstream of a potential TiSS. If only one of the previous criteria is met, but an ORF is found starting downstream of it, it is more likely that the potential TiSS is correct. Please note that we carefully assessed all potential TiSS, which only fulfilled one other criterion than (ix). For more details on iTiSS' peak calling (i-iv) see chapter 3. Thus, 9 criteria were utilized to confirm a potential TiSS. All identified TiSS were manually assessed and curated.

Afterwards, potential TiSS within a ±5 bp window were combined into a single TiSS. Consequently, a TiSS is defined by a single-nucleotide position including a ±5 bp window. The fidelity of this definition can be appreciated by the strong enrichment of the Inr motif even for the most weakly expressed viral transcripts.

The reported enrichment factors for dRNA-seq and cRNA-seq were calculated based on predicted TiSS in human rather than HSV-1. This was done to prevent undesired biases due to read-in caused by the extraordinary high number of overlapping transcripts in HSV-1. The predicted TiSS were ordered based on the number of reads starting at their respective positions. The median was then calculated over the 50 strongest and 10 strongest expressed TiSS for cRNA-seq and dRNA-seq, respectively.

Significance of the correlation between the presence of a TATA-box-like motif and the transcription strength of TiSS was calculated using Fisher's exact test. Here, the TiSS were ordered by their numbers of reads starting and sorted into three equally sized bins. For the bin containing the strongest TiSS as well as the bin containing the weakest TiSS, the number of all nucleotides between position -30 and -25 relative to the TiSS were summed up. For the parameters of the Fisher's exact test, the following sums were used a = T + A(strongest bin), b = C + G(strongest bin), c = T + A(weakest bin) and d = C + G(weakest bin).

We used our in-house tool PRICE [Erhard et al., 2018] version 1.0.1 to call ORFs separately for the two replicates of ribosome profiling data but pooling all samples from each replicate.

RNA-seq data were mapped using STAR [Dobin et al., 2013] version 2.5.3a using a combined reference index derived from Ensembl 90 and our final HSV-1 annotation.

We analyzed mass spectrometry data using MaxQuant [Cox and Mann, 2008] version 1.6.5.0. Spectra were matched against a combined database of proteins from Ensembl (version 75), and all ORFs identified by ribosome profiling. We used carbamidomethylation as fixed and acetylation (N-terminal) and oxidation at methionine as variable modifications. Peptides were filtered for 1% FDR using the target-decoy approach by MaxQuant.

The export indices of chromatin-associated RNA and cytoplasmic RNA were derived by computing their fold changes between the wild-type and the null mutant for ICP27 using the lfc R-package [Erhard, 2018]. Data handling and visualization was done using R including the ComplexHeatmap [Gu et al., 2016], circlize [Gu et al., 2014], ggseqlogo, ggplot2, reshape2, plyr, scales, ggforce, ggrepel, and the gridExtra packages.

To reveal the potential function or functional motifs of predicted protein sequences, we used sequence comparison, domain composition, structure prediction and motif searches [Dandekar et al., 2000, Gaudermann et al., 2006, Bencurova et al., 2018]. Sequence comparisons exploited Blast searches iteratively [Camacho et al., 2009] and identified catalytic as well as regulatory domains including predictions by the conserved domain database [Lu et al., 2020]. Predicted domain composition was verified using domain databank tools SMART [Letunic et al., 2015] and Prodom [Hernández et al., 2015]. Motif searches exploited Prosite regular expressions and profiles and the integrative protein signature database [Hunter et al., 2009]. As independent tests for resulting function assignments structure annotation for protein domains was done using AnDOM software [Schmidt et al., 2002] as well as homology predictions by SwissModel [Waterhouse et al., 2018]. Gene context methods were applied for unclear sequences related to non-viral sequences (STRING database [Szklarczyk et al., 2011]). In addition, Clusters of Orthologous Groups using the latest version (5.0) of the eggNOG tool with its 2502 virus strains provided independent annotation input [Huerta-Cepas et al., 2017].

4.5.10 Manual curation

Once the automatic scoring was done, potential TiSS with a score ¿3 were accepted into our final annotation following manual inspection. Furthermore, potential TiSS with a score of 2 scored by criterion (i)-(viii) were also considered bona fide TiSS. Nevertheless, they were all carefully inspected manually and all found to highly likely represent bona fide TiSS.

All remaining potential TiSS were manually curated by looking at the data in our viewer. In particular, we consider the orphan ORF TiSS criterion (ix) the weakest piece of evidence for a bona fide TiSS. For this reason, we removed TiSS that only fulfilled this and one other criterion, and only kept those that exhibited additional strong evidence (for instance a fold-change between 3 and 4 instead of the picked threshold of 4 in dRNA-seq). In addition, we had a close look at the nucleotide sequence at the TiSS looking for factors that could have impeded cloning or mapping of the respective reads, e.g. poly(C) or poly(G) stretches as well as repeat regions.

Information on the reasons for including each respective potential TiSS into the final HSV-1 genome annotation are included in Tab. B.1 under the column named *Justification*.

Finally, bona fide TiSS were automatically extended to the next poly(A)-site. Those were manually checked for potential poly(A)-read-through transcripts that were validated by our, PacBio or MinION data. The resulting transcripts were included into our final HSV-1 genome annotation. For heavily spliced genomic loci, we also considered the PacBio and

MinION data to annotate specific transcript isoforms.

4.5.11 Principles of the new nomenclature of HSV-1 transcripts and ORFs

- No previously annotated ORFs were renamed to avoid causing confusions with previous work. All viral ORFs and transcript mentioned in the 6th Edition of Fields of Virology were included.
- To differentiate all new ORFs from the previously reported ORFs, we labeled all previous ORFs as "coding sequences", e.g. UL1 CDS.
- We differentiate long (≥100aa; named "ORF") from short (3 - 99aa) viral ORFs.
- We differentiated five different kinds of sORFs. These include upstream open reading frames ("uORFs"), upstream overlapping ORFs ("uoORFs"), internal ORFs ("iORFs") and downstream ORFs ("dORFs"). In addition, sORFs, which are expressed from transcripts not containing any large ORF were named "sORFs", e.g. UL34.5 sORF 1 and 2.
- Translation of "uORFs" both starts and terminates upstream of a large ORF. A transcript can have multiple uORFs (e.g. UL14 uORF 1 and 2). In case a transcript does not encode any ORF >100aa, all short ORFs it encodes are labeled "sORFs", e.g. UL30.5 sORF 1 and UL30.5 sORF 2.
- In contrast to uORFs, uoORFs overlap with the main ORF expressed from the respective transcript.
- Internal ORFs (iORFs) are located within the coding sequence of large ORFs but expressed in a different frame. In principle, two scenarios can explain their translation.
  - They can be translated by ribosomes, which have missed the TaSS of the main ORF (e.g. UL20 iORF) and thus initiate translation at the iORF.
  - They can result from alternative independent transcripts initiating downstream of the respective TaSS of the main ORF, e.g. UL53 iORF RNA #2. iORFs were thus not labeled as "orphan".
- Finally, a small number of downstream ORFs (dORFs) were annotated. These represent sORFs located downstream of large ORFs, which could not be explained by an independent transcript, e.g. UL39.6 dORF 1 and 2 downstream of UL39.6 ORF. Their translation may result from ribosomes re-initiating after completing the translation of the large ORF located further upstream. Therefore, they were not labeled as "orphan". However, in most cases it is equally likely that they are translated from yet unidentified viral transcripts.

- In principle, novel viral transcripts, ORFs and sORFs can all result in the introduction of a new viral gene identifier, e.g. UL28.5.
  - Any novel large viral ORF, e.g. UL36.5 ORF, was given a new identifier unless it was overlapping with another large ORF. In the rare case that two overlapping viral ORFs (translated from different frames) were obviously expressed from the same transcript, these were named A and B, e.g. UL40.7A ORF and UL40.7B ORF as well as TRL2 CDS and TRL2A ORF.
  - For viral transcripts to be given a new identifier, this required a transcription start site (TiSS) >500 nucleotides upstream of the closest other transcript, e.g. UL54.5 RNA (orphan).
  - Any sORF >20aa in length that could not be attributed to another viral gene as a either uORF, uoORF, iORF or dORF was given a new identifier, e.g. UL27.5 sORF 1.
- Numbering of new identifiers was defined based on the location of the TiSS or TaSS in relation to the neighboring previously annotated genes (x and x+1) on either strand. In case multiple new identifiers were required between two annotated genes, the most strongly expressed gene was named x.5, the neighboring ones x.4 and x.6. As annotations of additional genes by previous studies did not all follow the same rules in regards to neighboring genes, we tried to choose the best possible numbering for each locus.
- Usage of alternative transcription start sites is a very common phenomenon in the HSV-1 genome. Many of the additional transcriptions contain additional uORFs and thereby explain their expression. As such, we commonly observed >1 distinct TiSS within a window of 250 nt up- or downstream of the transcript of a given locus. The main TiSS was defined by the highest cRNA-seq or dRNA-seq peak. Within a window of +/-10 nt, no additional TiSS were annotated. TiSS identified by cRNA-seq, dRNA-seq and PacBio commonly matched perfectly at single nucleotide level.
- Any transcript that did not contain an ORF within its first 500 nucleotides (nt) was labeled as "orphan", e.g. UL54.5 RNA (orphan).
- Any ORF or sORF for which no transcript could be identified that explained its translation within the transcript's first 500 nt was labeled as "orphan", e.g. US11.5 ORF (orphan).
- Additional transcripts initiating upstream of the main transcript were labeled "*"+"number" with higher numbers reflecting increasing distance to the main TiSS, e.g. UL24 RNA *1.
- Transcripts initiating downstream of the main transcript were labeled "#"+"number" with higher numbers reflecting increasing distance to the TiSS of the main transcript, e.g. UL41 RNA #1 and UL41 RNA #2.

- Transcript experiencing alternative splicing were labeled as “iso1, iso2...”.
- The annotation of uORFs was based on the most prominent transcript of the respective locus, e.g. UL6 uORF. Alternative TiSS commonly explained the expression of additional uORFs, e.g. UL6 RNA *1 explained UL6 uORF RNA *1.
- Transcripts with retained introns were labeled as “i”+”number”, e.g. IRL2 RNA i1. The respective ORF variants were labeled accordingly, e.g. IRL2 ORF RNA i1
- N-terminal extensions of ORFs were labeled with “*1”, e.g. UL50 CDS *1. In case of a second, longer N-terminal extension this was labeled “*2”, e.g. UL50 CDS *2. All N-terminal extensions of previously identified proteins initiated from non-AUG start codons. Both the start codon and the length of the extension are indicated in brackets, e.g. US3 CDS *1 (includes 23 aa N-terminal extension initiating from CUG).
- N-terminal truncations of ORFs were labeled with “#”+”number”, e.g. UL37.6 ORF #1. We did not observe any more than 1 truncated version of a given ORF.
- ORFs, sORFs and transcripts expressed from the repeat regions of the viral genome where named accordingly, e.g. IRL2.5 ORF and TRL2.5 ORF. We did not differentiate the three other possible orientations of the unique long and unique short regions.

Chapter 5

Dissecting newly transcribed and old RNA using GRAND-SLAM

Motivation: *A recent paper by [Herzog et al., 2017] introduced an improved sequencing method to metabolically label newly synthesized RNA termed SLAM-seq. Compared to the earlier model, where newly synthesized RNA was physically separated from old RNA, SLAM-seq introduced T→C mismatches only found in newly synthesized RNA. Consequently, new from old RNA must be separated in the data using a computational approach. Although their sequencing technique was groundbreaking, they only used a fairly simple method to count the number of T→C mismatches. However, as I already indicated in section 2.1.6, SLAM-seq experiments hold much more information than just the number of T→C. By statistically inferring the actual distribution of new to total reads, we hypothesized that we are able to deduce an actual new to total ratio (NTR) for each gene, which finally led to this paper and the implementation of our tool GRAND-SLAM.*

Publication: *This chapter has been published in Bioinformatics [Jürges et al., 2018] and was presented at the Intelligent Systems for Molecular Biology (ISMB) 2019 conference in Chicago, USA by me in the proceedings track. For this dissertation I made minor changes to the layout and some corrections to the text.*

Individual author contributions: *See Appendix D*

5.1 Abstract

Global quantification of total RNA is used to investigate steady state levels of gene expression. However, being able to differentiate pre-existing RNA (that has been synthesized prior to a defined point in time) and newly transcribed RNA can provide invaluable information e.g. to estimate RNA half-lives or identify fast and complex regulatory processes. Recently, new techniques based on metabolic labeling and RNA-seq have emerged that allow to quantify new and old RNA: Nucleoside analogs are incorporated into newly transcribed RNA and are made detectable as point mutations in mapped reads. However, relatively

infrequent incorporation events and significant sequencing error rates make the differentiation between old and new RNA a highly challenging task. We developed a statistical approach termed GRAND-SLAM that, for the first time, allows to estimate the proportion of old and new RNA in such an experiment. Uncertainty in the estimates is quantified in a Bayesian framework. Simulation experiments show our approach to be unbiased and highly accurate. Furthermore, we analyze how uncertainty in the proportion translates into uncertainty in estimating RNA half-lives and give guidelines for planning experiments. Finally, we demonstrate that our estimates of RNA half-lives compare favorably to other experimental approaches and that biological processes affecting RNA half-lives can be investigated with greater power than offered by any other method. GRAND-SLAM is freely available for non-commercial use at http://software.erhard-lab.de; R scripts to generate all figures are available at zenodo (doi: https://doi.org/10.5281/zenodo.1162340).

5.2 Introduction

Gene expression is a highly dynamic process and determined by the interplay of RNA transcription, processing and decay [Schwanhausser et al., 2011]. High-throughput techniques such as microarray and next generation sequencing (NGS) have become standard tools to quantify gene expression on the level of total RNA. However, knowing the amount of total RNA for each gene at the time of cell lysis does not provide information to distinguish between the processes that constitute gene expression. For instance, when gene expression changes between some treatment and control condition are investigated, differences between total RNA levels can arise due to the treatment affecting transcription, processing or decay. Moreover, if changes after a short period of time (e.g. 1 h after infection by a virus) are of interest, considering total RNA levels can be heavily misleading [Marcinowski et al., 2012].

To resolve these issues, powerful biochemical approaches have been developed in recent years. Most successfully, newly transcribed RNA can be metabolically labeled using nucleoside analogs such as 4-thiouridine (4sU) in living cells. After RNA extraction, labeled RNA can be biochemically separated from pre-existing, unlabeled RNA by thiol-specific biotinylation. Both fractions, in addition to total RNA, can be quantified using microarrays or RNA sequencing. This has allowed to precisely measure RNA half-lives [Dölken et al., 2008], monitor RNA splicing [Windhager et al., 2012] or investigate extremely short-lived RNAs [Schwalb et al., 2016] or complex regulatory processes [Rabani et al., 2014]. However, the biochemical separation step is laborious and error-prone, and requires large amounts of RNA. Moreover, imperfect biochemical separation may introduce severe bias and bioinformatic analysis such as data normalization is highly challenging [Uvarovskii and Dieterich, 2017].

Recently, three studies introduced an alternative approach to differentiate between new and old RNA: SLAM-seq [Herzog et al., 2017], Timelapse-seq [Schofield et al., 2018] and TUC-seq [Riml et al., 2017] directly visualize labeled RNA by sequencing: After labeling

by 4sU and extraction of RNA, chemical agents are used to convert 4sU to cytosine analogs. The sample is sequenced without prior separation, and old and new RNA can be differentiated on the basis of specific T to C mismatches of reads mapped to the reference transcriptome. Importantly, the accuracy of this bioinformatic separation strongly depends on the error rates of sequencing and the 4sU incorporation rates. Even with very long periods of labeling (24h) and high concentrations of 4sU (100 lM), no more than one in 40 uridines is substituted by 4sU [Dölken et al., 2008, Herzog et al., 2017]. Thus, only a small fraction of sequencing reads will contain more than one conversion. Moreover, the error rates of modern NGS dropped below 0.1%, but still give rise to many reads with T to C mismatches. Thus, it is not possible to decide with certainty for each individual read whether it originated from a new or an old RNA molecule.

Therefore, the computational approach termed SLAM-DUNK [Herzog et al., 2017] utilizes all observed T to C mismatches of reads mapped to a gene, and subtracts the observed mismatches from a control experiment without 4sU labeling. These corrected conversion rates were used to compute RNA half-lives in pulse-chase experiments: Efficient 4sU incorporation is achieved by long periods of labeling, followed by wash-out of free 4sU and monitoring the drop of corrected conversion rates over several time points. In addition, labeling for 3 h and 12 h was sufficient to reveal changes of RNA half-life induced by microRNAs and N6 adenosine methylation of the mRNAs in differential experiments, e.g. by knocking out an essential factor for microRNA biogenesis and comparing corrected conversion rates between knock-out and wild-type cells.

Here, we expand on this methodology and present the computational approach *Globally Refined Analysis of Newly transcribed RNA and Decay rates using SLAM-seq* (GRAND-SLAM) that allows to infer the proportion and the corresponding posterior distribution of new and old RNA for each gene in a single SLAM-seq experiment. Compared to the corrected conversion approach, it provides five major advantages: First, no control experiment is needed. Second, a single labeling experiment (as compared to a pulse-chase timecourse) is in principle sufficient to estimate RNA half-lives. Naturally, more experiments increase the accuracy of the estimate. Third, by directly utilizing the posterior distributions, estimated half-lives are more accurate. Fourth, the variance of the posterior distribution, or, alternatively, the size of credible intervals, provide an internal quality control for each gene and experiment. Finally, and most importantly, knowing the proportion of new RNA for each gene allows to investigate fast regulatory processes such as induced by virus infection, which is not possible when only knowing corrected conversion rates.

5.3 Approach

Our approach is based on a binomial mixture model (Fig. 1):

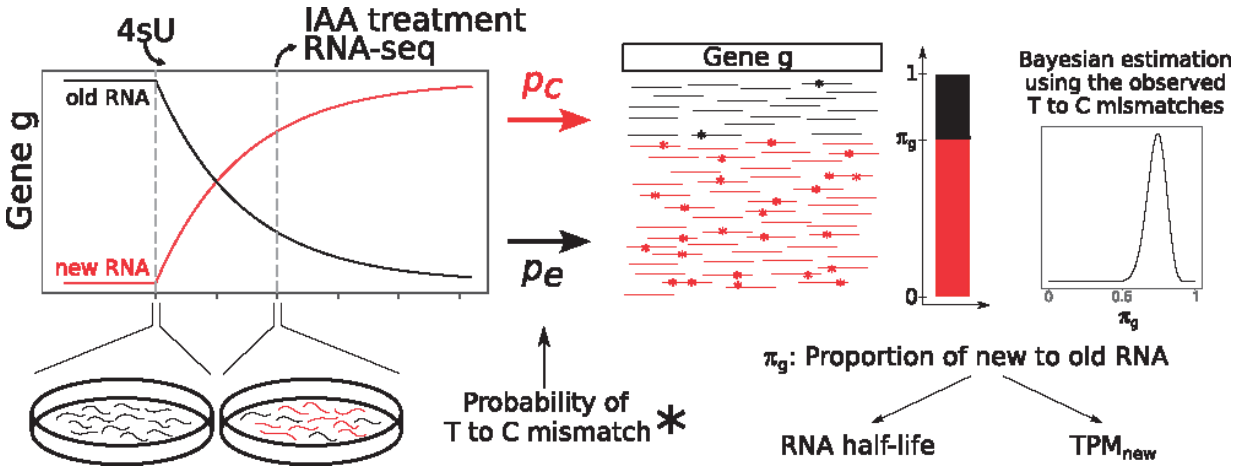


Figure 5.1: GRAND-SLAM overview. After a period of labeling with 4-thiouridine (4sU), RNA is extracted from cells, treated with iodoacetamide (IAA) and sequenced. Shown is a theoretical timecourse of the abundances of new and old RNA for a gene g. IAA converts incorporated 4sU into cytosine analogs with an overall rate p_c (including the incorporation rate, conversion rate and error rate), and uridines are sequenced as cytosine with an error rate p_e. Based on the observed mismatches from T to C, the proportion of new to old RNA of gene g, π_g, can be estimated using Bayesian inference. Estimates of π_g can be transformed into estimates of the gene's RNA half-life or relative abundance measures.

$$P(y; p_e, p_c, n, \pi_g) = (1-\pi_g)B(y; n, p_e) + \pi_g B(y; n, p_c) \tag{5.1}$$

$$B(k; n, p) = \binom{n}{k} \cdot p^k (1-p)^{n-k} \tag{5.2}$$

The sufficient statistics for this model are the number of observed T to C mismatches (y) for each read mapped to a genomic region containing n thymines within a gene g. If π_g is the fraction of newly transcribed RNA among all RNAs of gene g, p_e is the average T to C mismatch rate in unlabeled RNA and p_c is the average mismatch rate in labeled RNA, then observed mismatches for a read are either due to a binomial distribution with success probability p_e (with probability $1-\pi_g$) or a binomial distribution with success probability p_c (with probability π_g).

Thus, our goal is to estimate all π_g using Bayesian inference for observed data $\mathbf{y} = y_1, ..., y_m$ and $\mathbf{n} = n_1, ..., n_m$ (i.e. for each read, how many of the potential T to C mismatches have been observed):

$$f(\pi_g; \mathbf{y}, \mathbf{n}, p_e, p_c) = \frac{\prod_i P(y_i; p_e, p_c, n, \pi_g) \cdot b(\pi_g; \alpha, \beta)}{P(\mathbf{y})} \tag{5.3}$$

For the sake of simplicity, we use a beta prior with density function b and hyperparameters α and β. As we have no prior knowledge on the proportion for each gene, we here use the

uninformative uniform prior with $\alpha = \beta = 1$. The integral $P(\mathbf{y}) = \int_0^1 \prod_i P(y_i; p_e, p_c, n, \pi'_g) \cdot b(\pi'_g; \alpha, \beta) d\pi'_g$ is computed numerically.
Here, we assume p_e and p_c to be constant throughout a sample. Thus, before solving equation 5.3 for each gene, we estimate p_e and p_c based on the data from all genes.

5.4 Materials and methods

5.4.1 Sufficient statistics

The sufficient statistics for parameter estimation are collected in a matrix $A^{(g)}$. Each entry $a^{(g)}_{k,n}$ is the number of reads mapped to a genomic region within gene g containing n thymines with k observed T to C mismatches. We only consider reads consistently mapped to a known transcript (i.e. matching all intron boundaries). Alternatively, for the SLAM-seq experiments [Herzog et al., 2017] where 3' ends of transcripts are sequenced, we only consider reads mapped to the 3' regions defined by SLAM-DUNK [Herzog et al., 2017]. In addition we identify and exclude potential SNPs defined as thymines where more than half of the reads covering it show a mismatch.

5.4.2 Estimating p_e

In principle, p_e can be directly estimated from either spike-in RNAs in the same sample, or using an additional sample without 4sU labeling (no4sU sample) by counting T to C mismatches. However, utilizing an additional experiment may lead to a bad estimate for the 4sU sample of interest, as a broad range of p_e values is observed already in available no4sU samples (see Fig. 5.4C). However, we noticed that the other eleven error rates (one of the nucleotides to any of the other three) are highly correlated to the T to C error rate in the no4sU samples. Thus, we trained a linear regression model to predict the T to C error rate from the other error rates. Manual feature selection revealed that the T to A and T to G error rates alone provided sufficient prediction performance in a leave-one-out cross validation for the data sets used in this study. Consequently, we used the linear regression model based on these two features here (see Fig. 5.4D), but our implementation can handle any linear regression model or, alternatively, estimates from spike-in RNA.

5.4.3 Estimating p_c

To estimate p_c, we first compute $A = (a_{k,n}) = \sum_g A^{(g)}$. Since $y = 0$ is the mode of both component distributions of the binomial mixture model, standard approaches to estimate p_e and p_c based on the expectation maximization (EM) algorithm failed. However, as we can assume that $p_e < p_c$, there is a certain k where only a minor fraction of reads with at least k T to C mismatches originates from the p_e component. This k can be computed for each n such that less than 1% of the observed reads with $\geq k$ mismatches is expected to

originate from unlabeled RNA. Thus, for each n and k we compute

$$e_{k,n} = B(k; n, p_e) \cdot \sum_{k'} a_{k',n} \tag{5.4}$$

and exclude (k, n) if $e_{k,n} > 0.01 a_{k,n}$. More than 99% of the remaining $a_{k,n}$ originate from the p_c component, allowing to estimate p_c using an EM algorithm that treats the excluded $X = \{(k_1, n_1), ...\}$ as missing data. If enough reads (we used 10,000 reads as threshold) remain, which was the case in all data sets but the 45 min labeling experiments from [Herzog et al., 2017], p_c can be estimated with sufficient precision. Otherwise, our implementation stops with an error. Importantly, this will only happen when extremely few labeled RNA was in the sample.

The E step consists of replacing excluded read counts by their expected values given the current estimate $p_c^{(t)}$:

$$a_{k,n}^{(t+1)} = \frac{\sum_{(k',n)\notin X} B(k; n, p_c^{(t)}) \cdot a_{k',n}}{\sum_{(k',n)\notin X} B(k'; n, p_c^{(t)})} \tag{5.5}$$

The M step computes a better estimate for p_c as

$$p_c^{(t+1)} = \frac{\sum_{k,n} k a_{k,n}^{(t+1)}}{\sum_{k,n} n a_{k,n}^{(t+1)}} \tag{5.6}$$

We noticed that running the EM algorithm led to extremely slow convergence rates. Thus, we use the following bisection scheme instead: For the search interval $[l, r]$ (starting with $l = 0$ and $r = 1$), we set $p^{(t)} = \frac{l+r}{2}$ and compute $p^{(t+1)}$ by a single EM iteration. If $p^{(t+1)} < p^{(t)}$, we continue with the search interval $[l, p^{(t)}]$, otherwise with $[p^{(t)}, r]$. We stop if $r - l < 10^{-8}$.

5.4.4 Estimating the posterior

In principle, we compute the integral by dividing $[0, 1]$ into k equally sized intervals and employ Newton-Cotes quadrature using the trapezoidal rule. This also gives straightforward access to any credible interval. To allow for relatively small k even for potentially extremely narrow posterior distributions f, we first identify the mode m of f by numerically maximizing

$$g(\pi_g; \mathbf{y}, \mathbf{n}, p_e, p_c) = \sum_i \log\left(P(y_i; p_e, p_c, n, \pi_g)\right) \cdot \log b(\pi_g; \alpha, \beta) \tag{5.7}$$

Then, we identify the values $l < m$, where $f(l) = 10^{-3} f(m)$ and $h > m$ where $f(h) = 10^{-3} f(m)$ by bisection. The interval $[l, h]$ contains most of the probability mass, so we use this interval for the numerical integration.

Finally, we noticed that the posterior distribution for any gene g closely resembles a beta distribution with density b_g. Importantly, having a closed-form representation for the posterior is important for subsequent steps. Therefore we fit parameters α_g and β_g by numerically minimizing the sum of squares computed between f and b_g for all Newton-Cotes points.

5.4.5 Estimating RNA half-life

For the abundance a of an RNA with transcription rate σ and decay rate δ, the change over time is modeled by the following differential equation:

$$\frac{da}{dt} = \sigma - \delta a(t) \tag{5.8}$$

With an initial abundance a_0, this has the following closed-form solution:

$$a(t) = \left(a_0 - \frac{\sigma}{\delta}\right) e^{-t\delta} + \frac{\sigma}{\delta} \tag{5.9}$$

Setting the initial abundance to zero for newly synthesized RNA and to the steady state for pre-existing RNA, we obtain the following functions for the abundance of new RNA a_{new} and old RNA a_{pre}:

$$a_{new}(t) = -\frac{\sigma}{\delta} e^{-t\delta} + \frac{\sigma}{\delta} \tag{5.10}$$

$$a_{pre}(t) = \frac{\sigma}{\delta} e^{-t\delta} \tag{5.11}$$

Thus, at any time t, the proportion of new to old RNA is

$$\pi(t) = \frac{a_{new}(t)}{a_{new}(t) + a_{pre}(t)} = 1 - e^{-t\delta} \tag{5.12}$$

This can be used to transform the decay rate into a proportion π at time t and vice-versa:

$$\delta_t(\pi) = -\frac{1}{t} \log(1 - \pi) \tag{5.13}$$

$$\pi_t(\delta) = 1 - e^{-t\delta} \tag{5.14}$$

Hence, for gene g if at any time t, the proportion of new and old RNA is an approximately beta distributed random variable $\mathcal{P}^{(t)} \sim Beta(\alpha, \beta)$ with density function $b_g(\pi; \alpha, \beta)$, the density function $d(\delta; \alpha, \beta)$ of the distribution of the transformed random variable $\mathcal{D}^{(t)} = \delta_t(\mathcal{P}^{(t)})$ can be found by substitution:

$$d(\delta; \alpha, \beta) = b_g\left(\pi_t(\delta); \alpha, \beta\right) \frac{d\pi_t}{d\delta} \tag{5.15}$$

$$= \frac{t}{B(\alpha, \beta)} \left(1 - e^{-t\delta}\right)^{\alpha - 1} \cdot e^{-t\beta\delta} \tag{5.16}$$

Thus, if several approximate posterior beta densities defined by $(\alpha_1, \beta_1), ..., (\alpha_n, \beta_n)$ for proportion parameters measured at times $t_1, ..., t_n$ are given, the maximum a posteriori estimator for the decay rate δ can be found by numerically maximizing:

$$l(\delta) = \sum_i (\alpha_i - 1) \log(1 - e^{-t_i \delta}) - t_i \beta_i \delta \tag{5.17}$$

Finally, the estimated decay rate δ can be transformed into an estimate of the RNA half-life λ by

$$\lambda = \frac{\log(2)}{\delta} \tag{5.18}$$

5.4.6 Simulation

We utilized the available SLAM-seq data from [Herzog et al., 2017] to determine realistic parameters for simulation. Specifically, we downloaded the processed table of a random sample (GSM2666852) from GEO and converted the *CPM* (read counts per million) into a read count distribution for genes by multiplying all CPM values by 20 million. Next, we downloaded the table containing half-lives estimated from their pulse-chase experiment and applied equations 5.18 and 5.14 to derive a realistic distribution of new to old proportions for a putative experiment with $3h$ 4sU labeling.

Data for Fig. 5.2 were simulated by the following procedure: We simulated as many genes as in the read count distribution. For each gene, we randomly sampled a read count from this distribution and a π_g from the proportion distribution (except for Fig. 5.2B, where we set $\pi_g = 0.5$ for all genes). For each read, we first sampled the total number of thymines n from a binomial distribution with parameters L (read length, here $L = 50$ as in the available experiments) and u (thymine content, here we set $u = 0.3$ as computed from the 3' end investigated by [Herzog et al., 2017]. Then, we determined whether this read originated from a new RNA (with probability π_g) or old RNA (with probability $1 - \pi_g$). Finally, the number of T to C mismatches k was drawn from a binomial distribution with parameters n and either p_e or p_c (here we set $p_e = 1 \times 10^{-4}$ and $p_c = 0.023$, compare Fig. 5.4).

Reads for Fig. 5.3 were generated in a similar manner, but here we directly selected a random read location from the gene 3' regions defined by [Herzog et al., 2017] and generated mismatches accordingly (all twelve possible mismatches with rate p_e or p_e when appropriate). Here, a fixed $\pi_g = 0.2$ was used. Read locations were either directly written to a read mapping file, or sequences were generated and written to fastq files.

5.4.7 Read mapping

To map simulated reads or available SLAM-seq data we used STAR 2.5.3a [Dobin et al., 2013] with default parameters against a reference genome prepared from the murine gen-

omic sequence and gene annotation from Ensembl version 90. We also mapped the simulated reads using NGM [Sedlazeck et al., 2013], which is utilized by SLAM-DUNK [Herzog et al., 2017] and can be parameterized specifically for SLAM-seq samples. For NGM we used the same parameters as used by SLAM-DUNK with the exception that we had to increase the gap penalty parameters since GRAND-SLAM was not able to handle the format how Indels were reported by NGM. Of note, for the simulated data there were no true Indels. We handled multimapping reads by fractional counts (e.g. if a read maps to three locations on the genome equally well, there is 1/3 of a read at each location).

5.5 Results

5.5.1 GRAND-SLAM

Metabolic labeling followed by RNA-seq in principle allows to quantify both pre-existing (i.e. before labeling) and newly transcribed RNA. In the SLAM-seq protocol [Herzog et al., 2017], RNA is labeled using 4-thiouridine, which is converted into a cytosin analog using iodoacetamide (IAA). Thus, libraries can readily be prepared for sequencing, and pre-existing and newly transcribed RNA can be distinguished based on observed T to C mismatches of reads mapped to the reference genome.

However, it is not possible to determine with certainty, whether an observed read originated from a newly transcribed or pre-existing RNA molecule: Sequencing errors produce T to C mismatches also on reads from old RNA, and because of relatively infrequent 4sU incorporations ($\sim 2\%$ of all uridines are replaced [Dölken et al., 2008, Herzog et al., 2017]), a substantial fraction of reads from new RNA will not have a T to C mismatch. Of note, only a minority will contain more than one T to C mismatch. Nevertheless, based on all reads mapped to a gene, it is possible to statistically infer the proportion of new and old RNA.

To this end, we developed a statistical model based on a binomial mixture model (see Figure 5.1). We assume that the number of observed T to C mismatches for a read is generated by one of two binomial distributions. One corresponds to old RNA and its success probability parameter is the average T to C error rate. The other models new RNA and its parameter is the combined error and incorporation rate. Naturally, the mixture parameter of the model corresponds to the proportion of new and old RNA.

We are not only interested in computing a point estimate of the proportion, but similarly to our previous work [Erhard and Zimmer, 2015], we also compute the posterior distribution on this parameter. This is of great interest here, as the accuracy of the estimator greatly depends on the number of reads mapped to a gene, and the difference between the conversion and error rates. Thereby, the size of credible intervals provide a potent quality measure for SLAM-seq experiments.

5.5.2 Validation by simulation

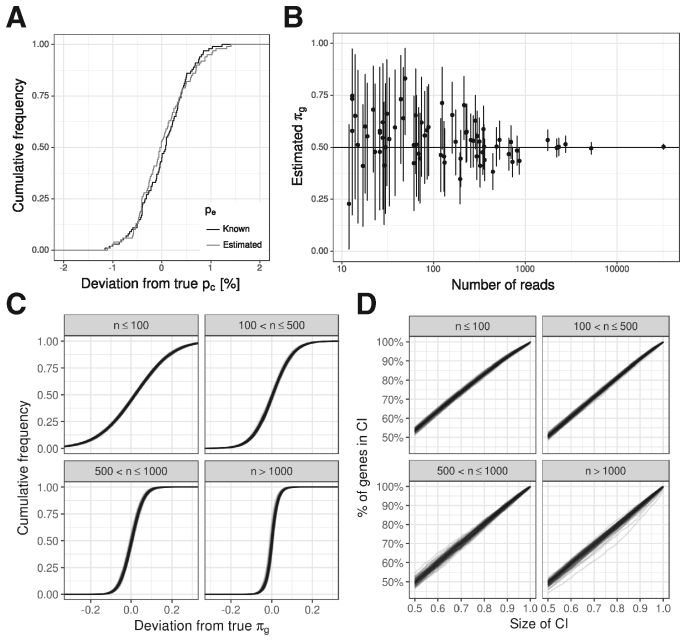


Figure 5.2: Validation by simulation. (**A**) Estimation accuracy for the conversion rate p_c is shown as the deviation of the estimated value from the true value in percentage of the true value. The error rate p_e must be known to estimate p_c. Either the true error rate (Known p_e) is supplied to the algorithm, or the true error rate plus a normally distributed error (according to parameters inferred from the *no4sU* experiments from [Herzog et al., 2017], see Fig. 5.4C; Estimated p_e) (**B**) 90% credible intervals and the posterior means for the proportion parameter π_g are shown for 70 randomly sampled simulated genes. Here, the true p_c and p_e have been supplied to estimate. (**C**) Cumulative distributions of the absolute deviation from the true proportion are shown for all 18.917 simulated genes split by their read count n. Distributions for 100 simulations are overlayed. (**D**) The percentage of genes within equal-tailed credible intervals (CI; x axis) is shown for all 18.917 simulated genes split by their read count n. As in (C) 100 simulations are overlayed.

The first step of our method is to estimate the conversion and error rate parameters p_c and p_e. Both may vary between samples, but we assume them to be constant for all genes from a single sample. Therefore, we use all reads from a sample to estimate p_e and p_c. Because both probabilities are relatively small, standard techniques for estimation on the binomial mixture model failed. However, if p_e is known, it is possible to estimate p_c by an EM algorithm (see Methods for details). Estimating p_e is more problematic, as it depends on an accurate estimate for π (the overall proportion of new and old RNA in the sample), which, in turn, depends on accurate estimates for p_c and p_e. Again, standard techniques based on EM algorithms failed. However, in principle, p_e can be experimentally determined by spiking-in unlabeled RNA before IAA treatment. Alternatively, p_e can be measured in additional experiments without 4sU labeling (no4sU sample). The problem

with the approach based on no4sU samples is that measurements vary between samples and an externally measured value may not be accurate enough for precisely estimating p_c and π_g (the proportion of new and old RNA for gene g) for each gene g. However, we noticed that the twelve different error rates were highly correlated between samples. Thus, T to C error rates can be estimated from the other error rates, which are measured in SLAM-seq experiments (see Methods for details).

Thus, our first check was how accurately p_c could be estimated if p_e is known (e.g. measured by RNA spike-ins) or if p_e is estimated using additional no4sU samples. To this end, we simulated a hundred data sets with realistic values of p_e and p_c. Then, we either supplied the true p_e for estimating p_c, or a slightly deviating p_e (based on observed deviations in the no4sU data sets from [Herzog et al., 2017]). The estimates of p_c were highly accurate (less than 1% deviation; see Figure 5.2A). Importantly, this was the case when the true p_e was used and when a slightly deviating p_e was used.

Next, we tested how well the individual gene proportions π_g could be estimated when p_c and p_e are known. Estimates were not biased, and always within the expected bounds given by credible intervals (see Fig. 5.2B). Finally, we expanded our simulations on a realistic scenario, i.e. p_c and p_e were estimated for simulated data, and then the π_g were estimated based on p_c and p_e. Again, estimates were not biased, and especially for genes with many reads, highly accurate (less than 0.05 absolute deviation; see Fig. 5.2C). Moreover, the number of genes within any credible interval exactly matched the expected number in all cases. This means that observed deviations are not due to errors in the process of estimation, but are because of insufficient data. Thus, computed credible intervals provide a potent mean to judge the quality of a data set and the estimates for all genes.

5.5.3 Influence of read mapping

So far, we directly simulated numbers (k_i, n_i), i.e. k_i T to C mismatches were observed for n_i thymines in read i. Even if read mapping has high sensitivity and specificity in finding the right location for all reads, correct read mapping is crucial especially for reads with one or more T to C mismatches. In [Herzog et al., 2017], the authors extended their own read mapping software NGM [Sedlazeck et al., 2013] specifically for the purpose of mapping SLAM-seq reads. Therefore, by generating sequencing reads in-silico, we tested whether read mapping by a standard tool (STAR; [Dobin et al., 2013]) or NGM affected our method.

First, we compared how well error and conversion rates could be estimated when read locations were directly written into read mapping files or mapped with STAR or NGM. Interestingly, read mapping resulted in significantly reduced estimates for both p_e and p_c (see Fig. 5.3A and B), indicating that indeed a substantial number of reads with simulated mismatches was either not mapped at all, mapped to more than one location or mapped to a wrong location. Of note, STAR and NGM read mappings were affected by this to a highly similar degree. However, this does neither introduce bias into estimating gene proportions π_g (see Fig. 5.3C), nor does it affect the size of credible intervals (see Fig. 5.3D). In

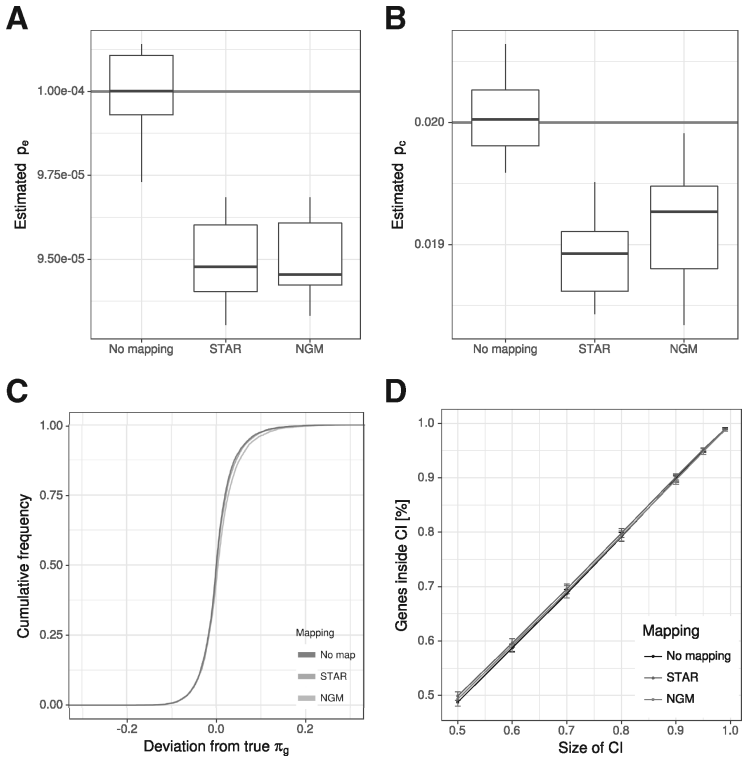


Figure 5.3: Influence of read mapping. (**A** and **B**) We simulated ten data sets of reads and either used the true locations of the reads (*No mapping*) as input for GRAND-SLAM, or a fastq file for STAR or NGM. Here, the distributions of the estimated error rates (A) and conversion rates (B) are shown. The true values are indicated. Read mapping with both STAR and NGM led to slightly, but significantly biased estimates. (**C**) The cumulative distribution of the absolute deviation from the true proportion is shown for reliably quantified genes (at least 100 reads). In spite of underestimated rates, read mapping effects on estimating the proportion are negligible. (**D**) The percentage of genes within equal-tailed credible intervals (CI; x axis) is shown. Read mapping does not affect the accuracy of credible intervals. Error bars indicate the standard deviation of the ten simulations.

summary, there is room for improving read mapping for SLAM-seq, but our method is robust enough to handle reads mapped even by widely used standard read mapping tools.

5.5.4 Evaluation of mESC datasets

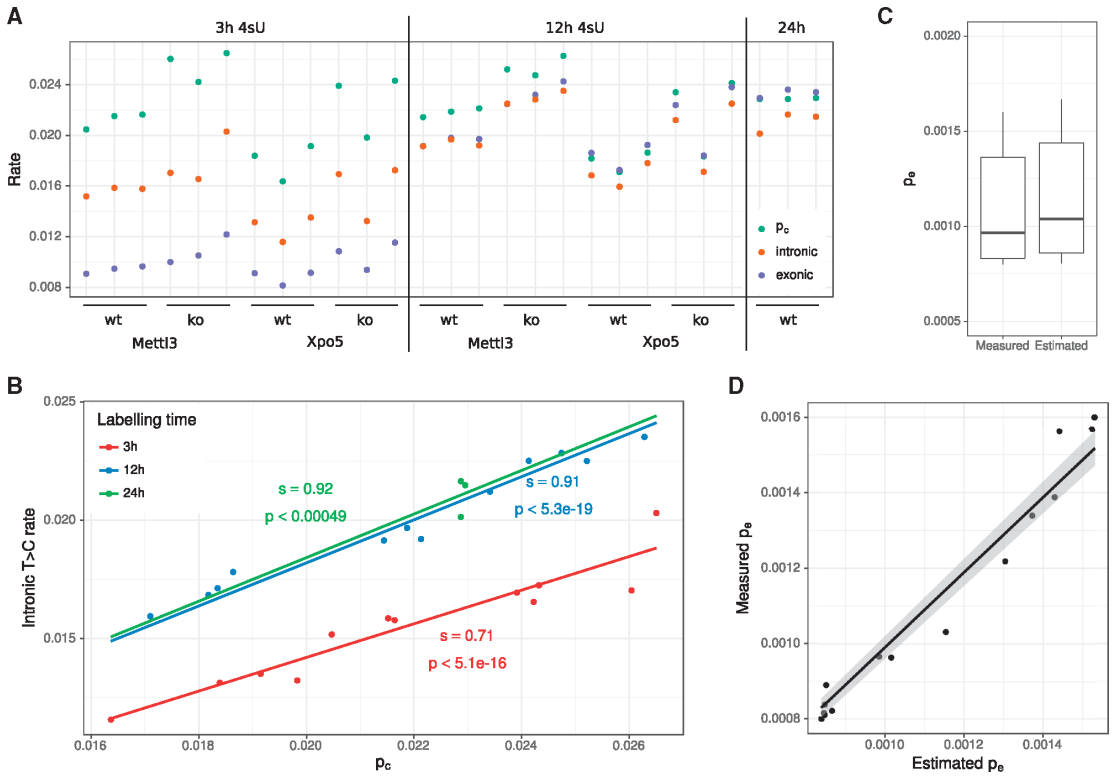


Figure 5.4: Evaluation of mESC data. (**A**) For all SLAM-seq experiments from [Herzog et al., 2017], the estimated conversion rate p_c is compared to the intronic and exonic T to C mismatch rates. (**B**) Linear regression analysis of p_c against the intronic T to C mismatch rate. Slopes (s) and p values are indicated. For all three regressions $r^2 > 0.99$. (**C**) The distribution of the error rate p_e as measured in the 15 *no4sU* samples is compared to the estimated error rates in the 27 4sU samples (see (A)). (**D**) p_e can be predicted by linear regression of the other error rates. In the *no4sU* samples, p_e can be directly measured by counting T to C mismatches. This shows the results of a leave-one-out cross validation in the no4sU samples comparing the predictions (x axis) against the measured values (y axis).

[Herzog et al., 2017] conducted several SLAM-seq experiments on murine embryonic stem cells (mESCs) with different periods of labeling (45 minutes, 3 hours, 12 hours and 24 hours). We examined the conversion and error rates p_c and p_e estimated by GRAND-SLAM for each of these experiments. For the 45min experiments, p_c could not be estimated because too few reads had more than one T to C conversion. In such a case,

our implementation prints a warning. Thus, we excluded these experiments from further analyses.

We first compared the estimated conversion rate with the observed T to C mismatch rates from exonic and intronic reads for all samples (see Fig. 5.4A). Estimated conversion rates were spread around slightly above 0.02 in all cases, and were not correlated to the period of labeling. Especially for the 3h samples, both exonic and intronic mismatch rates were substantially lower than the estimated conversion rates and were correlated to the period of labeling. For exons, this was expected since a substantial fraction of the total mature RNA is older than 3h. Interestingly, albeit to a lesser extent, we also observed this for intronic RNA, which is believed to be quickly degraded after splicing [Windhager et al., 2012]. The fact that the T to C mismatch rate is significantly higher after 12h of labeling than after 3h of labeling is indicative for frequent intron retention, or that at least some introns are relatively long-lived.

Intronic RNA was excluded from estimating conversion rates, but there was nevertheless a high correlation ($r^2 > 0.99$) of intronic T to C mismatch rates with estimated conversion rates. This indicates that conversion rates were estimated very accurately, and that a certain amount of intronic RNA is older than 3, 12 or 24 hours. Regression analysis revealed these amounts to be 70%, 91% and 92% in mESCSs, respectively (see Fig. 5.4B).

In [Herzog et al., 2017], 15 samples have also been measured without 4sU labeling. For these all RNA is by definition old, and the mixture model reduces to a model with a single binomial component. Thus, p_e is directly measured in these samples. Interestingly, the measured p_e varied between 0.8×10^{-3} and 1.6×10^{-3} (see Fig. 5.4C). Thus, taking such a measured p_e for another sample where the sample specific p_e is some value within this range can lead to biased estimates of the new to old proportions π_g. This can be circumvented by either directly measuring p_e in each sample using RNA spike-ins, or by employing a linear regression based estimation of p_e: We noticed that between the no4sU samples other error rates (e.g. T to A) were highly correlated to T to C error rates. Thus, we trained a linear regression model in the no4sU samples to estimate T to C error rates. Of note, estimated T to C error rates from the samples with 4sU were in the same range as observed error rates in the no4sU samples (see Fig. 5.4C), and the T to C error rates in the no4sU samples could be predicted with high accuracy (see Fig. 5.4D).

5.5.5 Estimating RNA half-life

The proportion π_g of new and old RNA after some period of labeling t can be transformed into the RNA half-life λ_g (see Fig. 5.5A and Methods for details). The functions f_t transforming π_g into λ_g vary greatly for different values of t. Naturally, very short labeling periods (e.g. $t = 1/2h$) can resolve short RNA half-lives (e.g. $\lambda_g < 1h$) very accurately, but small differences in π_g result in large deviations of λ_g for genes with long half-life (see Fig. 5.5B).

To analyze the variance in estimating RNA half-lives using GRAND-SLAM, we first theoretically considered an experiment with typical parameters as observed in the data sets

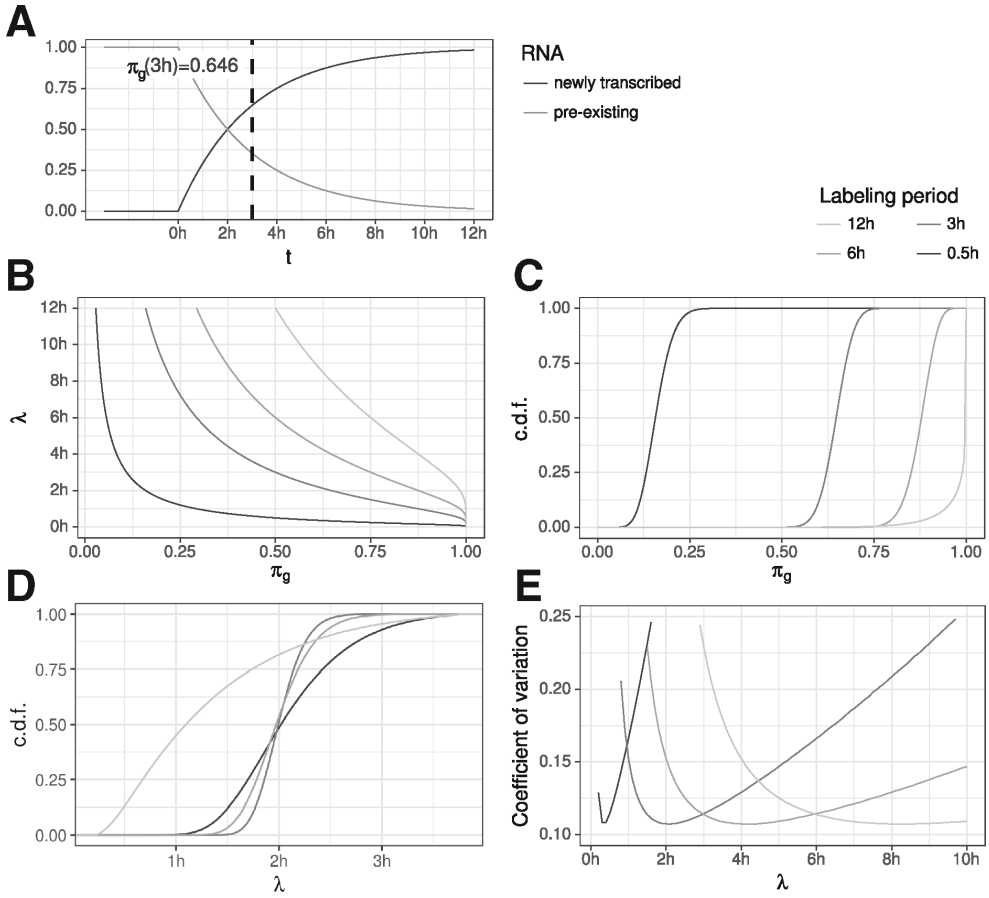


Figure 5.5: RNA half-life (**A**) The proportion of new and old RNA of a gene g at any time t, $\pi_g(t)$, is directly related to its RNA half-life (here, $2h$). (**B**) The functions are shown that transform the proportion π_g to the RNA half-life for different periods of labeling (see common legend of subfigures B to E at top right corner). (**C**) Posterior distributions of a theoretical gene g with 1000 reads and an RNA half-life of $2h$ for the four different periods of labeling. (**D**) These posterior distributions for π_g translate to specific posterior distributions on λ, with the one for $t = 3h$ being the most precise one. (**E**) Coefficients of variation (standard deviation divided by the mean) of the posterior distributions on λ for theoretical genes with RNA half-lives between 0 and $10h$.

of [Herzog et al., 2017], and a gene g with 1000 reads with a half-life of $\lambda_g = 2h$. For different labeling periods, this gives rise to specific posterior distributions on the proportion parameter π_g (see Fig. 5.5C) which can be transformed into posterior distributions on the estimated RNA half-life (see Fig. 5.5D). Interestingly, the estimate due to 3h labeling is the most precise, followed by 6h, 0.5h and 12h. We expanded this analysis to genes with different RNA half-lives, and computed the coefficient of variation (CV) of the posterior distribution of the estimated half-lives (see Fig. 5.5E). The CV is the standard deviation divided by the mean and therefore describes the expected relative deviation. The CV varied greatly depending on the labeling period, with short labeling periods generally most precise for genes with short RNA half-life. In addition, each labeling period has a range of true RNA half-lives where it is most precise and it extremely imprecise for too long or short-lived genes. E.g. with 3h labeling, estimation precision deteriorates for genes with a half-life below half an hour or longer than 8h. Thus, to precisely estimate the whole range of RNA half-lives in an experiment, several samples with different labeling periods are necessary as well as a method that automatically weighs the contributions of each sample to the overall estimate based on the varying variances. This can be achieved by maximum a posteriori (MAP) estimation of the RNA decay rate (see Methods for details).

5.5.6 RNA half-lives for mESCs

[Herzog et al., 2017] estimated RNA half-lives by pulse-chase experiments: To achieve sufficient labeling, cells were supplied with 4sU for 24h. After that 4sU was washed out and the drop of conversions was monitored for several time points via SLAM-seq. RNA half-lives were then estimated by fitting an exponential decay model using least squares. These experiment are relatively laborious and introduce the wash-out efficiency as an additional source for potential bias. Furthermore, the least squares fitting does not respect the varying precision of estimating different half-lives with different labeling periods. For comparison, RNA half-lives were also determined using actinomycin D (ActD) treatment and monitoring the drop of RNA levels over time using RNA-seq.

In addition to the pulse-chase and ActD estimates, we used the 3h or 12h labeling data or their combination to estimate RNA half-lives for mESCs using our maximum a posteriori approach ($MAP_3, MAP_{12}, MAP_{comb}$). The correlation coefficients computed over all genes also utilized for comparison in [Herzog et al., 2017] showed that MAP_{comb} performed equally well ($R \approx 0.7$) as the pulse-chase experiments in reproducing the ActD estimates (see Fig. 5.6A). MAP_3 resulted in a similarly high correlation but the MAP_{12} estimates showed worse correlation ($R \approx 0.46$). For genes with ultra-short ($< 2h$) and short ($< 3h$) RNA half-lives however, the correlation of the pulse-chase experiment was poor ($R \approx 0$ and $R \approx 0.26$, respectively), and significantly better for MAP_3 and MAP_{comb} ($R > 0.59$ and $R > 0.49$). For genes with longer RNA half-lives, correlations were generally poor, but MAP_{12} provided the highest correlations (see Fig. 5.6D).

Furthermore, MAP_{comb} always resulted in correlation coefficients (computed for the comparison to the pulse-chase or the ActD experiment) that were close to the better of MAP_3

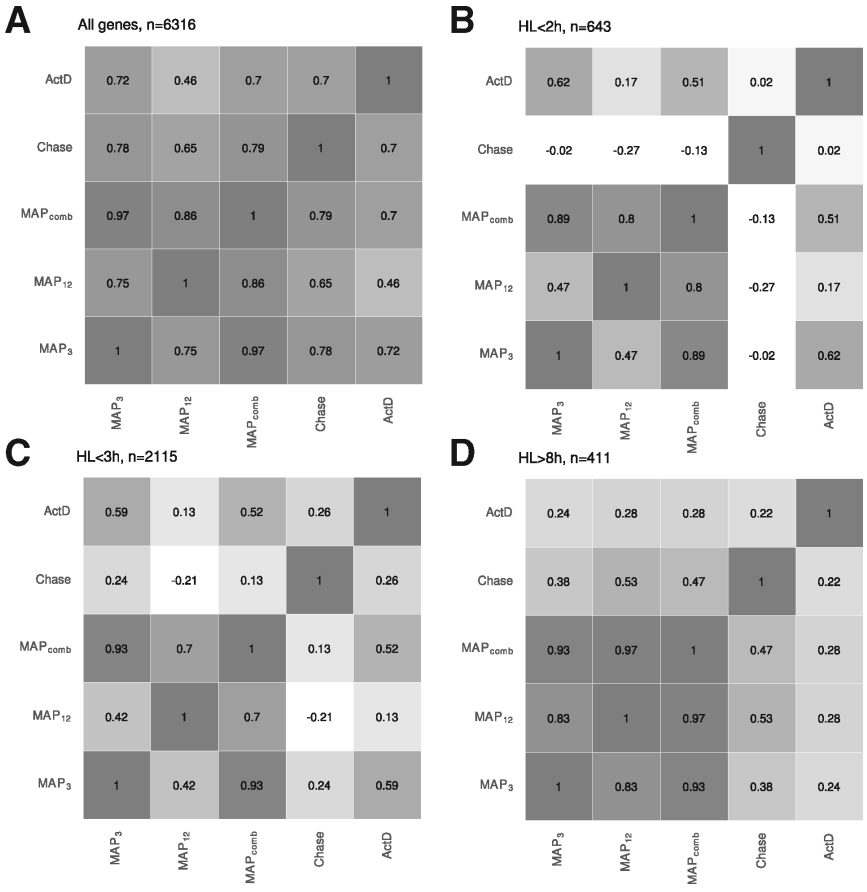


Figure 5.6: Pearson's correlation coefficient for RNA half-lifes (**A**) For $n = 6.316$ genes, the correlation between any pair of methods estimating RNA half-lives was computed. MAP_3, MAP_{12} and MAP_{comb} are the maximum a posterior estimators of GRAND-SLAM computed on the $3h$, $12h$ samples or both. Chase is the exponential decay model fit of [Herzog et al., 2017] on the pulse-chase experiments. ActD is the exponential decay model fit for the actinomycin D experiment.(**B-D**) Correlation coefficients for different subsets of genes split according to ultra-short RNA half-life, short RNA half-life and long RNA half-life.

or MAP_{12}. This indicates that the MAP estimation of the RNA half-live effectively weighs the different precisions obtained for measuring with different labeling periods.

5.5.7 Differential analysis of RNA half-life changes

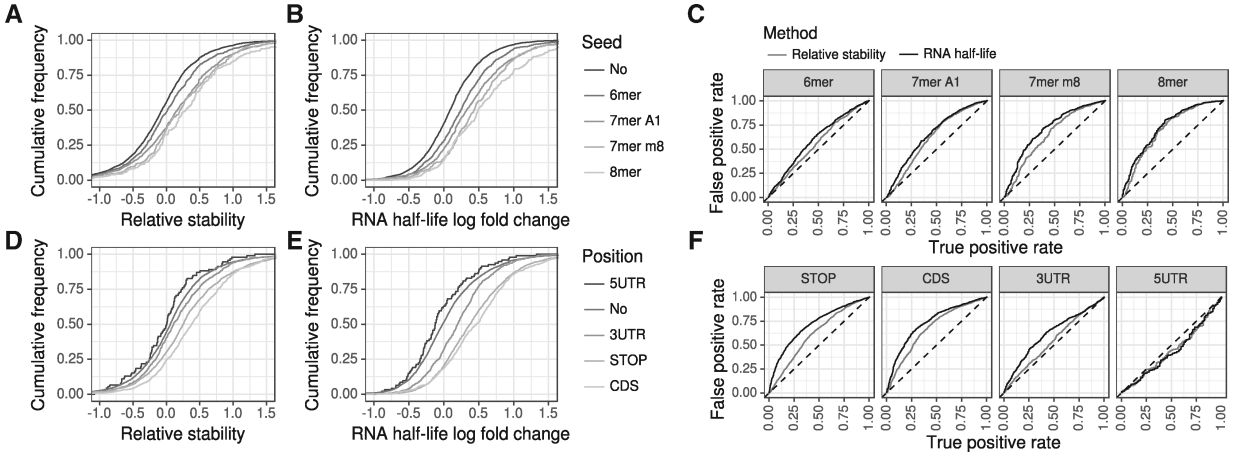


Figure 5.7: Differential analysis (**A-B**) We repeated the microRNA target prediction analysis of [Herzog et al., 2017]. Relative stability values (A) are computed from the corrected conversion counts from the $3h$ and $12h$ experiments, RNA half-life $\log_2$ fold changes using GRAND-SLAM (B). The RNA half-lives show a stronger enrichment of targets upon Xpo5 knock-out for all seed types. (C) We performed ROC analyses by treating predicted microRNA targets as the *true* objects, and mRNAs without seed as the *false* objects. Then, either the relative stability or RNA half-life $\log_2$ fold change was taken as prediction score. MicroRNA target predictions agreed better with RNA half-lives than with relative stabilities for all four seed types. (**D-F**) We repeated the m6A modification analyses from [Herzog et al., 2017]. As expected, no enrichment upon Mettl3 knock-out was found for mRNAs with m6A in the 5' UTR. For all other mRNA locations defined in [Batista et al., 2014], the RNA half-lives show a substantially stronger enrichment of m6A containing mRNAs.

Several factors are known to affect the stability of specific RNAs. Most prominently, microRNAs are small RNAs expressed by virtually all eukaryotic cells, that (imperfectly) basepair to cognate sites in mRNAs. Thereby, they guide the RNA induced silencing complex (RISC) to target mRNAs, which leads to translational repression and induces RNA decay [Bartel, 2009, Jonas and Izaurralde, 2015]. By knocking-out Exportin-5 (Xpo5), an essential factor in the biogenesis of the most abundant family of microRNAs in mESCs (miR-291a), repression of mRNA targets of this family is reduced, effectively prolongating their RNA half-life.

In [Herzog et al., 2017], this has been analyzed by considering the *relative RNA stability* computed by comparing the (no4sU corrected) T to C mismatch rates for wild-type (wt) and Xpo5 knock-out (ko) cells. This indeed revealed different sets of predicted microRNA

targets to have increased RNA stabilities (see Fig. 5.7A). Using GRAND-SLAM we were able to compute RNA half-lives for both conditions (wt and ko) and compute their $\log_2$ fold changes (see Fig. 5.7B). Comparing the distributions of RNA stabilities with RNA half-live $\log_2$ fold changes already indicates that the latter is a better measure to capture the action of the microRNAs. Indeed, receiver operating characteristic (ROC) analysis revealed RNA half-life $\log_2$ fold change to better agree with predicted microRNA targets than relative RNA stability. However, the overall difference between genes predicted to be a microRNA target and the remaining genes is generally poor, presumably due to secondary effects of knocking down Xpo5 or the difficulty in predicting microRNA targets [Ritchie et al., 2009].

Another cellular mechanism affecting RNA stability is N^6 adenosine methylation (m6A) [Meyer and Jaffrey, 2014, Yue et al., 2015]. It has been shown that m6A at specific mRNA locations induces mRNA degradation [Wang et al., 2014]. m6A modification is performed by the protein complex N6-adenosine-methyltransferase. Thus, by knocking out its 70 kDa subunit (Mettl3), genes affected by m6A mediated degradation that have been experimentally determined in mESCs [Batista et al., 2014] are de-repressed. Similarly to the microRNA analyses, relative RNA stabilities can reveal this (see Fig. 5.7A), but RNA half-lives computed by GRAND-SLAM reveal substantially more differences between targets and non-targets (see Fig. 5.7B and C).

5.6 Discussion

The most successful experimental technique to discriminate between newly transcribed and old RNA is based on metabolic labeling of RNA. To this end, non-toxic nucleoside analogs are introduced into living cells, which are then readily incorporated into newly transcribed RNAs. Previously, before measurement using microarrays or next generation sequencing labeled and unlabeled RNA was separated via thiol-specific biotinylation [Dölken et al., 2008, Rabani et al., 2014]. The novel approach published recently [Herzog et al., 2017, Schofield et al., 2018, Riml et al., 2017] replaced this biochemical separation with a bioinformatic separation: Nucleoside analogs are chemically converted into distinct nucleoside types and therefore in principle distinguishable based on observed mismatches. However, with incorporation rates of $\sim$2%, the discrimination between labeled and unlabeled RNA is highly challenging.

Here, we presented a statistical method to precisely estimate the proportion of new and old RNA in such experiments. This is based on a binomial mixture model, where the number of observed, experiment-specific mismatches is generated from one of two binomial distributions for reads from labeled and unlabeled RNA molecules. The output of our method is the full posterior distribution of the proportion of new and old RNA. This posterior is narrow for genes with many reads and for experiments with high incorporation and low sequencing error rates. Thus, it provides a straight-forward mean for quality control.

In addition to sufficient incorporation rates that must be achieved, additional considera-

tions for planning such experiments are important: [Herzog et al., 2017] used single-end sequencing with read length 50bp on a Illumina HiSeq 2500. With the four-color chemistry, the HiSeq 2500 provides the smallest error rates possible today. Sequencing devices with two-color chemistry should be avoided due to significantly greater error rates ($\sim$ 10x). Longer reads are generally preferable, as the probability of catching a modified nucleotide increases with longer reads.

Furthermore, paired-end sequencing would provide two significant advantages over single-end reads: First, error rates can be estimated from the other read, since 4sU converted to cytosine results in a T to C mismatch in the first read, and in an A to G mismatch in the second read. Second, especially if RNA is strongly fragmented, read pairs overlap. All nucleotides in the overlapping part are sequenced twice, making the differentiation of true conversions from sequencing errors much easier: The probability for two independent sequencing errors of the same nucleotide is negligible small. I.e. in such situations, it is possible to decide with almost certainty that a read pair originated from a newly transcribed RNA molecule.

The estimation of sample specific error rates is a crucial component of our method. Here, this has been solved by the observation that error rates were correlated, which we could exploit by a linear regression model. The model was trained on available samples that have not been treated with 4sU, and could predict T to C error rates with sufficient accuracy. A potent alternative would be to use RNA spike-ins such as the ERCC mix [Jiang et al., 2011] in each sample. This way, error rates could be directly estimated by counting mismatches on the ERCC RNAs.

We have shown that the estimated proportions of new and old RNA can be used to compute precise RNA half-lives. Importantly, this (and all other estimates of RNA half-life based on metabolic labeling) heavily relies on the incorporation rate of the nucleoside analog to be constant over time. Considering that they have to cross cell membranes, the cytoplasm and the nuclear membrane to increase their concentration in the nucleus, we expect this assumption to be problematic. 4sU needs time to accumulate, and methods are needed to measure this effectively reduced time of labeling to be considered in estimating RNA half-lives.

We uncovered that using SLAM-seq, short half-lives can be resolved more precisely with short periods of labeling. For such it is difficult to achieve enough 4sU incorporation for our method to estimate the conversion rate. Thus, labeling periods and 4sU concentrations should be carefully tested, potentially in a cell type specific manner.

5.7 Conclusion

SLAM-seq experiments provide an exciting new technique to access newly transcribed RNA for obtaining RNA half-lives or investigating fast and complex regulatory processes. However, tailored computational analyses approaches for such high-throughput experiments are an essential factor for the success of any study employing SLAM-seq. Here, we provide the

first statistical method that is able to precisely delineate the quantities of newly transcribed RNA for each gene and discriminate it from pre-existing RNA before labeling.

Chapter 6

Integrative multi-omics reveals principles of gene regulation and pervasive transcription of transient RNAs in the human cytomegalovirus genome

Motivation: *In the previous chapters I introduced my two tools iTiSS and GRAND-SLAM. With GRAND-SLAM, we are able to estimate NTRs for genes in SLAM-seq experiments and consequently are able to accurately differentiate between actively transcribed genes and genes that are currently switched off independently of the total RNA levels. However, in HCMV, multiple transcripts overlap each other, i.e. they use the same Poly(A)-site with different TiSS. This makes it impossible to accurately estimate NTRs for each individual transcript, which, however, is important, as they might employ different kinetics. Hence, we came up with a new sequencing protocol, combining the TiSS-profiling approach dRNA-seq and the metabolic labeling approach SLAM-seq. As for HCMV a transcriptome annotation is completely missing, I had to combine my two tools iTiSS and GRAND-SLAM, which, in turn, would provide us with accurate NTRs on a per transcript level. With this technique we set out to answer the question on how HCMV regulates its genes throughout the course of the infection on a transcriptomic level.*

Publication: *A shorter version of this article has been submitted for publication in Nature Microbiology. This longer version is available in similar form on BioRxiv.*

Individual author contributions: *See Appendix D*

6.1 Abstract

Human cytomegalovirus was initially thought to express about 200 viral proteins that are expressed in a cascade of immediate early, early and late genes. Since then, systems biology approaches uncovered hundreds of additional viral proteins and microproteins and suggested thousands of viral sites of transcription initiation. Despite all available data, the connection between transcription and protein expression and how viral genes are regulated remains poorly understood. Here, we decipher the regulation of lytic HCMV gene expression in primary fibroblasts employing transcription start site profiling combined with metabolic RNA labeling as well as integrative computational analysis of previously published big data. This confirmed the expression of >2,600 viral transcripts and explained the complex kinetics of viral protein expression by cumulative effects of multiple transcription start sites per viral open reading frame, translation of incoming virion-associated RNA and differences in viral protein stability. Finally, we reveal pervasive non-productive transcription within the HCMV genome. Our findings explain conflicting results of previous studies and provide a unified model of HCMV gene expression.

6.2 Introduction

Human Cytomegalovirus (HCMV) is a prevalent member of the herpesvirus family and responsible for life-threatening disease in the immunocompromised. It is also the most common congenitally transmitted virus, affecting about 0.2% to 2.2% of all newborns [Yinon et al., 2010, Stagno et al., 1986, Pultoo et al., 2000]. A substantial part of the 236 kb HCMV genome is devoted to reprogram its host's cells and subvert intrinsic and adaptive host immunity [Griffiths and Reeves, 2021]. About a decade ago, HCMV was thought to encode roughly 200 gene products [Davison et al., 2003, Murphy et al., 2003]. In a ground-breaking study from 2012, hundreds of novel viral gene products were identified using ribosome profiling (Ribo-seq) [Stern-Ginossar et al., 2012]. Improved computational analysis of the respective Ribo-seq data subsequently refined the viral translatome and identified hundreds of additional viral open reading frames (ORFs) raising the number of HCMV ORFs translated during lytic infection of fibroblasts to ¿1,000 [Erhard et al., 2018]. Similar findings were subsequently reported for KSHV [Arias et al., 2014], EBV [Bencun et al., 2018] and HSV-1 [Whisnant et al., 2020]. The majority of the newly identified ORFs are short (sORFs) and may either encode for functional microproteins or, similar to their cellular counterparts, regulate viral gene expression by tuning translation of the larger viral ORFs [Železnjak et al., 2019, Hinnebusch et al., 2016, Starck et al., 2016, Young and Wek, 2016, Cabrera-Quio et al., 2016, Chu et al., 2015].

Like in all other herpesviruses, HCMV gene expression is regulated in a cascade of immediate early (IE), early (E) and late (L) genes. This classification is based on the use of chemical inhibitors of protein synthesis (Cycloheximide) and viral DNA replication (e.g., phosphonacetic acid (PAA) or Ganciclovir (GCV)). The viral IE2 protein serves as a transactivator and translation of IE2 is required for the transcription of early genes. Late gene

expression is dependent on viral genome replication and the activity of the viral late transcription factor (LTF; a complex of proteins encoded by UL49, UL79, UL87, UL91, UL92 and UL95 [Hiroki et al., 2011, Li et al., 2021]). LTF binds to TATT sequences in viral promoters instead of canonical TATA-boxes [Gruffat et al., 2016, Sarisky and Hayward, 1996, Malone et al., 1990]. In contrast to this functional classification, recent mass spectrometry data defined five temporal classes characterized by distinct protein expression patterns along the infection cycle [Weekes et al., 2014]. The regulatory principles behind the complex temporal kinetics beyond the IE, E and L classes of proteins remain poorly understood. Recent precision nuclear run-on of capped RNA fragments (PRO-cap) experiments identified 7,478 viral transcription start sites regions (TSR) to be active at 96 h post infection in primary fibroblasts [Parida et al., 2019]. This can only in part be attributed to the activity of the viral IE2 protein [Li et al., 2020] and LTF [Li et al., 2021]. While omics approaches thus drastically expanded the repertoire of HCMV's mRNAs and proteins, their biogenesis, kinetic regulation and function remain poorly understood.

To address these conflicting findings, we first combined two transcription start site (TSS) profiling approaches (dRNA-seq [Sharma and Vogel, 2014] and STRIPE-seq [Policastro et al., 2020]) with metabolic labelling of RNA (SLAM-seq [Herzog et al., 2017]) to obtain an accurate, detailed and time-resolved map of transcriptional activity with transcript isoform resolution in the HCMV genome. We identified 2,668 viral TSS that give rise to stable RNAs. Integrative analysis of our TSS data, PROcap, Ribo-seq and proteomics data subsequently revealed that the temporal kinetics of viral gene expression are governed by the combined effects of (i) activation of TATT box promoters before the onset of genome replication and signal amplification thereafter (ii) translation of incoming virion-associated viral mRNAs [Bresnahan and Shenk, 2000, Terhune et al., 2004], (iii) combined transcriptional output by multiple viral TSS with distinct kinetics per ORF, and (iv) differences in viral protein and RNA stability. Finally, we identify extensive pervasive transcription of the HCMV genome that does not result in stable viral mRNAs and does not contribute to the viral translatome. In summary, our study provides a unifying model of HCMV gene expression that explains many of the surprising results of previous studies and identifies pervasive transcription as a common feature of this large DNA virus and its human host.

6.3 Results

6.3.1 Establishing bona-fide TSS of stable viral transcripts

To resolve the time kinetics of HCMV's transcriptome, we infected primary human foreskin fibroblasts (HFFs) with HCMV (strain TB40E-Lisa [Tomasec et al., 2005]). After labeling newly transcribed RNA with 4-thiouridine (4sU) for one hour at multiple different timepoints post infection (Fig. 6.1A), we performed transcription start site (TSS) profiling using two different protocols (STRIPE-seq [Policastro et al., 2020] and dRNA-seq [Whisnant et al., 2020, Sharma and Vogel, 2014]). Before reverse transcription, 4sU incorporated into newly synthesized RNA was alkylated into a cytosine analog using iod-

oacetamide (IAA) [Herzog et al., 2017]. A systematic analysis of known bona-fide cellular TSS indicated excellent signal to noise ratios of on average 992 (TSS/downstream) and 4,350 (TSS/upstream) for dRNA-seq and 156 (TSS/downstream) and 179 (TSS/upstream) for STRIPE-seq, respectively (Fig. 6.1B). Replicate experiments for all time points were strongly correlated on the viral genome (R>0.75 for time points after 2h; R=0.68 for 2h; see Fig. 6.2B/C), demonstrating overall high reproducibility. To resolve the kinetics of RNA expression by our metabolic labeling strategy, sufficiently frequent (>2%) incorporation of 4sU is required [Jürges et al., 2018]. Known short-lived cellular RNAs such as MYC exhibited frequent T→C substitutions in contrast to long-lived RNAs such as GAPDH (Fig. 6.2D/E), suggesting strong incorporation of 4sU. Furthermore, GRAND-SLAM analysis revealed significantly more T→C substitutions than other substitutions (Fig. 6.1C and Fig. 6.2F/G), as well as background T→C substitutions in unlabeled RNA of <0.07%. The 4sU incorporation frequencies in labeled RNA were estimated to be >2.7% (Appendix C.1), which is sufficient to accurately quantify newly synthesized and pre-existing RNA for TSS with dozens of reads [Jürges et al., 2018]. In summary, these analyses demonstrate the high quality of our TSS profiling data.

We used our tool iTiSS [Jürges et al., 2021] to call viral TSS for each time point for both TSS profiling protocols. In line with our previous observations in the HSV-1 genome [Whisnant et al., 2020] and the human genome [Jürges et al., 2021], both dRNA-seq and STRIPE-seq resulted in substantial amounts of putative TSS that were only reproducibly observed by one protocol but not the other (Fig. 6.1D, Fig. 6.2A). 649/2,606 (24.9%) of the TSS observed by both techniques could be validated by 3rd generation sequencing data [Balázs et al., 2018], but merely 214/2,217 (9.65%) and 384/5,633 (6.82%) of TSS identified only by dRNA-seq or STRIPE-seq, respectively. Of note, due to its limited sequencing depth, 3rd generation RNA-seq reliably captures only the most strongly expressed transcripts. While we cannot exclude that at least some of the viral TSS observed only by one technique represent bona fide viral TSS, we nevertheless decided to excluded them from further analyses.

We used GRAND-SLAM [Jürges et al., 2018] to calculate new to total RNA ratios (NTRs) for each identified TSS. Global analysis of NTRs of human and HCMV TSS showed declining NTRs throughout the infection for cellular TSS (Fig. 6.2H). This was not observed for HCMV TSS. Instead, the overall number of reads mapping to HCMV TSS increased throughout the infection for both TSS profiling data sets, except for a transient peak at 6 h post infection, and a slight decline at 96 h. Late in infection, >70% of newly synthesized RNA was viral, which was reproducibly observed in STRIPE-seq as well as dRNA-seq (Fig 6.1E). To facilitate integrative analyses and visualization, we implemented and provide an HCMV genome browser containing our TSS profiling data as well as publicly available Ribo-seq, 2nd- and 3rd-generation sequencing data, and a web-based platform to visualize time courses of all data sets (see BioRxiv).

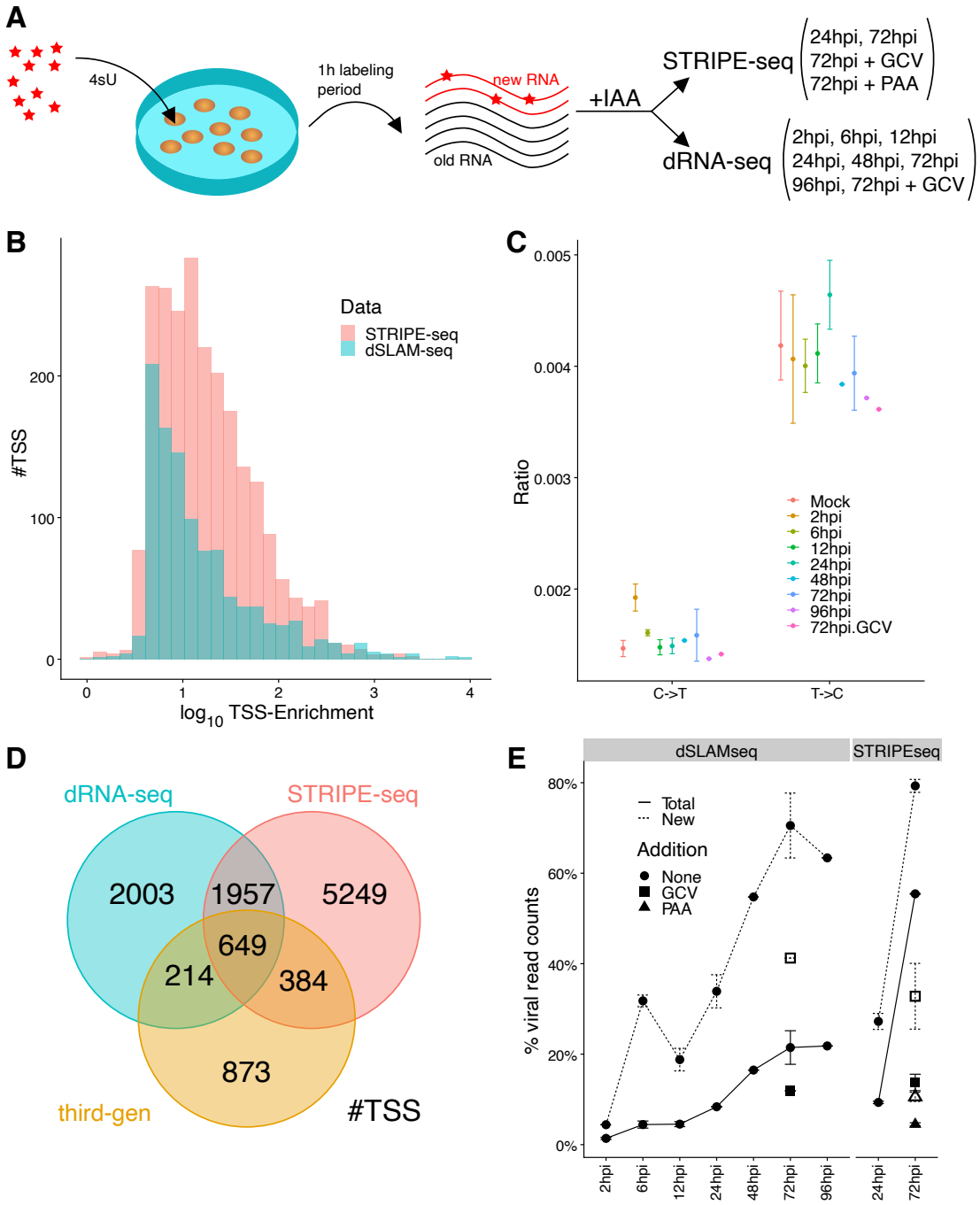

A
4sU
1h labeling period
new RNA
old RNA
+IAA
STRIPE-seq
24hpi, 72hpi
72hpi + GCV
72hpi + PAA
dRNA-seq
2hpi, 6hpi, 12hpi
24hpi, 48hpi, 72hpi
96hpi, 72hpi + GCV
B
Data
STRIPE-seq
dSLAM-seq
#TSS
log10 TSS-Enrichment
C
Ratio
Mock
2hpi
6hpi
12hpi
24hpi
48hpi
72hpi
96hpi
72hpi.GCV
C->T
T->C
D
dRNA-seq
STRIPE-seq
2003
1957
5249
649
214
384
873
third-gen
#TSS
E
dSLAMseq
STRIPEseq
Total
New
Addition
None
GCV
PAA
% viral read counts
2hpi
6hpi
12hpi
24hpi
48hpi
72hpi
96hpi

Figure 6.1: **A)** A schematic overview of our sequencing experiment, which was conducted for two independent cell samples. Cells were exposed to 4-thiouridine (4sU) for one hour prior to harvesting RNA at the indicated time points. Sequencing libraries were prepared using two different TSS-profiling approaches (STRIPE-seq [Policastro et al., 2020] and dRNA-seq [Sharma and Vogel, 2014]; one for each sample), after converting 4sU that was incorporated into newly synthesized RNA into a cytosine analog using iodoacetamide (IAA). **B)** TSS enrichment (read count at the TSS +/-1 bp divided by the read count 100 bp downstream of the TSS) at annotated cellular TSS in the dRNA-seq (red) and STRIPE-seq (blue) sample. **C)** Cytosine to Thymine (C→T) mismatches and Thymine to Cytosine (T→C) mismatches observed for each timepoint in the dRNA-seq sample. Mismatch rates were calculated using GRAND-SLAM [Jürges et al., 2018]. Error bars depict the minimum and maximum of the two replicates per timepoint if available. Points represent the respective means. See also Fig. 6.2F-G for all mismatch types. **D)** Venn-Diagram depicting the overlap of predicted TSS in dRNA-seq, STRIPE-seq and third-generation sequencing approaches using iTiSS [Jürges et al., 2021]. **E)** Relative amounts of total (solid line) or new (dotted line) viral RNA among overall total or new RNA for dRNA-seq and STRIPE-seq, respectively. Ganciclovir (GCV) and phosphonoacetic acid (PAA) treated samples are indicated.

6.3.2 Distinct promoter sequence motifs govern viral gene expression at different time points

Except for TATA and TATT motifs [Gruffat et al., 2016, Perera, 2000], as well as NF-κB binding sites in the major immediate early promoter (MIEP) [Sambucetti et al., 1989, Cherrington and Mocarski, 1989], little is known about cis-regulatory motifs driving the expression from viral promoters. Our collection of viral TSS represents a comprehensive, quantitative, and time-resolved view on transcription initiation in the HCMV genome. Of note, our metabolic labeling strategy and our in-house tool GRAND-SLAM [Jürges et al., 2018] enabled us to analyze transcriptional activity during defined 1 h windows of labeling in addition to global changes in total RNA levels. This provided us with the opportunity to screen for novel sequence motifs enriched in HCMV promoters that are active at specific time points without bias due to RNA half-lives in the order of many hours. First, we aligned all viral promoter sequences at the corresponding TSS and computed sequence logos for the -100 to +10 region. For each time point, we first focused on the 500 most strongly expressed TSS in either human or HCMV and assigned them to 5 equally sized bins according to their transcriptional activity estimated by new RNA (Appendix C.2). This revealed a strong PyPu motif directly at viral TSSs corresponding to the well-known initiator (Inr) element [Haberle and Stark, 2018] as well as TA-rich regions ≈30bp upstream of the TSS, likely corresponding to TATA or TATT (TATW) boxes (Fig. 6.3A, Fig. 6.4A). Of note, the PyPu motif was detectable also for the set of most weakly expressed TSS at each time point thereby demonstrating that these indeed represent bona-fide TSS rather than false positive peaks from our TSS profiling data. The Inr element as well as TATA-

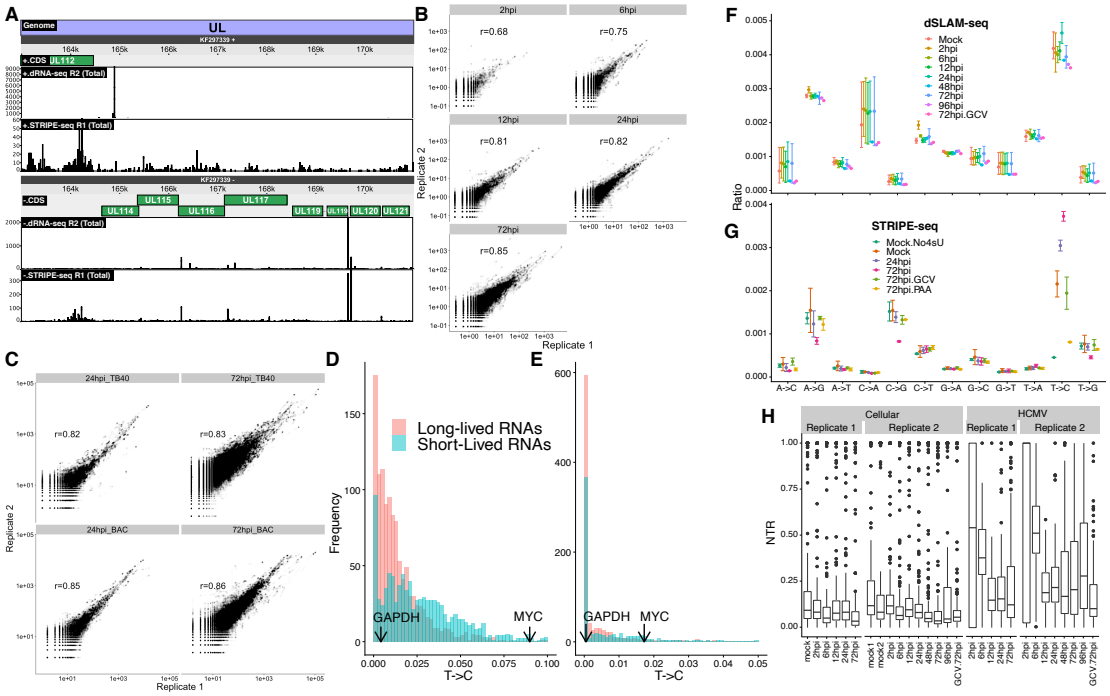


Figure 6.2: **A)** A region of the HCMV genome depicted in our genome browser. The upper half shows the plus strand, the lower half the minus strand. Canonical ORFs are depicted in green. Data shown is the total 5' read count over all time points for dRNA-seq (replicate 2) and STRIPE-seq (replicate 1), respectively. **B,C)** Pairwise correlation analysis of the read counts per nucleotide in the HCMV genome for individual replicates per time point in the dRNA-seq (B) or STRIPE-seq (C) data. Pearson coefficients are indicated. **D,E)** Comparison of long-lived (red) and short-lived (blue) RNAs based on the predicted new-to-total RNA ratios (NTRs) in the dRNA-seq (D) and STRIPE-seq (E) data. NTRs were estimated using GRAND-SLAM [Jürges et al., 2018]. The NTRs for GAPDH and MYC are indicated. **F,G)** The ratios of all occurring nucleotide mismatches in our dRNA-seq and STRIPE-seq (G) experiment, respectively. Mismatch ratios were calculated using GRAND-SLAM [Jürges et al., 2018]. **H)** Distribution of NTRs throughout the time course of infection for the dRNA-seq experiment separated into cellular NTRs (left) and viral NTRs (right).

boxes are bound by different components of the transcription factor II D (TFIID). The TATT motif is recognized by the viral LTF [Gruffat et al., 2016]. Similar motifs were observed for TSS of cellular protein coding transcripts (Fig. 6.3A). When analyzing total RNA, the TATW-boxes were less prominent, especially in cellular and at the viral 2 h timepoint (Fig. 6.4B), highlighting the importance of analyzing transcriptional activity rather than transcript levels.

Interestingly, the TATW boxes were mostly present in strong promoters whereas PyPu motifs were also detectable for weakly used TSS. We confirmed this by counting occurrences of canonical TATA and TATT sequences, as well as PyPu motifs in the top 500 promoters from each time point. PyPu motifs were significantly more frequent for viral TSS than for cellular TSS (Fig. 6.3B, p-value $< 2.03^{-4}$; Fisher's Exact Test). For viral TSS PyPu occurrences were uniformly distributed among the top 500 TSS at each time point, whereas in the host genome, the strongest promoters had less frequent PyPu (p value $= 3.8^{-6}$, Fisher's exact test). This indicates that viral transcription uniformly utilizes PyPu-dependent initiation, whereas initiation for the most strongly transcribed cellular genes is more diverse. TATW boxes occurred more frequently in the promoters of the top 100 most strongly expressed genes (HCMV: p-value $< 10^{-3}$, cellular: p-value $< 10^{-5}$; Fisher's Exact Test) in both cellular and viral genomes. Of note, canonical TATT only became active at the 48 h or 72 h post infection (hpi) time points. This indicates that TATW boxes are the main determinants of the strength of viral and cellular expression.

TATA boxes are known to be located 25-33 bp upstream of the TSS [Vo ngoc et al., 2017]. Our nucleotide precision TSS profiling data confirmed this for both cellular as well as HCMV genes (Fig. 6.3C). Interestingly, however, the location of TATT boxes was shifted 2-3 bp further upstream (p-value $< 10^{-5}$; Fisher's Exact Test). This suggests that the viral LTF is structurally distinct from TFIID.

For a comprehensive analysis of promoter sequences, we curated the binding motifs and the distances relative to the TSS of all known core promoter elements including TATA, TATT and PyPu (Appendix C.3). To enable a quantitative comparison of viral and cellular promoters, we computed the enrichment over the expected number of occurrences in random sequences. As suggested by the analyses above, the PyPu element was more frequent in HCMV than in cellular promoters, and there was no change throughout infection (Fig. 6.3D). Early in infection TATA-boxes had a slightly stronger enrichment in cellular than in viral promoters. Interestingly, for cellular promoters, this dropped down to viral levels between the 12- and 48-hour timepoint and even lower late in infection (p-value < 0.012; Fisher's Exact Test). This decline started at 6 h post infection and was also observed under GCV treatment. This suggests that TFIID is sequestered to viral genomes already in the E phase of infection. This effect is even stronger later in infection in a genome replication dependent manner. By contrast, TATT motifs were not enriched in cellular promoters, and TATT containing viral promoters became much more active in the late phase, which was dependent on genome replication (Fig. 6.3D). The only other statistically significant motif in viral promoters was the BREd element, which is located directly downstream of TATA boxes [Sandelin et al., 2007] (Fig. 6.4C). However, BREd only occurred in 14 out of

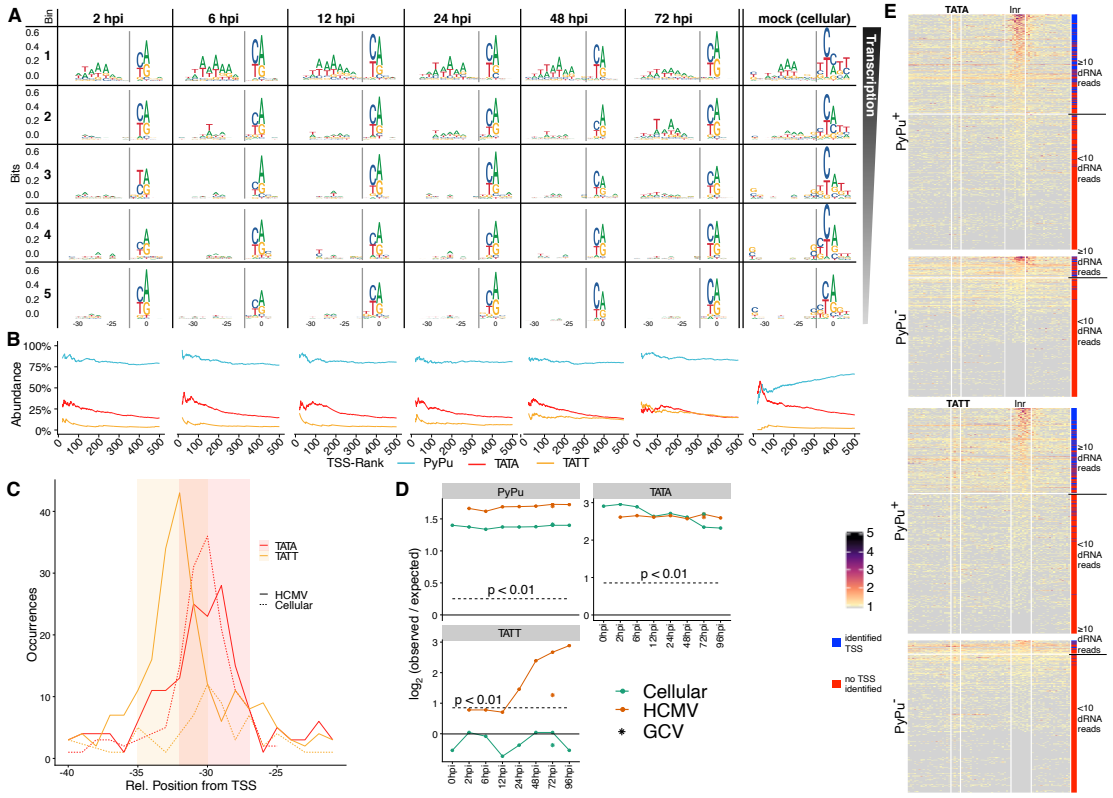


Figure 6.3: **A)** Sequence logos showing the -32 to -23 bp and the -2 to +2 bp of the 500 most highly expressed TSS during each time point (from left to right) in our dRNA-seq dataset divided into 5 bins according to expression strength (from top to bottom). See also Fig. 6.4A-B for the full +/- 100 bp around the TSS. **B)** Line plot showing the percentage of promoters containing correctly positioned (Appendix C.3) PyPu, TATA and TATT motifs among the top n TSS on the x axis. Shown are the TSS ranked from 20 to 500. **C)** Position distribution of TATA (red) and TATT (yellow) motifs relative to the TSS for cellular (dotted line) and HCMV (solid line). The distance between the motif start and the TSS is taken. **D)** Log2 odds of observed vs. expected occurrences of correctly positioned PyPu, TATA and TATT motifs. The critical region ($p < 0.01$, binomial test) is indicated as a dashed line. **E)** Heatmaps showing the total 5'-read counts over all timepoints around all TATA-motifs (top) and TATT-motifs (bottom) found in the HCMV genome. The TATW-motifs as well as the expected region of the TSS are indicated. For each row the bar on the right indicates whether a TSS was identified in the respective region by iTiSS (blue) or not (red). The heatmap is sorted based on the number of reads inside the Inr-region. The point at which the Inr-region contains ≥ 10 reads is indicated.

the 500 viral promoters (6 hpi). For cellular promoters, we also identified the TCT motif (Fig. 6.4C), which is known to replace the Inr-element for many genes involved in translation. Taken together, these analyses show that HCMV evolution favored strong promoters transcribed via canonical initiation mediated by TFIID or its own LTF complex.

Having established to which degree TATW-boxes and Inr elements govern viral gene expression, we next asked whether these core promoters are also sufficient. To this end, we collected all occurrences of these motifs (n=970, TATA; n=973 TATT) on both strands of the HCMV genome and analyzed our TSS profiling sequencing reads pooled from all time points (Fig. 6.3E). Of all TATW boxes upstream of Inr elements, only 42% (TATA, 260 out of 613) and 38% (TATT, 224 out of 590) showed signs of transcription initiation (one position with > 10 reads inside the expected Inr location) in our data. Thus, a core promoter consisting of a TATW box and potentially the Inr element is not sufficient for stable transcription. To test whether additional sequence elements define effective and ineffective core promoters in the HCMV genome, we conducted differential motif searches. These analyses revealed a TA-rich extension of 2 bp downstream of TATT as well as TATA (Fig. 6.4D). However, besides that only two spurious motifs were found, which were randomly located either upstream or downstream of TATW (Fig. 6.4D). We concluded that elements other than simple sequence motifs around TATW boxes are necessary to define sufficiency of core promoters for viral transcription.

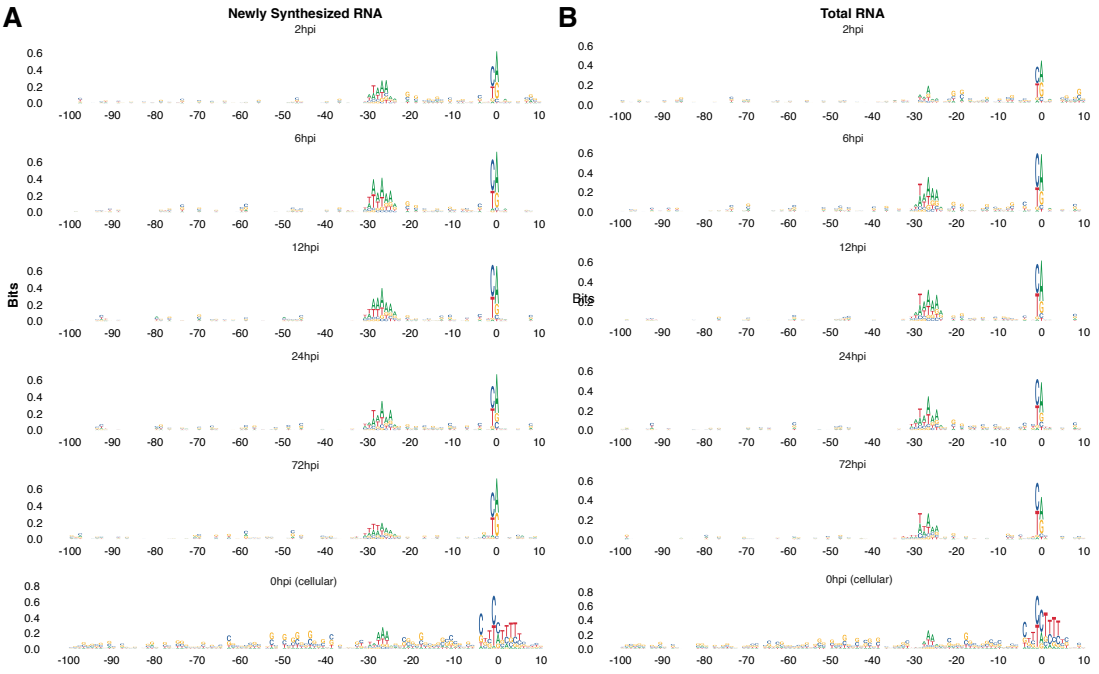
Next, we used MEME [Bailey and Elkan, 1994] to discover putative transcription factor binding sites other than core promoter sequences in the HCMV genome. We performed these analyses for different sequence windows around the TSS as well as running it on individual bins. We furthermore performed these analyses also after excluding TSS that are located in close proximity to other TSS, which might mislead the motif discovery algorithm. In our control sets of human promoters, we identified the TATA-box as well as the TCT-promoter and the GC-box, all of which are known cellular promoter sequences [Vo ngoc et al., 2017, Blake et al., 1990, Parry et al., 2010] (Fig. 6.4E). However, none of these analyses resulted in any viral motif that did not resemble TATW boxes (Fig. 6.4E). Thus, either there is no sequence-specific transcription factor that drives the expression of a sufficiently large set of viral promoters, or they only bind at enhancers that activate transcription via long-range interactions.

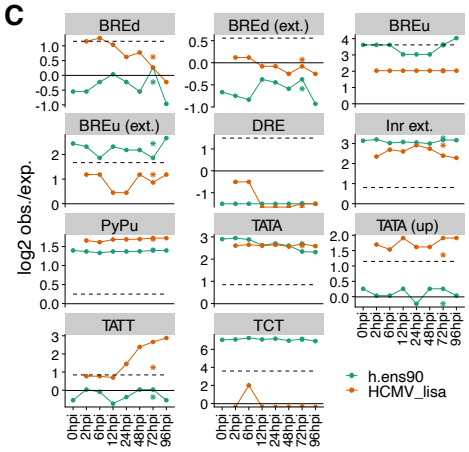
Searching for known motifs of transcription factors can improve the sensitivity. We therefore evaluated all known transcription factors (TF) for binding sites in viral promoters using TFM-Explorer [Tonon et al., 2010]. These analyses revealed 9 significantly enriched TF binding motifs in the top 100 promoters of HCMV at different time points (Appendix C.4). In addition to the TATA-binding protein (36 sites, p-value: 8.09^{-10}), the most strongly enriched TF was HIF-1α (34 sites, p-value: 1.18^{-10}), which has been described previously to be induced in the first few hours of HCMV infection [McFarlane et al., 2011] and to act as an antiviral host factor [Wise et al., 2021]. The enrichment of HIF-1α target sites in viral promoters indicates that HCMV usurped this host defense mechanism to regulate the expression of dozens of viral genes. Another top hit in our list was MYC (18 sites, p-value: 4.04^{-7}), which has been described to be activated by IE1 and IE2 [Hagemeier et al., 1992].

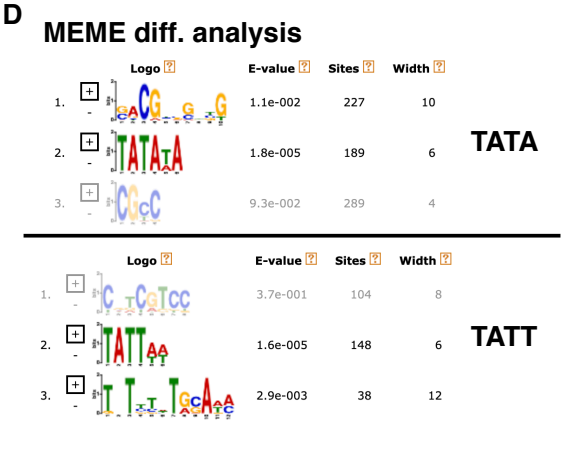
Interestingly, MYC target sites were not enriched before 6 hpi, consistently with our work in murine cytomegalovirus [Marcinowski et al., 2012]. This is consistent with a model that HCMV activates MYC by its IE proteins to enhance the expression of several viral genes [Amati et al., 2001]. In summary, extensive sequence analyses of our TSS data revealed viral transcription initiation to depend uniformly on the PyPu element, that the strength of transcriptional activity throughout the course of infection is predominantly governed by TATW boxes, and that HCMV usurped specific cellular TFs including HIF-1α and MYC to facilitate the expression of specific viral genes.

6.3.3 TATT-boxes define early-late transcription

Our analysis of core promoter elements indicated that strong viral transcription is initiated from TATT promoters at late time points and depends on viral genome replication. However, TATT promoters were also already significantly enriched for the top 500 TSS at early time points and also after blocking viral genome replication (Fig. 6.4D). To confirm this, we performed metagene analyses anchored at TATW-boxes with a TSS identified inside the respective downstream Inr-window (Fig. 6.5A). Downstream of such TATA boxes, levels of metabolically labeled reads increased substantially within the expected window, and the extent of this increase resembled the overall strength of viral gene expression per time point. By contrast, for TATT-boxes the timepoints could be divided into 3 classes: Before 12 hpi, read levels at the TSS did not significantly exceed the background signal, indicating that TATT boxes do not play a role for the initiation of transcription at these early infection time points. In contrast, TATT boxes at late time points (after 48 hpi) showed the same marked increase at the TSS as TATA boxes. Importantly, we observed similar results in 24 hpi. Of note, when considering total RNA, metagene plots did not show this pattern (Fig. 6.6A), again highlighting the importance of our metabolic labeling approach. The relatively high levels of reads at the TSS for TATT-TSS without genome replication (24 hpi and 72hpi+GCV) indicated that either a small number of TATT-TSS are strongly active already at 24 hpi independent on genome replication, or that a large number of TATT-TSS already exhibit weak transcriptional activity before genome replication. To investigate the dependency on genome replication for individual TSS, we defined the "TSS true late score" (TLS) as the percentage of normalized TSS reads in the 72 hpi+GCV sample compared to the 72h-GCV sample (Fig. 6.5B). Interestingly, 50% of TATA-TSS had a TLS between 24.5% and 191% (25% and 75% percentiles). This was only partly due to inaccurate quantification from the limited amounts of reads, as the percentiles still were 14.9% and 336% when only considering TATA-TSS with at least 50 reads in the 72 hpi-GCV sample. This could not be explained by incomplete inhibition of viral genome replication at 72 hpi by GCV, as we observed similar results when using the 24 hpi sample to compute the TLS instead of the 72 hpi+GCV sample (TLS percentiles 9.1% and 147%; Fig. 6.6B). We concluded that, transcription is already initiated from TATT promoters without genome replication, but that transcriptional activity is strongly amplified as viral DNA replication provides more template DNA.


A
Newly Synthesized RNA
B
Total RNA
2hpi
6hpi
12hpi
24hpi
72hpi
0hpi (cellular)
Bits


C
BREd
BREd (ext.)
BREu
BREu (ext.)
DRE
Inr ext.
PyPu
TATA
TATA (up)
TATT
TCT
log2 obs./exp.
0hpi
2hpi
6hpi
12hpi
24hpi
48hpi
72hpi
96hpi
h.ens90
HCMV_lisa


D
MEME diff. analysis
Logo
E-value
Sites
Width
1. 1.1e-002 227 10
2. 1.8e-005 189 6
3. 9.3e-002 289 4
TATA
1. 3.7e-001 104 8
2. 1.6e-005 148 6
3. 2.9e-003 38 12
TATT

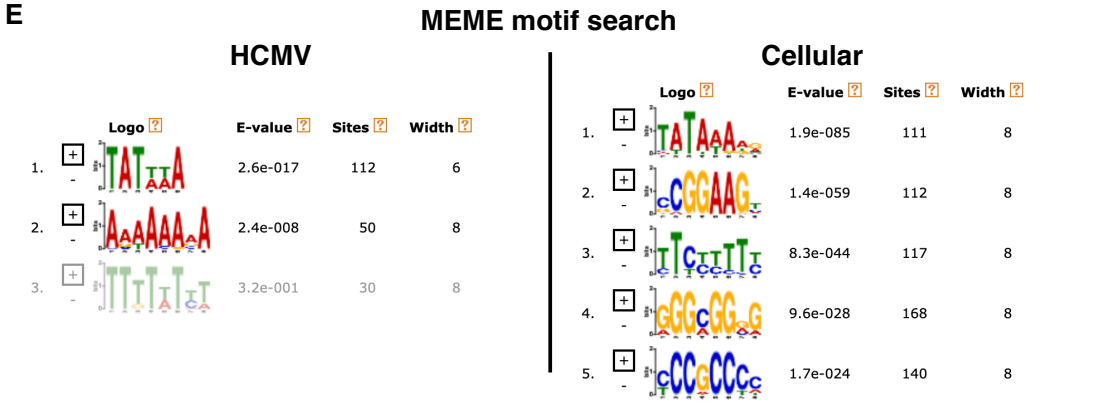

E
MEME motif search
HCMV
Logo
E-value
Sites
Width
1. 2.6e-017 112 6
2. 2.4e-008 50 8
3. 3.2e-001 30 8
Cellular
1. 1.9e-085 111 8
2. 1.4e-059 112 8
3. 8.3e-044 117 8
4. 9.6e-028 168 8
5. 1.7e-024 140 8

Figure 6.4: **A,B)** Sequence logos of the 100 most strongly expressed TSS of each time point based on newly synthesized RNA (A) or total RNA (B), respectively. For cellular only the mock (0 hpi) timepoint is shown. Sequence logos range from 100 bp upstream to 100 bp downstream of the respective TSS. **C)** Log2 odds of observed vs. expected occurrences of correctly positioned promoter elements (see Appendix C.3). The critical region ($p< 0.01$, binomial test) is indicated as a dashed line. **D)** Results of the differential motif analysis using MEME. Shown are the three top scoring (lowest E-value) motifs for the TATA- and TATT-associated TSS, respectively. **E)** All significantly enriched motifs found using MEME in the HCMV or cellular genome, respectively.

Interestingly, the TLS for TATA-TSS were much more variable than for TATT. For 21 out of 70 TATA-TSS with at least 20 reads, GCV treatment substantially increased RNA levels at 72 hpi (TLS > 150%, Fig 6.5C, Fig. 6.6C, which did not change when restricting to strongly transcribed TSS (Fig. 6.5B). These TSS all have in common that they are strongly expressed early and downregulated later. Their TLS > 150% suggests that this inhibition is dependent on viral genome replication. Furthermore, 26 out of the 70 TATA-TSS had a TLS < 30%, i.e. were sensitive to GCV treatment (Fig. 6.5D, Fig. 6.6D). This strong dependence on genome replication was unexpected as late kinetics are believed to depend on the viral LTF binding TATT. However, when we analyzed the positioning of TATA and TATT boxes relative to the TSS, we observed a subset of viral but not cellular TATA-TSS to be located more than 33 bp upstream of the TSS, i.e. further away from the TSS than expected (Fig. 6.3C). Interestingly, these largely are TATA-TSS with TLS < 30% (Fig. 6.5E). This indicates that promiscuous binding by the viral LTF to TATA boxes contributes significantly to the expression from the respective TSS. Interestingly, 17 of the TATA-TSS have TLS < 15% suggesting complete dependence on genome replication. Of note, there were no differences in sequence logos for TATA boxes with high or low TLS (Fig. 6.6E). In summary, our data show that LTF binding to TATT and potentially dozens of TATA motifs define early late transcription.

6.3.4 Virion-associated RNAs are less efficiently translated than de novo transcribed viral RNAs

When considering total RNA in the metagene analysis above, TATT-boxes appear to show a substantial increase in read coverage at the TSS also for the 2 hpi time point and, to a lesser extent, at 6 hpi (Fig. 6.6A). We hypothesized that this enrichment is due to incoming virion associated RNA and not because of transcriptional activity driven by these promoters early in infection. We first concentrated on examples of TSS located upstream of ORFs that are not translated early [Weekes et al., 2014] (including UL72, UL97, UL35, UL50, UL76, UL85, UL100). All the respective TSS were expressed with true late kinetics (TLS < 10%). Interestingly, we found substantial levels of RNA in the first hours of infection (Fig. 6.7A). However, our metabolic labeling data indicated that no RNA was transcribed

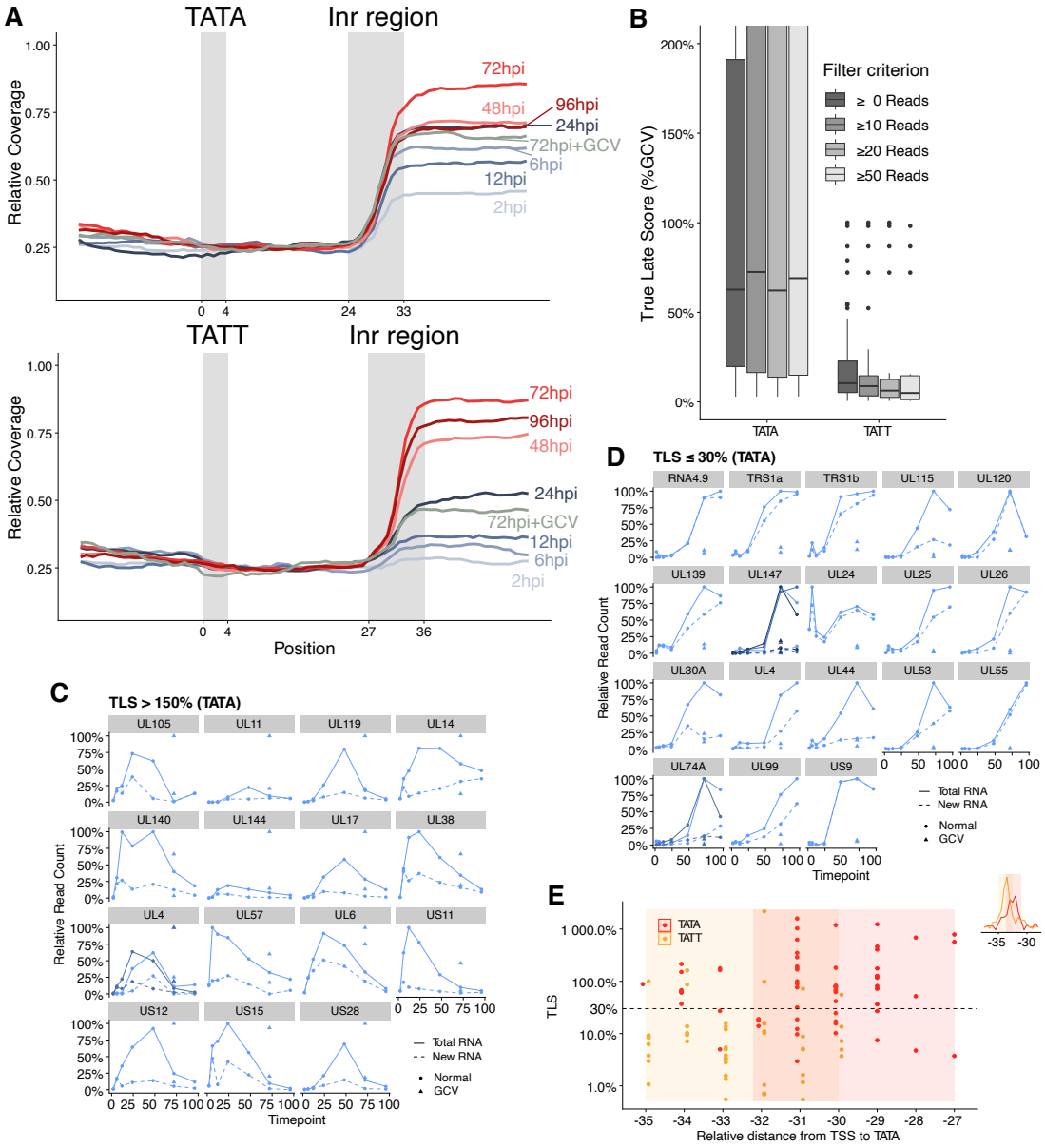

A
TATA
Inr region
Relative Coverage
72hpi
96hpi
48hpi
24hpi
72hpi+GCV
6hpi
12hpi
2hpi
TATT
Inr region
Position
B
True Late Score (%GCV)
Filter criterion
≥ 0 Reads
≥10 Reads
≥20 Reads
≥50 Reads
TATA
TATT
C
TLS > 150% (TATA)
UL105
UL11
UL119
UL14
UL140
UL144
UL17
UL38
UL4
UL57
UL6
US11
US12
US15
US28
Relative Read Count
Total RNA
New RNA
Normal
GCV
Timepoint
D
TLS ≤ 30% (TATA)
RNA4.9
TRS1a
TRS1b
UL115
UL120
UL139
UL147
UL24
UL25
UL26
UL30A
UL4
UL44
UL53
UL55
UL74A
UL99
US9
Relative Read Count
Timepoint
E
TATA
TATT
TLS
Relative distance from TSS to TATA

Figure 6.5: **A)** Metagene plots for each timepoint centered around all TATA or TATT-motifs with identified TSS in the HCMV genome. Only reads with at least 2 T→C mismatches are considered. The respective TATW-motif as well as the region of the expected TSS are indicated. See Fig. 3A for a metagene plot involving all reads. **B)** Boxplots of the True Late Score (TLS) for all TATA- or TATT-TSS, TSS with ≥ 10 reads, TSS with ≥ 20 reads and TSS with ≥ 50 reads, respectively **C,D)** Abundance profiles for all TSS with at least 20 reads and a canonical ORF located within 1,000bp downstream for total RNA (solid line) and newly synthesized RNA (dashed line). GCV treated samples are indicated as a triangle. For each TSS, the upper triangle depicts total RNA the lower newly synthesized RNA. (C) shows the abundance profiles for TSS with a TLS$\geq$ 150% and (D) for TSS with a TLS $\geq$ 30%. **E)** The TLS for TATA-associated TSS is shown according to the relative positions of the TATA-motifs to their respective TSS. The expected locations for TATA and TATT-motifs are indicated by the respective shading of the background. The smaller upper right Figure depicts the locations of viral TATA- and TATT-boxes relative to the TSS. It is identical to Figure 6.3C without the cellular data.

during the first hours from the respective TSSs. Moreover, the total RNA levels for those decreased during the first hours of infection, providing independent evidence against their de novo transcription in the IE or E phase of infection. Since these RNAs were also absent in uninfected cells, we concluded that the RNA observed early during infection for these genes came into the cells with the infecting virus particles as described previously [Terhune et al., 2004, E. et al., 2000]. To analyze a larger set of virion-associated RNAs, we focused on TATT promoters. Above, we established that these generally follow late kinetics and do not give rise to transcription in the first hours of infection. Nevertheless, TATT RNAs were among the viral genes with the highest total RNA expression levels at 2 hpi (Fig. 6.7B). This can be explained by the fact that, among all TATW promoters for the 2 hpi time point, the vast majority (50 out of 53) of TATT genes and also a significant part of TATA genes (48 out of 75) did not show any sign of reproducible transcriptional activity (Fig. 6.7C). Furthermore, TATT genes generally show no transcriptional activity before the 24h time point, were sensitive to GCV, showed a decline in total RNA levels during the first few hours of infection, and a sharp increase late in infection (Fig. 6.7D). Thus, the TATT promoters defined by our data represent a bona-fide set of late TSS, and the majority of transcripts of TATT genes expressed in the first few hours of infection are translated from virion associated RNA.

To test whether virion-associated transcripts are productively translated, and to quantify their translational efficiency, we compared our TSS profiling data to Ribo-seq data [Stern-Ginossar et al., 2012]. In order to circumvent any potential noise, we only considered TSS that were identified during the 2 hpi timepoint and are located in front of canonical large ORFs. We found 28 TATA-associated and 17 TATT-associated TSS that passed our stringent filtering. Analyzing these revealed that the translational efficiencies defined as the translation strength per mRNA of TATA-associated TSS was significantly higher than that

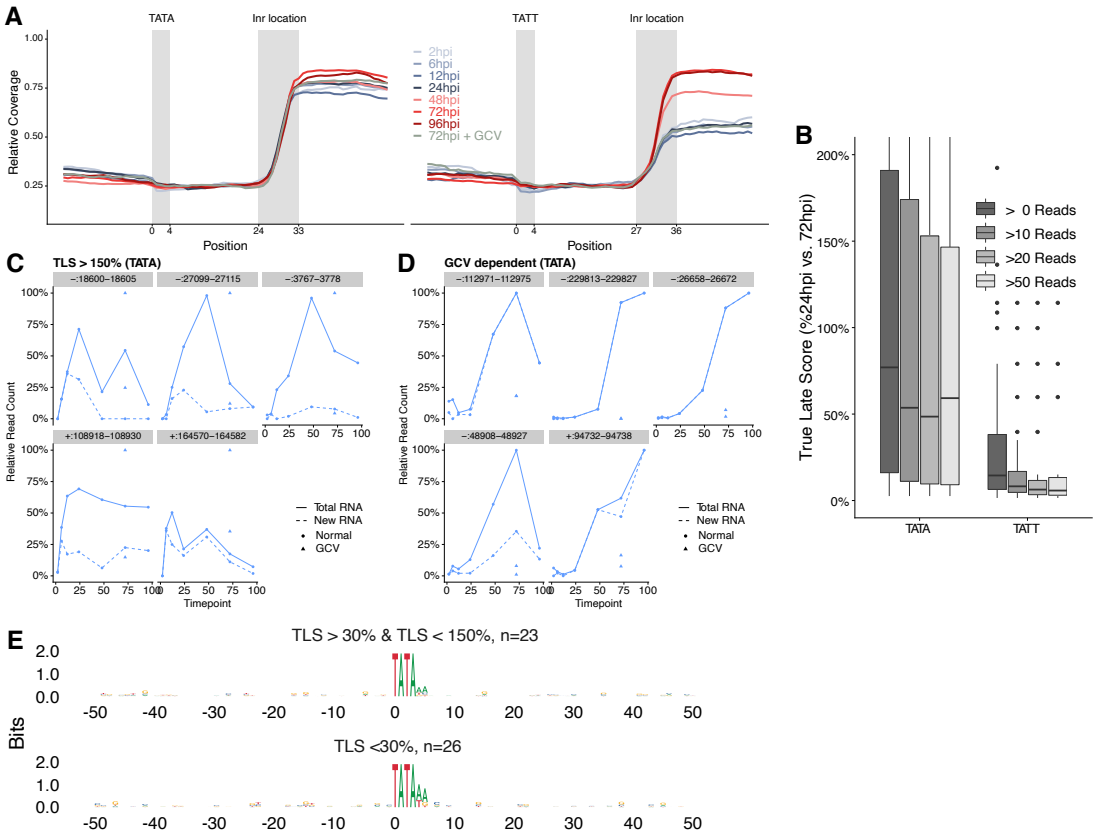


Figure 6.6: **A)** Metagene plots for each timepoint centered around all TATA or TATT-motifs with identified TSS in the HCMV genome . All reads are considered. The respective TATW-motif as well as the region of the expected TSS are indicated. **B)** Boxplots of the True Late Score (TLS) calculated using the 24hpi sample instead of the GCV treated sample. TLS are shown for TATA- and TATT-associated TSS, respectively. TSS were filtered by read counts as indicated. **C,D)** Abundance profiles for all TSS with at least 20 reads and without a canonical ORF located within 1,000bp downstream for total RNA (solid line) and newly synthesized RNA (dashed line). GCV treated samples are indicated as a triangle. For each TSS, the upper triangle depicts total RNA the lower newly synthesized RNA. (C) shows the abundance profiles for TSS with a TLS $\geq$ 150% and D for TSS with a TLS $<$ 30%. **E)** Sequence logos for TATA-associated GCV independent (top) or dependent (bottom) TSS, respectively. Sequence logos are centered around the TATA-motif.

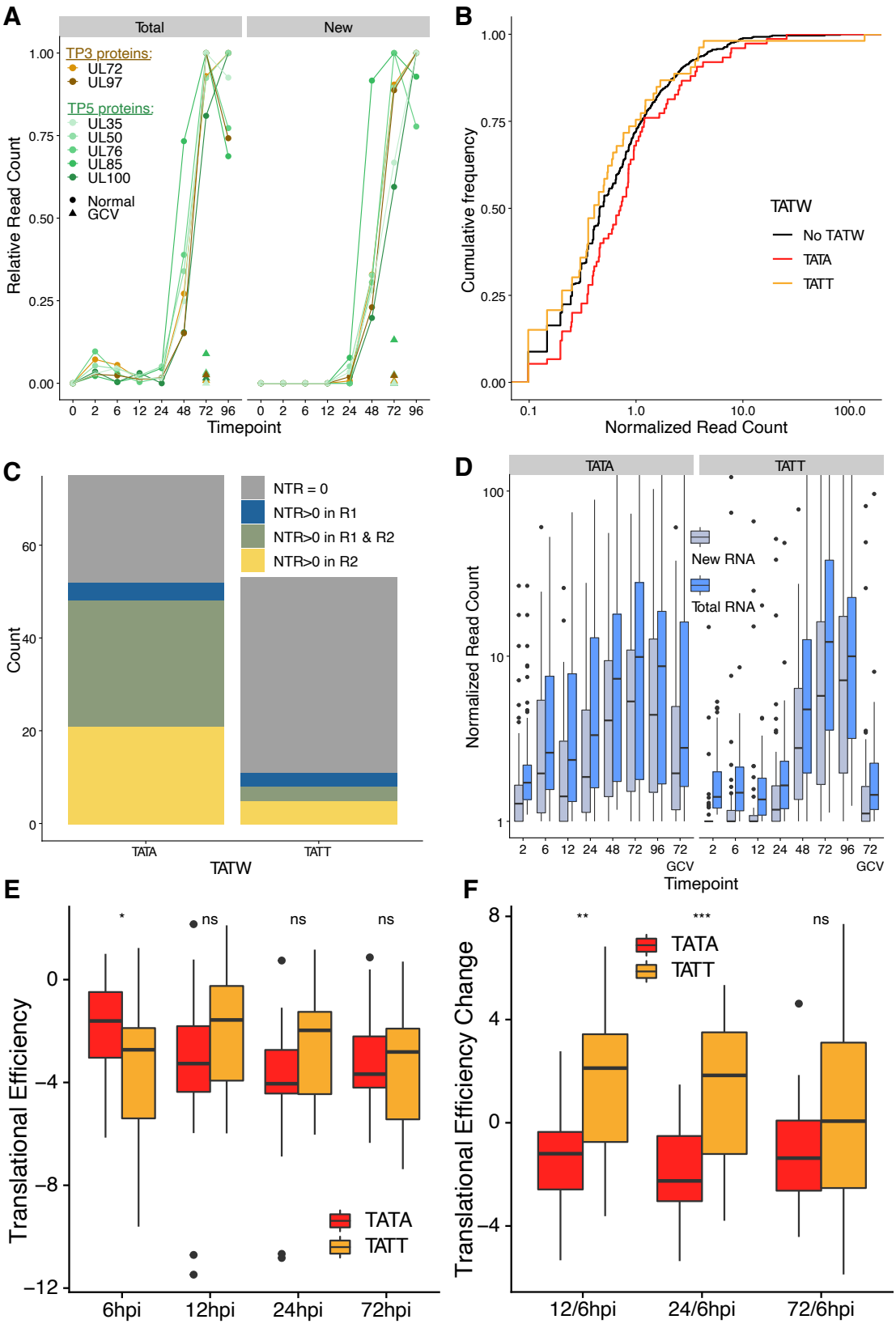

A
Total
New
TP3 proteins:
UL72
UL97
TP5 proteins:
UL35
UL50
UL76
UL85
UL100
Normal
GCV
Relative Read Count
Timepoint
B
Cumulative frequency
Normalized Read Count
TATW
No TATW
TATA
TATT
C
NTR = 0
NTR>0 in R1
NTR>0 in R1 & R2
NTR>0 in R2
Count
TATA
TATT
TATW
D
TATA
TATT
New RNA
Total RNA
Normalized Read Count
GCV
Timepoint
E
Translational Efficiency
6hpi
12hpi
24hpi
72hpi
TATA
TATT
F
Translational Efficiency Change
12/6hpi
24/6hpi
72/6hpi
TATA
TATT

Figure 6.7: **A)** Abundance profiles for total (left) or newly synthesized (right) RNA of manually selected TSS that are classified as late genes by Weekes et al [Weekes et al., 2014] (Tp3 & Tp5). Non-treated and GCV-treated samples are indicated. **B)** Empirical cumulative distribution function of normalized read counts for viral TATA (red), TATT (yellow) and non-TATW associated TSS at the 2 hpi timepoint. **C)** Number of TATA or TATT associated TSS found in the 2hpi timepoint that show transcriptional activity in both (green) replicates, only the first replicate (blue), only the second replicate (yellow) or none of the replicates (grey). **D)** Boxplots showing the normalized read counts for total (blue) and new (gray) RNA over the course of infection for TATA (left) or TATT (right) associated TSS identified at the 2 hpi timepoint. **E)** Translational efficiency (translation / transcription) for TATA and TATT associated TSS identified in the 2 hpi timepoint. Only TSS in front of canonical large ORFs were selected that were the most strongly expressed ones as well as being TATW-associated. For the translation the Ribo-seq codon count throughout the respective ORF in our cycloheximide treated sample was taken (additionally normalized by length). Statistical significance according to a Mann-Whitney-U test is indicated (*: $p< 0.05$; the p-values were: 6hpi, $p = 0.033$; 12hpi, $p = 0.06$, 24hpi, $p = 0.68$, 72hpi, $p = 0.82$). **F)** Change in translational efficiency between the 6 hpi and the other three timepoints for TATA and TATT associated TSS identified in the 2 hpi timepoint. Only TSS in front of canonical large ORFs were selected that were the most strongly expressed ones as well as being TATW-associated. For the translation the Ribo-seq codon count at the start codon of the respective ORF in our cycloheximide treated sample was taken. Statistical significance according to a Mann-Whitney-U test is indicated (**: $p< 0.01$, ***: $p< 1.0^{-3}$; the p-values were: 12/6hpi, $p = 0.001$; 24/6hpi, $p = 5.9^{-4}$, 72/6hpi, $p = 0.23$).

of TATT-associated TSS during the first six hours of infection (Fig. 6.7E, Mann-Whitney-U Test p-value: 6 hpi = 0.03). Moreover, we observed significant differences in the way these translational efficiencies changed during infection (Fig. 6.7F). TATA-associated TSS became less efficiently translated at late times of infection, whereas TATT-associated TSS efficiencies increased drastically. This was also confirmed by our LTM-treated sample (Fig. 6.8A/B). We concluded that virion-associated RNAs are successfully imported into newly infected cells, however, they are not as efficiently translated as de novo transcribed viral RNAs.

6.3.5 Integrative analysis predicts post-translational regulation during HCMV infection

The IE, E and L kinetic classes defined HCMV protein expression, but are largely dependent on gene regulation exerted at the level of transcription. To close this gap, we performed integrative analysis of genome-wide data on transcriptional activity (TSS profiling, new RNA), mRNA levels (TSS profiling, total RNA), translational activity (Ribo-seq

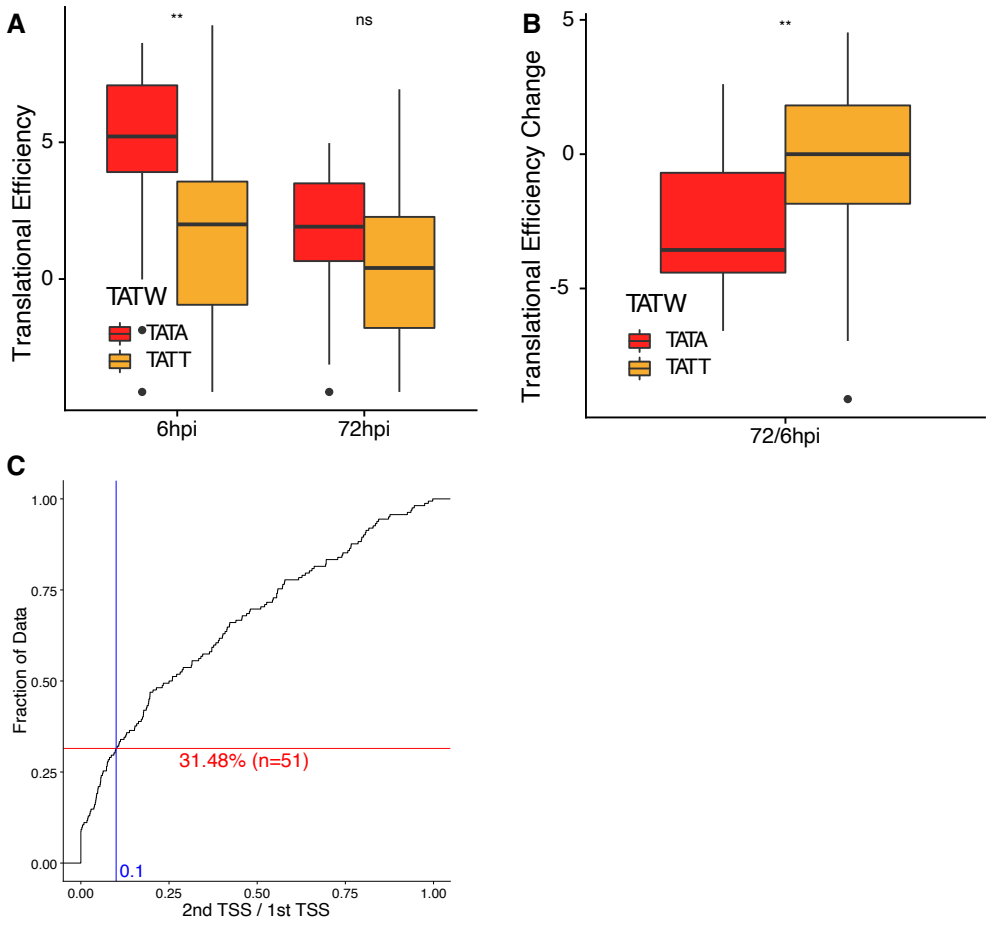


Figure 6.8: **A)** Translational efficiency (translation / transcription) for TATA and TATT associated TSS identified in the 2 hpi timepoint. Only TSS in front of canonical large ORFs were selected that were the most strongly expressed ones as well as being TATW-associated. For the translation the Ribo-seq codon count at the start codon of the respective ORF in our LTM treated sample was taken. Statistical significance according to a Mann-Whitney-U test is indicated (**: $p < 1.0^{-2}$; the p-values were: 6hpi, $p = 0.001$; 72hpi, $p = 0.24$). **B)** Change in translational efficiency between the 6 hpi and the 72 hpi timepoint for TATA and TATT associated TSS identified in the 2 hpi timepoint. Only TSS in front of canonical large ORFs were selected that were the most strongly expressed ones as well as being TATW-associated. For the translation the Ribo-seq codon count at the start codon of the respective ORF in our LTM treated sample was taken. Statistical significance according to a Mann-Whitney-U test is indicated (**: $p < 1.0^{-2}$; p-value for 72/6hpi: 0.006). **C)** The empirical cumulative distribution function for the fold change of the TSS with the second highest read count (subdominant TSS) per ORF by the first one (dominant TSS). If only one TSS was present for a given ORF the fold-change was set to 0. The number of ORFs without a subdominant TSS that exceeds read levels of 10% of the dominant TSS are indicated.

[Stern-Ginossar et al., 2012]) and protein levels (mass spectrometry [Weekes et al., 2014]). It is important to note that most of the protein coding ORFs can be translated from multiple mRNA isoforms that arise from the wealth of viral TSS. We therefore identified all TSS located within a 1 kb window upstream of each protein coding ORF. This defined at least one TSS for 162 out of the 168 (96.43%) canonical ORFs. The TSS with the highest normalized read count (total RNA) in any sample was called the dominant TSS, and we discarded all subdominant TSSs with an expression that did not exceed 10% of their dominant TSS. 51 ORFs had no subdominant TSS, and the expression of 49 of all other proteins was governed by multiple TSS (expression of strongest subdominant TSS exceeded 50% of the dominant, Fig. 6.8C). First, we computed the total mRNA profiles per ORF as the sum of read counts from all corresponding TSS per time point. Interestingly, the mRNA profiles for the five different temporal protein (Tp) classes defined by Weekes et al. [Weekes et al., 2014] closely resembled their protein abundance profiles (Fig. 6.9A). Of note, Tp1-5 could largely be characterized solely by the change from early (6 hpi for Tp1 and 12 hpi otherwise) to late (48 hpi) and from late (48 hpi) to true late (96 hpi) expression (Fig. 6.9B). In contrast to Tp3 and Tp5, protein levels decline for Tp1, Tp2 and Tp4 after 48 hpi. From 12 hpi to 48 hpi, Tp4 protein levels increase, Tp2 levels stay almost constant, and decline for Tp1. Characterizing the five temporal classes by two parameters (fold change from early to late, and from late to true late) enabled us to quantitatively compare regulation of proteins and their corresponding transcripts. Interestingly, early regulation from 12 hpi to 48 hpi of the temporal protein classes was fully reflected in the RNA levels (Fig. 6.9C). After 48 hpi, downregulation of Tp1, Tp2 and Tp4 proteins was also paralleled on the RNA level. By contrast, the increase in RNA levels from 48 hpi to 96 hpi was much less pronounced than for proteins for Tp3 and Tp5 (Fig. 6.9D). This suggests that RNA levels largely have reached steady state levels around 48 hpi, whereas protein levels are below steady state still at 96 hpi for Tp3 and Tp5. We concluded that the mRNA kinetics inferred from our TSS data are highly consistent with published proteomics data, and that differences can be explained by differing half-lives of RNAs and proteins.

6.3.6 A subset of HCMV proteins is translated from multiple mRNAs with distinct kinetics

We next asked whether there are differences among the dominant and subdominant TSS for individual ORFs. For this we clustered all TSS according to their new RNA profiles into five temporal RNA (Tr) classes and compared these Tr clusters with the corresponding Tp clusters defined by Weekes et al. [Weekes et al., 2014]. Strikingly, except for differences at 96 hpi for Tp3/Tr3 and Tp5/Tr5, the averages of Tr clusters closely resembled the Tp cluster averages (Fig. 6.10A). Upstream of the majority but not all ORFs (68%) we could identify at least one TSS with equivalent Tr and Tp classes (Fig. 6.10B). Interestingly, however, the Tr class of the dominant TSS per ORF did only recapitulate the Tp class in 42% of the cases (Fig. 6.10B). Of the 79 ORFs with at least one subdominant TSS, 63 had

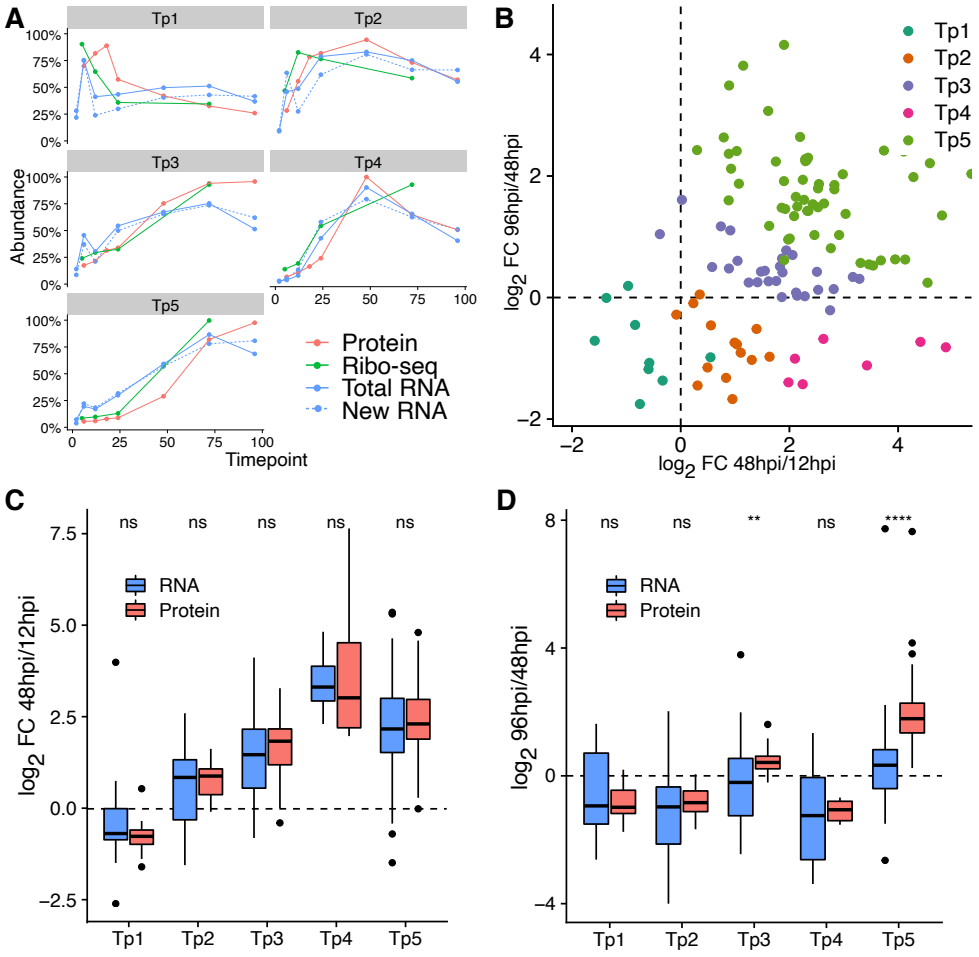


Figure 6.9: **A)** For all temporal protein (Tp) clusters defined by Weekes et al. [Weekes et al., 2014], the relative protein abundance (red), the translation strength according to Ribo-seq (green), RNA levels according to dRNA-seq (blue, solid lines) and transcription strength according to new RNA in dRNA-seq (blue, dashed lines) are shown. All TSS within at most 1kbp upstream of the ORF were considered. **B)** Scatterplot comparing the fold changes between the 48 and 12 hpi timepoint (x-axis) and the 96 and 48 hpi timepoint (y-axis) for protein abundances. Proteins are colored based on their assigned Tp-cluster. 2 outlier points with x or y values > 7 from Tp 4&5 are omitted. **C)** Comparison of the 48hpi vs. 12hpi fold changes for TSS expression and protein abundances for each of the five Tp-clusters. For none of the Tp-clusters the difference is significant (Mann-Whitney-U test; p-values: Tp1=0.63, Tp2=0.82, Tp3=0.35, Tp4=0.75, Tp5=0.61) **D)** Comparison of the 96hpi vs. 48hpi fold changes for TSS expression and protein abundances for each of the five Tp-clusters. Only for clusters Tp3 and Tp5 the differences are significant (Mann-Whitney-U test; **: $p < 1.0^{-2}$, ****: $p < 1.0^{-4}$; p-values: Tp1=0.83, Tp2=0.55, Tp3=0.006, Tp4=1.0, Tp5=2.2^{-11}).

TSS classified into more than one Tr profile (Fig. 6.11A/B). We concluded that, for the majority of proteins, the temporal kinetics cannot be explained by an equivalent kinetic behavior of a single viral TSS on the level of transcription, but that multiple TSS with distinct kinetics govern the protein expression throughout infection.

Alternative TSS for the same protein differ in the length of the corresponding 5' untranslated regions. Therefore, and potentially due to distinct sets of uORFs, each TSS might give rise to a distinct translational efficiency (TE) for the protein coding ORF. To evaluate the impact of multiple TSS on translation, we developed a model to estimate TSS-specific TEs (see Methods) and used these TEs to predict the total translation for each time point from the TSS data. For 20 out of the 63 ORFs that had subdominant TSS with different Tr, a single TSS was sufficient to accurately predict the Ribo-seq signal (Fig. 6.10C). However, 34 ORFs could only be predicted accurately with two TSS and for 9 ORFs 3 TSS were necessary. This included UL19 (Fig. 6.10D). The longer isoform belongs to Tr4, includes a uORF (Fig. 6.12), and is estimated by our model to be translated 2x as efficient as the shorter isoform that lacks the uORF (Tr3). Interestingly, our model predicts the translation to be strongest at the 48 hpi time point that is missing in the Ribo-seq data, which is consistent with the proteomics data (Fig. 6.7D). Furthermore, explaining the translation kinetics of the ORF of the viral Fc-gamma receptor-like protein UL119 also requires two TSS with distinct kinetics. The longer isoform, which follows Tr5 kinetics, is predicted to be translated 5x as efficient as the shorter isoform (Tr4) by our model (Fig. 6.11C). Interestingly, the Ribo-seq data clearly shows that the longer isoform gives rise to translation of an N-terminally extended UL119 protein initiated from a non-canonical AUA codon late in infection (Fig. 6.10E). In summary, the distinct transcriptional regulation of multiple TSS governs the effective translation of dozens of HCMV ORFs, which in turn governs the temporal expression profile and kinetics for their respective proteins.

6.3.7 Non-productive, pervasive transcription in HCMV infection

Recently, Parida et al. [Parida et al., 2019] utilized PRO-seq and PRO-cap experiments to identify over 7,000 transcription start site regions (TSRs) active at 96 hpi in the HCMV genome. We mapped the previously identified PRO-cap TSRs to the TB40E-Lisa reference genome, which left us with 6985 TSS (see Methods). Interestingly, we recovered only 946 of these at the 96 hpi time point in our TSS data, and 1,712 when pooling the data of all time points (Fig. 6.13A). Heatmap analysis showed that this was not simply a consequence of lower sensitivity due to differing read depths between these data sets, as there was no correlation between the PRO-cap read count and the identification of a TSS in our data (Fig. 6.13B). Furthermore, we reproducibly identified $> 1,100$ TSS that had not been detected by PRO-cap. Importantly, these TSS nevertheless showed the characteristic TATW-boxes and PyPu motifs confirming that they indeed represent bona fide TSS (Fig. 6.14A).

It is important to note that both dRNA-seq and STRIPE-seq identify TSS of stable viral

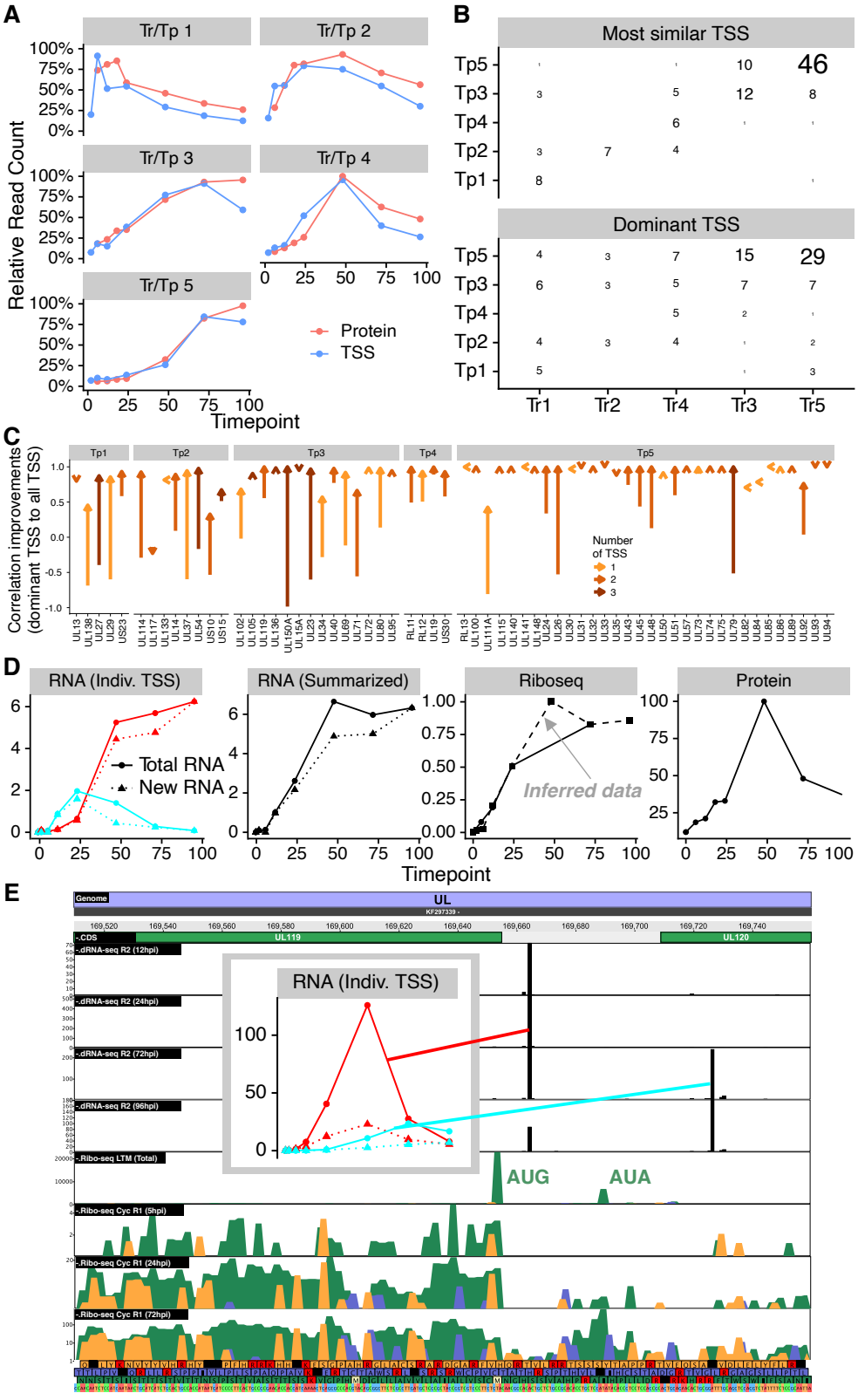

A
Tr/Tp 1
Tr/Tp 2
Tr/Tp 3
Tr/Tp 4
Tr/Tp 5
Relative Read Count
Timepoint
Protein
TSS
B
Most similar TSS
Dominant TSS
Tp5
Tp3
Tp4
Tp2
Tp1
Tr1
Tr2
Tr4
Tr3
Tr5
C
Correlation improvements (dominant TSS to all TSS)
Number of TSS
D
RNA (Indiv. TSS)
RNA (Summarized)
Riboseq
Protein
Total RNA
New RNA
Inferred data
Timepoint
E
Genome
UL
CDS
UL119
UL120
RNA (Indiv. TSS)
AUG
AUA

Figure 6.10: **A)** Comparison of the average profile for protein abundances in the five temporal protein (Tp) classes defined by Weekes et al. [Weekes et al., 2014] and the average RNA profiles measured by dRNA-seq in the five temporal RNA (Tr) classes. **B)** Comparison of the Tp classification and the corresponding Tr classification of the corresponding TSS. The numbers indicate the number of ORFs belonging to a specific combination of Tr and Tp class. In the upper panel, the TSS with the most similar temporal profile was considered. In the lower panel, the dominant TSS was used. **C)** Depiction of the improvements of the correlation between the transcription (TSS-profiling data) and translation (Ribo-seq data) by using only the dominant TSS (lower part of the arrow) or the strongest TSS of each Tr class (upper part of the arrow) for each gene. The colors show the number of available distinct TSS by their Tr class. Arrows pointing upwards depict an improvement of the correlation by using more than just the dominant TSS, arrows pointing downwards show a decrease in correlation when using multiple TSS and arrows pointing to the left depict no change whether only the dominant TSS or all TSS are used. **D)** Individual TSS expression (top left), summarized TSS expression (top right), Ribo-seq expression (bottom left) and protein abundance (top right) for the UL19 protein. Total RNA (solid line) and newly synthesized RNA levels (dotted line) are indicated. For the Ribo-seq expression, the solid line indicates the measured expression, and the dashed line indicates the inferred data based on our modelling approach (see Methods), including the 48 hpi timepoint, which is not present in the Ribo-seq data. **E)** The two TSS of UL119 are shown in the genome browser with their respective individual transcription abundance profiles. The canonical start codon (AUG) for UL119 is indicated as well as the non-canonical start codon (AUA) of a potential N-terminal extension of UL119. Tracks show from top to bottom: Tracks show from top to bottom: The canonical large ORFs denoted as CDS (coding sequences), the dRNA-seq 5'-counts for the timepoints 12, 24, 72 and 96 hpi, the totalized codon counts for all three frames of the LTM treated Ribo-seq sample (linear scale) and finally the codon counts for all three frames of the cycloheximide treated Ribo-seq samples for the timepoints 5, 24 and 72hpi (log-scale).

transcripts that accumulate during infection whereas PRO-cap measures nascent transcription initiation at a define time of infection. Indeed, metagene analysis in the human genome identified thousands of PROMPTs [Cherrington and Mocarski, 1989], corresponding to transient transcription initiation events antisense upstream to TSS of protein-coding genes that were visible in PRO-cap but not our dRNA or STRIPE-seq data (Fig. 6.13C). We thus asked whether PROcap-only TSS might represent pervasive transcription. To address this, we first assessed whether TATW-motifs associated with PyPu elements with no corresponding TSS detectable by dRNA-seq or STRIPE-seq (Fig. 6.3E) might be associated with instable viral transcripts only detectable by PRO-cap. Indeed, $> 52\%$ (208 of 398) of the respective TATA-motifs and $> 51\%$ (214 of 419) of the TATT-motifs had at least 10 reads in the PROcap data (Fig. 6.13D and Fig. 6.14B). The respective TSS thus presumably represent pervasive HCMV transcription, e.g. promoter or enhancer RNAs.

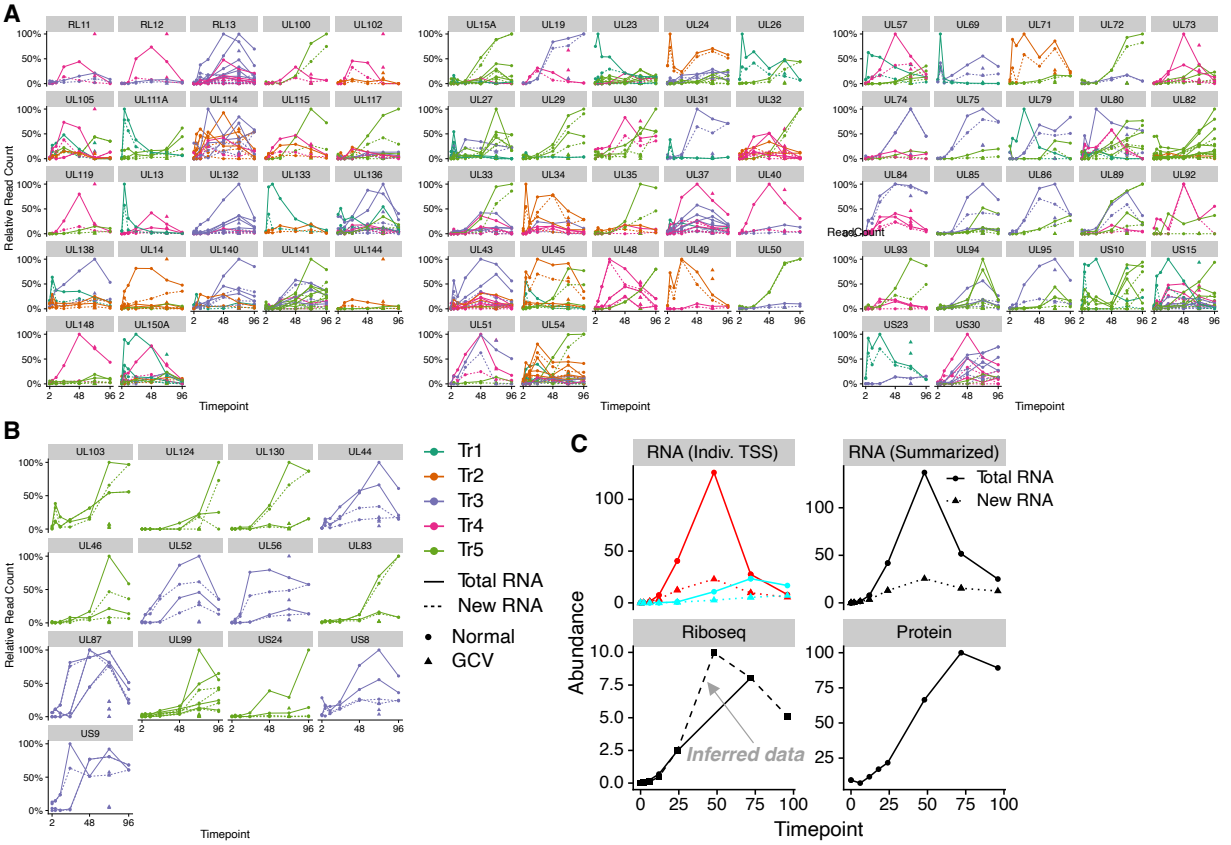


Figure 6.11: **A)** Transcription abundance profiles for all ORFs (n=66) with at least one subdominant TSS clustered into a different Tr-class than the dominant TSS. Individual TSS are colored based on their Tr-class. **B)** Transcription abundance profiles of all ORFs (n=13) with at least one subdominant TSS, where all TSS are clustered into the same Tr-class. **C)** Individual TSS expression (top left), summarized TSS expression (top right), Ribo-seq expression (bottom left) and protein abundance (top right) for the UL119 protein. Total RNA (solid line) and newly synthesized RNA levels (dotted line) are indicated. For the Ribo-seq expression, the solid line indicates the measured expression, and the dashed line indicates the inferred data based on our modelling approach (see Methods), including the 48 hpi timepoint, which is not present in the Ribo-seq data.

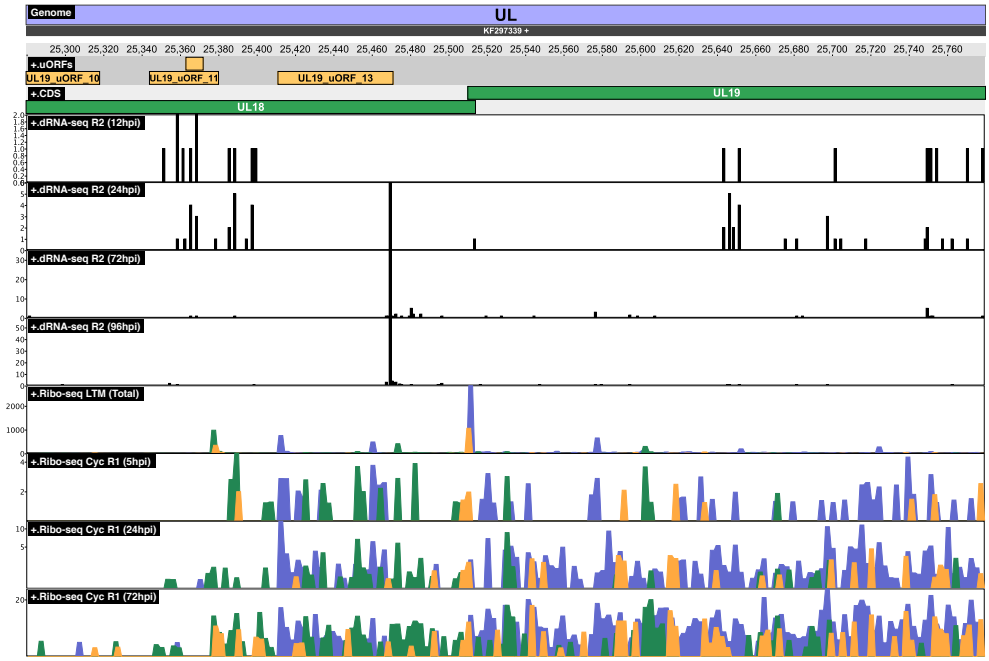


Figure 6.12: The two TSS of UL19 are shown in the genome browser. The more upstream TSS exhibits a dispersed distribution of initiation, whereas the second TSS exhibits focused initiation. Codon counts for the respective three frames are shown. Tracks show from top to bottom: The predicted uORFs by PRICE [Erhard et al., 2018], the canonical large ORFs denoted as CDS (coding sequences), the dRNA-seq 5'-counts for the timepoints 12, 24, 72 and 96 hpi, the totalized codon counts for all three frames of the LTM treated Ribo-seq sample (linear scale) and finally the codon counts for all three frames of the cycloheximide treated Ribo-seq samples for the timepoints 5, 24 and 72hpi (log-scale).

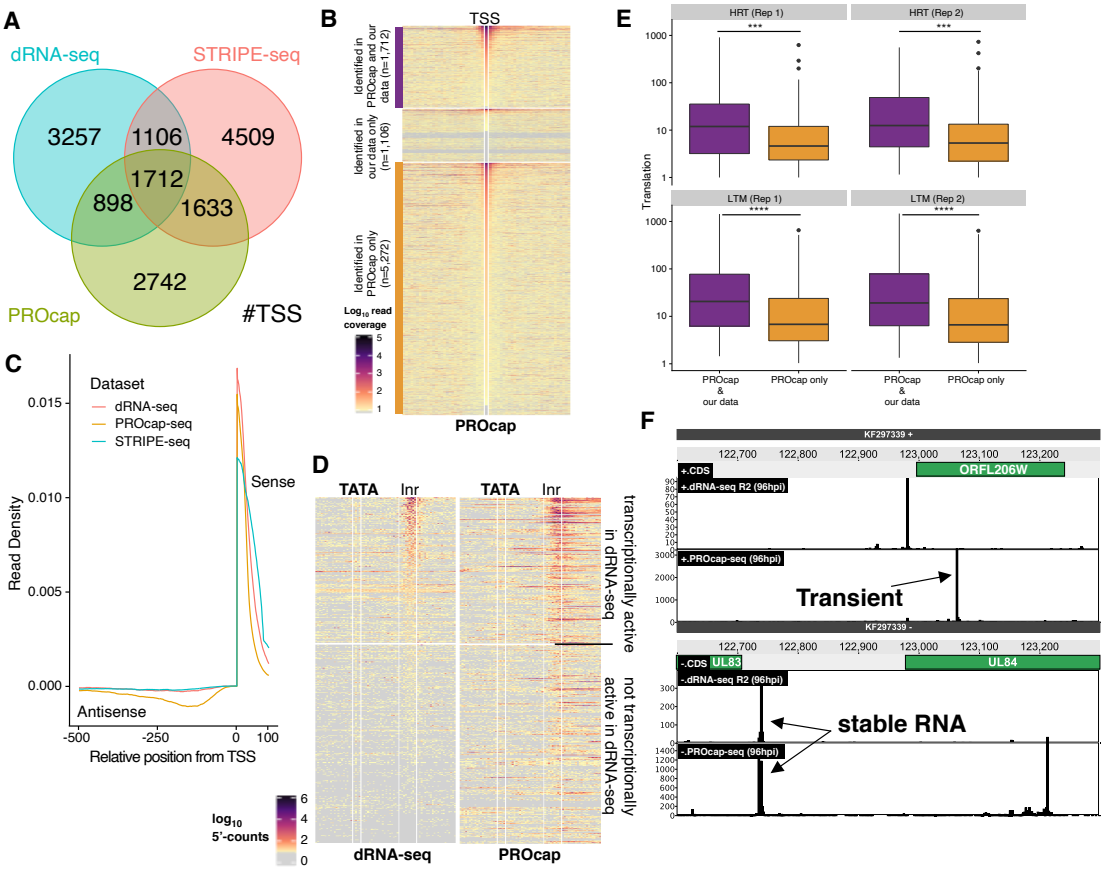


Figure 6.13: **A)** Venn-Diagram depicting the overlap of TSS found in PROcap with the TSS found in dRNA-seq and STRIPE-seq, respectively. TSS in PROcap were predicted using TSRFinder [Parida et al., 2019] on reads mapped to the Towne strain of HCMV and subsequently mapped to the Lisa strain (see Methods). TSS in dRNA-seq and STRIPE-seq were predicted using iTiSS [Jürges et al., 2021]. **B)** Heatmaps of the read densities for +/- 30bp around TSS from the 96 hpi timepoint in the PROcap (left) and dRNA-seq (right) data. The upper part shows TSS identified in both PROcap and dRNA-seq data, the middle part TSS identified in dRNA-seq only, and the lower part TSS only identified in the PROcap data. The rows in each part are sorted by the read counts in the PROcap data. **C)** Metagene plot for dRNA-seq (red), PROcap (yellow) and STRIPE-seq (blue) of cellular TSS found at annotated protein coding transcripts. Downstream of the TSS read counts are shown for the sense strand, upstream of the TSS the read counts are shown for the antisense strand. **D)** Heatmaps showing the total 5'-read counts over all timepoints in the dRNA-seq data (left) and the total 5'-read counts of the 96 hpi timepoint in the PROcap data (right) around all TATA-motifs found in the HCMV genome. The TATA-motifs as well as the expected region of the TSS are indicated. The heatmap is sorted based on the number of reads inside the Inr-region of the dRNA-seq data. The point at which the Inr-region contains ≥ 10 reads in the dRNA-seq dataset is indicated.

Figure 6.13: **E)** Comparison of the translation rates downstream of TSS identified in both data sets (PROcap and dRNA-seq) or only in PROcap inferred from the harringtonine (HRT) or lactimidomicin (LTM) treated Ribo-seq samples. PROcap only TSS were selected to match the expression strength of TSS identified in both data sets. The differences are significant (Mann-Whitney-U test; ***: p< 1.0^{-3}, ****: p< 1.0^{-4}; p-values: HRT (Rep 1)=1.2^{-4}, HRT (Rep 2)=1.2^{-4}, LTM (Rep 1)=6.5^{-6}, LTM (Rep 2)=5.7^{-6}) **F)** UL83/UL84 locus in the genome browser. TSS giving rise to transient and stable RNAs are indicated. Tracks show from top to bottom: The dRNA-seq 5'-counts of the 96 hpi timepoint on the plus strand, the 5'-counts of the 96 hpi timepoint of the PROcap-seq data on the plus strand, the dRNA-seq 5'-counts of the 96 hpi timepoint on the minus strand and finally the 5'-counts of the 96 hpi timepoint of the PROcap-seq data on the minus strand.

In total ≈ 70% of all TATW motifs with corresponding PyPu motifs initiate transcription indicating that the presence of these two motifs in defined proximity is indeed sufficient to initiate transcription but that additional motifs or signals are required to promote efficient transcription elongation of stable viral transcripts.

Only stable viral transcripts detectable by dRNA-seq and STRIPE-seq in total RNA should contribute to the viral translatome. We thus utilized the available Ribo-seq data to analyze translation initiation at the first AUG start codon downstream of individual TSS (see Methods). Indeed, PROcap-only TSS showed significantly weaker translation initiation rates (p-value < 1.3^{-4}, Mann-Whitney-U-Test) downstream compared to TSSs from stable viral transcripts (p-value > 0.36, Mann-Whitney-U-Test) for equally well transcribed TSS in the PRO-cap data (Fig. 6.13E and Fig. 6.14C,D).

Of the 5,272 PROcap-only TSS in the HCMV genome, i.e., candidate sites for transient transcription, 3,268 (62%) were within 100-350 bp antisense upstream of an identified TSS. Jointly, these 250 bp antisense upstream regions cover 83.9% of the HCMV genome. For example, ≈ 300bp upstream of the TSS for UL83, we identified a PRO-cap peak with read levels similar to the UL83 TSS (n = 2,183 reads for the TSS, n = 3,233 reads for the PROMPT; Fig. 6.13D). While the TSS was well covered with dRNA-seq reads (n = 398), only a single dRNA-seq read was located at the PROMPT. Interestingly, an additional TSS was clearly visible 60 bp upstream of the PROMPT that was well represented in the dRNA-seq data (n = 92 reads) and corresponds to the mRNA for the non-canonical ORF ORFL206W [Stern-Ginossar et al., 2012], highlighting the complexity of transcription initiation in the HCMV genome.

Finally, we hypothesized that at least a small fraction of pervasive transcription should be captured by dRNA/STRIPE-seq. In this case, reads matching to the respective TSS should predominantly be labeled with 4sU and show significantly higher U→C conversion rates. We thus first collected all dRNA-seq and STRIPE-seq reads that coincided with strong (≥1,000 reads) PROcap-only TSS. We included the set of dRNA-seq and STRIPE-seq reads 25 bp further upstream to evaluate the influence of background. Overall, we found 221 reads in our dRNA-seq and 45 reads in our STRIPE-seq data at PRO-cap only sites,

and only 1 background read in dRNA-seq and 6 in STRIPE-seq. At PRO-cap only sites, 33.6% of the dRNA-seq reads had at least one T→C conversion. Considering read-length and U→C conversion rates, this corresponded to a new-to-total RNA ratio (NTR) of 0.57 to 1.0 with a mean NTR of 0.78. We concluded that a large number of PRO-cap-only TSS (presumably > 50%) represent bona-fide sites of transcription initiation giving rise to short-lived viral RNAs that do not contribute to the viral translatome and thus resemble viral pervasive transcription.

6.4 Discussion

The temporal kinetics of the HCMV proteome during lytic infection have been extensively studied over the past decades using both inhibitors of translation or genome replication as well as quantitative mass spectrometry. However, the corresponding quantitative behavior throughout infection of the corresponding mRNAs remained largely elusive. Second generation sequencing-based RNA-seq experiments could not reliably differentiate between overlapping transcript isoforms present in large numbers in the densely packed HCMV genome [Lurain et al., 2006, Lee et al., 2016]. Third generation sequencing approaches are, in principle, able to sequence full-length mRNAs, but still lack the throughput to be able to identify mRNAs over the whole dynamic range of viral gene expression spanning multiple orders of magnitude [Martí-Carreras and Maes, 2019]. Moreover, transcript isoforms with alternative TSS cannot be distinguished from truncated RNAs during library preparation [Sessegolo et al., 2019]. We solved these problems by performing TSS profiling time course experiments. As demonstrated previously [Whisnant et al., 2020], combining multiple TSS profiling protocols was important to minimize false positive identifications due to protocol specific artifacts. Moreover, here we combined TSS profiling with metabolic labeling of RNA to identify TSS that are turned off during the lytic infection cycle, which would otherwise not be identifiable due to RNA half lives of many hours.

Hallmarks of strong mammalian promoters are TATA boxes as well as either the PyPu element (Inr) or the TCT motif [Blake et al., 1990, Parry et al., 2010]. The TCT motif is a cis-regulatory RNA element (known as 5'-TOP motif), which predominantly occurs in RNAs encoding for translation factors and ribosomal proteins. Available data suggest that, upon mTOR inhibition, the LARP1 protein binds to the 5'TOP motif to inhibit translation of the corresponding downstream ORFs [Philippe et al., 2020]. Cellular stress such as virus infection inhibits mTOR activity, and we speculate that viral evolution favored PyPu initiation over the TCT motif to circumvent LARP1-dependent inhibition of viral RNAs. Moreover, viral promoters contained TATA boxes to a similar extent and at the same positions as strong cellular promoters, and had the same positive correlation with expression strength. Interestingly, cellular but not viral TATA box promoters were inhibited over the course of infection. This inhibition occurred in two marked drops, one in between 6 and 12 hpi, and a second in between 24 and 72 hpi. This suggests that HCMV recruits TFIID to the viral genome already in the early phase, an effect that is amplified upon viral genome replication. During the late phase of infection, many active promoters contain

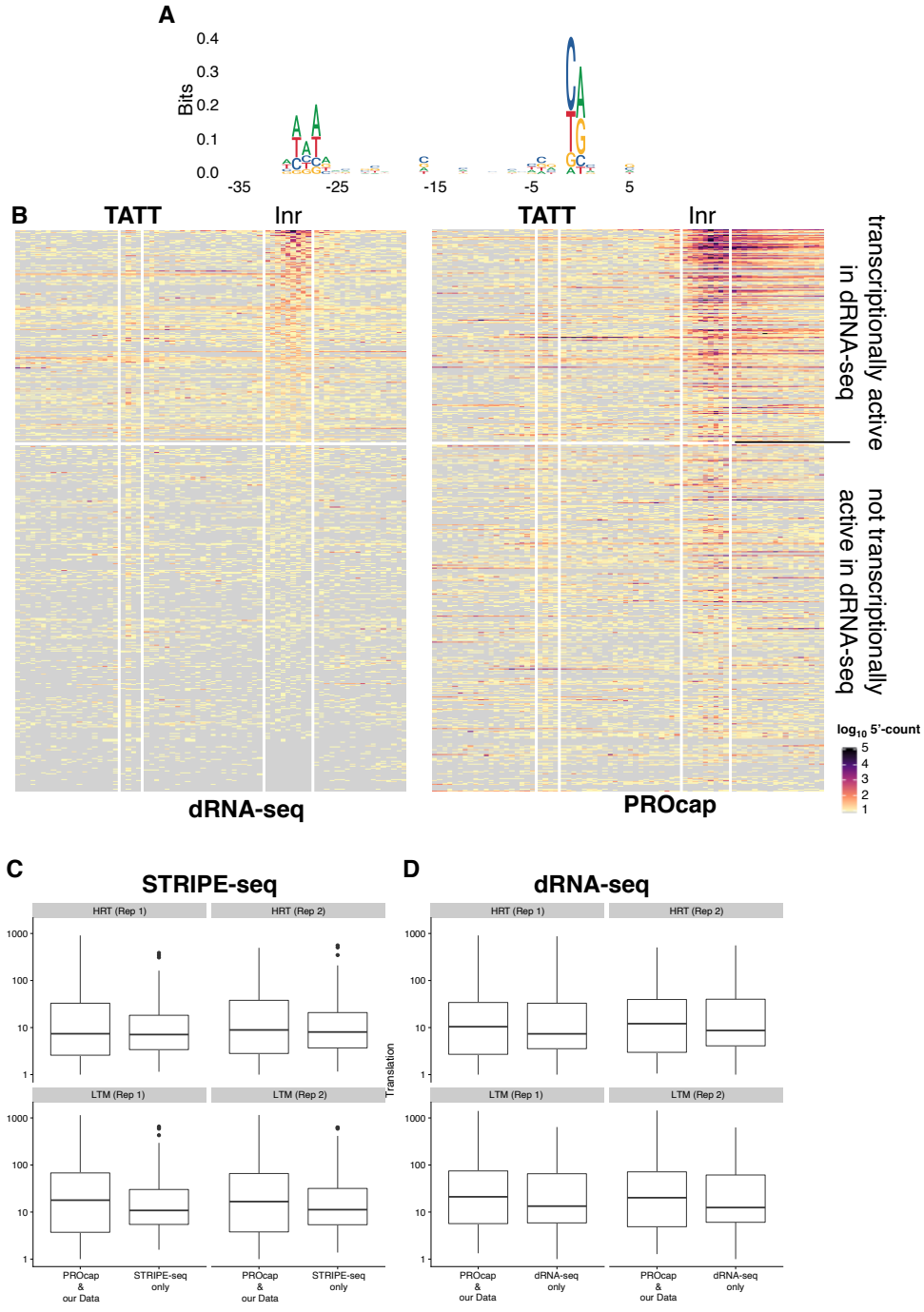

A
Bits
0.4
0.3
0.2
0.1
0.0
-35
-25
-15
-5
5
B
TATT
Inr
TATT
Inr
transcriptionally active in dRNA-seq
not transcriptionally active in dRNA-seq
log10 5'-count
5
4
3
2
1
dRNA-seq
PROcap
C
STRIPE-seq
HRT (Rep 1)
HRT (Rep 2)
LTM (Rep 1)
LTM (Rep 2)
Translation
1000
100
10
1
PROcap & our Data
STRIPE-seq only
D
dRNA-seq
HRT (Rep 1)
HRT (Rep 2)
LTM (Rep 1)
LTM (Rep 2)
PROcap & our Data
dRNA-seq only

Figure 6.14: **A)** Sequence logo for the top 100 most strongly transcribed TSS validated by both (dRNA-seq & STRIPE-seq) datasets, but not by PROcap-seq. **B)** Heatmaps showing the total 5'-read counts over all timepoints in the dRNA-seq data (left) and the total 5'-read counts of the 96 hpi timepoint in the PROcap data (right) around all TATT-motifs found in the HCMV genome. The TATT-motifs as well as the expected region of the TSS are indicated. The heatmap is sorted based on the number of reads inside the Inr-region of the dRNA-seq data. The point at which the Inr-region contains ≥ 10 reads in the dRNA-seq dataset is indicated. **C)** Comparison of the translation rates downstream of TSS identified in both data sets (PROcap and dRNA-seq) or only in dRNA-seq inferred from the harringtonine (HRT) or lactimidomicin (LTM) treated Ribo-seq samples. dRNA-seq only TSS were selected to match the expression strength of TSS idenfied in both data sets. The differences are not significant (Mann-Whitney-U test; p-values: HRT (Rep 1)=0.89, HRT (Rep 2)=0.74, LTM (Rep 1)=0.37, LTM (Rep 2)=0.39). **D)** Comparison of the translation rates downstream of TSS identified in both data sets (PROcap and STRIPE-seq) or only in STRIPE-seq inferred from the harringtonine (HRT) or lactimidomicin (LTM) treated Ribo-seq samples. STRIPE-seq only TSS were selected to match the expression strength of TSS idenfied in both data sets. The differences are not significant (Mann-Whitney-U test; p-values: HRT (Rep 1)=0.97, HRT (Rep 2)=0.77, LTM (Rep 1)=0.47, LTM (Rep 2)=0.47).

TATT boxes. TATT promoters exhibit the same correlation with expression strength as TATA boxes and occur in similar frequencies among the promoters of strongly expressed viral late genes. Interestingly, however, their position is shifted 2 bp away from the TSS. This indicates that the LTF, which has been shown to bind to TATT sequences [Davis et al., 2015], is structurally distinct to TFIID.

Previously, viral late genes have been subdivided into early-late and true-late genes. Early-late genes are expressed weakly before viral genome replication, while true late genes fully depend on viral genome replication. Several lines of evidence indicate that already early-late gene expression prior to viral DNA replication is mediated by TATT: Firstly, TATT boxes were already significantly enriched among the top 500 promoters at 24 hpi, i.e. before viral genome replication is strongly initiated, and under GCV treatment. Secondly, the TATT-dependent TSS exhibit a substantial accumulation of TSS profiling reads already at 12 hpi and under GCV treatment. Finally, the TLS score per gene showed that most TATT-TSS are already weakly expressed at 12 hpi and also under GCV treatment. Moreover, we also found TATA promoters with the same kinetic behavior as TATT promoters. The analysis of the positional shift of these TATA boxes suggested that promiscuous binding of LTF and not TFIID results in their early late kinetics. Both findings, early-late TATT and TATA boxes are consistent with recent PROseq data upon depletion of LTF [Li et al., 2021]. In contrast to PROseq, our data highlights that the corresponding transcripts are indeed stable and that this phenomenon is not restricted to transiently transcribed RNAs.

A minuscule percentage of TATA sequences in the human genome represents an actual

promoter that drives expression. One major reason for that is that many of the TATA sequences occur in regions of densely packed chromatin. In the HCMV genome the situation is different: While the formation of chromatin has been shown, it is believed to be not as compact as in the human genome [Sinclair, 2010]. Our analysis reflects this belief, as, although, only about 40% of all TATW motifs gave rise to stable transcription, additional 30% showed signs of transient transcription initiation events. Conclusively, most of HCMV's genome seems to be accessible and TATW-boxes with downstream PyPu-elements driving initiation events. However, the mechanism, which governs either stable or transient transcription remains unclear. We found evidence of additional sequence elements in proximity to TATW boxes, but these were not able to fully differentiate between active and inactive TATW.

In addition to measuring transcriptional activity instead of RNA levels, our metabolic labeling approach enabled us to accurately quantify virion-associated RNA. RNA incoming with virus particles has previously been shown to be translated [Bresnahan and Shenk, 2000]. However, whether virion-associated RNA is as efficiently translated as de novo transcribed viral RNA remained unclear. Efficient translation depends on effective formation of polysomes that in turn depend on a number of RNA binding proteins including the cap binding protein and poly(A) binding proteins that are partly loaded onto the mRNA during processing [Hinnebusch and Lorsch, 2012]. Whether or not the incoming RNAs efficiently enter polysomes is unclear. Integrative analysis with Ribo-seq data suggested that the translational efficiency of virion-associated RNA is indeed 2-8x lower than for de novo transcribed mRNA in the same sample, arguing against efficient loading of ribosomes.

Overall, 1,331 of the 2,668 TSS we identified were located upstream of one of the 168 canonical ORFs. We found 79 of the canonical HCMV ORFs to be translated from multiple transcript isoforms with distinct TSS. Interestingly, the kinetics of the dominant TSS matched the kinetics of the protein only for 70 TSS and the kinetics of protein levels could only be explained by combining translation from more than one transcript isoform for 41 proteins. For these proteins, expression during the different phases of infection was governed by distinct TSS. This does not only imply two independent promoters that exhibit specific regulation over time, but also the potential of additional post-transcriptional differential regulation, since the translational efficiency of an mRNA can be influenced by its 5'-untranslated region (5'-UTR) [Barbosa et al., 2013]. Furthermore, for UL119 we uncovered a switch from a shorter transcript isoform expressed in the early phase to a longer transcript isoform in the late phase. This switch resulted in the expression of an N-terminally extended protein isoform late in infection. Interestingly, UL119 belongs to the viral Fc-gamma receptor-like proteins and enters the secretory pathway via a canonical signal peptide. The 12 amino acid long N-terminal extension does not disrupt the signal peptide according to SignalP 5.0 predictions (Fig. 6.15A/B). Nevertheless, the signal peptide sequence is believed to dictate secretion efficiency [Kober et al., 2013]. We therefore speculate that the N-terminal extension represents a mechanism to fine-tuning the abundance of UL119 at the cell surface late in infection.

Recently published PROcap data suggested over 7,000 TSRs in the HCMV genome [Parida

et al., 2019], i.e. a site of transcription initiation every 65 nucleotides on average on both strands of the 230 kb genome. Integration with our TSS profiling data revealed that the majority of the 7,000 PROcap TSS represent initiation events presumably represent pervasive transcription that does not give rise to stable viral mRNAs and does not contribute to the viral translatome (as depicted by Ribo-seq). Such transient transcription events are well known to occur in the host genome in antisense direction at promoters (PROMPTs) or at enhancers in both directions (eRNAs). Although, we also see such distinct transient initiation events in HCMV (e.g. TATW-boxes with downstream PyPu), overall, transient initiation events are much more pervasive in viral as opposed to transient initiation at very defined loci in the host. This is most likely a consequence of differences in the chromatin structure of viral and host genome.

We provide access to our unified model of viral gene expression on the genomic level via a custom genome browser, as well as on a descriptive level via an interactive web interface displaying the quantitative behavior of transcription, RNA levels, translation and protein levels over time. These tools will facilitate further studies of HCMV gene expression as well as their regulation.

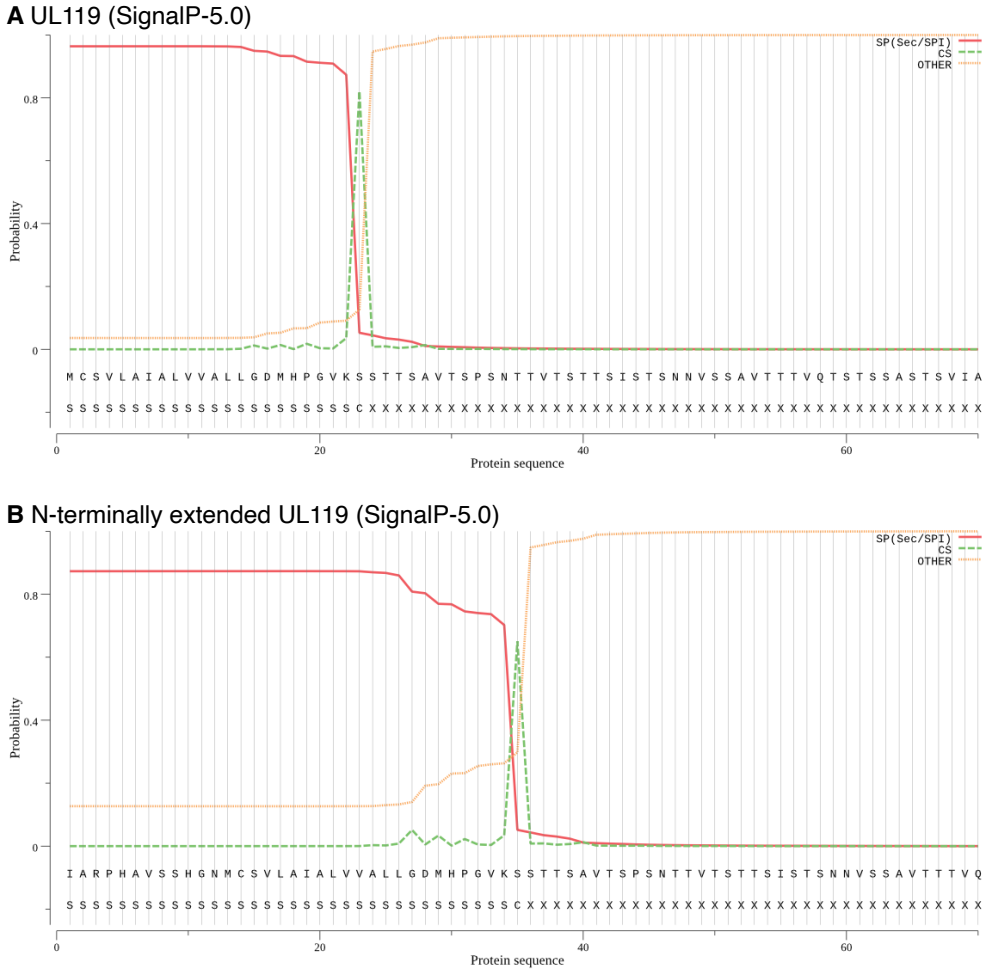


Figure 6.15: **A,B)** Predictions of the locations of cleavage sites in the location of signaling peptides in (A) the canonical UL119 and (B) the N-terminally extended protein UL119. Plots were generated using SingalP 5.0 [Hagemeier et al., 1992].

6.5 Methods

6.5.1 Cell culture, viruses, and infections

Primary human foreskin fibroblasts (HFFs) were cultured in DMEM (Dulbecco's Modified Eagle Medium) supplemented with 10% FBS (fetal bovine serum). HFFs were seeded on six well plates and infected with HCMV strains – TB40E Lisa and BAC2 (AD169VarL) at MOI:10. 4sU labeling at 500uM was conducted 1 hour prior to lysis at respective time points. Treatment with GCV (ganciclovir) and PAA (phosphonoacetic acid) was performed at 12hpi using 50 μM and 250 μg/mL respectively.

6.5.2 RNA extraction and TSS profiling

Cells were lysed in 1mL Trizol and RNA extraction was performed with Zymo research Direct-Zol^{TM} RNA Miniprep Plus kit as per manufacturer's instructions. Prior to TSS profiling, 4sU labelled RNA was subject to alkylation (SLAM) as previously described [Herzog et al., 2017]. Briefly, RNA was treated with 10mM Iodoacetamide-50% DMSO solution at 50°C for 15 minutes in 50mM PBS (phosphate buffered saline). The reaction was quenched using excess DTT. RNA was purified using Qiagen RNeasy® Minelute® Cleanup kit and eluted in nuclease free water.

TSS profiling was conducted using dRNA-Seq and STRIPE-Seq. dRNA-Seq was conducted as described previously [Sharma and Vogel, 2014] with minor modifications introduced by the Core Unit Systems Medicine (Würzburg)[Železnjak et al., 2019]. Briefly, isolated RNA was subject to 3' dephosphorylation prior to enzymatic treatment with Xrn-1 nuclease, which led to strong enrichment of 5' capped RNA. De-capping was conducted to mediate adaptor ligation using RppH (NEB) followed by library preparation using NEBNext® Multiplex Small RNA Library Prep (Illumina). Pair-ended 2x75 bp sequencing was performed on the NextSeq 500 (Illumina) at Core Unit Systems Medicine.

6.5.3 STRIPE-seq with SLAM-seq

STRIPE-seq libraries were prepared from 160 ng of iodoacetamide treated RNA according to the protocol described in Policastro et al [Policastro et al., 2020]. Briefly, RNA was subjected to terminator exonuclease treatment, template switching reverse transcription followed by PCR amplification and size selection. Libraries were then sequenced with the paired-end mode for 100 cycles on the DNBSEQ-G400 platform at BGI Tech Solutions, Hong Kong.

6.5.4 Data analysis, statistics, and reproducibility

In the following, the bioinformatical steps taken to analyze the data and produce the Figures are explained in detail. Additionally, we provide all the respective scripts and source code at Zenodo to provide full information about the exact parameters used for the individual programs as well as enabling the full reproduction of all our analysis starting from the raw data.

For second generation transcriptomic sequencing data, adapter sequences were trimmed using Trimmomatic [Bolger et al., 2014] (v. 0.39). Resulting reads with a length <18 were discarded. The remaining reads were mapped to the ribosomal RNA of homo sapiens using bowtie2 [Langmead and Salzberg, 2012] (v. 2.3.0) with standard parameters. Only the unmappable reads were subsequently mapped against the combined reference genome of homo sapiens (Ensembl hg38 version 90) and HCMV (strain TB40E-Lisa, GenBank accession number KF297339.1; strain Towne for PROcap only, GenBank accession number FJ616285.1) using STAR [Dobin et al., 2013] in paired-end mode. Samtools [Li et al., 2009] (v. 1.9) was used to sort and index the final BAM-file. For the subsequent analysis steps the gedi toolkit (https://github.com/erhard-lab/gedi) was used. The BAM-files were converted into CIT-files, which can combine multiple BAM-files into a single file, considerably saving memory and access-times. STRIPE-seq data as well as PROcap-seq data contain unique molecular identifiers (UMIs). These were accounted for and removed using the same mismatch correction algorithm as UMItools [Smith et al., 2017]. For STRIPE-seq, all reads having a 5'-softclip of length > 3 were removed [Policastro et al., 2020]. Further, to reduce error-rates in STRIPE-seq, we only considered reads with at least one additional duplicate. In case of mismatching nucleotides between duplicates, a majority-vote was used to determine the correct one.

Third generation sequencing reads were trimmed and oriented based on the presence of the 5'-adapter and poly(A)-tail. The remaining reads were mapped using GMAP [Wu and Watanabe, 2005]. The consecutive indexing, sorting and conversion into CIT-files was identical to second generation sequencing data as described above.

For the Ribo-seq data, adapter sequences were searched for using minion. Adapters were trimmed using reaper (minion and reaper are from the kraken v. 15.065 toolkit [Davis et al., 2013]). Bowtie [Langmead et al., 2009] (v. 1.2) was used to first align against the homo sapiens rRNA genome and secondly, remaining unaligned reads against the homo sapiens (Ensembl hg38 version 90) genome. Sorting, indexing and conversion into CIT-files remains the same as for the aforementioned sequencing data.

For the calculation of the enrichment at TSS, reads +/1 bp around the TSS were collected as well as reads in the 100 bp up- and downstream window. A read was considered in one of these three windows, if and only if its 5'-end was located inside it. Read counts of the respective windows were normalized and the down- and upstream enrichment was calculated by dividing the TSS-read count by the up- or downstream window read count, respectively. The final enrichment value for each dataset is derived by the median of all the TSS enrichment values.

Correlation between datasets was calculated by extracting the number of reads per bp in the HCMV genome. The read counts per bp were normalized by the total amount of mappable reads per dataset and multiplied by 1,000,000. Finally, the Pearson correlation coefficient of the $\log_{10}$ transformed read counts per bp between replicates was calculated. Positions with 0 reads were discarded.

Thymine to Cytosine (T→C) conversion rates as well as error rates and new to total RNA rates (NTRs) for TSS were estimated using GRAND-SLAM [Jürges et al., 2018]. Only reads, which 5'-ends were inside a predicted TSS were considered. As GRAND-SLAM does not offer this functionality, we had to add this feature manually. To decrease error-rates even further, only the overlapping parts of read-pairs were considered for the dRNA-seq sample. For STRIPE-seq, all positions were used as they are already error-corrected during the deduplication process (see above).

Transcription start sites (TSS) were called using iTiSS [Jürges et al., 2021] (v. 1.3) with the SPARSE_PEAK module. Reads of the individual replicates per timepoint were pooled. Subsequently, for each dataset, TSS were merged with a +/- 10 bp windows using iTiSS' sub-program TiSSMerger2. Finally, only TSS validated by both (STRIPE-seq and dRNA-seq) datasets were considered bona-fide TiSS. Since early timepoints for our STRIPE-seq sample were not available, we additionally included TSS from the 2 hpi and 6 hpi timepoints of dRNA-seq, if they were validated by at least 3 of the 4 replicates.

Sequence logos were generated using the R-package 'ggseqlogo' [Wagih, 2017]. TSS per timepoint were sorted based on their total read count or newly synthesized read count (NTR·totalRNA) and in descending order, respectively. For the prediction of promoter sequences, MEME [Bailey and Elkan, 1994] (v. 5.3.3) was used with *-minw 4, -maxw 15, -objfun de*, and *-mod zoops* parameters. We used multiple different window sizes around the TSS (-50 to 10, -100 to 10, -200 to 10, -10 to 100, -100 to 100) as well as running MEME only on the top 100 and top 500 most strongly expressed TSS. We furthermore performed these analyses also after excluding TSS that are located in close proximity to other TSS, which might mislead the motif discovery algorithm.

For each TSS we searched for position specific promoters. The promoters we searched for as well as their respective sequences and windows relative to the TSS are as follows (of note: the TSS has position 0): TATA-box (Sequence: TATA, Window: -32 to -24), TATT-box (Sequence: TATT, Window: -35 to -27), PyPu (Sequence: YR, Window: -1 to 0), Inr extended (Sequence: BBCABW, Window: -3, 3), BRE upstream (Sequence: SSRCGCC, Window: -38 to -32), BRE downstream (Sequence: RTDKKKK, Window: -24 to -17), DRE (Sequence: WATCGATW, Window: -100 to -1), TCT (Sequence: YYCTTTYY, Window: -2 to 6). The occurrence per TSS-rank plot (Fig. 2B) was generated by sorting the top 500 TSS in descending order based on their newly synthesized read count per time point. Then, in descending order the fraction of observed PyPu, TATA or TATT, respectively, up to the current rank is calculated and plotted. As for the first few ranks the fraction alternates drastically, the first 20 ranks were discarded per timepoint. To test the differences in PyPu occurrence frequencies between HCMV and cellular promoters, we compared the top 500 viral TSS for each timepoint against the top 500 cellular TSS

during the mock timepoint. For each comparison the Fisher's Exact test was used with a=number of viral TSS with PyPu, b=number of viral TSS without PyPu, c=number of cellular TSS with PyPu, d=number of cellular TSS without PyPu. For the comparison of PyPu and TATW-box occurrences between strongly expressed TSS and lower expressed TSS, we binned the TSS for each timepoint based on the expression strength into 5 bins, each containing 100 TSS. Then, we compared the first bin (top 100 most strongly expressed TSS) with the 5th bin by using the Fisher's Exact Test with a=number of TSS with PyPu or TATW, respectively, in bin 1, b= number of TSS without PyPu or TATW, respectively, in bin 1, c= number of TSS with PyPu or TATW, respectively, in bin 2, d= number of TSS without PyPu or TATW, respectively, in bin 2. To test the presence of a shift of location of the TATT-box compared to the TATA-box in virus, we used the Fisher's Exact Test with a=number of TATA-boxes starting between 33 and 35 bp upstream of the TSS, b=number of TATA-boxes starting between 27 and 29 bp upstream of the TSS, c=number of TATT-boxes starting between 33 and 35 bp upstream of the TSS, d=number of TATT-boxes starting between 27 and 29 bp upstream of the TSS. To test the decrease if TATA occurrences in human over the course of infection, the Fisher's Exact Test was used with a=number of TSS with TATA during the mock timepoint, b=number of TSS without TATA during the mock timepoint, c=number of TSS with TATA during the 96 hpi timepoint, d=number of TSS without TATA during the 96hpi timepoint.

For the generation of the TATW-heatmaps we used the tool MetagenePlot (`https://github.com/erhard-lab/MetagenePlot`). TATW-boxes were centered based on their first 'T'. The 5'-counts of all reads over all time points were pooled and are depicted in the heatmap. A TATW-box was considered to be in front of a TSS if a TSS was found 24 to 32 bp downstream of it (starting from the first 'T') for TATA and 27 to 35 bp downstream for TATT, respectively.

The True Late Score (TLS) was calculated for each TSS by dividing the sum of the normalized read counts of both replicates of the 72 hpi timepoint for the sample treated with GCV by the sum of the normalized read counts of both replicates of the 72 hpi timepoint for the sample treated without GCV.

Metagene plots were also generated by using the tool MetagenePlot. Here, we used all the TATW-boxes that had a TSS identified downstream (see above). The tool calculates the read coverage for each position. Then, each row (TATW-box and down- and upstream region) is normalized by the total read count. Subsequently, for each position in the metagene analysis, the mean normalized read coverage is plotted. Further, the metagene plots of all timepoints were combined by normalizing them such that the mean read coverage value of the region between the TATW-box and the supposed TSS-region is fixed at 0.25.

PRICE [Erhard et al., 2018] (v. 1.0.3b) was used to estimate the codon counts (number of reads assigned to a specific codon) indicating the positions of the Ribosome during translation. The codon counts per codon were normalized by the total codon count times 1,000,000.

Translational efficiencies of virion associated RNA were calculated once by obtaining the codon count throughout the coding sequence of the respective ORF for our cycloheximide

treated samples and additionally normalizing it by the ORF's length. For our samples treated with lactimidimycin, we only obtained the codon count at the respective ORF's start codon. Further, in both cases, the obtained codon count is divided by the read count at the corresponding TSS. Only ORFs were considered, where the most strongly transcribed TSS in a 1kb upstream window at any timepoint was either TATA- or TATT-associated.

For each TSS the next protein-coding downstream ORF found in a 1000 bp window and annotated in HCMV strain TB40E-Lisa (GenBank accession number KF297339.1) was considered to be its respective transcribed ORF. If no ORF was found, the TSS was not assigned any ORF. For each ORF, the assigned TSS with the maximal normalized read count at any timepoint was considered its dominant TSS. Any TSS, where the normalized read count during any timepoint did not exceed 10% of the dominant TSS for the respective ORFs were discarded.

Protein abundances were obtained from Table S2 of Weekes et al. [Weekes et al., 2014] and their associated temporal classes from Table S6C. Table S6C were missing some temporal classifications, which we manually added based on their Data S1 plots. Comparison of transcriptional expression and the protein abundances was performed for all proteins, where at least one TSS was found for. Per protein the read counts over all timepoints were normalized by the maximum read count. Then, the mean read-count per timepoint over all TSS of the respective temporal class (taken from the associated protein) are plotted.

The transcription classes (Tr) were assigned to the TSS by first clustering the dominant TSS for each ORF and subsequently using the resulting centromeres to cluster the remaining TSS. For each dominant TSS the read counts of all timepoints were normalized by dividing them by the maximum read count of any timepoint for the respective TSS (i.e. the timepoint with the highest read count is subsequently 1). Then, the KMeans++ clustering algorithm [Arthur and Vassilvitskii, 2007] was used on the 6, 12, 48, 72 and 96 hpi timepoint with k=5 to cluster the TSS into 5 transcriptional classes. The clustering was performed multiple times which always resulted in similar profiles. Consequently, in order to assure a deterministic clustering to ease subsequent automatic analysis steps, we set the seed to 42 for the KMeans++ algorithm.

To compute the correlation of the temporal Ribo-seq profile with the combined TSS profiles, we used non-linear least squares regression (nnls package version 1.4) to fit a model predicting the overall translational activity of an ORF based on the translational efficiency of each of its TSS. In this model, the Ribo-seq signal of the main ORF M_j in sample j is related to the dRNA-seq signal $u_{i,j}$ of TSS i in sample j and the translation rate R_i of the main ORF from TSS i via $M_j = \sum_i u_{i,j} R_i$. When there are more samples than TSS, this is an overdetermined system of linear equations and the non-negative coefficients R_i can be estimated by regression. The correlation was then computed for M_j vs. $\sum_i u_{i,j} R_i$.

TSS were predicted as described by Parida et al., by first aligning the PROcap-seq data to the Towne strain of HCMV (GenBank accession number FJ616285.1), removing duplicates using the provided UMIs and then running their tool TSRFinder on it [Parida et al., 2019]. As described in their paper, we merged consecutive TSRs and took the position with the highest read count as the TSS. To make these TSS comparable to our data, which is

mapped to TB40E-Lisa, we proceeded by extracting the sequences +/- 40 bp around the TSS and mapping them to our reference genome using STAR with the "*–alignEndsType EndToEnd*" option.

For the comparison of translation rates of TSS validated by both our data and PROcap-seq data versus TSS validated only by PROcap-seq data, we sorted both kinds of TSS (PROcap-seq 96hpi and dRNA-seq 96hpi) in descending order. Then, from top to bottom, we pickled TSS pairs of both sets with similar read counts (+/-5% difference per pair). TSS, which are in close proximity (1,000 bp) to an already picked TSS were discarded (i.e. if multiple TSS are in close proximity, only the one with the highest read count was considered). Translation rates of TSS were calculated by using the codon-counts calculated using the tool PRICE [Erhard et al., 2018] in the Ribo-seq data treated with either harringtonine (HRT) or lactimidomycin (LTM), respectively. Here, similar to TSS-profiling, only the start sites of translations are covered. Subsequently, for each TSS, the translational efficiency is calculated by dividing the codon count of the codon with the highest codon count within a 1000 bp downstream window by the read count of the respective TSS.

Chapter 7

Conclusion and Outlook

Throughout this work I implemented many programs and scripts that all aimed at analyzing numerous different high-throughput sequencing datasets. As demonstrated in the previous chapters, it is important to reevaluate already existing tools whether they consider all the potential biases and leverage all the information embedded in different datasets. If not, interesting discoveries might be overlooked or, even worse, the data might be misinterpreted. Further, previous chapters also unraveled the benefits of combining the information of several sequencing experiments in order to validate individual findings as well as capturing at least parts of the whole systems biology.

I analyzed data from HSV-1 as well as HCMV by applying the tools implemented throughout this thesis. My work mainly focused on the transcriptional landscape of these herpesviruses. With iTiSS (see chapter 3 as well as [Jürges et al., 2021]) I implemented a tool for the detection of TiSS providing a novel TiSS-prediction procedure based on local read-enrichment. I demonstrated that this new approach greatly surpasses simple count based approaches such as TSRFinder [Parida et al., 2019] as well as read-cluster based approaches such as CAGEr [Haberle et al., 2015] and CAGEfightR [Thodberg et al., 2019] in integrated studies. With new TiSS-profiling sequencing methods that have emerged as recently as last year [Policastro et al., 2020], this field is very active. By making iTiSS open-source and therefore freely available, the program can help the scientific community to accurately pinpoint TiSS in a wide variety of TiSS-profling datasets.

In chapter 4 (see also [Whisnant et al., 2020]), iTiSS was further applied to help annotating the full transcriptional landscape of HSV-1. More importantly, however, we outlined the advantages of combining the information content from multiple different sequencing experiments to help to validate their individual findings. For instance, ORFs need to have an associated transcript. Consequently, ORFs predicted from our Ribo-seq data were used to validate TiSS predicted from TiSS-profiling datasets and vice versa. In total we analyzed and combined over 7 different datasets in this single study. A challenge that directly follows and is mostly overlooked by other studies is the accessibility of the resulting data. Sequencing datasets in particular often suffer from low accessibility, as they are very big in size and simply visualizing their mapped reads expects the whole interpretation process to

be done by the person looking at it. We overcame this problem by choosing to implement our own viewer, displaying processed as well as raw data all at once. Users can then add or remove information that they are not interested in with simple clicks. This viewer now displays the revised annotation of the HSV-1 genome and is now serving as the fundamental basis for any future transcriptomic study on this virus.

GRAND-SLAM [Jürges et al., 2018] is the second tool I implemented in this work. In chapter 5 (see also [Jürges et al., 2018]), I demonstrated that with its new approach GRAND-SLAM provides absolute quantifications of new to total RNA for the very first time. On top, it also outperformed previous approaches for relative quantification. With its multiple updates over the recent years GRAND-SLAM is able to accurately estimate NTRs in SLAM-seq experiments. It was further applied on single cell data from MCMV infected mouse cells. Here, GRAND-SLAM helped to decipher the heterogeneity in transcriptomes between individual cells in a study published in Nature [Erhard et al., 2019].

In the future, GRAND-SLAM might also be applied to 3rd-gen sequencing data. The new HiFi-sequencing technique recently introduced by PacBio offers improved error rates for long-read sequencing data [Wenger et al., 2019]. Up until now, high error rates are the main reason that long-read sequencing is not compatible with metabolic labeling experiments such as SLAM-seq, as these error rates make it impossible to differentiate between them and intended nucleotide conversions from incorporated nucleotides. However, with the new HiFi-sequencing technique, this might change. Being able to apply the SLAM-seq library preparation steps for long-read sequencing would open up new possibilities. First, this would combine three different sequencing techniques, namely 5'-sequencing, 3'-sequencing and SLAM-seq, rendering transcriptom annotation projects much more cost effective. Second, with the additional length per read (or full transcripts in this case), dissection between newly synthesized and old RNA would be much easier, as more thymines would be available per read. And lastly, this could further provide information about the transcriptional speed of individual transcripts, by observing the amounts of T→C mismatches along the whole reads.

In chapter 6, we combined the SLAM-seq protocol [Herzog et al., 2017] with the TiSS-profiling method dRNA-seq [Sharma and Vogel, 2014]. For the analysis, I did the same by combining the two tools introduced in this work (iTiSS & GRAND-SLAM). This approach enabled us to accurately determine NTRs for individual transcripts, which, in turn, enlightened us about HCMV's transcription regulation processes. Further, I demonstrated, once again, the importance of employing a systems biology approach by not only being fixated on our own data alone, but also incorporating publicly available data giving us insights over other aspects (translation, protein levels, ...) from the cell. With the ever increasing amounts of publicly available data, this systems biology approach will become a necessity in the future.

Finally, although we are still far away from understanding the whole systems biology of cells, this work demonstrated that by using the massive amounts of publicly available data in an integrated approach along with the data specifically generated for a particular scientific question, we can at least already take a step in this direction.

Appendix A

iTiSS

Table A.1: All TSS shown here are identified in all datasets by iTiSS. The **identified** column depicts whether an annotated TSS is present in our ground truth (1=yes, 0=no).

Reference	MaxTSS	TSR	CAGE	dRNA-seq	cRNA-seq	PROcap	MaxTSS-Score	TSR-Score	Identified
1+	163069360	163069358-163069363	1	1	1	1	3	4	1
11+	9428765	9428763-9428768	1	1	1	1	4	4	1
11+	10305072	10305070-10305101	1	1	1	1	4	4	1
12+	6867530	6867528-6867543	1	1	1	1	4	4	1
12+	27710831	27710830-27710848	1	1	1	1	3	4	1
12+	49741556	49741553-49741563	1	1	1	1	4	4	1
12+	57694140	57694140-57694163	1	1	1	1	3	4	1
12+	132031223	132031220-132031225	1	1	1	1	3	4	1
13-	44436853	44436840-44436859	1	1	1	1	4	4	1
13-	75537840	75537838-75537850	1	1	1	1	4	4	1
13+	45120504	45120497-45120512	1	1	1	1	3	4	1
14-	22919139	22919138-22919153	1	1	1	1	3	4	1
14-	67359803	67359801-67359807	1	1	1	1	4	4	1
14-	67695758	67695744-67695773	1	1	1	1	4	4	1
14-	94081201	94081195-94081208	1	1	1	1	3	4	1
14+	20455225	20455222-20455237	1	1	1	1	4	4	1
14+	20469409	20469402-20469425	1	1	1	1	4	4	1
14+	34546713	34546706-34546722	1	1	1	1	3	4	1
14+	49586578	49586575-49586630	1	1	1	1	4	4	1
14+	90397028	90397027-90397049	1	1	1	1	3	4	1
15+	39581078	39581075-39581093	1	1	1	1	4	4	1
15+	44711516	44711510-44711520	1	1	1	1	3	4	1
15+	78540443	78540436-78540459	1	1	1	1	4	4	1
15+	80779370	80779365-80779376	1	1	1	1	3	4	1
16+	70346904	70346904-70346921	1	1	1	1	3	4	1
17+	7577954	7577952-7577955	1	1	1	1	3	4	1
17+	16381090	16381085-16381111	1	1	1	1	4	4	1
18-	49491345	49491341-49491348	1	1	1	1	3	4	1
18+	12308246	12308236-12308273	1	1	1	1	4	4	1
18+	36129902	36129884-36129912	1	1	1	1	3	4	1
19+	572596	572592-572616	1	1	1	1	3	4	1
19+	2476126	2476122-2476127	1	1	1	1	4	4	1
19+	12938608	12938604-12938616	1	1	1	1	4	4	1
19+	55385931	55385930-55385939	1	1	1	1	4	4	1
21-	34526821	34526806-34526824	1	1	1	1	3	4	1
21+	8212570	8212565-8212581	1	1	1	1	4	4	1
21+	8256779	8256774-8256790	1	1	1	1	4	4	1
21+	31659689	31659684-31659706	1	1	1	1	4	4	1
3-	170870194	170870190-170870198	1	1	1	1	4	4	1
3+	186784797	186784797-186784817	1	1	1	1	4	4	1
4-	120066826	120066790-120066829	1	1	1	1	2	4	1
4+	1974635	1974631-1974636	1	1	1	1	3	4	1
4+	6640683	6640682-6640692	1	1	1	1	3	4	1
4+	56978895	56978864-56978897	1	1	1	1	3	4	1
4+	108620575	108620566-108620580	1	1	1	1	4	4	1
4+	108650630	108650620-108650637	1	1	1	1	4	4	1
5-	134176949	134176946-134176953	1	1	1	1	4	4	1
5+	55160179	55160170-55160192	1	1	1	1	4	4	1
5+	73498447	73498441-73498448	1	1	1	1	4	4	1
5+	86617940	86617936-86617958	1	1	1	1	4	4	1
5+	138753420	138753420-138753436	1	1	1	1	3	4	1
5+	140711274	140711271-140711276	1	1	1	1	4	4	1
5+	140726156	140726156-140726158	1	1	1	1	3	4	1
5+	151771953	151771950-151771960	1	1	1	1	4	4	1
6-	159693240	159693240-159693245	1	1	1	1	3	4	1
6+	27133042	27133037-27133076	1	1	1	1	4	4	1
6+	36678713	36678708-36678725	1	1	1	1	4	4	1
7-	73578578	73578563-73578580	1	1	1	1	4	4	1
7+	74174355	74174351-74174358	1	1	1	1	3	4	1
8-	119638838	119638833-119638845	1	1	1	1	4	4	1

Reference	MaxTSS	TSR	CAGE	dRNA-seq	cRNA-seq	PROcap	MaxTSS-Score	TSR-Score	Identified
10+	68332063	68332058-68332069	1	1	1	1	4	4	0
12-	124914649	124914641-124914655	1	1	1	1	4	4	0
12+	14774401	14774401-14774407	1	1	1	1	3	4	0
12+	120438117	120438111-120438126	1	1	1	1	4	4	0
13-	100540267	100540262-100540277	1	1	1	1	2	4	0
13+	21140540	21140540-21140576	1	1	1	1	3	4	0
13+	43023589	43023587-43023626	1	1	1	1	3	4	0
14+	60397728	60397724-60397732	1	1	1	1	3	4	0
14+	88979095	88979087-88979139	1	1	1	1	3	4	0
16+	57299943	57299942-57299949	1	1	1	1	3	4	0
18-	47176318	47176304-47176366	1	1	1	1	3	4	0
18+	12702971	12702967-12703044	1	1	1	1	2	4	0
18+	59899995	59899992-59900021	1	1	1	1	4	4	0
2-	231464483	231464483-231464494	1	1	1	1	4	4	0
21-	32393004	32393003-32393025	1	1	1	1	2	4	0
22+	40951377	40951373-40951388	1	1	1	1	4	4	0
3+	22381739	22381739-22381743	1	1	1	1	4	4	0
4-	113214418	113214371-113214422	1	1	1	1	3	4	0
4-	119300616	119300608-119300643	1	1	1	1	4	4	0
5+	150094281	150094279-150094302	1	1	1	1	4	4	0
5+	181207747	181207746-181207754	1	1	1	1	3	4	0
6+	26020450	26020450-26020466	1	1	1	1	3	4	0
6+	26156329	26156322-26156354	1	1	1	1	4	4	0
6+	26158124	26158120-26158127	1	1	1	1	3	4	0
6+	26251613	26251609-26251660	1	1	1	1	4	4	0
6+	26577098	26577095-26577100	1	1	1	1	3	4	0
6+	27139281	27139277-27139327	1	1	1	1	4	4	0
6+	27158126	27158119-27158134	1	1	1	1	4	4	0
6+	27824091	27824090-27824115	1	1	1	1	4	4	0
6+	27893424	27893421-27893468	1	1	1	1	4	4	0
6+	28941048	28941044-28941066	1	1	1	1	4	4	0
6+	44247100	44247092-44247106	1	1	1	1	4	4	0
7+	101127103	101127100-101127106	1	1	1	1	4	4	0
8+	66113363	66113357-66113382	1	1	1	1	4	4	0

Appendix B

Integrative function genomics decodes herpes simplex virus 1

Table B.1: List of all identified Transcripts including their location, name, gene cluster, and information about which dataset or method identified its TiSS.

Name	ID	Gene	Cluster	Strand	TiSS	TTS	Length	Exons	(i)	(ii)	(iii)	(iv)	(v)	(vi)	(vii)	(viii)	(ix)	Num.Criteria	Justification
TRL1 RNA (spliced)	TRL1_RNA	TRL1	TRL1-TRL2	+	408	5716	4408	3	1	1	0	0	1	0	0	1	1	5	
TRL1.5 RNA (spliced)	TRL1.5_RNA	TRL1.5	TRL1-TRL2	+	1750	5716	3066	3	1	1	0	0	0	0	1	1	1	5	
TRL2 RNA iso1 (with retained intron 1)	TRL2_RNA_iso1	TRL2	TRL1-TRL2	+	2117	5716	3464	2	1	1	0	0	1	1	1	1	1	7	
TRL2 RNA (spliced)	TRL2_RNA	TRL2	TRL1-TRL2	+	2117	5716	2699	3	1	1	0	0	1	1	1	1	1	7	
TRL2 RNA iso2 (NAG-NAG, spliced)	TRL2_RNA_iso2	TRL2	TRL1-TRL2	+	2117	5716	2696	3	1	1	0	0	1	1	1	1	1	7	
TRL3 RNA (orphan)	TRL3_RNA	TRL3	TRL3-UL3	+	8447	10969	2523	1	1	0	0	0	0	0	0	1	0	2	
UL1 RNA	UL1_RNA	UL1	TRL3-UL3	+	9245	10969	1725	1	1	1	1	0	1	1	1	1	1	8	
UL2 RNA	UL2_RNA	UL2	TRL3-UL3	+	9739	10969	1231	1	1	1	1	1	1	1	1	0	1	8	
UL3 RNA *1 (PacBio; read-through transcript)	UL3_RNA_*1	UL3	TRL3-UL3	+	10736	11736	1001	1	0	0	0	0	0	1	0	0	1	2	
UL3 RNA	UL3_RNA	UL3	TRL3-UL3	+	10974	11736	763	1	1	0	0	0	1	1	0	1	1	5	
UL3 RNA #1 (orphan)	UL3_RNA_#1	UL3	TRL3-UL3	+	11543	11736	194	1	1	0	0	0	0	0	0	0	0	1	Peak in both dRNA-seq replicates. Accumulation of reads within +/-5 nt of the TiSS in both cRNA-seq replicates.
UL4.5 RNA *1	UL4.5_RNA_*1	UL4.5	UL4.5-UL7	+	11784	18069	6286	1	1	0	0	0	0	0	0	0	0	1	Strong peak in both dRNA-seq replicates as well as in one of the two cRNA-seq replicates. Increase in downstream read-coverage in both cRNA-seq replicates.
UL4.5 RNA (spliced)	UL4.5_RNA	UL4.5	UL4.5-UL7	+	12210	18069	5317	2	1	0	0	0	0	0	0	0	1	2	
UL4.5 RNA iso1 (with retained intron)	UL4.5_RNA_iso1	UL4.5	UL4.5-UL7	+	12210	18069	5860	1	1	0	0	0	0	0	0	0	1	2	
UL5.5 RNA	UL5.5_RNA	UL5.5	UL4.5-UL7	+	13390	18069	4680	1	1	0	0	0	0	0	0	0	1	2	
UL5.5 RNA #1	UL5.5_RNA_#1	UL5.5	UL4.5-UL7	+	13508	18069	4562	1	1	0	0	0	0	0	1	0	1	3	
UL5.6 RNA	UL5.6_RNA	UL5.6	UL4.5-UL7	+	14012	18069	4058	1	0	1	0	0	0	0	0	0	1	2	
UL5.7 RNA	UL5.7_RNA	UL5.7	UL4.5-UL7	+	14576	18069	3494	1	0	0	0	0	0	0	0	0	1	1	Strong peak in both dRNA-seq replicates. Required for expression of UL5.7 ORF.
UL6 RNA *2	UL6_RNA_*2	UL6	UL4.5-UL7	+	14929	18069	3141	1	1	0	0	0	1	0	1	0	0	3	
UL6 RNA *1	UL6_RNA_*1	UL6	UL4.5-UL7	+	15013	18069	3057	1	0	0	0	0	1	0	0	0	1	2	
UL6 RNA	UL6_RNA	UL6	UL4.5-UL7	+	15051	18069	3019	1	1	1	0	0	1	0	0	0	1	4	
UL6 RNA #1 (encodes UL6 iORF)	UL6_RNA_#1	UL6	UL4.5-UL7	+	15747	18069	2323	1	0	0	0	0	0	0	0	0	1	1	Strong peak in both dRNA-seq replicates. Required for expression of UL6 iORF.
UL6 RNA #2 (encodes UL6 iORF)	UL6_RNA_#2	UL6	UL4.5-UL7	+	15850	18069	2220	1	0	0	0	0	1	0	0	0	1	2	
UL7 RNA *1 (PacBio only; orphan)	UL7_RNA_*1	UL7	UL4.5-UL7	+	16433	18069	1637	1	0	0	0	0	0	1	0	0	0	1	Multiple PacBio reads starting at the same position accompanied by a weak increase in read densities in the 4sU-seq data.
UL7 RNA	UL7_RNA	UL7	UL4.5-UL7	+	16956	18069	1114	1	1	1	1	0	1	1	1	1	1	8	
UL7.5 RNA	UL7.5_RNA	UL7.5	UL7.5-UL10	+	18401	24667	6267	1	0	0	0	0	0	0	0	0	1	1	Weak peak in both cRNA-samples. Required for expression of UL7.5 uORF and UL7.5 ORF
UL8.4 RNA	UL8.4_RNA	UL8.4	UL7.5-UL10	+	20437	24667	4231	1	0	0	0	0	0	0	1	1	0	2	
UL10 RNA	UL10_RNA	UL10	UL7.5-UL10	+	22943	24667	1725	1	1	1	1	0	0	1	1	1	1	7	
UL10 RNA #1	UL10_RNA_#1	UL10	UL7.5-UL10	+	23029	24667	1639	1	1	1	0	1	1	1	1	0	0	6	
UL13.5 RNA	UL13.5_RNA	UL13.5	UL13.5-UL15	+	27043	34849	7807	1	1	0	0	0	0	0	0	0	0	1	Strong peak in one cRNA-seq replicate and weak peak in the other cRNA-seq and both dRNA-seq replicates. Required for expression of UL13.5 ORF.
UL15 RNA iso1 (with retained intron 1)	UL15_RNA_iso1	UL15	UL13.5-UL15	+	28801	34849	6049	1	1	0	1	0	1	1	1	0	1	6	
UL15 RNA (spliced)	UL15_RNA	UL15	UL13.5-UL15	+	28801	34849	2463	2	1	0	1	0	1	1	1	0	1	6	
UL19.5 RNA	UL19.5_RNA	UL19.5	UL19.5-UL21	+	39645	43712	4068	1	1	0	0	0	0	0	0	0	0	1	Strong peak in both dRNA-seq replicates. Required for expression of UL19.5 sORF.
UL20.4 RNA	UL20.4_RNA	UL20.4	UL19.5-UL21	+	40822	43712	2891	1	1	0	0	0	0	0	0	1	1	3	
UL21 RNA (encodes N-terminal extension of UL21)	UL21_RNA	UL21	UL19.5-UL21	+	41862	43712	1851	1	1	1	1	0	1	1	0	1	1	7	
UL21 RNA #1 (does not encode N-terminal extension of UL21)	UL21_RNA_#1	UL21	UL19.5-UL21	+	41991	43712	1722	1	1	0	0	0	0	0	0	0	1	2	
UL22.5 RNA	UL22.5_RNA	UL22.5	UL22.5-UL26.5	+	43939	48768	4830	1	1	0	0	0	0	0	0	0	1	2	
UL23.5 RNA	UL23.5_RNA	UL23.5	UL22.5-UL26.5	+	46606	48768	2163	1	0	0	0	0	0	0	0	1	1	2	
UL24 RNA *1	UL24_RNA_*1	UL24	UL22.5-UL26.5	+	47407	48768	1362	1	1	1	0	0	1	1	1	1	1	7	
UL24 RNA	UL24_RNA	UL24	UL22.5-UL26.5	+	47671	48768	1098	1	1	0	0	0	1	1	0	0	1	4	
UL24.5 RNA	UL24.5_RNA	UL24.5	UL22.5-UL26.5	+	48073	48768	696	1	1	0	0	0	1	1	1	1	1	6	
UL25 RNA (read-through transcript)	UL25_RNA	UL25	UL22.5-UL26.5	+	48630	52793	4164	1	1	1	1	0	1	1	0	0	1	6	
UL26 RNA	UL26_RNA	UL26	UL22.5-UL26.5	+	50663	52793	2131	1	1	1	0	0	1	1	0	0	1	5	
UL26.5 RNA	UL26.5_RNA	UL26.5	UL22.5-UL26.5	+	51634	52793	1160	1	1	1	1	0	1	1	0	1	1	7	
UL27.5 RNA	UL27.5_RNA	UL27.5	UL27.5-UL30	+	54922	66582	11661	1	1	0	0	0	0	0	0	1	1	3	
UL27.6 RNA	UL27.6_RNA	UL27.6	UL27.5-UL30	+	56133	66582	10450	1	1	0	0	0	0	0	1	0	1	3	
UL28.4 RNA	UL28.4_RNA	UL28.4	UL27.5-UL30	+	57711	66582	8872	1	1	0	0	0	0	1	0	0	1	3	
UL30 RNA *1 (=Ori-L RNA)	UL30_RNA_*1	UL30	UL27.5-UL30	+	62456	66582	4127	1	0	0	0	0	0	1	0	0	1	2	
UL30 RNA	UL30_RNA	UL30	UL27.5-UL30	+	62604	66582	3979	1	0	1	1	0	1	1	1	1	1	7	
UL31.5 RNA *1	UL31.5_RNA_*1	UL31.5	UL31.5-UL35	+	67063	70963	3901	1	1	0	0	0	0	0	0	0	1	2	
UL31.5 RNA	UL31.5_RNA	UL31.5	UL31.5-UL35	+	67480	70963	3484	1	1	0	0	0	0	0	0	0	1	2	
UL31.6 RNA	UL31.6_RNA	UL31.6	UL31.5-UL35	+	67768	70963	3196	1	0	0	0	0	0	0	0	0	0	0	Peak in one cRNA-seq replicate. Required for expression of UL31.6 ORF.
UL32.5 RNA	UL32.5_RNA	UL32.5	UL31.5-UL35	+	68144	70963	2820	1	0	0	0	0	0	0	0	0	1	1	Weak peak in both cRNA-seq replicates. Required for expression of UL32.5 uORF.

UL32.6 RNA	UL32.6_RNA	UL32.6	UL31.5-UL35	+	68621	70963	2343	1	0	1	0	0	0	0	1	0	1	3	
UL33 RNA	UL33_RNA	UL33	UL31.5-UL35	+	69064	70963	1900	1	1	1	1	0	1	1	1	1	1	8	
UL34 RNA	UL34_RNA	UL34	UL31.5-UL35	+	69441	70963	1523	1	1	1	1	0	1	1	0	1	1	7	
UL35 RNA	UL35_RNA	UL35	UL31.5-UL35	+	70488	70963	476	1	1	1	0	0	1	1	1	1	1	7	
UL36.4 RNA	UL36.4_RNA	UL36.4	UL36.4-UL38	+	73207	86039	12833	1	0	0	0	0	0	0	0	0	1	1	Strong peak in one cRNA-seq replicate. Strong increase in coverage downstream of the TiSS in both cRNA-seq replicates. Required for expression of UL36.4 uORF.
UL36.5 RNA	UL36.5_RNA	UL36.5	UL36.4-UL38	+	74860	86039	11180	1	0	0	0	0	0	0	1	0	1	2	
UL36.6 RNA (may start further upstream at 79448 or 79373 according to PacBio)	UL36.6_RNA	UL36.6	UL36.4-UL38	+	79564	86039	6476	1	0	0	0	0	0	0	1	0	0	1	Strong peak in both dRNA-seq replicates. Required for expression of UL36.6 uORF and UL36.6 ORF
UL36.6 RNA iso1 (spliced)	UL36.6_RNA_iso1	UL36.6	UL36.4-UL38	+	79564	86039	6269	2	0	0	0	0	0	0	1	0	0	1	Strong peak in both dRNA-seq replicates. Required for expression of UL36.6 uORF and UL36.6 ORF
UL36.6 RNA iso2 (NAGNAG, spliced)	UL36.6_RNA_iso2	UL36.6	UL36.4-UL38	+	79564	86039	6272	2	0	0	0	0	0	0	1	0	0	1	Strong peak in both dRNA-seq replicates. Required for expression of UL36.6 uORF and UL36.6 ORF
UL38 RNA	UL38_RNA	UL38	UL36.4-UL38	+	84393	86039	1647	1	1	1	1	0	1	1	0	1	1	7	
UL39 RNA	UL39_RNA	UL39	UL39-UL40	+	86216	91000	4785	1	1	1	1	1	1	1	0	0	1	7	
UL40 RNA	UL40_RNA	UL40	UL39-UL40	+	89773	91000	1228	1	1	1	0	0	1	1	1	0	1	6	
UL40.5 RNA *1	UL40.5_RNA_*1	UL40.5	UL40.5-UL42	+	91011	94657	3647	1	1	0	0	0	0	1	0	0	1	3	
UL40.5 RNA iso1 (spliced, encodes UL40.5 iORF)	UL40.5_RNA_iso1	UL40.5	UL40.5-UL42	+	91011	94657	2373	3	1	0	0	0	0	1	0	0	1	3	
UL40.5 RNA	UL40.5_RNA	UL40.5	UL40.5-UL42	+	91076	94657	3582	1	1	1	0	0	0	0	0	0	1	3	
UL40.5 RNA #1 (encodes UL40.5 iORF)	UL40.5_RNA_#1	UL40.5	UL40.5-UL42	+	91296	94657	3362	1	1	0	0	0	0	1	0	0	1	3	
UL40.6 RNA	UL40.6_RNA	UL40.6	UL40.5-UL42	+	91523	94657	3135	1	1	0	0	0	0	1	0	0	1	3	
UL40.6 RNA iso1 (spliced, encodes UL40.7 dORF)	UL40.6_RNA_iso1	UL40.6	UL40.5-UL42	+	91523	94657	2152	2	1	0	0	0	0	1	0	0	1	3	
UL40.7 RNA	UL40.7_RNA	UL40.7	UL40.5-UL42	+	91994	94657	2664	1	1	0	0	0	0	1	0	0	1	3	
UL42 RNA	UL42_RNA	UL42	UL40.5-UL42	+	92934	94657	1724	1	1	1	1	0	1	1	1	1	1	8	
UL43 RNA	UL43_RNA	UL43	UL43	+	94764	96085	1322	1	0	1	0	1	0	0	0	0	1	3	
UL44 RNA	UL44_RNA	UL44	UL44-UL45	+	96173	98704	2532	1	1	1	1	0	0	1	0	0	1	5	
UL44 RNA iso1 (spliced)	UL44_RNA_iso1	UL44	UL44-UL45	+	96173	98704	2306	2	1	1	1	0	0	1	0	0	1	5	
UL45 RNA	UL45_RNA	UL45	UL44-UL45	+	97953	98704	752	1	1	0	0	1	1	1	1	0	1	6	
UL46.5 RNA	UL46.5_RNA	UL46.5	UL46.5-UL50	+	101286	108175	6890	1	0	0	0	0	0	0	0	0	0	0	Peak in both cRNA-seq replicates. Required for expression of UL46.5 sORF 1 and 2.
UL47.5 RNA (orphan)	UL47.5_RNA	UL47.5	UL46.5-UL50	+	103415	108175	4761	1	1	0	0	0	0	0	1	0	0	2	
UL50 RNA	UL50_RNA	UL50	UL46.5-UL50	+	106829	108175	1347	1	0	1	1	0	1	1	1	1	1	7	
UL52 RNA	UL52_RNA	UL52	UL52-UL53	+	108985	113465	4481	1	1	1	0	0	0	1	0	0	1	4	
UL53 RNA	UL53_RNA	UL53	UL52-UL53	+	111761	113465	1705	1	1	1	0	0	1	1	0	1	1	6	
UL53 RNA #1	UL53_RNA_#1	UL53	UL52-UL53	+	111943	113465	1523	1	1	0	0	0	1	1	0	0	1	4	
UL53 RNA #2 (encodes UL53 iORF)	UL53_RNA_#2	UL53	UL52-UL53	+	112511	113465	955	1	1	0	0	0	0	1	1	0	1	4	
UL54 RNA	UL54_RNA	UL54	UL54-UL55	+	113598	115302	1705	1	0	1	1	0	1	1	0	0	1	5	
UL54.4 RNA (PacBio; read-through; orphan)	UL54.4_RNA	UL54.4	UL54-UL55	+	115027	116122	1096	1	1	0	0	1	0	1	0	0	0	3	
UL55 RNA	UL55_RNA	UL55	UL54-UL55	+	115444	116122	679	1	0	1	0	0	1	1	0	1	1	5	
IRL2.5 RNA *2 (=LAT 0.7kb; =pri-miRNA-H1)	IRL2.5_RNA_*2	IRL2.5	IRL2.5	+	117948	118731	784	1	1	0	0	0	0	1	0	0	0	2	
IRL2.5 RNA *1 (=LAT 0.7kb; =pri-miRNA-H1)	IRL2.5_RNA_*1	IRL2.5	IRL2.5	+	117978	118731	754	1	1	0	0	0	0	0	0	0	0	1	Strong peak in both dRNA-seq and one cRNA-seq replicate. Weak peak in the other cRNA-seq replicate. 0.7kb LAT.
IRL2.5 RNA (=pri-miRNA-H1)	IRL2.5_RNA	IRL2.5	IRL2.5	+	118139	118731	593	1	1	0	0	0	0	0	0	1	1	3	
IRL2.5 RNA #1 (orphan)	IRL2.5_RNA_#1	IRL2.5	IRL2.5	+	118617	118731	115	1	1	0	0	0	0	0	0	1	0	2	
LAT iso1 (spliced)	LAT_iso1	LAT_gene	LAT-ORF-O	+	118806	127171	6410	2	0	0	0	0	0	0	0	0	0	0	Latency associated transcript (LAT)
LAT	LAT	LAT_gene	LAT-ORF-O	+	118806	127171	8366	1	0	0	0	0	0	0	0	0	0	0	Latency associated transcript (LAT)
pri-miRNA-H2/H7/H8/H3/H4/H5 (initiates within LAT)	pri-miRNA-H2/H7/H8/H3/H4/H5	pri-miRNA-H2/H7/H8/H3/H4/H5	LAT-ORF-O	+	121163	127171	6009	1	0	1	0	0	0	0	1	0	0	2	
pri-miRNA-H7/H8/H3/H4/H5 (initiates within LAT)	pri-miRNA-H7/H8/H3/H4/H5	pri-miRNA-H7/H8/H3/H4/H5	LAT-ORF-O	+	122774	127171	4398	1	1	0	0	0	0	0	1	1	0	3	
ORF-O/P RNA (=pri-miRNA-H3/H4/H5, inhibited by ICP4)	ORF-O/P_RNA	ORF-O/P	LAT-ORF-O	+	125048	127171	2124	1	0	1	0	0	0	0	0	0	1	2	
US1 RNA (spliced)	US1_RNA	US1	US1-US1.5	+	132129	133967	1671	2	1	1	1	1	1	1	0	0	0	6	
US1.5 RNA (Carter & Roizman, JVI 1996)	US1.5_RNA	US1.5	US1-US1.5	+	133083	133967	885	1	0	0	0	0	0	0	0	0	1	1	Observed by Carter & Roizman, JVI 1996. No evidence for this TiSS in our data.
US2.5 RNA	US2.5_RNA	US2.5	US2.5-US4	+	134622	137540	2919	1	0	1	0	0	0	0	0	0	1	2	
US3 RNA (also encodes US3.5 CDS)	US3_RNA	US3	US2.5-US4	+	134966	137540	2575	1	1	1	0	0	0	0	1	0	1	4	
US3 RNA #1 (also encodes US3.5 CDS)	US3_RNA_#1	US3	US2.5-US4	+	135122	137540	2419	1	0	0	0	0	0	0	0	1	1	2	
US3 RNA #2 (also encodes US3.5 CDS)	US3_RNA_#2	US3	US2.5-US4	+	135189	137540	2352	1	0	0	0	0	1	1	0	0	1	3	
US4 RNA	US4_RNA	US4	US2.5-US4	+	136734	137540	807	1	1	1	0	1	1	1	1	1	1	8	
US5 RNA	US5_RNA	US5	US5-US7	+	137625	141041	3417	1	1	1	1	1	1	1	0	0	1	7	
US6 RNA	US6_RNA	US6	US5-US7	+	138345	141041	2697	1	1	1	1	0	1	1	0	1	1	7	
US7 RNA	US7_RNA	US7	US5-US7	+	139699	141041	1343	1	1	1	1	0	1	1	0	0	1	6	
US8 RNA	US8_RNA	US8	US8-US9	+	141171	143694	2524	1	1	1	1	0	1	1	0	0	1	6	
US8.5 RNA (encodes US8A)	US8.5_RNA	US8.5	US8-US9	+	142629	143694	1066	1	1	1	0	1	1	1	1	0	1	7	
US9 RNA	US9_RNA	US9	US8-US9	+	143251	143694	444	1	1	0	0	0	1	1	1	0	1	5	
Ori-S RNA (Hubenthal-Voss,Starr,Roizman JVI 1987; encode Ori-S CDS)	Ori-S_RNA	Ori-S	Ori-S-TRS1	+	145907	151026	5120	1	0	0	0	0	0	0	0	0	1	1	Observed by Hubenthal-Voss,Starr,Roizman JVI 1987. Weak peak as well as increase in read coverage downstream of TiSS in both cRNA-seq replicates.

TRS1 RNA	TRS1_RNA	TRS1	Ori-S-TRS1	+	146766	151026	4261	1	1	1	0	0	0	1	0	0	0	3	
TRS1 RNA #1 (orphan, may encode truncated isoform of TRS1 CDS)	TRS1_RNA_#1	TRS1	Ori-S-TRS1	+	147355	151026	3672	1	0	1	0	1	0	0	0	0	0	2	
TRL2.5 RNA #1 (orphan)	TRL2.5_RNA_#1	TRL2.5	TRL2.5	-	7755	7641	115	1	1	0	0	0	0	0	0	1	0	2	
TRL2.5 RNA (=LAT 0.7kb; =pri-miRNA-H1)	TRL2.5_RNA	TRL2.5	TRL2.5	-	8233	7641	593	1	1	0	0	0	0	0	0	1	1	3	
TRL2.5 RNA *1 (=LAT 0.7kb; =pri-miRNA-H1)	TRL2.5_RNA_*1	TRL2.5	TRL2.5	-	8394	7641	754	1	1	0	0	0	0	0	0	0	0	1	Strong peak in both dRNA-seq and one cRNA-seq replicate. Weak peak in the other sRNA-seq replicate.
TRL2.5 RNA *2 (=pri-miRNA-H1)	TRL2.5_RNA_*2	TRL2.5	TRL2.5	-	8424	7641	784	1	1	0	0	0	0	1	0	0	0	2	
UL4 RNA	UL4_RNA	UL4	UL4-UL6.6	-	12494	11733	762	1	1	0	0	0	1	1	0	0	1	4	
UL4.6 RNA (orphan; PacBio only)	UL4.6_RNA	UL4.6	UL4-UL6.6	-	13518	11733	1786	1	0	0	0	0	0	1	1	0	0	2	
UL5 RNA	UL5_RNA	UL5	UL4-UL6.6	-	15581	11733	3849	1	0	1	0	0	1	1	0	1	1	5	
UL5 RNA *1	UL5_RNA_*1	UL5	UL4-UL6.6	-	15617	11733	3885	1	1	0	0	0	0	1	0	0	1	3	
UL6.4 RNA	UL6.4_RNA	UL6.4	UL4-UL6.6	-	16006	11733	4274	1	1	0	0	0	0	0	0	0	1	2	
UL6.5 RNA	UL6.5_RNA	UL6.5	UL4-UL6.6	-	16723	11733	4991	1	0	0	0	0	0	0	1	0	1	2	
UL6.6 RNA	UL6.6_RNA	UL6.6	UL4-UL6.6	-	17358	11733	5626	1	1	0	0	0	0	0	0	0	1	2	
UL8 RNA	UL8_RNA	UL8	UL8-UL9.5	-	20679	18187	2493	1	1	1	0	0	0	1	1	1	0	1	6
UL8.5 RNA	UL8.5_RNA	UL8.5	UL8-UL9.5	-	22194	18187	4008	1	1	0	0	0	0	0	1	0	1	3	
UL8.5 RNA *1 (Baradaran et al 1994; TiSS not observed)	UL8.5_RNA_*1	UL8.5	UL8-UL9.5	-	22350	18187	4164	1	0	0	0	0	0	1	0	0	1	2	
UL9 RNA (Baradaran et al 1994; TiSS not observed)	UL9_RNA	UL9	UL8-UL9.5	-	23380	18187	5194	1	0	0	0	0	0	0	0	1	1	2	
UL9.5 RNA	UL9.5_RNA	UL9.5	UL8-UL9.5	-	24631	18187	6445	1	1	1	0	0	0	0	0	0	1	3	
UL11 RNA	UL11_RNA	UL11	UL11-UL14	-	25115	24774	342	1	1	0	0	0	0	0	0	0	1	2	
UL11.5 RNA (orphan)	UL11.5_RNA	UL11.5	UL11-UL14	-	25498	24774	725	1	1	0	1	0	1	1	0	0	0	4	
UL12.5 RNA	UL12.5_RNA	UL12.5	UL11-UL14	-	26615	24774	1842	1	1	1	1	0	0	1	0	0	1	5	
UL12 RNA	UL12_RNA	UL12	UL11-UL14	-	27045	24774	2272	1	1	1	0	0	1	1	1	0	1	6	
UL13 RNA	UL13_RNA	UL13	UL11-UL14	-	28684	24774	3911	1	1	0	0	0	1	0	0	1	1	4	
UL14 RNA #1 (does not encode UL14 uORF 1 & 2)	UL14_RNA_#1	UL14	UL11-UL14	-	29126	24774	4353	1	1	0	0	0	1	0	0	0	0	2	
UL14 RNA	UL14_RNA	UL14	UL11-UL14	-	29249	24774	4476	1	1	0	0	0	1	1	0	0	1	4	
UL16 RNA	UL16_RNA	UL16	UL16-UL17	-	31607	30146	1462	1	1	1	0	0	1	1	0	1	0	5	
UL17 RNA #1	UL17_RNA_#1	UL17	UL16-UL17	-	33626	30146	3481	1	1	0	0	0	0	0	0	0	1	2	
UL17 RNA	UL17_RNA	UL17	UL16-UL17	-	33814	30146	3669	1	1	1	0	0	1	0	1	1	1	6	
UL17 RNA *1	UL17_RNA_*1	UL17	UL16-UL17	-	34038	30146	3893	1	1	0	0	0	0	0	0	0	1	2	
UL18 RNA	UL18_RNA	UL18	UL18-UL20.6	-	36247	35003	1245	1	1	1	1	0	1	1	1	0	1	7	
UL19 RNA	UL19_RNA	UL19	UL18-UL20.6	-	40766	35003	5764	1	1	1	0	0	1	1	0	1	1	6	
UL20 RNA	UL20_RNA	UL20	UL18-UL20.6	-	41613	35003	6611	1	1	1	0	0	1	1	1	0	1	6	
UL20.6 RNA	UL20.6_RNA	UL20.6	UL18-UL20.6	-	43562	35003	8560	1	0	1	0	0	0	0	0	0	0	1	Strong peak in both cRNA-seq replicates and notable increase in coverage downstream of TiSS. Required for expression of UL26.6 uORF and UL20.6 ORF.
UL22 RNA	UL22_RNA	UL22	UL22	-	46581	43814	2768	1	1	1	0	0	1	1	0	0	1	5	
UL23 RNA	UL23_RNA	UL23	UL23-UL25.5	-	47909	46604	1306	1	1	1	0	0	1	1	1	1	1	7	
UL25.5 RNA	UL25.5_RNA	UL25.5	UL23-UL25.5	-	50556	46604	3953	1	1	0	0	0	0	0	0	0	0	1	Strong peak in both dRNA-seq replicates. Required for expression of UL25.5 ORF.
UL27 RNA	UL27_RNA	UL27	UL27-UL28	-	56080	53032	3049	1	1	1	1	0	1	1	0	1	1	7	
UL28 RNA	UL28_RNA	UL28	UL27-UL28	-	58385	53032	5354	1	1	0	0	0	0	0	0	0	1	2	
UL29.5 RNA	UL29.5_RNA	UL29.5	UL29.5-UL30.6	-	60551	58390	2162	1	0	1	0	0	0	0	0	0	1	2	
UL29 RNA	UL29_RNA	UL29	UL29.5-UL30.6	-	62313	58390	3924	1	1	1	0	0	1	1	1	1	0	6	
UL30.5 RNA #1	UL30.5_RNA_#1	UL30.5	UL29.5-UL30.6	-	63320	58390	4931	1	1	0	0	0	0	0	0	0	1	2	
UL30.5 RNA	UL30.5_RNA	UL30.5	UL29.5-UL30.6	-	63652	58390	5263	1	1	0	0	0	0	0	0	0	1	2	
UL30.6 RNA	UL30.6_RNA	UL30.6	UL29.5-UL30.6	-	65674	58390	7285	1	0	1	0	0	0	0	0	0	0	1	Strong peak in both cRNA-seq replicates as well as a weak kinetic changes (increase late in infection) in both replicates.
UL31 RNA	UL31_RNA	UL31	UL31-UL32	-	67457	66362	1096	1	1	1	0	1	1	1	1	1	1	8	
UL32 RNA	UL32_RNA	UL32	UL31-UL32	-	69219	66362	2858	1	1	1	0	0	1	1	0	0	1	5	
UL36 RNA #1 (orphan, may encode truncated isoform of UL36 CDS)	UL36_RNA_#1	UL36	UL36	-	74802	70958	3845	1	0	1	0	0	0	0	1	0	0	2	
UL36 RNA	UL36_RNA	UL36	UL36	-	80471	70958	9514	1	1	1	0	0	0	0	0	0	0	2	
UL37 RNA #1 (initiates downstream of TaSS of UL37 uoORF)	UL37_RNA_#1	UL37	UL37-UL39.6	-	84146	80690	3457	1	1	0	0	0	1	0	0	0	1	3	
UL37 RNA	UL37_RNA	UL37	UL37-UL39.6	-	84214	80690	3525	1	0	1	0	0	0	1	1	1	1	5	
UL37.5 RNA #1	UL37.5_RNA_#1	UL37.5	UL37-UL39.6	-	85202	80690	4513	1	0	1	0	0	0	0	0	0	1	2	
UL37.5 RNA	UL37.5_RNA	UL37.5	UL37-UL39.6	-	85413	80690	4724	1	0	1	0	0	0	0	0	0	1	2	
UL37.6 RNA #1	UL37.6_RNA_#1	UL37.6	UL37-UL39.6	-	85847	80690	5158	1	0	0	0	0	0	0	1	0	1	2	
UL37.6 RNA	UL37.6_RNA	UL37.6	UL37-UL39.6	-	86020	80690	5331	1	1	0	0	0	0	0	1	0	1	3	
UL39.4 RNA	UL39.4_RNA	UL39.4	UL37-UL39.6	-	88271	80690	7582	1	1	0	0	0	0	0	0	0	1	2	

UL39.5 RNA	UL39.5_RNA	UL39.5	UL37-UL39.6	-	88885	80690	8196	1	1	0	0	0	0	0	0	0	1	Strong peak in both dRNA-seq replicates. Required for expression of UL39.5 ORF.
UL39.6 RNA	UL39.6_RNA	UL39.6	UL37-UL39.6	-	90061	80690	9372	1	1	0	0	0	0	0	0	0	1	Strong peak in both dRNA-seq replicates. Required for expression of UL39.6 ORF.
UL41 RNA #2 (encodes UL41 iORF)	UL41_RNA_#2	UL41	UL41	-	91843	91084	760	1	1	0	0	0	1	1	0	0	1	4	
UL41 RNA #1 (encodes UL41 iORF)	UL41_RNA_#1	UL41	UL41	-	91876	91084	793	1	1	0	0	0	0	1	0	0	1	3	
UL41 RNA (VHS)	UL41_RNA	UL41	UL41	-	92753	91084	1670	1	1	1	0	0	1	1	0	0	1	5	
UL46 RNA (VP11/12)	UL46_RNA	UL46	UL46-UL47	-	100992	98708	2285	1	1	1	1	1	1	1	1	1	1	9	
UL47 RNA #1 (VP13/14)	UL47_RNA_#1	UL47	UL46-UL47	-	103229	98708	4522	1	1	0	0	0	0	0	0	0	1	2	
UL47 RNA (VP13/14)	UL47_RNA	UL47	UL46-UL47	-	103309	98708	4602	1	1	1	0	0	1	1	0	0	1	5	
UL48 RNA #1 (VP16)	UL48_RNA_#1	UL48	UL48	-	105150	103512	1639	1	1	0	0	0	0	1	1	0	1	4	
UL48 RNA (VP16)	UL48_RNA	UL48	UL48	-	105257	103512	1746	1	1	1	1	0	1	1	0	0	1	6	
UL49 RNA	UL49_RNA	UL49	UL49-UL50.6	-	106538	105440	1099	1	1	1	0	1	1	1	1	1	1	8	
UL49.5 RNA (=UL49A)	UL49.5_RNA	UL49.5	UL49-UL50.6	-	107126	105440	1687	1	1	1	0	0	1	1	0	1	1	6	
UL50.5 RNA	UL50.5_RNA	UL50.5	UL49-UL50.6	-	107582	105440	2143	1	1	0	0	0	0	0	1	0	1	3	
UL50.6 RNA	UL50.6_RNA	UL50.6	UL49-UL50.6	-	108170	105440	2731	1	1	0	0	0	0	0	0	0	0	1	Strong peak in both dRNA-seq replicates. Required for expression of UL50.6 sORF.
UL51 RNA #1	UL51_RNA_#1	UL51	UL51-UL54.5	-	109167	108259	909	1	1	0	0	0	1	1	0	0	0	3	
UL51 RNA	UL51_RNA	UL51	UL51-UL54.5	-	109302	108259	1044	1	1	1	0	0	1	1	1	1	1	7	
UL52.4 RNA	UL52.4_RNA	UL52.4	UL51-UL54.5	-	109735	108259	1477	1	0	0	0	0	0	0	0	0	1	1	Strong peak in both dRNA-seq replicates. Required for expression of UL52.4 sORF.
UL52.5 RNA	UL52.5_RNA	UL52.5	UL51-UL54.5	-	112131	108259	3873	1	1	1	0	0	0	1	1	0	1	5	
UL53.5 RNA	UL53.5_RNA	UL53.5	UL51-UL54.5	-	113089	108259	4831	1	1	0	0	0	0	0	1	1	1	4	
UL54.5 RNA (orphan)	UL54.5_RNA	UL54.5	UL51-UL54.5	-	115845	108259	7587	1	0	1	0	0	0	0	0	0	1	Strong peak in both cRNA-seq replicates as well as a weak kinetic change (increase late in infection) in both replicates.
UL56 RNA	UL56_RNA	UL56	UL56-IRL3	-	117083	116178	906	1	0	1	0	0	1	1	1	1	1	6	
IRL3 RNA (orphan)	IRL3_RNA	IRL3	UL56-IRL3	-	117925	116178	1748	1	1	0	0	0	0	0	0	1	0	2	
AL-RNA (Perng et al., JVI 2002, not observed)	AL-RNA	AL-RNA	AL-RNA	-	118958	117998	961	1	0	0	0	0	0	0	0	0	0	0	Observed by Perng et al., JVI 2002. No evidence for this TiSS in our data.
IRL2 RNA iso1 (with retained intron 1)	IRL2_RNA_iso1	IRL2	IRL2-IRL1	-	124255	120656	3464	2	1	1	0	0	1	1	1	1	1	7	
IRL2 RNA iso2 (NAG-NAG, spliced)	IRL2_RNA_iso2	IRL2	IRL2-IRL1	-	124255	120656	2696	3	1	1	0	0	1	1	1	1	1	7	
IRL2 RNA (spliced)	IRL2_RNA	IRL2	IRL2-IRL1	-	124255	120656	2699	3	1	1	0	0	1	1	1	1	1	7	
IRL1.5 RNA (spliced)	IRL1.5_RNA	IRL1.5	IRL2-IRL1	-	124622	120656	3066	3	1	1	0	0	0	0	1	1	1	5	
IRL1 RNA (spliced)	IRL1_RNA	IRL1	IRL2-IRL1	-	125964	120656	4408	3	1	1	0	0	0	0	0	1	1	4	
IRS1 RNA #1 (orphan, may encode truncated isoform of IRS1 CDS)	IRS1_RNA_#1	IRS1	IRS1	-	130840	127169	3672	1	0	1	0	1	0	0	0	0	0	2	
IRS1 RNA	IRS1_RNA	IRS1	IRS1	-	131429	127169	4261	1	1	1	0	0	0	1	0	0	0	3	
US2 RNA	US2_RNA	US2	US2-US5.5	-	135305	134020	1286	1	1	1	0	0	1	1	1	1	0	6	
US3.6 RNA	US3.6_RNA	US3.6	US2-US5.5	-	136667	134020	2648	1	1	0	0	0	0	0	0	0	1	2	
US5.1 RNA (Jovasevic & Roizman Virol J 2010. not observed in TiSS or PacBio data; orphan)	US5.1_RNA	US5.1	US2-US5.5	-	137986	134020	3967	1	0	0	0	0	0	0	0	0	0	0	Observed by Jovasevic & Roizman Virol J 2010. No evidence for this TiSS in our data.
US5.5 RNA (orphan)	US5.5_RNA	US5.5	US2-US5.5	-	138910	134020	4891	1	1	0	0	0	0	0	0	0	1	2	
US10 RNA	US10_RNA	US10	US10-US12	-	145168	144122	1047	1	1	0	0	0	1	1	0	0	1	4	
US11 RNA	US11_RNA	US11	US10-US12	-	145461	144122	1340	1	1	0	0	0	1	1	1	1	1	6	
US12 RNA (spliced)	US12_RNA	US12	US10-US12	-	146066	144122	1777	2	1	1	1	0	1	1	0	0	0	5	

Table B.2: List of all possible splicing events with the corresponding read numbers spanning a splice-junction and not spanning it at the 5'- and 3'-end for 4sU-seq, total RNA, RNA from subcellular fraction in wild-type and ΔICP27 mutant and from data from Tang et al [Tang et al., 2019]. Furthermore, the type indicates if it was already known, if it was identified by PacBio sequencing or by Illumina sequencing. ESR = Exon-spanning reads, UEIR = Upstream exon-intron reads, DIER = Downstream intron-exon reads, SR= subcellular RNA fractions

Intron Position	Affected gene(s)	ESR (d27-1) [Tang et al., 2019]	ESR (Kos) [Tang et al., 2019]	Type	UEIR (4sU-Seq)	ESR (4sU-Seq)	DIER (4sU-Seq)	UEIR (total-RNA)	ESR (total-RNA)	DIER (total-RNA)	UEIR (WT:SR)	ESR (WT:SR)	DIER (WT:SR)	UEIR (DICP27:SR)	ESR (DICP27:SR)	DIER (DICP27:SR)
JN555585:-122487-122623	IRL1,RNA	2903	4677	Known	1748	8176	870	898	1214	154	155	183	60	280	595	72
JN555585:-122290-124055	IRL1,RNA	16873	28617	Known	4144	32944	3789	5839	19125	2588	447	3022	188	617	6786	742
JN555585:-146633-146821	US12,RNA	79081	66113	Known	655	20662	151	906	4665	49	108	50241	10475	89	42030	4438
JN555585:-119464-121420	LAT,RNA	4	429	Known	5223	1656	2302	602	766	882	3	0	23	224	0	1065
JN555585:+:123375-123543	US1,RNA	30190	30732	Known	2763	84762	1811	621	22672	1163	978	17590	104	1236	29631	1129
JN555585:+:2018-3083	TRL2,RNA_iso	.	.	Known	4288	30836	3224	5839	18967	2410	441	2996	153	617	6670	632
JN555585:+:30049-33635	UL13,RNA	3573	2098	Known	3484	5224	2796	884	1329	820	1415	6357	1886	3294	3384	3501
JN555585:+:3750-3886	TRL2,RNA_i1	2903	4677	Known	2069	7752	690	1001	1199	140	159	110	36	374	563	44
JN555585:+:122484-122623	TRL2,RNA_iso	1868	3475	NAGNAG	7748	5103	8880	998	786	1305	155	0	232	280	0	716
JN555585:+:3750-3889	TRL2,RNA_iso	1868	3475	NAGNAG	2069	4872	8663	776	786	1355	159	0	154	374	0	639
JN555585:+:78883-80087	UL36,RNA	.	.	NAGNAG	2252	180	2318	209	1605	1075	1442	1	1442	2357	3	3698
JN555585:+:78883-80090	UL36,RNA	8	5	PacBio con-firmed	2252	126	2319	144	1883	1075	1490	0	3255	2357	1	3820
JN555585:+:91121-91412	UL40,5,RNA	28	870	PacBio con-firmed	2185	771	4171	1312	972	1920	1946	0	3255	2189	1	4466
JN555585:+:91551-92534	UL40,6,RNA	6	3387	PacBio con-firmed	2322	738	6146	2491	1111	3700	4134	0	5479	7750	0	8913
JN555585:+:97724-97960	UL44,RNA	173	215	PacBio con-firmed	68273	3743	79886	18853	550	15525	1776	0	1124	71312	1	50506
JN555585:-122290-123590	IRL2,RNA_iso	5	1045	Illumina only	5325	233	3789	20516	637	2588	257	0	188	3117	0	742
JN555585:-122290-124572	IRL1,5,RNA	0	46	Illumina only	6323	103	3789	3819	140	2588	299	0	188	3117	0	742
JN555585:-124179-124572	IRL1,5,RNA	15	78	Illumina only	6323	269	5725	3819	451	2875	299	0	1491	1325	0	6922
JN555585:-124179-124572	IRL1,5,RNA	.	.	Illumina only	6323	97	49049	3819	110	24789	299	0	3221	1325	0	6205
JN555585:-124263-124572	IRL1,5,RNA	.	.	Illumina only	6323	59	7909	3819	22	4737	299	0	222	1325	0	1817
JN555585:-132612-133234	US2,RNA	.	.	Illumina only	3144	107	1705	3706	118	1929	391	0	35	8933	0	5772
JN555585:-132612-134195	US2,RNA	.	.	Illumina only	17593	42	1705	4735	58	1929	2191	0	35	12352	0	5772
JN555585:-132612-137654	US5,1,RNA	.	.	Illumina only	8120	20	1705	3677	63	1929	392	0	397	8847	0	5772
JN555585:-133408-134195	US2,RNA	.	.	Illumina only	17593	65	4628	4735	126	4823	2191	0	463	12352	0	9610
JN555585:-134246-134195	US2,RNA	.	.	Illumina only	17593	72	27590	4735	15	6586	2191	0	1418	12352	0	16543
JN555585:-134490-135924	US2,RNA	24	.	Illumina only	17363	246	9507	3589	69	2366	4094	0	3124	5103	0	4022
JN555585:-134701-135214	US2,RNA	.	.	Illumina only	28137	175	18666	6859	105	2307	3880	0	2439	16553	0	15009
JN555585:-134701-137654	US5,1,RNA	.	.	Illumina only	8120	117	18666	3677	118	2307	392	0	2439	8847	0	15009
JN555585:-137812-138717	US5,5,RNA	.	.	Illumina only	2065	50	3111	6747	113	2196	1005	0	397	6965	0	5596
JN555585:-137812-138980	US5,5,RNA	.	.	Illumina only	1201	291	3111	3052	144	2196	347	0	397	3564	0	5596
JN555585:-137812-139121	US5,5,RNA	.	.	Illumina only	331	69	3111	1398	72	2196	212	0	397	2852	0	5596
JN555585:-137812-139386	US5,5,RNA	.	.	Illumina only	1136	49	3111	1959	97	2196	398	0	397	3879	0	5596
JN555585:-137812-139736	US5,5,RNA	.	.	Illumina only	1460	36	3111	2066	41	2196	475	0	397	3880	0	5596
JN555585:-16987-17576	UL6,6,RNA	.	.	Illumina only	4207	140	2294	2296	103	1543	1368	0	1350	4619	0	3446
JN555585:-76255-140944	US10,RNA	.	.	Illumina only	210306	381	24051	92001	33	3157	119073	0	1143	122186	0	23171
JN555585:-76467-76517	UL36,RNA	.	.	Illumina only	9633	1247	10509	1488	93	1516	647	0	668	9583	0	11285
JN555585:-87147-87308	UL39,4,RNA	.	.	Illumina only	1001	96	1515	1579	394	1600	341	0	441	2167	0	2787
JN555585:-87147-88413	UL39,5,RNA	.	.	Illumina only	1543	28	1515	1847	82	1600	664	0	441	3427	0	2787
JN555585:-87728-88413	UL39,5,RNA	.	.	Illumina only	1543	183	2843	1847	487	3347	664	0	863	3427	0	5231
JN555585:-88551-88815	UL39,5,RNA	.	.	Illumina only	2813	127	2574	3011	356	2560	949	0	732	4323	0	3855
JN555585:-88551-90068	UL39,6,RNA	.	.	Illumina only	2406	44	2574	3004	115	2560	333	0	732	3697	0	3855
JN555585:-89100-89406	UL39,6,RNA	.	.	Illumina only	1628	9	2273	2206	40	2354	637	0	862	2663	0	3322
JN555585:-89100-90068	UL39,6,RNA	.	.	Illumina only	2406	8	2273	3004	59	2354	333	0	862	3697	0	3322
JN555585:-90553-90889	.	.	.	Illumina only	3346	39	3759	2689	50	3192	639	0	708	3625	0	5050
JN555585:+:113428-113787	UL53,RNA_#2	.	.	Illumina only	16269	265	50891	4617	77	17672	1370	0	9258	11699	0	29575
JN555585:+:11898-12972	UL4,5,RNA_*1	.	.	Illumina only	2006	46	5219	3562	70	3305	1553	0	868	5290	0	7155
JN555585:+:119464-119661	LAT	.	.	Illumina only	5223	40	7672	602	22	2352	3	0	70	224	0	1266
JN555585:+:12189-12972	UL4,5,RNA_*1	5	96	Illumina only	5476	251	5219	3725	322	3305	1814	0	868	6697	0	7155
JN555585:+:12429-12972	UL4,5,RNA	14	381	Illumina only	5304	776	5219	4785	1128	3305	1229	0	868	7977	1	7155
JN555585:+:12847-12972	UL4,5,RNA	.	.	Illumina only	8561	61	5219	6225	73	3305	1307	0	868	8947	0	7155
JN555585:+:132356-132543	US1,RNA	171	102	Illumina only	81045	707	1811	21394	146	1163	18809	0	104	28521	0	1129
JN555585:+:134732-135342	US2,5,RNA	.	.	Illumina only	124	33	10165	125	34	3849	320	0	4942	692	0	11797
JN555585:+:135234-135342	US2,RNA_#1	133	24	Illumina only	5636	113	10165	2720	94	3849	3063	8	4942	6353	35	11797
JN555585:+:137812-138980	US5,RNA	1	36	Illumina only	8915	39	110587	3030	6	17445	3394	0	22942	10360	0	60532
JN555585:+:137812-138736	US5,RNA	.	.	Illumina only	8915	5	41278	3030	5	9564	3394	0	19032	10360	0	56062
JN555585:+:139062-139201	US6,RNA	249	3	Illumina only	104969	66	112897	14216	20	12440	19331	52	22415	57789	1	65994
JN555585:+:32455-33635	UL15,RNA	.	.	Illumina only	4112	25	6616	2183	19	2413	2452	0	8294	5740	0	5916
JN555585:+:33486-33635	UL15,RNA	169	283	Illumina only	2816	215	2796	891	106	820	1837	0	1886	3558	0	3501
JN555585:+:41710-42246	UL20,4,RNA	.	.	Illumina only	1483	778	2796	674	228	820	1464	0	1886	3071	0	3501
JN555585:+:45744-46676	UL22,5,RNA	.	.	Illumina only	6821	206	15138	2571	53	6903	402	0	2119	5948	0	14019
JN555585:+:46772-47680	UL23,5,RNA	.	.	Illumina only	735	101	2563	562	187	1645	425	0	637	1960	0	3036
JN555585:+:46772-48032	UL23,5,RNA	.	.	Illumina only	2199	84	4190	1371	58	3158	989	0	2526	4119	0	5093
JN555585:+:46772-48275	UL23,5,RNA	.	.	Illumina only	2199	160	7323	1371	108	2413	989	0	2873	4119	0	5425
JN555585:+:48251-48813	UL24,5,RNA	7	246	Illumina only	2199	331	8160	1371	323	2602	989	0	2339	4119	2	4654
JN555585:+:60650-61892	UL28,4,RNA	.	.	Illumina only	14963	204	38103	3688	213	9929	7093	0	4873	8113	0	17121
JN555585:+:63718-63964	UL30,RNA	.	.	Illumina only	2567	29	3003	2500	48	820	184	0	181	4397	0	3968
JN555585:+:66650-81892	UL36,6,RNA	.	.	Illumina only	51395	163	34611	24470	78	20702	29259	0	23432	30455	0	24301
JN555585:+:79961-80087	UL40,RNA	.	.	Illumina only	2907	50	2318	1819	63	1605	1305	0	1442	3149	0	3698
JN555585:+:90943-91412	UL40,RNA	.	.	Illumina only	50462	72	4171	5672	39	1920	21552	0	3255	30270	1	4466
JN555585:+:91078-91412	UL40,5,RNA	4	369	Illumina only	2094	403	4171	1307	382	1920	2078	0	3255	1978	0	4466
JN555585:+:97843-97960	UL44,RNA	.	.	Illumina only	53277	96	79886	11450	28	15525	596	0	1124	35327	0	50506
JN555585:+:100018-106091	ul49,5-49 sp	.	.	PacBio n.c.	43424	26	38842	12443	2	11018	15533	0	11453	58271	0	47711
JN555585:+:78852-79742	AST-2,4,3	.	.	PacBio n.c.	794	3	1385	246	2	811	168	0	606	664	0	1801
JN555585:+:79581-79742	AST-2,4,1	.	.	PacBio n.c.	612	5	1385	380	0	811	260	0	606	1552	0	1801
JN555585:+:91121-91517	ul41-42 c sp3	1	91	PacBio n.c.	2185	18	4327	1312	22	3875	1946	0	3787	2189	0	8462
JN555585:+:91551-97948	ul41-45 c sp	.	.	PacBio n.c.	2322	1	81919	2491	0	15900	4134	0	1063	7750	0	48681
JN555585:+:92981-94004	ul41-42 c sp 1	.	.	PacBio n.c.	36524	15	97052	18168	1	45695	12814	0	20133	28403	0	60194
JN555585:+:92981-94146	ul41-42 c sp2	.	.	PacBio n.c.	36524	0	84731	18168	0	43579	12814	0	15232	28403	0	64023

Table B.3: List of all identified ORFs. All ORFs of the previous reference annotation are labeled as CDS (coding sequence).

Name	ID	Type	Gene	Transcript	Strand	Length (aa)	Start codon	Stop codon	TaSS	Stop	Location
IRL1 CDS (ICP34.5)	IRL1_CDS	CDS	IRL1	IRL1_RNA	-	248	AUG	UAA	125860	125114	JN555585-:125114-125861
IRL1.5 ORF	IRL1.5_ORF	ORF	IRL1.5	IRL1.5_RNA	-	106	AUG	UAA	124602	124282	JN555585-:124282-124603
IRL1A sORF (uoORF of RL1; 93aa initiating from AUG)	IRL1A_sORF	sORF	IRL1	IRL1_RNA	-	93	AUG	UGA	125906	125625	JN555585-:125625-125907
IRL2 CDS (ICP0; IE110)	IRL2_CDS	CDS	IRL2	IRL2_RNA	-	775	AUG	UAA	124111	120883	JN555585-:120883-122487|122623-123290|124055-124112
IRL2 CDS iso1 (RL2 CDS with different C-terminus due to intron retention of intron 1)	IRL2_CDS_iso1	CDS	IRL2	-	-	72	AUG	UAG	124111	123893	JN555585-:123893-124112
IRL2 CDS iso2 (NAG-NAG)	IRL2_CDS_iso2	CDS	IRL2	-	-	774	AUG	UAA	124111	120883	JN555585-:120883-122484|122623-123290|124055-124112
IRL2 dORF RNA iso1 (orphan, located within intron 1)	IRL2_dORF_RNA_iso1	dORF	IRL2	IRL2_RNA_iso1	-	42	AUG	UGA	123731	123603	JN555585-:123603-123732
IRL2.5 ORF (TaSS unclear due to repeat regions upstream)	IRL2.5_ORF	ORF	IRL2.5	IRL2.5_RNA	+	122	AUC	UAA	118253	118621	JN555585+:118253-118622
IRL2A ORF	IRL2A_ORF	ORF	IRL2	IRL2_RNA	-	181	ACG	UGA	124227	122917	JN555585-:122917-123290|124055-124228
IRS1 CDS (ICP4; IE175;)	IRS1_CDS	CDS	IRS1	IRS1_RNA	-	1298	AUG	UAA	131130	127234	JN555585-:127234-131131
ORF-O CDS (TaSS probably 76nt upstream of ORF-P initiating from ACG start codon; no evidence of frame-shift)	ORF-O_CDS	CDS	ORF-O/P	ORF-O/P_RNA	+	271	ACG	UAG	125111	125926	JN555585+:125111-125927
ORF-P CDS	ORF-P_CDS	CDS	ORF-O/P	ORF-O/P_RNA	+	233	AUG	UAG	125187	125888	JN555585+:125187-125889
Ori-S CDS (Hubenthal-Voss,Starr,Roizman JVI 1987)	Ori-S_CDS	CDS	Ori-S	Ori-S_RNA	+	330	AUG	UAG	145940	146932	JN555585+:145940-146933
UL1 CDS (Glycoprotein L)	UL1_CDS	CDS	UL1	UL1_RNA	+	224	AUG	UAA	9337	10011	JN555585+:9337-10012
UL1 uORF	UL1_uORF	uORF	UL1	UL1_RNA	+	19	GUG	UAG	9272	9331	JN555585+:9272-9332
UL10 CDS (Glycoprotein M; may include N-terminal extension but TaSS unclear)	UL10_CDS	CDS	UL10	UL10_RNA	+	473	AUG	UAG	23204	24625	JN555585+:23204-24626
UL10 uORF (TaSS unclear; could also be 23012 or 23039)	UL10_uORF	uORF	UL10	UL10_RNA	+	26	ACG	UGA	23033	23113	JN555585+:23033-23114
UL11 CDS (Tegument protein)	UL11_CDS	CDS	UL11	UL11_RNA	-	96	AUG	UAA	25091	24801	JN555585-:24801-25092
UL12 CDS (Exonuclease activity)	UL12_CDS	CDS	UL12	UL12_RNA	-	626	AUG	UGA	26887	25007	JN555585-:25007-26888
UL12 uoORF (TaSS unclear, could also be at 26971)	UL12_uoORF	uoORF	UL12	UL12_RNA	-	23	AUC	UAG	26943	26872	JN555585-:26872-26944
UL12.5 CDS (Exonuclease activity; truncated isoform of UL12)	UL12.5_CDS	CDS	UL12.5	UL12.5_RNA	-	500	AUG	UGA	26509	25007	JN555585-:25007-26510
UL13 CDS (Serine-threonince protein kinase)	UL13_CDS	CDS	UL13	UL13_RNA	-	518	AUG	UGA	28502	26946	JN555585-:26946-28503
UL13.5 ORF	UL13.5_ORF	ORF	UL13.5	UL13.5_RNA	+	141	AUG	UAG	27646	28071	JN555585+:27646-28072
UL13.5 ORF #1 (truncated isoform of 85aa of UL13.5 ORF initiating at 27814 from AUG)	UL13.5_ORF_#1	ORF	UL13.5	UL13.5_RNA	+	85	AUG	UAG	27814	28071	JN555585+:27814-28072
UL14 CDS (Tegument protein)	UL14_CDS	CDS	UL14	UL14_RNA	-	219	AUG	UGA	28915	28256	JN555585-:28256-28916
UL14 uORF 1	UL14_uORF_1	uORF	UL14	UL14_RNA	-	21	GUG	UAA	29220	29155	JN555585-:29155-29221
UL14 uORF 2	UL14_uORF_2	uORF	UL14	UL14_RNA	-	32	AUG	UGA	29126	29028	JN555585-:29028-29127
UL15 CDS (Terminase; Processing and packaging DNA)	UL15_CDS	CDS	UL15	UL15_RNA	+	735	AUG	UGA	29020	34813	JN555585+:29020-30049|33635-34814
UL15 iORF 1 (located in intron; orphan)	UL15_iORF_1	iORF	UL15	-	+	26	CUG	UAG	33181	33261	JN555585+:33181-33262
UL15 iORF 2 (located in intron; orphan)	UL15_iORF_2	iORF	UL15	-	+	73	AUG	UAG	33303	33524	JN555585+:33303-33525
UL15 iORF 3 (located in exon 2; orphan)	UL15_iORF_3	iORF	UL15	UL15_RNA	+	46	AUG	UGA	33696	33836	JN555585+:33696-33837
UL15 ORF RNA iso1 (UL15 CDS with different C-terminus due to intron retention)	UL15_ORF_RNA_iso1	ORF	UL15	UL15_RNA_iso1	+	550	AUG	UAG	29020	30672	JN555585+:29020-30673
UL15 uoORF	UL15_uoORF	uoORF	UL15	UL15_RNA	+	38	GUG	UAA	28974	29090	JN555585+:28974-29091
UL15 uORF	UL15_uORF	uORF	UL15	UL15_RNA	+	29	GUG	UGA	28862	28951	JN555585+:28862-28952
UL15.4 sORF 1 (orphan; possible alternative TaSS at 30021, AUG)	UL15.4_sORF_1	sORF	UL15.4	-	-	65	AUG	UAG	30071	29874	JN555585-:29874-30072
UL15.4 sORF 2 (orphan)	UL15.4_sORF_2	sORF	UL15.4	-	-	58	AUG	UAG	29834	29658	JN555585-:29658-29835
UL15.5 CDS (orphan; Capside associated protein)	UL15.5_CDS	CDS	UL15.5	-	+	293	AUG	UGA	33932	34813	JN555585+:33932-34814
UL16 CDS (Tegument protein)	UL16_CDS	CDS	UL16	UL16_RNA	-	373	AUG	UAA	31295	30174	JN555585-:30174-31296
UL17 CDS (Processing and packaging)	UL17_CDS	CDS	UL17	UL17_RNA	-	703	AUG	UAG	33497	31386	JN555585-:31386-33498
UL17 uoORF (TaSS unclear)	UL17_uoORF	uoORF	UL17	UL17_RNA	-	11	AUC	UGA	33529	33494	JN555585-:33494-33530
UL17 uORF	UL17_uORF	uORF	UL17	UL17_RNA	-	49	CUG	UAG	33780	33631	JN555585-:33631-33781
UL17 uORF RNA *1	UL17_uORF_RNA_*1	uORF	UL17	UL17_RNA_*1	-	27	AUG	UAA	33984	33901	JN555585-:33901-33985
UL17.4 sORF 1 (orphan)	UL17.4_sORF_1	sORF	UL17.4	-	-	59	AUG	UAA	34380	34201	JN555585-:34201-34381
UL17.4 sORF 2 (orphan, contains 36aa N-terminal extension of UL17 uORF 1 RNA *1 initiating from AUG)	UL17.4_sORF_2	sORF	UL17.4	-	-	67	AUG	UAA	34104	33901	JN555585-:33901-34105

UL17.5 ORF (orphan)	UL17.5_ORF	ORF	UL17.5	-	-	96	AUG	UAG	34752	34462	JN555585-:34462-34753
UL17.5 uoORF (orphan)	UL17.5_uoORF	uoORF	UL17.5	-	-	14	AUG	UAA	34787	34743	JN555585-:34743-34788
UL18 CDS (VP23; Capsid protein)	UL18_CDS	CDS	UL18	UL18_RNA	-	318	AUG	UAA	36051	35095	JN555585-:35095-36052
UL18 uoORF	UL18_uoORF	uoORF	UL18	UL18_RNA	-	74	AUG	UGA	36152	35928	JN555585-:35928-36153
UL18 uORF	UL18_uORF	uORF	UL18	UL18_RNA	-	11	AUG	UAA	36205	36170	JN555585-:36170-36206
UL19 CDS (VP5; Major capsid protein)	UL19_CDS	CDS	UL19	UL19_RNA	-	1374	AUG	UAA	40528	36404	JN555585-:36404-40529
UL19 CDS *1 (includes 11 aa N-terminal extension initiating from AUC)	UL19_CDS_*1	CDS	UL19	UL19_RNA	-	1385	AUC	UAA	40561	36404	JN555585-:36404-40562
UL19.4 sORF (orphan)	UL19.4_sORF	sORF	UL19.4	-	+	6	AUG	UAG	39336	39356	JN555585+:39336-39357
UL19.5 sORF (may be part of new ORF initiating at 39420)	UL19.5_sORF	sORF	UL19.5	UL19.5_RNA	+	68	ACG	UAA	39969	40175	JN555585+:39969-40176
UL2 CDS (Uracil-DNA glycosylase)	UL2_CDS	CDS	UL2	UL2_RNA	+	334	AUG	UGA	9884	10888	JN555585+:9884-10889
UL20 CDS (Membrane protein)	UL20_CDS	CDS	UL20	UL20_RNA	-	222	AUG	UAA	41488	40820	JN555585-:40820-41489
UL20 iORF (ORF initiates 13nt downstream of UL20 from AUG)	UL20_iORF	iORF	UL20	UL20_RNA	-	59	AUG	UGA	41475	41296	JN555585-:41296-41476
UL20 uORF (corresponds to C-terminal part of UL20.5 CDS)	UL20_uORF	uORF	UL20	UL20_RNA	-	21	AUG	UAA	41599	41534	JN555585-:41534-41600
UL20.4 dORF 1	UL20.4_dORF_1	dORF	UL20.4	UL20.4_RNA	+	21	AUG	UGA	41324	41389	JN555585+:41324-41390
UL20.4 dORF 2 (alternative TaSS at next AUG 6nt downstream at 41487)	UL20.4_dORF_2	dORF	UL20.4	UL20.4_RNA	+	75	AUG	UGA	41481	41708	JN555585+:41481-41709
UL20.4 ORF	UL20.4_ORF	ORF	UL20.4	UL20.4_RNA	+	109	AUG	UAG	40845	41174	JN555585+:40845-41175
UL20.5 CDS (orphan)	UL20.5_CDS	CDS	UL20.5	-	-	160	AUG	UAA	42016	41534	JN555585-:41534-42017
UL20.6 ORF	UL20.6_ORF	ORF	UL20.6	UL20.6_RNA	-	181	AUG	UAG	42816	42271	JN555585-:42271-42817
UL20.6 uORF	UL20.6_uORF	uORF	UL20.6	UL20.6_RNA	-	60	AUG	UAG	43311	43129	JN555585-:43129-43312
UL21 CDS (Tegument protein)	UL21_CDS	CDS	UL21	UL21_RNA	+	535	AUG	UAA	42074	43681	JN555585+:42074-43682
UL21 CDS *1 (includes 65 aa N-terminal extension initiating from CUG)	UL21_CDS_*1	CDS	UL21	UL21_RNA	+	600	CUG	UAA	41879	43681	JN555585+:41879-43682
UL22 CDS (Glycoprotein H)	UL22_CDS	CDS	UL22	UL22_RNA	-	838	AUG	UAA	46382	43866	JN555585-:43866-46383
UL22 uORF 1 (=uoORF for UL22 uORF 2)	UL22_uORF_1	uORF	UL22	UL22_RNA	-	30	CUG	UGA	46556	46464	JN555585-:46464-46557
UL22 uORF 2 (possible alternative TaSS downstream at 46468, GUG)	UL22_uORF_2	uORF	UL22	UL22_RNA	-	35	AUU	UAG	46524	46417	JN555585-:46417-46525
UL22.5 sORF	UL22.5_sORF	sORF	UL22.5	UL22.5_RNA	+	51	AUG	UAG	44020	44175	JN555585+:44020-44176
UL23 CDS (Thymidine kinase; Peripheral to DNA replication)	UL23_CDS	CDS	UL23	UL23_RNA	-	376	AUG	UGA	47802	46672	JN555585-:46672-47803
UL23 CDS *1 (includes 15 aa N-terminal extension initiating from GUG)	UL23_CDS_*1	CDS	UL23	UL23_RNA	-	391	GUG	UGA	47847	46672	JN555585-:46672-47848
UL23 uORF 1	UL23_uORF_1	uORF	UL23	UL23_RNA	-	5	CUG	UAA	47898	47881	JN555585-:47881-47899
UL23 uORF 2 (uoORF for N-terminal extension of UL23; C-terminal 7 aa of UL23.6 ORF)	UL23_uORF_2	uORF	UL23	UL23_RNA	-	7	GUG	UGA	47867	47844	JN555585-:47844-47868
UL23.5 ORF	UL23.5_ORF	ORF	UL23.5	UL23.5_RNA	+	133	AUG	UGA	46954	47355	JN555585+:46954-47356
UL23.5 uORF	UL23.5_uORF	uORF	UL23.5	UL23.5_RNA	+	54	UUG	UAA	46655	46819	JN555585+:46655-46820
UL23.6 ORF (orphan)	UL23.6_ORF	ORF	UL23.6	-	-	154	AUG	UGA	48308	47844	JN555585-:47844-48309
UL24 CDS (Exonuclease activity)	UL24_CDS	CDS	UL24	UL24_RNA	+	269	AUG	UGA	47737	48546	JN555585+:47737-48547
UL24 uORF 1 RNA *1	UL24_uORF_1_RNA_*1	uORF	UL24	UL24_RNA_*1	+	29	AUG	UAG	47414	47503	JN555585+:47414-47504
UL24 uORF 2 RNA *1	UL24_uORF_2_RNA_*1	uORF	UL24	UL24_RNA_*1	+	13	AUG	UAA	47504	47545	JN555585+:47504-47546
UL24 uORF 3 RNA *1	UL24_uORF_3_RNA_*1	uORF	UL24	UL24_RNA_*1	+	9	GUG	UAG	47629	47658	JN555585+:47629-47659
UL24.5 CDS (truncated isoform of UL24)	UL24.5_CDS	CDS	UL24.5	UL24.5_RNA	+	148	AUG	UGA	48100	48546	JN555585+:48100-48547
UL25 CDS (Processing and packaging DNA; Capside Protein)	UL25_CDS	CDS	UL25	UL25_RNA	+	580	AUG	UAG	48813	50555	JN555585+:48813-50556
UL25.4 ORF (orphan)	UL25.4_ORF	ORF	UL25.4	-	-	241	AUG	UGA	49438	48713	JN555585-:48713-49439
UL25.5 ORF	UL25.5_ORF	ORF	UL25.5	UL25.5_RNA	-	127	AUG	UAG	50110	49727	JN555585-:49727-50111
UL26 CDS (P40; VP24; VP22A; Capsid protein)	UL26_CDS	CDS	UL26	UL26_RNA	+	635	AUG	UGA	50809	52716	JN555585+:50809-52717
UL26.5 CDS (truncated isoform of UL26)	UL26.5_CDS	CDS	UL26.5	UL26.5_RNA	+	329	AUG	UGA	51727	52716	JN555585+:51727-52717
UL27 CDS (Glycoprotein B)	UL27_CDS	CDS	UL27	UL27_RNA	-	904	AUG	UGA	55794	53080	JN555585-:53080-55795
UL27 CDS *1 (includes 43 aa N-terminal extension initiating from ACG)	UL27_CDS_*1	CDS	UL27	UL27_RNA	-	947	ACG	UGA	55923	53080	JN555585-:53080-55924
UL27.5 sORF	UL27.5_sORF	sORF	UL27.5	UL27.5_RNA	+	48	AUG	UGA	54980	55126	JN555585+:54980-55127
UL27.6 ORF	UL27.6_ORF	ORF	UL27.6	UL27.6_RNA	+	214	AUG	UGA	56357	57001	JN555585+:56357-57002
UL27.6 uORF	UL27.6_uORF	uORF	UL27.6	UL27.6_RNA	+	5	AUG	UAA	56145	56162	JN555585+:56145-56163

UL28 CDS (ICP18.5; Processing and packaging)	UL28_CDS	CDS	UL28	UL28_RNA	-	785	AUG	UAG	58159	55802	JN555585-:55802-58160
UL28 uORF	UL28_uORF	uORF	UL28	UL28_RNA	-	38	GUG	UAG	58335	58219	JN555585-:58219-58336
UL28.4 sORF	UL28.4_sORF	sORF	UL28.4	UL28.4_RNA	+	53	AUG	UAG	57749	57910	JN555585+:57749-57911
UL28.6 ORF (orphan)	UL28.6_ORF	ORF	UL28.6	-	+	348	AUG	UAA	60902	61948	JN555585+:60902-61949
UL29 CDS (ICP8; Major DNA-binding protein)	UL29_CDS	CDS	UL29	UL29_RNA	-	1196	AUG	UGA	62053	58463	JN555585-:58463-62054
UL29.5 CDS (truncated isoform of UL29)	UL29.5_CDS	CDS	UL29.5	UL29.5_RNA	-	681	AUG	UGA	60508	58463	JN555585-:58463-60509
UL3 CDS	UL3_CDS	CDS	UL3	UL3_RNA	+	224	AUG	UAA	10990	11664	JN555585+:10990-11665
UL3 uORF RNA *1	UL3_uORF_RNA_*1	uORF	UL3	UL3_RNA_*1	+	5	AUG	UAA	10940	10957	JN555585+:10940-10958
UL30 CDS (DNA polymerase)	UL30_CDS	CDS	UL30	UL30_RNA	+	1235	AUG	UGA	62806	66513	JN555585+:62806-66514
UL30 uORF	UL30_uORF	uORF	UL30	UL30_RNA	+	10	AUG	UAA	62682	62714	JN555585+:62682-62715
UL30.5 sORF 1	UL30.5_sORF_1	sORF	UL30.5	UL30.5_RNA	-	37	AUG	UAA	63533	63420	JN555585-:63420-63534
UL30.5 sORF 2	UL30.5_sORF_2	sORF	UL30.5	UL30.5_RNA	-	34	AUG	UGA	63350	63246	JN555585-:63246-63351
UL30.5 sORF RNA #1 (only comprises the C-terminal 4aa of UL30.5 sORF 2; initiated from AUG)	UL30.5_sORF_RNA_#1	sORF	UL30.5	UL30.5_RNA_#1	-	4	AUG	UGA	63260	63246	JN555585-:63246-63261
UL30.6 sORF	UL30.6_sORF	sORF	UL30.6	UL30.6_RNA	-	42	AUG	UAA	64379	64251	JN555585-:64251-64380
UL31 CDS (Nuclear matrix protein)	UL31_CDS	CDS	UL31	UL31_RNA	-	306	AUG	UAG	67378	66458	JN555585-:66458-67379
UL31.5 ORF	UL31.5_ORF	ORF	UL31.5	UL31.5_RNA	+	103	AUG	UAG	67510	67821	JN555585+:67510-67822
UL31.5 uORF 1 RNA *1	UL31.5_uORF_1_RNA_*1	uORF	UL31.5	UL31.5_RNA_*1	+	17	AUG	UAG	67146	67199	JN555585+:67146-67200
UL31.5 uORF 2 RNA *1	UL31.5_uORF_2_RNA_*1	uORF	UL31.5	UL31.5_RNA_*1	+	23	AUG	UAA	67358	67429	JN555585+:67358-67430
UL31.6 ORF (C-terminal part is identical to UL32.5 ORF)	UL31.6_ORF	ORF	UL31.6	UL31.6_RNA	+	298	AUG	UGA	68034	68930	JN555585+:68034-68931
UL32 CDS (Processing and packaging)	UL32_CDS	CDS	UL32	UL32_RNA	-	596	AUG	UGA	69161	67371	JN555585-:67371-69162
UL32.5 ORF (TaSS unclear)	UL32.5_ORF	ORF	UL32.5	UL32.5_RNA	+	147	AUG	UGA	68487	68930	JN555585+:68487-68931
UL32.5 uORF	UL32.5_uORF	uORF	UL32.5	UL32.5_RNA	+	12	AUG	UAG	68389	68427	JN555585+:68389-68428
UL32.6 ORF	UL32.6_ORF	ORF	UL32.6	UL32.6_RNA	+	79	AUG	UGA	68761	69000	JN555585+:68761-69001
UL33 CDS (Processing and packaging DNA)	UL33_CDS	CDS	UL33	UL33_RNA	+	130	AUG	UGA	69160	69552	JN555585+:69160-69553
UL33 CDS *1 (includes 21 aa N-terminal extension initiating from ACG)	UL33_CDS_*1	CDS	UL33	UL33_RNA	+	151	CUG	UGA	69097	69552	JN555585+:69097-69553
UL34 CDS (Inner nuclear membrane protein)	UL34_CDS	CDS	UL34	UL34_RNA	+	275	AUG	UAA	69632	70459	JN555585+:69632-70460
UL34.5 sORF 1 (orphan)	UL34.5_sORF_1	sORF	UL34.5	-	-	28	AUG	UAG	70540	70454	JN555585-:70454-70541
UL34.5 sORF 2 (orphan)	UL34.5_sORF_2	sORF	UL34.5	-	-	68	AUG	UGA	70137	69931	JN555585-:69931-70138
UL35 CDS (VP26; Capsid protein)	UL35_CDS	CDS	UL35	UL35_RNA	+	112	AUG	UGA	70565	70903	JN555585+:70565-70904
UL36 CDS (Omega peptidase activity; Tegument protein)	UL36_CDS	CDS	UL36	UL36_RNA	-	3129	AUG	UAG	80437	71048	JN555585-:71048-80438
UL36 CDS *1 (includes 10 aa N-terminal extension initiating from AUG)	UL36_CDS_*1	CDS	UL36	UL36_RNA	-	3139	AUG	UAG	80467	71048	JN555585-:71048-80468
UL36.4 ORF (possible truncated isoform initiating from AUGs at 73767 and 73929)	UL36.4_ORF	ORF	UL36.4	UL36.4_RNA	+	105	AUG	UGA	73725	74042	JN555585+:73725-74043
UL36.4 uORF	UL36.4_uORF	uORF	UL36.4	UL36.4_RNA	+	80	AUG	UAG	73356	73598	JN555585+:73356-73599
UL36.5 ORF	UL36.5_ORF	ORF	UL36.5	UL36.5_RNA	+	249	AUG	UGA	74887	75636	JN555585+:74887-75637
UL36.6 ORF (possible truncated isoform initiating from AUG at 80316)	UL36.6_ORF	ORF	UL36.6	UL36.6_RNA	+	100	AUG	UAG	80274	80576	JN555585+:80274-80577
UL36.6 uORF	UL36.6_uORF	uORF	UL36.6	UL36.6_RNA	+	13	AUG	UAA	79833	79874	JN555585+:79833-79875
UL37 CDS (Tegument protein; Capsid assembly)	UL37_CDS	CDS	UL37	UL37_RNA	-	1123	AUG	UAA	84083	80712	JN555585-:80712-84084
UL37 uoORF	UL37_uoORF	uoORF	UL37	UL37_RNA	-	70	CUG	UAG	84196	83984	JN555585-:83984-84197
UL37 uoORF *1 (orphan, N-terminal 23aa extension of UL37 uoORF initiating from AUG)	UL37_uoORF_*1	uoORF	UL37	-	-	93	AUG	UAG	84265	83984	JN555585-:83984-84266
UL37.3 ORF (orphan; possible 7aa N-terminal extension initiating from AUG at 81763)	UL37.3_ORF	ORF	UL37.3	-	+	104	AUG	UAG	81784	82098	JN555585+:81784-82099
UL37.4 ORF (orphan, TaSS unclear: most likely at 83320 from AUG)	UL37.4_ORF	ORF	UL37.4	-	+	265	AUG	UAG	83320	84117	JN555585+:83320-84118
UL37.4 uORF (orphan, probable uORF of the UL37.4 ORF)	UL37.4_uORF	uORF	UL37.4	-	+	80	CUG	UGA	82960	83202	JN555585+:82960-83203
UL37.5 ORF (could also initiate from second AUG directly following first AUG)	UL37.5_ORF	ORF	UL37.5	UL37.5_RNA	-	102	AUG	UAG	84853	84545	JN555585-:84545-84854

UL37.5 uORF 1	UL37.5_uORF_1	uORF	UL37.5	UL37.5_RNA	-	27	AUG	UGA	85397	85314	JN555585-:85314-85398
UL37.5 uORF 2	UL37.5_uORF_2	uORF	UL37.5	UL37.5_RNA	-	19	AUG	UAG	85227	85168	JN555585-:85168-85228
UL37.5 uORF 3 (=uORF of UL37.5 RNA #1; ends directly in front of TaSS of UL37.5 ORF)	UL37.5_uORF_3	uORF	UL37.5	UL37.5_RNA	-	53	CUG	UAG	85015	84854	JN555585-:84854-85016
UL37.6 ORF	UL37.6_ORF	ORF	UL37.6	UL37.6_RNA	-	94	AUG	UGA	85878	85594	JN555585-:85594-85879
UL37.6 ORF #1 (translated from UL37.6 RNA #1 initiating with AUG; lacks the first 24aa of UL37.6 ORF)	UL37.6_ORF_#1	ORF	UL37.6	UL37.6_RNA	-	70	AUG	UGA	85806	85594	JN555585-:85594-85807
UL37.6 uoORF	UL37.6_uoORF	uoORF	UL37.6	UL37.6_RNA	-	14	AUG	UGA	85919	85875	JN555585-:85875-85920
UL38 CDS (VP19C; Capsid protein; DNA maturation)	UL38_CDS	CDS	UL38	UL38_RNA	+	465	AUG	UGA	84530	85927	JN555585+:84530-85928
UL39 CDS (ICP6; RR-1; Ribonucleotide reductase; large subunit)	UL39_CDS	CDS	UL39	UL39_RNA	+	1137	AUG	UGA	86442	89855	JN555585+:86442-89856
UL39 CDS *1 (includes 38 aa N-terminal extension initiating from UUG)	UL39_CDS_*1	CDS	UL39	UL39_RNA	+	1175	UUG	UGA	86328	89855	JN555585+:86328-89856
UL39.4 sORF (C-terminal 23aa of UL39.5 ORF, initiating from AUG)	UL39.4_sORF	sORF	UL39.4	UL39.4_RNA	-	23	AUG	UAG	88258	88187	JN555585-:88187-88259
UL39.5 ORF	UL39.5_ORF	ORF	UL39.5	UL39.5_RNA	-	138	AUG	UAG	88603	88187	JN555585-:88187-88604
UL39.6 dORF 1	UL39.6_dORF_1	dORF	UL39.6	UL39.6_RNA	-	6	AUG	UGA	89122	89102	JN555585-:89102-89123
UL39.6 dORF 2	UL39.6_dORF_2	dORF	UL39.6	UL39.6_RNA	-	88	AUG	UAG	89044	88778	JN555585-:88778-89045
UL39.6 ORF	UL39.6_ORF	ORF	UL39.6	UL39.6_RNA	-	102	ACG	UGA	89539	89231	JN555585-:89231-89540
UL4 CDS	UL4_CDS	CDS	UL4	UL4_RNA	-	199	AUG	UAG	12422	11823	JN555585-:11823-12423
UL4.5 iORF (located in intron)	UL4.5_iORF	iORF	UL4.5	-	+	39	AUG	UAG	12616	12735	JN555585+:12616-12736
UL4.5 ORF	UL4.5_ORF	ORF	UL4.5	UL4.5_RNA	+	122	AUG	UGA	12235	13146	JN555585+:12235-12429|12972-13147
UL40 CDS (RR-2; Ribonucleotride reductase; small subunit)	UL40_CDS	CDS	UL40	UL40_RNA	+	340	AUG	UGA	89924	90946	JN555585+:89924-90947
UL40.5 iORF (alternative TaSS 2aa downstream initiating from AUG at 91425)	UL40.5_iORF	iORF	UL40.5	UL40.5_RNA	+	25	AUG	UAA	91419	91496	JN555585+:91419-91497
UL40.5 ORF (TaSS unclear, probably located further upstream)	UL40.5_ORF	ORF	UL40.5	UL40.5_RNA	+	130	AUG	UAG	91271	91663	JN555585+:91271-91664
UL40.5 uORF	UL40.5_uORF	uORF	UL40.5	UL40.5_RNA	+	4	AUU	UAG	91088	91102	JN555585+:91088-91103
UL40.6 sORF 1	UL40.6_sORF_1	sORF	UL40.6	UL40.6_RNA	+	18	CUG	UGA	91722	91778	JN555585+:91722-91779
UL40.6 sORF 2	UL40.6_sORF_2	sORF	UL40.6	UL40.6_RNA	+	22	AUG	UAG	91847	91915	JN555585+:91847-91916
UL40.7 dORF (translated from UL40.6 spliced RNA)	UL40.7_dORF	dORF	UL40.7	UL40.7_RNA	+	35	AUG	UAA	92544	92651	JN555585+:92544-92652
UL40.7A ORF	UL40.7A_ORF	ORF	UL40.7	UL40.7_RNA	+	104	AUG	UAA	92123	92437	JN555585+:92123-92438
UL40.7B ORF (in different frame than UL40.7A, overlapping C-terminal half)	UL40.7B_ORF	ORF	UL40.7	UL40.7_RNA	+	71	AUG	UAA	92265	92480	JN555585+:92265-92481
UL41 CDS (VHS; Tegument protein)	UL41_CDS	CDS	UL41	UL41_RNA	-	489	AUG	UAG	92635	91166	JN555585-:91166-92636
UL41 iORF RNA #1 (translated from UL41 RNA #1 & #2)	UL41_iORF_RNA_#1	iORF	UL41	UL41_RNA_#1	-	52	AUG	UAA	91818	91660	JN555585-:91660-91819
UL41 uoORF	UL41_uoORF	uoORF	UL41	UL41_RNA	-	9	AUU	UGA	92646	92617	JN555585-:92617-92647
UL41 uORF 1 (overlaps UL41 uoORF)	UL41_uORF_1	uORF	UL41	UL41_RNA	-	26	AUA	UGA	92722	92642	JN555585-:92642-92723
UL41 uORF 2 (iORF of UL41 uORF 1)	UL41_uORF_2	uORF	UL41	UL41_RNA	-	2	ACG	UAG	92676	92668	JN555585-:92668-92677
UL42 CDS (DNA polymerase processivity factor)	UL42_CDS	CDS	UL42	UL42_RNA	+	488	AUG	UGA	93111	94577	JN555585+:93111-94578
UL43 CDS (Membrane protein)	UL43_CDS	CDS	UL43	UL43_RNA	+	417	AUG	UGA	94798	96051	JN555585+:94798-96052
UL43.5 iORF (orphan)	UL43.5_iORF	iORF	UL43.5	-	-	25	AUG	UAG	93317	93240	JN555585-:93240-93318
UL43.5 ORF (orphan; TaSS unclear)	UL43.5_ORF	ORF	UL43.5	-	-	124	UUG	UGA	93454	93080	JN555585-:93080-93455
UL43.5 uORF (orphan)	UL43.5_uORF	uORF	UL43.5	-	-	16	CUG	UGA	93571	93521	JN555585-:93521-93572
UL43.6 ORF (orphan)	UL43.6_ORF	ORF	UL43.6	-	-	311	AUG	UGA	95715	94780	JN555585-:94780-95716
UL43.6 uORF (orphan; may have alternative TaSS further upstream)	UL43.6_uORF	uORF	UL43.6	-	-	6	AUG	UGA	95807	95787	JN555585-:95787-95808
UL44 CDS (Glycoprotein C)	UL44_CDS	CDS	UL44	UL44_RNA	+	511	AUG	UAA	96311	97846	JN555585+:96311-97847
UL44 iORF	UL44_iORF	iORF	UL44	UL44_RNA	+	105	AUG	UAG	96519	96836	JN555585+:96519-96837
UL44 ORF RNA iso1 (altered C-terminus)	UL44_ORF_RNA_iso1	ORF	UL44	UL44_RNA_iso1	+	489	AUG	UGA	96311	98006	JN555585+:96311-97724|97950-98007

UL44.5 ORF (orphan)	UL44.5_ORF	ORF	UL44.5	-	-	162	AUG	UAG	96765	96277	JN555585-:96277-96766
UL45 CDS (Membrane protein; C-type lectin)	UL45_CDS	CDS	UL45	UL45_RNA	+	172	AUG	UGA	98032	98550	JN555585+:98032-98551
UL46 CDS (VP11; VP12; Tegument protein)	UL46_CDS	CDS	UL46	UL46_RNA	-	718	AUG	UGA	100952	98796	JN555585-:98796-100953
UL46.5 sORF 1	UL46.5_sORF_1	sORF	UL46.5	UL46.5_RNA	+	4	AUG	UAG	101948	101962	JN555585+:101948-101963
UL46.5 sORF 2	UL46.5_sORF_2	sORF	UL46.5	UL46.5_RNA	+	74	AUG	UAA	102230	102454	JN555585+:102230-102455
UL47 CDS (Vp13; VP14; Tegument protein)	UL47_CDS	CDS	UL47	UL47_RNA	-	693	AUG	UAA	103116	101035	JN555585-:101035-103117
UL48 CDS (VP16; Alpha-TIF; Tegument protein; Virion maturation)	UL48_CDS	CDS	UL48	UL48_RNA	-	490	AUG	UAG	105079	103607	JN555585-:103607-105080
UL49 CDS (VP22; Tegument protein)	UL49_CDS	CDS	UL49	UL49_RNA	-	301	AUG	UGA	106391	105486	JN555585-:105486-106392
UL49 uORF	UL49_uORF	uORF	UL49	UL49_RNA	-	17	AUC	UAA	106482	106429	JN555585-:106429-106483
UL49.5 CDS (UL49A; Glycoprotein N)	UL49.5_CDS	CDS	UL49.5	UL49.5_RNA	-	91	AUG	UGA	106993	106718	JN555585-:106718-106994
UL49.5 uORF	UL49.5_uORF	uORF	UL49.5	UL49.5_RNA	-	25	ACG	UGA	107092	107015	JN555585-:107015-107093
UL5 CDS (ATP binding; helicate-primase complex-associated protein)	UL5_CDS	CDS	UL5	UL5_RNA	-	882	AUG	UAA	15131	12483	JN555585-:12483-15132
UL5 CDS *1 (includes 9 aa N-terminal extension initiating from GUG)	UL5_CDS_*1	CDS	UL5	UL5_RNA	-	891	GUG	UAA	15158	12483	JN555585-:12483-15159
UL5 uORF (=C-terminal 47aa of US6.4 sORF 2)	UL5_uORF	uORF	UL5	UL5_RNA	-	47	GUG	UAA	15518	15375	JN555585-:15375-15519
UL5.5 dORF (TaSS only 3nt downstream of TiSS UL6 RNA*1)	UL5.5_dORF	dORF	UL5.5	UL5.5_RNA	+	30	AUG	UAA	14931	15023	JN555585+:14931-15024
UL5.5 iORF 1	UL5.5_iORF_1	iORF	UL5.5	UL5.5_RNA	+	11	AUG	UAG	13708	13743	JN555585+:13708-13744
UL5.5 iORF 2	UL5.5_iORF_2	iORF	UL5.5	UL5.5_RNA	+	50	AUG	UAG	13750	13902	JN555585+:13750-13903
UL5.5 iORF 3	UL5.5_iORF_3	iORF	UL5.5	UL5.5_RNA	+	24	AUG	UAG	13918	13992	JN555585+:13918-13993
UL5.5 ORF (TaSS unclear; ORF validated by two independent peptides)	UL5.5_ORF	ORF	UL5.5	UL5.5_RNA	+	407	CUG	UAA	13680	14903	JN555585+:13680-14904
UL5.5 uORF	UL5.5_uORF	uORF	UL5.5	UL5.5_RNA	+	3	AUG	UAG	13402	13413	JN555585+:13402-13414
UL5.6 dORF (orphan)	UL5.6_dORF	dORF	UL5.6	UL5.6_RNA	+	63	AUG	UAG	14389	14580	JN555585+:14389-14581
UL5.6 ORF (located within UL5.5 ORF, different frame)	UL5.6_ORF	ORF	UL5.6	UL5.6_RNA	+	112	AUG	UGA	14026	14364	JN555585+:14026-14365
UL5.7 ORF	UL5.7_ORF	ORF	UL5.7	UL5.7_RNA	+	124	AUG	UGA	14698	15072	JN555585+:14698-15073
UL50 CDS (dUTPase)	UL50_CDS	CDS	UL50	UL50_RNA	+	371	AUG	UAG	107010	108125	JN555585+:107010-108126
UL50 CDS *1 (includes 10 aa N-terminal extension intitating from CUG)	UL50_CDS_*1	CDS	UL50	UL50_RNA	+	381	CUG	UAG	106980	108125	JN555585+:106980-108126
UL50 CDS *2 (includes 23 aa N-terminal extension intitating from ACG)	UL50_CDS_*2	CDS	UL50	UL50_RNA	+	394	ACG	UAG	106941	108125	JN555585+:106941-108126
UL50.5 sORF 1	UL50.5_sORF_1	sORF	UL50.5	UL50.5_RNA	-	51	AUG	UGA	107401	107246	JN555585-:107246-107402
UL50.5 sORF 2 (TaSS unclear; shares C-terminal part with UL49.5 uORF 1)	UL50.5_sORF_2	sORF	UL50.5	UL50.5_RNA	-	68	AUA	UGA	107221	107015	JN555585-:107015-107222
UL50.6 sORF	UL50.6_sORF	sORF	UL50.6	UL50.6_RNA	-	47	AUG	UGA	107743	107600	JN555585-:107600-107744
UL51 CDS (Tegument protein)	UL51_CDS	CDS	UL51	UL51_RNA	-	244	AUG	UAA	109011	108277	JN555585-:108277-109012
UL51 uoORF (TaSS unclear)	UL51_uoORF	uoORF	UL51	UL51_RNA	-	50	ACG	UGA	109100	108948	JN555585-:108948-109101
UL51 uORF (TaSS unclear)	UL51_uORF	uORF	UL51	UL51_RNA	-	15	GUG	UAA	109163	109116	JN555585-:109116-109164
UL51.5 sORF 1 (orphan)	UL51.5_sORF_1	sORF	UL51.5	-	+	11	AUG	UAG	108309	108344	JN555585+:108309-108345
UL51.5 sORF 2 (orphan)	UL51.5_sORF_2	sORF	UL51.5	-	+	51	AUG	UGA	108545	108700	JN555585+:108545-108701
UL52 CDS (DNA helicase/primase complex protein)	UL52_CDS	CDS	UL52	UL52_RNA	+	1058	AUG	UGA	109048	112224	JN555585+:109048-112225
UL52.4 sORF (may include N-terminal 4aa extension initiating from CUG)	UL52.4_sORF	sORF	UL52.4	UL52.4_RNA	-	88	AUG	UAG	109544	109278	JN555585-:109278-109545
UL52.5 sORF 1	UL52.5_sORF_1	sORF	UL52.5	UL52.5_RNA	-	12	AUG	UAA	112057	112019	JN555585-:112019-112058
UL52.5 sORF 2	UL52.5_sORF_2	sORF	UL52.5	UL52.5_RNA	-	54	AUG	UAG	111881	111717	JN555585-:111717-111882
UL53 CDS (Glycoprotein K)	UL53_CDS	CDS	UL53	UL53_RNA	+	338	AUG	UGA	112179	113195	JN555585+:112179-113196
UL53 iORF RNA #2	UL53_iORF_RNA_#2	iORF	UL53	UL53_RNA_#2	+	82	AUG	UGA	112532	112780	JN555585+:112532-112781
UL53 uORF	UL53_uORF	uORF	UL53	UL53_RNA	+	16	AUG	UAG	111821	111871	JN555585+:111821-111872
UL53.5 sORF 1	UL53.5_sORF_1	sORF	UL53.5	UL53.5_RNA	-	55	AUG	UGA	112936	112769	JN555585-:112769-112937
UL53.5 sORF 2	UL53.5_sORF_2	sORF	UL53.5	UL53.5_RNA	-	38	CUG	UAA	112273	112157	JN555585-:112157-112274
UL54 CDS (ICP27; IE63; Transcriptional regulation)	UL54_CDS	CDS	UL54	UL54_RNA	+	512	AUG	UAG	113734	115272	JN555585+:113734-115273
UL54 CDS *1 (includes 38 aa N-terminal extension initiating from ACG)	UL54_CDS_*1	CDS	UL54	UL54_RNA	+	550	ACG	UAG	113620	115272	JN555585+:113620-115273

UL56 CDS (Membrane protein)	UL56_CDS	CDS	UL56	UL56_RNA	-	234	AUG	UAA	116925	116221	JN555585-:116221-116926
UL56 uORF	UL56_uORF	uORF	UL56	UL56_RNA	-	37	AUG	UAG	117070	116957	JN555585-:116957-117071
UL6 CDS (Virion portal protein)	UL6_CDS	CDS	UL6	UL6_RNA	+	676	AUG	UGA	15130	17160	JN555585+:15130-17161
UL6 iORF RNA #1	UL6_iORF_RNA_#1	iORF	UL6	UL6_RNA_#1	+	69	AUG	UAA	15888	16097	JN555585+:15888-16098
UL6 uORF	UL6_uORF	uORF	UL6	UL6_RNA	+	12	UUG	UAG	15066	15104	JN555585+:15066-15105
UL6 uORF RNA *2	UL6_uORF_RNA_*2	uORF	UL6	UL6_RNA_*2	+	31	AUG	UGA	14977	15072	JN555585+:14977-15073
UL6.4 sORF 1	UL6.4_sORF_1	sORF	UL6.4	UL6.4_RNA	-	13	AUG	UAG	15869	15828	JN555585-:15828-15870
UL6.4 sORF 2 (may include N-terminal 28aa extension initiating from GUG)	UL6.4_sORF_2	sORF	UL6.4	UL6.4_RNA	-	89	AUG	UAA	15644	15375	JN555585-:15375-15645
UL6.5 sORF 1	UL6.5_sORF_1	sORF	UL6.5	UL6.5_RNA	-	9	AUG	UAG	16637	16608	JN555585-:16608-16638
UL6.5 sORF 2	UL6.5_sORF_2	sORF	UL6.5	UL6.5_RNA	-	8	AUG	UAG	16427	16401	JN555585-:16401-16428
UL6.5 sORF 3	UL6.5_sORF_3	sORF	UL6.5	UL6.5_RNA	-	54	AUG	UAG	16382	16218	JN555585-:16218-16383
UL6.5 sORF 4	UL6.5_sORF_4	sORF	UL6.5	UL6.5_RNA	-	73	AUG	UGA	16199	15978	JN555585-:15978-16200
UL6.6 ORF	UL6.6_ORF	ORF	UL6.6	UL6.6_RNA	-	133	AUG	UAA	17313	16912	JN555585-:16912-17314
UL6.7 sORF 1 (orphan)	UL6.7_sORF_1	sORF	UL6.7	-	-	6	AUG	UAG	17919	17899	JN555585-:17899-17920
UL6.7 sORF 2 (orphan)	UL6.7_sORF_2	sORF	UL6.7	-	-	43	AUG	UGA	17858	17727	JN555585-:17727-17859
UL7 CDS (Tegument protein)	UL7_CDS	CDS	UL7	UL7_RNA	+	296	AUG	UGA	17135	18025	JN555585+:17135-18026
UL7 CDS *1 (includes 15 aa N-terminal extension initiating from CUG)	UL7_CDS_*1	CDS	UL7	UL7_RNA	+	311	CUG	UGA	17090	18025	JN555585+:17090-18026
UL7.5 ORF	UL7.5_ORF	ORF	UL7.5	UL7.5_RNA	+	190	AUG	UAA	18885	19457	JN555585+:18885-19458
UL7.5 uORF	UL7.5_uORF	uORF	UL7.5	UL7.5_RNA	+	72	AUG	UAA	18643	18861	JN555585+:18643-18862
UL8 CDS (helicate-primase complex-associated protein)	UL8_CDS	CDS	UL8	UL8_RNA	-	750	AUG	UGA	20476	18224	JN555585-:18224-20477
UL8 iORF	UL8_iORF	iORF	UL8	UL8_RNA	-	115	AUG	UGA	20414	20067	JN555585-:20067-20415
UL8 uORF	UL8_uORF	uORF	UL8	UL8_RNA	-	7	AUG	UGA	20517	20494	JN555585-:20494-20518
UL8.3 ORF (orphan)	UL8.3_ORF	ORF	UL8.3	-	+	121	AUG	UAG	19923	20288	JN555585+:19923-20289
UL8.4 ORF (may include a small truncated isoform of 23aa initiating at 21143 from ACG)	UL8.4_ORF	ORF	UL8.4	UL8.4_RNA	+	158	AUG	UAG	20738	21214	JN555585+:20738-21215
UL8.5 CDS (ATP binding; truncated isoform of UL9; Bahradaran et al, 1994)	UL8.5_CDS	CDS	UL8.5	UL8.5_RNA	-	487	AUG	UAA	22167	20704	JN555585-:20704-22168
UL9 CDS (ATP binding; Replication origin-binding protein)	UL9_CDS	CDS	UL9	UL9_RNA	-	851	AUG	UAA	23259	20704	JN555585-:20704-23260
UL9.4 dORF (orphan)	UL9.4_dORF	dORF	UL9.4	-	+	22	AUG	UGA	22786	22854	JN555585+:22786-22855
UL9.4 ORF (orphan)	UL9.4_ORF	ORF	UL9.4	-	+	118	AUG	UAG	22355	22711	JN555585+:22355-22712
UL9.5 iORF 1	UL9.5_iORF_1	iORF	UL9.5	UL9.5_RNA	-	17	AUG	UGA	24210	24157	JN555585-:24157-24211
UL9.5 iORF 2	UL9.5_iORF_2	iORF	UL9.5	UL9.5_RNA	-	17	AUG	UAG	24087	24034	JN555585-:24034-24088
UL9.5 ORF	UL9.5_ORF	ORF	UL9.5	UL9.5_RNA	-	472	AUG	UGA	24304	22886	JN555585-:22886-24305
UL9.5 uoORF	UL9.5_uoORF	uoORF	UL9.5	UL9.5_RNA	-	29	AUG	UGA	24333	24244	JN555585-:24244-24334
UL9.5 uORF	UL9.5_uORF	uORF	UL9.5	UL9.5_RNA	-	27	AUG	UAA	24552	24469	JN555585-:24469-24553
US1 CDS (ICP22; IE63; Viral replication)	US1_CDS	CDS	US1	US1_RNA	+	420	AUG	UGA	132646	133908	JN555585+:132646-133909
US1 CDS *1 (includes N-terminal extension initiating from GUG, 8 aa, or even from GUG, 21 aa)	US1_CDS_*1	CDS	US1	US1_RNA	+	428	GUG	UGA	132622	133908	JN555585+:132622-133909
US1.5 CDS (truncated isoform of US1)	US1.5_CDS	CDS	US1.5	US1.5_RNA	+	250	AUG	UGA	133156	133908	JN555585+:133156-133909
US10 CDS (Tegument protein)	US10_CDS	CDS	US10	US10_RNA	-	312	AUG	UAG	145099	144161	JN555585-:144161-145100
US11 CDS (Vmw21; RNA binding protein)	US11_CDS	CDS	US11	US11_RNA	-	161	AUG	UAG	145250	144765	JN555585-:144765-145251
US11 CDS *1 (includes 4aa N-terminal extension initiating with GUG)	US11_CDS_*1	CDS	US11	US11_RNA	-	165	GUG	UAG	145262	144765	JN555585-:144765-145263
US11.5 ORF (orphan)	US11.5_ORF	ORF	US11.5	-	+	126	AUG	UGA	145298	145678	JN555585+:145298-145679
US12 CDS (ICP47; IE12)	US12_CDS	CDS	US12	US12_RNA	-	88	AUG	UGA	145581	145315	JN555585-:145315-145582
US2 CDS (Tegument protein)	US2_CDS	CDS	US2	US2_RNA	-	291	AUG	UAG	134930	134055	JN555585-:134055-134931
US2.5 ORF (TaSS unclear)	US2.5_ORF	ORF	US2.5	US2.5_RNA	+	133	UUG	UAA	134767	135168	JN555585+:134767-135169
US2.6 ORF (orphan; probable alternative TaSS from first downstream AUG)	US2.6_ORF	ORF	US2.6	-	-	315	AUG	UAG	135970	135023	JN555585-:135023-135971
US3 CDS (Serine/threonine-protein kinase)	US3_CDS	CDS	US3	US3_RNA	+	481	AUG	UGA	135224	136669	JN555585+:135224-136670

US3 CDS *1 (includes 23 aa N-terminal extension initiating from CUG)	US3_CDS_*1	CDS	US3	US3_RNA	+	504	CUG	UGA	135155	136669	JN555585+:135155-136670
US3.5 CDS (truncated isoform of US3)	US3.5_CDS	CDS	US3	US3_RNA	+	405	AUG	UGA	135452	136669	JN555585+:135452-136670
US3.6 sORF (C-terminal 81aa of US4.5 ORF initiating from AUG)	US3.6_sORF	sORF	US3.6	US3.6_RNA	-	80	AUG	UGA	136468	136226	JN555585-:136226-136469
US3.6 uORF (4aa, initiating from AUG)	US3.6_uORF	uORF	US3.6	US3.6_RNA	-	4	AUG	UAG	136569	136555	JN555585-:136555-136570
US4 CDS (Glycoprotein G)	US4_CDS	CDS	US4	US4_RNA	+	238	AUG	UAG	136746	137462	JN555585+:136746-137463
US4.5 ORF (orphan, TaSS unclear)	US4.5_ORF	ORF	US4.5	-	-	222	AUG	UGA	136894	136226	JN555585-:136226-136895
US5 CDS (Glycoprotein J)	US5_CDS	CDS	US5	US5_RNA	+	92	AUG	UAA	137733	138011	JN555585+:137733-138012
US5 CDS *1 (includes 16 aa N-terminal extension initiating from CUG)	US5_CDS_*1	CDS	US5	US5_RNA	+	108	CUG	UAA	137685	138011	JN555585+:137685-138012
US5 dORF	US5_dORF	dORF	US5	US5_RNA	+	37	AUG	UGA	138112	138225	JN555585+:138112-138226
US5.5 sORF	US5.5_sORF	sORF	US5.5	US5.5_RNA	-	20	AUG	UAG	138834	138772	JN555585-:138772-138835
US6 CDS (Glycoprotein D)	US6_CDS	CDS	US6	US6_RNA	+	394	AUG	UAG	138422	139606	JN555585+:138422-139607
US6 uORF	US6_uORF	uORF	US6	US6_RNA	+	4	AUC	UAA	138380	138394	JN555585+:138380-138395
US7 CDS (Glycoprotein I)	US7_CDS	CDS	US7	US7_RNA	+	390	AUG	UAG	139788	140960	JN555585+:139788-140961
US8 CDS (Glycoprotein E)	US8_CDS	CDS	US8	US8_RNA	+	550	AUG	UAA	141246	142898	JN555585+:141246-142899
US8 CDS *1 (includes 7aa N-terminal extension initiating from AUC)	US8_CDS_*1	CDS	US8	US8_RNA	+	557	AUC	UAA	141225	142898	JN555585+:141225-142899
US8.5 CDS (US8A)	US8.5_CDS	CDS	US8.5	US8.5_RNA	+	159	AUG	UAA	142747	143226	JN555585+:142747-143227
US9 CDS (Membrane protein)	US9_CDS	CDS	US9	US9_RNA	+	90	AUG	UAA	143316	143588	JN555585+:143316-143589
US9 uoORF	US9_uoORF	uoORF	US9	US9_RNA	+	18	AUC	UGA	143263	143319	JN555585+:143263-143320

Table B.4: Information about the function, localization and GO terms of identified ORFs. The column containing the sequence was omitted for better readability. The full table can be obtained online at `https://www.nature.com/articles/s41467-020-15992-5#Sec19`

Name	Length (aa)	Function	Localization	GO terms
CDS				
IRL1	248	Phosphatase-1 catalytic subunit binding region	endoplasmic reticulum	no GO terms
IRL2	775	disruption by symbiont of host cell PML body	endoplasmic reticulum membrane	F:metal ion binding
IRS1	1298	Herpesvirus ICP4-like protein N-terminal region		C:host cell nucleus; P:positive regulation of transcription, DNA-templated
Ori-S	330	no annotation		no GO terms
UL1	224	Herpesvirus glycoprotein L		no GO terms
UL11	96	anatomical structure morphogenesis	secreted	P:anatomical structure morphogenesis; C:viral tegument
UL12	626	exonuclease activity	nucleus	F:DNA binding; F:exonuclease activity
UL12.5	500	exonuclease activity	mitochondrion	F:DNA binding; F:exonuclease activity
UL13	518	protein serine/threonine kinase activity	cytoplasm	F:protein kinase activity; F:ATP binding; P:protein phosphorylation
UL14	219	Herpesvirus UL14-like protein	nucleus	no GO terms
UL15	735	Probable DNA packing protein, C-terminus		P:DNA packaging
UL15.5	293	Probable DNA packing protein, C-terminus		P:DNA packaging
UL16	373	Herpesvirus UL16/UL94 family	secreted	no GO terms
UL17	703	Herpesvirus UL17 protein	cytoplasm	P:DNA packaging; C:virion
UL18	318	Herpesvirus VP23 like capsid protein	nucleus	F:structural molecule activity; C:viral capsid
UL19	1374	Herpes virus major capsid protein	mitochondrion	F:structural molecule activity; C:viral capsid
UL2	334	base-excision repair	cytoplasm	F:uracil DNA N-glycosylase activity; P:DNA repair; P:base-excision repair
UL20	222	Herpesvirus egress protein UL20	endoplasmic reticulum membrane	P:viral life cycle
UL20.5	160	no annotation	mitochondrion	no IPS match
UL21	535	Herpesvirus UL21	cytoplasm	no GO terms
UL22	838	Herpesvirus glycoprotein H	endoplasmic reticulum membrane	no GO terms
UL23	376	ATP binding	nucleus; Deoxynucleoside kinase; Thymidine kinase	F:thymidine kinase activity; F:ATP binding; P:TMP biosynthetic process
UL24	269	exonuclease activity		no GO terms
UL25	580	viral genome packaging	cytoplasm	P:viral genome packaging; C:host cell nucleus
UL26	635	Assemblin (Peptidase family S21)	plasma membrane; Acting on peptide bonds (peptidases)	F:serine-type endopeptidase activity; P:proteolysis
UL27	904	Herpesvirus Glycoprotein B	plasma membrane	no GO terms
UL28	785	protein processing	nucleus	P:viral DNA genome packaging
UL28.5	681	ssDNA binding protein, Caspase-3/ Caspase-7 cleavage site	cytoplasm	F:single-stranded DNA binding; P:DNA replication; C:host cell nucleus

UL29	1196	ssDNA binding protein, Caspase-3/ Caspase-7 cleavage site	cytoplasm	*F:single-stranded DNA binding; P:DNA replication; C:host cell nucleus*
UL29 prefix	515	ssDNA binding protein, Caspase-3/ Caspase-7 cleavage site	cytoplasm	BLAST: ssDNA binding protein; MAPK interacting molecules; ER retention motif
UL3	224	Herpesvirus UL3 protein	nucleus	no GO terms
UL30	1235	DNA-directed DNA polymerase activity	Acting on ester bonds; Acting on ester bonds; DNA-directed DNA polymerase; Ribonuclease H	F:nucleotide binding; F:nucleic acid binding; F:DNA binding; F:DNA-directed DNA polymerase activity; F:RNA-DNA hybrid ribonuclease activity
UL31	306	Herpesvirus UL31-like protein	nucleus	P:viral budding from nuclear membrane; P:exit of virus from host cell nucleus by nuclear egress
UL32	596	Herpesvirus putative major envelope glycoprotein	cytoplasm	C:viral envelope
UL33	130	viral DNA genome packaging		P:viral DNA genome packaging
UL34	275	Herpesvirus virion protein U34		no GO terms
UL35	112	Herpesvirus UL35 family		C:viral capsid
UL36	3129	omega peptidase activity	Ubiquitinyl hydrolase 1	F:cysteine-type peptidase activity; F:NEDD8-specific protease activity; F:thiol-dependent ubiquitinyl hydrolase activity; P:viral DNA genome replication
UL37	1123	Herpesvirus UL37 tegument protein	cytoplasm	C:viral tegument; P:virion assembly
UL38	465	Herpesvirus capsid shell protein VP19C		F:DNA binding; P:viral capsid assembly
UL39	1137	Ribonucleotide reductase, all-alpha domain	Acting on CH or CH(2) groups; Ribonucleoside-diphosphate reductase	F:ribonucleoside-diphosphate reductase activity, thioredoxin disulfide as acceptor; F:ATP binding; P:DNA replication; P:oxidation-reduction process
UL4	199	Herpesvirus UL4 family	secreted	no GO terms
UL40	340	deoxyribonucleoside diphosphate metabolic process		P:deoxyribonucleotide biosynthetic process; F:oxidoreductase activity; P:oxidation-reduction process
UL41	489	induction by virus of catabolism of host mRNA		F:nuclease activity
UL42	488	DNA polymerase processivity factor (UL42)		F:DNA binding; P:DNA replication
UL43	417	Herpesvirus UL43 protein	chloroplast membrane	C:membrane; C:viral tegument
UL44	511	host cell junction	nucleus	no GO terms
UL45	172	UL45 protein	plasma membrane	no GO terms
UL46	718	Herpesvirus UL46 protein		P:regulation of transcription, DNA-templated
UL47	693	Herpesvirus UL47 protein		P:regulation of transcription, DNA-templated
UL48	490	Alpha trans-inducing protein (Alpha-TIF)	nucleus	F:DNA binding; P:regulation of transcription, DNA-templated

UL49	301	Herpesvirus UL49 tegument protein	nucleus	no GO terms
UL49.5	91	host cell Golgi membrane		no GO terms
UL5	882	ATP binding	cytoplasm; Nucleoside-triphosphate phosphatase	F:helicase activity; F:ATP binding
UL50	371	dUTP diphosphatase activity	cytoplasm; dUTP diphosphatase; Acting on acid anhydrides	F:dUTP diphosphatase activity; P:dUTP metabolic process
UL51	244	Herpesvirus UL51 protein	cytoplasm	no GO terms
UL52	1058	Herpesviridae UL52/UL70 DNA primase	cytoplasm; DNA-directed RNA polymerase	F:DNA primase activity; P:bidirectional double-stranded viral DNA replication
UL53	338	positive regulation of syncytium formation by virus	plasma membrane	C:membrane
UL54	512	Herpesvirus transcriptional regulator family	nucleus	P:regulation of transcription, DNA-templated
UL55	186	Herpesvirus UL55 protein	secreted	P:viral life cycle
UL56	234	Herpesvirus UL56 protein	secreted	no GO terms
UL6	676	Herpesvirus UL6 like	nucleus	P:DNA packaging
UL7	296	Herpesvirus UL7 like	cytoplasm	no GO terms
UL8	750	Herpesvirus DNA helicase/primase complex associated protein		P:viral genome replication
UL8.5	487	ATP binding	mitochondrion	F:DNA replication origin binding; F:ATP binding; P:DNA replication
UL9	851	ATP binding	nucleus	F:DNA replication origin binding; F:ATP binding; P:DNA replication
US1	420	suppression by virus of host RNA polymerase II activity	nucleus	no GO terms
US1.5	250	suppression by virus of host RNA polymerase II activity	nucleus	no GO terms
US10	312	Gene 66 (IR5) protein	secreted	F:zinc ion binding
US11	161	suppression by virus of host RIG-I activity	nucleus	no GO terms
US12	88	suppression by virus of host TAP complex	secreted	P:evasion or tolerance of host defenses by virus
US2	291	US2 family	secreted	no GO terms
US3	481	protein serine/threonine kinase activity	cytoplasm	F:protein kinase activity; F:ATP binding; P:protein phosphorylation
US3.5	405	protein serine/threonine kinase activity	cytoplasm	F:protein kinase activity; F:ATP binding; P:protein phosphorylation
US4	238	host cell junction		no GO terms
US5	92	no annotation	secreted	P:suppression by virus of host apoptotic process
US6	394	host cell junction	plasma membrane	C:integral component of membrane
US7	390	host cell junction	plasma membrane	C:host cell
US8	550	host cell junction	plasma membrane	no GO terms
US8.5	159	host cell nucleolus	secreted	C:host cell nucleolus
US9	90	host cell smooth endoplasmic reticulum membrane		P:intracellular transport of virus

ORFs with N-terminal extensions (NTE)				
TRS1	1298	Herpesvirus ICP4-like protein N-terminal region		C:host cell nucleus; P:positive regulation of transcription, DNA-templated
UL10	473	Herpesvirus glycoprotein M		no GO terms; BLAST: glycoprotein M (gM) is an integral membrane protein
UL19	1385	Herpes virus major capsid protein		F:structural molecule activity; C:viral capsid
UL21	600	Herpesvirus UL21		no GO terms; BLAST: Tegument protein // The UL21 protein appears to be a dispensable component in herpesviruses *see Word-Doc.
UL23	391	ATP binding	Deoxynucleoside kinase; Thymidine kinase	F:thymidine kinase activity; F:ATP binding; P:TMP biosynthetic process
UL24.5	148	exonuclease activity		no GO terms; BLAST: The UL24 gene (product of the 24th ORF) is not essential for virus replication, and mutants with lesions in UL24 show a reduced ability to replicate in tissue culture and have reduced thymidine kinase activity, as the UL24 gene overlaps with thymidine kinase
UL26.5	329	Assemblin (Peptidase family S21)	Acting on peptide bonds (peptidases)	F:serine-type endopeptidase activity; P:proteolysis
UL27	947	Herpesvirus Glycoprotein B		no GO terms; BLAST: Envelope glycoprotein B
UL33	151	viral DNA genome packaging		P:viral DNA genome packaging
UL36	3139	Omega peptidase activity	Ubiquitinyl hydrolase 1	F:cysteine-type peptidase activity; F:NEDD8-specific protease activity; F:thiol-dependent ubiquitinyl hydrolase activity; P:viral DNA genome replication
UL39	1175	Ribonucleotide reductase, all-alpha domain	Acting on CH or CH(2) groups; Ribonucleoside-diphosphate reductase	F:ribonucleoside-diphosphate reductase activity, thioredoxin disulfide as acceptor; F:ATP binding; P:DNA replication; P:oxidation-reduction process
UL5	891	ATP binding	Nucleoside-triphosphate phosphatase	F:helicase activity; F:ATP binding
UL50	381	dUTP diphosphatase activity	dUTP diphosphatase; Acting on acid anhydrides	F:dUTP diphosphatase activity; P:dUTP metabolic process
UL54	550	Herpesvirus transcriptional regulator family		P:regulation of transcription, DNA-templated
UL7	311	Herpesvirus UL7 like		no GO terms; BLAST: Protein BBRF2 // Gene 42 protein // Protein UL7 bzw: functionally undefined proteins

US1	428	suppression by virus of host RNA polymerase II activity		no GO terms; BLAST: Transcriptional regulator ICP22 homolog
US11	165	suppression by virus of host RIG-I activity		no GO terms; BLAST: No PFAM-Numbers but repeats confirm a structure protein , see Word-Doc.
US3	504	protein serine/threonine kinase activity		F:protein kinase activity; F:ATP binding; P:protein phosphorylation
US5	108	dUTP diphosphatase activity		P:suppression by virus of host apoptotic process
US8	557	host cell junction		no GO terms; BLAST: Envelope glycoprotein E
All other ORFs				
IRL1A	93	hypothetic disoder protein, low identity to cadherin		*hypothetic protein, low identity to protein 4.1, AnDom on PDB: Capsid-Protein dengue Virus Helicase; Herpesvirus Glycoprotein B e:0,038; BLAST: No significant similarity found by exclude Herpesvirales; Reverse (C to N direction) of the classical MAPK docking motif; ER retention motif;*
IRL2 ORF RNA i1 (RL2 ORF with different C-terminus due to intron retention of intron 1)	72	hypothetic protein, low identity to protein 4.1, disruption by symbiont of host cell PML body		hypothetic disoder protein, low identity to cadherin
UL15 ORF RNA i1	550	Probable DNA packing protein, C-terminus		
IRL1.5 sORF	106		secreted	No clear hits
IRL2 dORF RNA i1 (orphan, located within intron 1)	42			
IRL2.5 ORF (TaSS unclear due to repeat regions upstream)	122			
IRL2A ORF	181	basic hydroxyproline-rich glycoprotein		ICP0B; hypothetical protein, basic hydroxyproline-rich glycoprotein DZ-HRGP; ICP0B; Psi-Blast: (10-6) α-0-Gen; alternate splicing RNA; regulatory protein (75% sim PNAS paper (Carter,K.L. and Roizman,B.); BLAST: No significant similarity found by exclude Herpesvirales, No PFAM-Numbers;Prosite: Proline rich Region; Proline-Directed Kinase (e.g. MAPK); ER retention motif
ORF-O (TaSS probably 76nt upstream of ORF-P initiating from ACG start codon; no evidence of frame-shift; orphan)	271		secreted	
ORF-P	233			No clear hits
UL1 uORF	19			No clear hits

UL10 uORF (TaSS unclear; could also be 23012 or 23039)	26		
UL12 uoORF (TaSS unclear, could also be at 26971)	23		
UL13.5 ORF	141		No clear hits
UL13.5 sORF #1 (truncated version of UL13.5 ORF initiating from AUG)	85		
UL14 uORF 1	21		No clear hits
UL14 uORF 2	32		Glucoamylase
UL15 iORF 1 (located in intron; orphan)	26		
UL15 iORF 2 (located in intron; orphan)	73		
UL15 iORF 3 (located in exon 2; orphan)	46		
UL15 uoORF	38		No clear hits
UL15 uORF	29		Feruloyl esterase domain; Carboxylesterase
UL15.4 sORF 1 (orphan; possible alternative TaSS at 30021, AUG)	65	secreted	
UL15.4 sORF 2 (orphan)	58	secreted	
UL17 uoORF (TaSS unclear)	11	secreted	
UL17 uORF	49	secreted	SCOP b.1.1.1: V set domains (antibody variable domain-like)
UL17 uORF RNA *1	27		hypothetical protein [Human alphaherpesvirus 1]
UL17.3 sORF (orphan, contains 36aa N-terminal extension of UL17 uORF 1 RNA *1 initiating from AUG)	67		
UL17.4 sORF (orphan)	59		
UL17.5 ORF (orphan)	96		
UL17.5 uoORF (orphan)	14		
UL18 uoORF	74	secreted	hypothetical protein TM1158 [Thermotoga maritima]; SCOP c.23.16.1: Family Class I glutamine amidotransferases (GAT)
UL18 uORF	11		No clear hits
UL19.4 sORF (orphan)	6		
UL19.5 sORF (may be part of new ORF initiating at 39420)	68		
UL20 iORF (ORF initiates 13nt downstream of UL20 from AUG)	59		
UL20 uORF (corresponds to C-terminal part of UL20.5 CDS)	21		
UL20.4 dORF 1	21		No clear hits
UL20.4 dORF 2 (alternative TaSS at next AUG 6nt downstream at 41487)	75		
UL20.4 ORF	109		No clear hits
UL20.6 ORF	181	secreted	hypothetical protein [Human alphaherpesvirus 1]
UL20.6 uORF	60	secreted	putative ATP-dependent RNA helicase, SCOP c.37.1.19: Family: Tandem AAA-ATPase domain
UL22 uORF 1 (=uoORF for UL22 uORF 2)	30		
UL22 uORF 2 (possible alternative TaSS downstream at 46468, GUG)	35		

UL22.5 sORF	51	hypothetical protein [Human alphaherpesvirus 1]
UL23 uORF 1	5	No clear hits
UL23 uORF 2 (uoORF for N-terminal extension of UL23; C-terminal 7 aa of UL23.6 ORF)	7	
UL23.5 ORF	133	hypothetical protein ACS92 08505 [Bacillus cereus]
UL23.5 uORF	54	N-acetylglucosamine-6-phosphate deacetylase; ron-responsive element binding protein 1, SCOP c.8.2.1; Family: LeuD-like
UL23.6 ORF (orphan)	154	
UL24 uORF 1 RNA *1	29	hypothetical protein A3H34 05385 [Betaproteobacteria bacterium RIFCSPLOWO2 02 FULL 67 19]
UL24 uORF 2 RNA *1	13	No clear hits
UL24 uORF 3 RNA *1	9	No clear hits
UL25.4 ORF (orphan)	241	
UL25.5 ORF	127	No clear hits
UL27.5 sORF	48	hypothetical protein [Human alphaherpesvirus 1]
UL27.6 ORF	214	RNHCP domain containing (bacterial protein family, Possible transcriptional regulatory function); NAD(P)/FAD-dependent oxidoreductase; glycosyltransferase family 2 protein; hypothetical protein AURANDRAFT 71710 [Aureococcus anophagefferens]; hypothetical protein BFL35 01000 [Clavibacter michiganensis]; hypothetical protein BU14 0724s0003 [Porphyra umbilicalis]
UL27.6 uORF	5	No clear hits
UL28 uORF	38	No clear hits
UL28.4 sORF	53	STAT homolgue, SCOP b.2.5.5: Family: STAT DNA-binding domain
UL28.6 ORF (orphan)	348	
UL3 uORF RNA *1	5	No clear hits
UL30 uORF	10	No clear hits
UL30.5 sORF 1	37	hypothetical protein [Human alphaherpesvirus 1]
UL30.5 sORF 2	34	No clear hits
UL30.5 sORF RNA #1 (only comprises the C-terminal 4aa of UL30.5 sORF 2; initiated from AUG)	4	
UL30.6 sORF	42	Metallochaperone MeaB; SCOP c.37.1.10: Family: Nitrogenase iron protein-like
UL31.5 ORF	103	No clear hits
UL31.5 uORF 1 RNA *1	17	Type III pantothenate kinase, CoaX
UL31.5 uORF 2 RNA *1	23	No clear hits

UL31.6 ORF (orphan, C-terminal part is identical to UL32.5 ORF)	298		
UL32.5 ORF (orphan; TaSS unclear)	147		
UL32.5 uORF (orphan)	12		
UL32.6 ORF	79	secreted	zinc finger and BTB domain containing 7A, isoform CRA c; one cut domain family member 2; protein enabled; hypothetical protein [Human alphaherpesvirus 1]; hypothetical protein VE03 05107 [Pseudogymnoascus sp. 23342-1-I1]
UL34.5 sORF 1 (orphan)	28		
UL34.5 sORF 2 (orphan)	68	secreted	
UL36.4 ORF (possible truncated versions initiating from AUGs at 73767 and 73929)	105		
UL36.4 uORF	80		hypothetical protein [Human alphaherpesvirus 1]
UL36.5 ORF (possible truncated version initiating from AUG at 75172)	249	endoplasmic reticulum membrane	
UL36.6 ORF (possible truncated version initiating from AUG at 80316)	100		
UL36.6 uORF	13		No clear hits
UL37 uoORF	70		No clear hits
UL37 uoORF *1 (orphan, N-terminal 23aa extension of UL37 uoORF initiating from AUG)	93		
UL37.3 ORF (orphan; possible 7aa N-terminal extension initiating from AUG at 81763)	104		
UL37.4 ORF (orphan, TaSS unclear: most likely at 83320 from AUG)	265		
UL37.4 uORF (orphan, probable uORF of the UL37.4 ORF)	80		
UL37.5 ORF (could also initiate from second AUG directly following first AUG)	102	secreted	
UL37.5 uORF 1	27		proline, histidine and glycine-rich protein 1; unchracterized protein Dere GG19648, isoform B
UL37.5 uORF 2	19		hypothetical protein [Defluviimonas aquaemixtae]
UL37.5 uORF 3 (=uORF of UL37.5 RNA #1; ends directly in front of TaSS of UL37.5 ORF)	53	secreted	
UL37.6 ORF	94		No clear hits
UL37.6 ORF #1 (translated from UL37.6 RNA #1 initiating with AUG; lacks the first 24aa of UL37.6 ORF)	70		
UL37.6 uoORF	14		No clear hits
UL39.4 sORF (C-terminal 23aa of UL39.5 ORF, initiating from AUG)	23		
UL39.5 ORF	138		hypothetical protein [Human alphaherpesvirus 1]
UL39.6 dORF 1	6		No clear hits
UL39.6 dORF 2	88	secreted	No clear hits

UL39.6 ORF	102	secreted	Thymidylate kinase
UL4.5 sORF 1	83	chloroplast	hypothetical protein [Human alphaherpesvirus 1]
UL4.5 sORF 2	39		No clear hits
UL4.5 sORF 3 (TaSS unclear)	44	chloroplast	
UL40.5 iORF (alternative TaSS 2aa downstream initiating from AUG at 91425)	25		
UL40.5 ORF (TaSS unclear, probably located further upstream)	130		
UL40.5 uORF	4		No clear hits
UL40.6 sORF 1	18		No clear hits
UL40.6 sORF 2	22	secreted	No clear hits
UL40.7 dORF	35		hypothetical protein [Human alphaherpesvirus 1]
UL40.7A ORF	104	secreted	DUF1236 domain-containing protein; tetratricopeptide repeat domain; uncharacterized protein LOC111598453; hypothetical protein FRACYDRAFT 244884 [Fragilariopsis cylindrus CCMP1102]
UL40.7B ORF (in different frame than UL40.7A, overlapping C-terminal half)	71		
UL41 iORF RNA #1 (translated from UL41 RNA #1 & #2)	52		
UL41 uoORF	9		No clear hits
UL41 uORF 1 (overlaps UL41 uoORF)	26		
UL41 uORF 2 (iORF of UL41 uORF 1)	2		
UL43.5 iORF (orphan)	25		
UL43.5 ORF (orphan)	124		
UL43.5 uORF (orphan)	16		
UL43.6 ORF (orphan)	311	secreted	
UL43.6 uORF (orphan; may have alternative TaSS further upstream)	6		
UL44 iORF	105	secreted	leucine-rich protein repeat extensin-like protein 3; hypothetical protein BS418 22440 [Cronobacter sakazakii]
UL44.5 ORF (orphan)	162		
UL46.5 sORF 1	4		No clear hits
UL46.5 sORF 2	74		hypothetical protein [Human alphaherpesvirus 1]
UL49 uORF	17		No clear hits
UL49 uORF *1 (includes N-terminal extension of at least 15 aa, TaSS unclear)	33		
UL49.5 uORF	25		No clear hits
UL5 uORF (=C-terminal 47aa of US6.4 sORF 2)	47		
UL5.5 dORF (TaSS only 3nt downstream of TiSS UL6 RNA*1)	30		
UL5.5 iORF 1	11		No clear hits
UL5.5 iORF 2	50		No clear hits
UL5.5 iORF 3	24		No clear hits
UL5.5 ORF (TaSS unclear; ORF validated by two independent peptides)	407		

UL5.5 uORF	3		No clear hits
UL5.6 dORF	63	secreted	No clear hits
UL5.6 ORF (located within UL5.5 ORF, different frame)	112	secreted	
UL5.7 ORF	124		UL6 virion protein; hypothetical protein [Human alphaherpesvirus 1]
UL50.5 sORF 1	51		No clear hits
UL50.5 sORF 2 (TaSS unclear; shares C-terminal part with UL49.5 uORF 1)	68		
UL50.6 sORF	47		hypothetical protein [Human alphaherpesvirus 1]
UL51 uoORF (TaSS unclear)	50	secreted	
UL51 uORF (TaSS unclear)	15		
UL51.5 sORF 1 (orphan)	11		
UL51.5 sORF 2 (orphan)	51	nucleus	
UL52.4 sORF (may include N-terminal 4aa extension initiating from CUG)	88		
UL52.5 sORF 1	12		No clear hits
UL52.5 sORF 2	54		hypothetical protein [Human alphaherpesvirus 1]
UL53 iORF RNA #2	82	secreted	HNH endonuclease; uncharacterized protein DUF222; hypothetical protein BGO96 01180 [Micrococcales bacterium 73-15]
UL53 uORF	16		No clear hits
UL53.5 sORF 1	55	secreted	No clear hits
UL53.5 sORF 2	38	secreted	urate oxidase
UL56 uORF	37		hypothetical protein [Human alphaherpesvirus 1]
UL6 iORF RNA #1	69	secreted	No clear hits
UL6 uORF	12		No clear hits
UL6 uORF RNA *1	31	secreted	hypothetical protein [Human alphaherpesvirus 1]
UL6.4 sORF 1	13		No clear hits
UL6.4 sORF 2 (may include N-terminal 28aa extension initiating from GUG)	89		
UL6.5 sORF 1	9		No clear hits
UL6.5 sORF 2	8		No clear hits
UL6.5 sORF 3	54	secreted	automated matches, SCOP b.40.4.0: Superfamily: Nucleic acid-binding proteins; Chaperone protein SycN
UL6.5 sORF 4	73	secreted	hypothetical protein [Human alphaherpesvirus 1]
UL6.6 ORF	133		hypothetical protein CAOG 02982 [capsaspora owczarzaki ATCC 30864]
UL6.7 sORF 1 (orphan)	6		
UL6.7 sORF 2 (orphan)	43	secreted	
UL7.5 ORF	190		Processive endocellulase
UL7.5 uORF	72		Creatine kinase

UL8 iORF	115		translation inititation factor IF-2; sensor histidine kinase; DUF4192 family protein; KS-AT-KR-ACP domain-containing polyene macrolide polyketide synthase/pimaricinolide synthase PimS2/candicidin polyketide synthase FscD [Streptomyces sp. SolWspMP-5a-2]
UL8 uORF	7		No clear hits
UL8.3 ORF (orphan)	121		
UL8.4 ORF (may include a small truncated version of 23aa initiating at 21143 from ACG)	158		
UL9.4 dORF (orphan)	22		
UL9.4 ORF (orphan)	118		
UL9.5 iORF 1	17		No clear hits
UL9.5 iORF 2	17		No clear hits
UL9.5 ORF	472	secreted	hypothetical protein [Actinomycetales bacterium JB111]
UL9.5 uoORF	29		No clear hits
UL9.5 uORF	27		No clear hits
US11.5 ORF (orphan)	126		
US2.5 ORF	133		serine/threonine protein kinase US3
US2.6 ORF (orphan; probable alternative TaSS from first downstream AUG)	315		
US3.6 sORF (C-terminal 81aa of US4.5 ORF initiating from AUG)	80		
US3.6 uORF (4aa, initiating from AUG)	4		
US4.5 ORF (orphan, TaSS unclear)	222		
US5 dORF	37	secreted	envelope glycoprotein J; hypothetical protein [Human alphaherpesvirus 1]
US5.5 sORF	20		No clear hits
US6 uORF	4		No clear hits
US9 uoORF	18		pUS8A(potential membran protein) ->https://www.openagrar.de/receive/fimport mods 00000193

Table B.5: List of ORFs identified by mass spectrometry including the number of peptides per ORF (HFF=human foreskin fibroblast data, HLF=human lung fibroblast data; All=any peptide within the protein/polypeptide are counted, Novel=only peptides outside of previously known proteins are counted).

ORF	Novel	Peptides (HFF): All	Peptide (HFF): Novel part	Peptides (HLF): All	Peptide (HLF): Novel part
IRL1.5 ORF	x	0	0	0	0
IRL1A sORF (uoORF of RL1; 93aa initiating from AUG)	x	0	0	0	0
IRL1 CDS (ICP34.5)		1	0	0	0
IRL2.5 ORF (TaSS unclear due to repeat regions upstream)	x	0	0	0	0
IRL2A ORF	x	0	0	0	0
IRL2 CDS (ICP0; IE110)		25	0	2	0
IRL2 CDS iso1 (RL2 CDS with different C-terminus due to intron retention of intron 1)	x	0	0	0	0
IRL2 CDS iso2 (NAGNAG)	x	0	0	0	0
IRL2 dORF RNA iso1 (orphan, located within intron 1)	x	0	0	0	0
IRS1 CDS (ICP4; IE175;)		36	0	22	0
ORF-O CDS (TaSS probably 76nt upstream of ORF-P initiating from ACG start codon; no evidence of frame-shift; orphan)	x	0	0	0	0
ORF-O CDS (TaSS probably 76nt upstream of ORF-P initiating from ACG start codon; no evidence of frame-shift)	x	0	0	0	0
ORF-P CDS (orphan)	x	0	0	0	0
ORF-P CDS	x	0	0	0	0
Ori-S CDS (Hubenthal-Voss,Starr,Roizman JVI 1987)	x	0	0	0	0
TRL1.5 ORF	x	0	0	0	0
TRL1A sORF (uoORF of RL1; 93aa initiating from AUG)	x	0	0	0	0
TRL1 CDS (ICP34.5)		1	0	0	0
TRL2.5 ORF (TaSS unclear due to repeat regions upstream)	x	0	0	0	0
TRL2A ORF	x	0	0	0	0
TRL2 CDS (ICP0; IE110)		25	0	2	0
TRL2 CDS iso1 (RL2 CDS with different C-terminus due to intron retention of intron 1)	x	0	0	0	0
TRL2 CDS iso2 (NAGNAG)	x	0	0	0	0
TRL2 dORF RNA iso1 (located within intron 1)	x	0	0	0	0
TRS1 CDS (ICP4; IE175; may include 21aa N-terminal extension initiating with GUG)		36	0	22	0
UL10 CDS (Glycoprotein M; may include N-terminal extension but TaSS unclear)		6	0	1	0
UL10 uORF (TaSS unclear; could also be 23012 or 23039)	x	0	0	0	0
UL11 CDS (Tegument protein)		0	0	0	0
UL12.5 CDS (Exonuclease activity; truncated isoform of UL12)	x	14	0	13	0
UL12 CDS (Exonuclease activity)		17	0	13	0
UL12 uoORF (TaSS unclear, could also be at 26971)	x	0	0	0	0
UL13.5 ORF #1 (truncated isoform of 85aa of UL13.5 ORF initiating at 27814 from AUG)	x	0	0	0	0
UL13.5 ORF	x	0	0	0	0
UL13 CDS (Serine-threonince protein kinase)		6	0	0	0
UL14 CDS (Tegument protein)		3	0	1	0
UL14 uORF 1	x	0	0	0	0
UL14 uORF 2	x	0	0	0	0
UL15.4 sORF 1 (orphan; possible alternative TaSS at 30021, AUG)	x	0	0	0	0
UL15.4 sORF 2 (orphan)	x	0	0	0	0
UL15.5 CDS (orphan; Capside associated protein)	x	0	0	0	0
UL15 CDS (Terminase; Processing and packaging DNA)		4	0	0	0
UL15 iORF 1 (located in intron; orphan)	x	0	0	0	0
UL15 iORF 2 (located in intron; orphan)	x	0	0	0	0
UL15 iORF 3 (located in exon 2; orphan)	x	0	0	0	0
UL15 ORF RNA iso1 (UL15 CDS with different C-terminus due to intron retention)	x	4	0	0	0
UL15 uoORF	x	0	0	0	0
UL15 uORF	x	0	0	0	0
UL16 CDS (Tegument protein)		5	0	5	0
UL17.4 sORF 1 (orphan)	x	0	0	0	0
UL17.4 sORF 2 (orphan, contains 36aa N-terminal extension of UL17 uORF 1 RNA *1 initiating from AUG)	x	0	0	0	0
UL17.5 ORF (orphan)	x	0	0	0	0
UL17.5 uoORF (orphan)	x	0	0	0	0
UL17 CDS (Processing and packaging)		12	0	6	0
UL17 uoORF (TaSS unclear)	x	0	0	0	0
UL17 uORF RNA *1	x	0	0	0	0
UL17 uORF	x	0	0	0	0
UL18 CDS (VP23; Capsid protein)		11	0	13	0
UL18 uoORF	x	0	0	0	0
UL18 uORF	x	0	0	0	0
UL19.4 sORF (orphan)	x	0	0	0	0
UL19.5 sORF (may be part of new ORF initiating at 39420)	x	0	0	0	0
UL19 CDS *1 (includes 11 aa N-terminal extension initiating from AUC)	x	44	0	46	0
UL19 CDS (VP5; Major capsid protein)		45	0	46	0
UL1 CDS (Glycoprotein L)		6	0	2	0
UL1 uORF	x	0	0	0	0
UL20.4 dORF 1	x	0	0	0	0
UL20.4 dORF 2 (alternative TaSS at next AUG 6nt downstream at 41487)	x	0	0	1	1
UL20.4 ORF	x	0	0	0	0
UL20.5 CDS (orphan)	x	0	0	0	0
UL20.6 ORF	x	0	0	0	0
UL20.6 uORF	x	0	0	0	0
UL20 CDS (Membrane protein)		0	0	0	0
UL20 iORF (ORF initiates 13nt downstream of UL20 from AUG)	x	0	0	0	0
UL20 uORF (corresponds to C-terminal part of UL20.5 CDS)	x	0	0	0	0
UL21 CDS *1 (includes 65 aa N-terminal extension initiating from CUG)	x	7	0	12	0
UL21 CDS (Tegument protein)		8	0	13	0
UL22.5 sORF	x	0	0	0	0
UL22 CDS (Glycoprotein H)		15	0	9	0
UL22 uORF 1 (=uoORF for UL22 uORF 2)	x	0	0	0	0
UL22 uORF 2 (possible alternative TaSS downstream at 46468, GUG)	x	0	0	0	0
UL23.5 ORF	x	0	0	1	1
UL23.5 uORF	x	0	0	0	0
UL23.6 ORF (orphan)	x	0	0	0	0
UL23 CDS *1 (includes 15 aa N-terminal extension initiating from GUG)	x	7	0	7	0
UL23 CDS (Thymidine kinase; Peripheral to DNA replication)		8	0	7	0
UL23 uORF 1	x	0	0	0	0
UL23 uORF 2 (uoORF for N-terminal extension of UL23; C-terminal 7 aa of UL23.6 ORF)	x	0	0	0	0
UL24.5 CDS (truncated isoform of UL24)	x	3	0	1	0

UL24 CDS (Exonuclease activity)		6	0	1	0
UL24 uORF 1 RNA *1	x	0	0	0	0
UL24 uORF 2 RNA *1	x	0	0	0	0
UL24 uORF 3 RNA *1	x	0	0	0	0
UL25.4 ORF (orphan)	x	0	0	1	1
UL25.5 ORF	x	0	0	0	0
UL25 CDS (Processing and packaging DNA; Capside Protein)		15	0	11	0
UL26.5 CDS (truncated isoform of UL26)		3	0	2	0
UL26 CDS (P40; VP24; VP22A; Capsid protein)		12	0	11	0
UL27.5 sORF	x	0	0	0	0
UL27.6 ORF	x	0	0	0	0
UL27.6 uORF	x	0	0	0	0
UL27 CDS *1 (includes 43 aa N-terminal extension initiating from ACG)	x	36	0	6	0
UL27 CDS (Glycoprotein B)		36	0	6	0
UL28.4 sORF	x	0	0	0	0
UL28.6 ORF (orphan)	x	0	0	0	0
UL28 CDS (ICP18.5; Processing and packaging)		3	0	1	0
UL28 uORF	x	0	0	0	0
UL29.5 CDS (truncated isoform of UL29)	x	16	0	23	0
UL29 CDS (ICP8; Major DNA-binding protein)		36	0	38	0
UL2 CDS (Uracil-DNA glycosylase)		4	0	2	0
UL30.5 sORF 1	x	0	0	0	0
UL30.5 sORF 2	x	0	0	0	0
UL30.5 sORF RNA #1 (only comprises the C-terminal 4aa of UL30.5 sORF 2; initiated from AUG)	x	0	0	0	0
UL30.6 sORF	x	1	1	0	0
UL30 CDS (DNA polymerase)		19	0	6	0
UL30 uORF	x	0	0	0	0
UL31.5 ORF	x	0	0	0	0
UL31.5 uORF 1 RNA *1	x	0	0	0	0
UL31.5 uORF 2 RNA *1	x	0	0	0	0
UL31.6 ORF (C-terminal part is identical to UL32.5 ORF)	x	0	0	0	0
UL31 CDS (Nuclear matrix protein)		8	0	2	0
UL32.5 ORF (TaSS unclear)	x	0	0	0	0
UL32.5 uORF	x	0	0	0	0
UL32.6 ORF	x	0	0	0	0
UL32 CDS (Processing and packaging)		1	0	3	0
UL33 CDS *1 (includes 21 aa N-terminal extension initiating from ACG)	x	2	0	0	0
UL33 CDS (Processing and packaging DNA)		2	0	0	0
UL34.5 sORF 1 (orphan)	x	0	0	0	0
UL34.5 sORF 2 (orphan)	x	0	0	0	0
UL34 CDS (Inner nuclear membrane protein)		10	0	2	0
UL35 CDS (VP26; Capsid protein)		2	0	1	0
UL36.4 ORF (possible truncated isoform initiating from AUGs at 73767 and 73929)	x	0	0	0	0
UL36.4 uORF	x	0	0	0	0
UL36.5 ORF	x	0	0	0	0
UL36.6 ORF (possible truncated isoform initiating from AUG at 80316)	x	0	0	0	0
UL36.6 uORF	x	0	0	0	0
UL36 CDS *1 (includes 10 aa N-terminal extension initiating from AUG)	x	54	0	26	0
UL36 CDS (Omega peptidase activity; Tegument protein)		54	0	26	0
UL37.3 ORF (orphan; possible 7aa N-terminal extension initiating from AUG at 81763)	x	0	0	0	0
UL37.4 ORF (orphan, TaSS unclear: most likely at 83320 from AUG)	x	0	0	0	0
UL37.4 uORF (orphan, probable uORF of the UL37.4 ORF)	x	0	0	0	0
UL37.5 ORF (could also initiate from second AUG directly following first AUG)	x	0	0	0	0
UL37.5 uORF 1	x	0	0	0	0
UL37.5 uORF 2	x	0	0	0	0
UL37.5 uORF 3 (=uORF of UL37.5 RNA #1; ends directly in front of TaSS of UL37.5 ORF)	x	0	0	0	0
UL37.6 ORF #1 (translated from UL37.6 RNA #1 initiating with AUG; lacks the first 24aa of UL37.6 ORF)	x	0	0	1	1
UL37.6 ORF	x	0	0	2	2
UL37.6 uoORF	x	0	0	0	0
UL37 CDS (Tegument protein; Capsid assembly)		24	0	24	0
UL37 uoORF *1 (orphan, N-terminal 23aa extension of UL37 uoORF initiating from AUG)	x	0	0	0	0
UL37 uoORF	x	0	0	0	0
UL38 CDS (VP19C; Capsid protein; DNA maturation)		12	0	8	0
UL39.4 sORF (C-terminal 23aa of UL39.5 ORF, initiating from AUG)	x	0	0	1	1
UL39.5 ORF	x	0	0	1	1
UL39.6 dORF 1	x	0	0	0	0
UL39.6 dORF 2	x	0	0	0	0
UL39.6 ORF	x	0	0	0	0
UL39 CDS *1 (includes 38 aa N-terminal extension initiating from UUG)	x	38	0	23	0
UL39 CDS (ICP6; RR-1; Ribonucleotide reductase; large subunit)		39	0	23	0
UL3 CDS		2	0	0	0
UL3 uORF RNA *1	x	0	0	0	0
UL40.5 iORF (alternative TaSS 2aa downstream initiating from AUG at 91425)	x	0	0	0	0
UL40.5 ORF (TaSS unclear, probably located further upstream)	x	0	0	0	0
UL40.5 uORF	x	0	0	0	0
UL40.6 sORF 1	x	0	0	0	0
UL40.6 sORF 2	x	0	0	0	0
UL40.7A ORF	x	0	0	0	0
UL40.7B ORF (in different frame than UL40.7A, overlapping C-terminal half)	x	0	0	0	0
UL40.7 dORF (translated from UL40.6 spliced RNA)	x	0	0	0	0
UL40 CDS (RR-2; Ribonucleotride reductase; small subunit)		8	0	7	0
UL41 CDS (VHS; Tegument protein)		6	0	2	0
UL41 iORF RNA #1 (translated from UL41 RNA #1 & #2)	x	0	0	0	0
UL41 uoORF	x	0	0	0	0
UL41 uORF 1 (overlaps UL41 uoORF)	x	0	0	0	0
UL41 uORF 2 (iORF of UL41 uORF 1)	x	0	0	0	0
UL42 CDS (DNA polymerase processivity factor)		18	0	20	0
UL43.5 iORF (orphan)	x	0	0	0	0
UL43.5 ORF (orphan; TaSS unclear)	x	0	0	0	0
UL43.5 uORF (orphan)	x	0	0	0	0
UL43.6 ORF (orphan)	x	0	0	0	0
UL43.6 uORF (orphan; may have alternative TaSS further upstream)	x	0	0	0	0
UL43 CDS (Membrane protein)		2	0	0	0
UL44.5 ORF (orphan)	x	0	0	0	0

UL44 CDS (Glycoprotein C)		7	0	7	0
UL44 iORF	x	1	1	0	0
UL44 ORF RNA iso1 (altered C-terminus)	x	0	0	0	0
UL45 CDS (Membrane protein; C-type lectin)		3	0	1	0
UL4.5 iORF (located in intron)	x	0	0	0	0
UL4.5 ORF	x	0	0	0	0
UL46.5 sORF 1	x	0	0	0	0
UL46.5 sORF 2	x	0	0	0	0
UL46 CDS (VP11; VP12; Tegument protein)		18	0	8	0
UL47 CDS (Vp13; VP14; Tegument protein)		25	0	17	0
UL48 CDS (VP16; Alpha-TIF; Tegument protein; Virion maturation)		10	0	10	0
UL49.5 CDS (UL49A; Glycoprotein N)		0	0	0	0
UL49.5 uORF	x	0	0	0	0
UL49 CDS (VP22; Tegument protein)		18	0	7	0
UL49 uORF	x	0	0	0	0
UL4 CDS		1	0	0	0
UL50.5 sORF 1	x	0	0	0	0
UL50.5 sORF 2 (TaSS unclear; shares C-terminal part with UL49.5 uORF 1)	x	0	0	0	0
UL50.6 sORF	x	0	0	0	0
UL50 CDS *1 (includes 10 aa N-terminal extension intitating from CUG)	x	14	0	11	0
UL50 CDS *2 (includes 23 aa N-terminal extension intitating from ACG)	x	14	0	11	0
UL50 CDS (dUTPase)		15	0	12	0
UL51.5 sORF 1 (orphan)	x	0	0	0	0
UL51.5 sORF 2 (orphan)	x	0	0	0	0
UL51 CDS (Tegument protein)		4	0	0	0
UL51 uoORF (TaSS unclear)	x	0	0	0	0
UL51 uORF (TaSS unclear)	x	0	0	0	0
UL52.4 sORF (may include N-terminal 4aa extension initiating from CUG)	x	0	0	0	0
UL52.5 sORF 1	x	0	0	0	0
UL52.5 sORF 2	x	0	0	0	0
UL52 CDS (DNA helicase/primase complex protein)		8	0	0	0
UL53.5 sORF 1	x	0	0	0	0
UL53.5 sORF 2	x	0	0	0	0
UL53 CDS (Glycoprotein K)		0	0	0	0
UL53 iORF RNA #2	x	0	0	0	0
UL53 uORF	x	0	0	0	0
UL54 CDS *1 (includes 38 aa N-terminal extension initiating from ACG)	x	12	0	12	0
UL54 CDS (ICP27; IE63; Transcriptional regulation)		12	0	12	0
UL55 CDS (Virion assembly)		1	0	0	0
UL5.5 dORF (TaSS only 3nt downstream of TiSS UL6 RNA*1)	x	0	0	0	0
UL5.5 iORF 1	x	0	0	0	0
UL5.5 iORF 2	x	0	0	0	0
UL5.5 iORF 3	x	0	0	0	0
UL5.5 ORF (TaSS unclear; ORF validated by two independent peptides)	x	1	1	1	1
UL5.5 uORF	x	0	0	0	0
UL56 CDS (Membrane protein)		3	0	0	0
UL5.6 dORF (orphan)	x	0	0	0	0
UL5.6 ORF (located within UL5.5 ORF, different frame)	x	0	0	0	0
UL56 uORF	x	0	0	0	0
UL5.7 ORF	x	0	0	0	0
UL5 CDS *1 (includes 9 aa N-terminal extension initiating from GUG)	x	9	0	0	0
UL5 CDS (ATP binding; helicate-primase complex-associated protein)		10	0	0	0
UL5 uORF (=C-terminal 47aa of US6.4 sORF 2)	x	0	0	0	0
UL6.4 sORF 1	x	0	0	0	0
UL6.4 sORF 2 (may include N-terminal 28aa extension initiating from GUG)	x	0	0	0	0
UL6.5 sORF 1	x	0	0	0	0
UL6.5 sORF 2	x	0	0	0	0
UL6.5 sORF 3	x	0	0	0	0
UL6.5 sORF 4	x	1	1	0	0
UL6.6 ORF	x	0	0	0	0
UL6.7 sORF 1 (orphan)	x	0	0	0	0
UL6.7 sORF 2 (orphan)	x	0	0	0	0
UL6 CDS (Virion portal protein)		8	0	1	0
UL6 iORF RNA #1	x	0	0	0	0
UL6 uORF RNA *2	x	0	0	0	0
UL6 uORF	x	0	0	0	0
UL7.5 ORF	x	0	0	0	0
UL7.5 uORF	x	0	0	0	0
UL7 CDS *1 (includes 15 aa N-terminal extension initiating from CUG)	x	4	0	0	0
UL7 CDS (Tegument protein)		5	0	0	0
UL8.3 ORF (orphan)	x	0	0	0	0
UL8.4 ORF (may include a small truncated isoform of 23aa initiating at 21143 from ACG)	x	0	0	0	0
UL8.5 CDS (ATP binding; truncated isoform of UL9; Bahradaran et al, 1994)	x	2	0	0	0
UL8 CDS (helicate-primase complex-associated protein)		5	0	0	0
UL8 iORF	x	0	0	0	0
UL8 uORF	x	0	0	0	0
UL9.4 dORF (orphan)	x	0	0	0	0
UL9.4 ORF (orphan)	x	0	0	0	0
UL9.5 iORF 1	x	0	0	0	0

UL9.5 iORF 2	x	0	0	0	0
UL9.5 ORF	x	0	0	0	0
UL9.5 uoORF	x	0	0	0	0
UL9.5 uORF	x	0	0	0	0
UL9 CDS (ATP binding; Replication origin-binding protein)		4	0	0	0
US10 CDS (Tegument protein)		7	0	0	0
US11.5 ORF (orphan)	x	0	0	0	0
US11 CDS *1 (includes 4aa N-terminal extension initiating with GUG)	x	6	0	0	0
US11 CDS (Vmw21; RNA binding protein)		7	0	0	0
US12 CDS (ICP47; IE12)		2	0	0	0
US1.5 CDS (truncated isoform of US1)	x	6	0	4	0
US1 CDS *1 (includes N-terminal extension initiating from GUG, 8 aa, or even from GUG, 21 aa)	x	7	0	4	0
US1 CDS (ICP22; IE63; Viral replication)		7	0	4	0
US2.5 ORF (TaSS unclear)	x	0	0	0	0
US2.6 ORF (orphan; probable alternative TaSS from first downstream AUG)	x	0	0	0	0
US2 CDS (Tegument protein)		5	0	2	0
US3.5 CDS (truncated isoform of US3)	x	6	0	1	0
US3.6 sORF (C-terminal 81aa of US4.5 ORF initiating from AUG)	x	0	0	0	0
US3.6 uORF (4aa, initiating from AUG)	x	0	0	0	0
US3 CDS *1 (includes 23 aa N-terminal extension initiating from CUG)	x	6	0	1	0
US3 CDS (Serine/threonine-protein kinase)		6	0	1	0
US4.5 ORF (orphan, TaSS unclear)	x	0	0	0	0
US4 CDS (Glycoprotein G)		1	0	0	0
US5.5 sORF	x	0	0	0	0
US5 CDS *1 (includes 16 aa N-terminal extension initiating from CUG)	x	0	0	0	0
US5 CDS (Glycoprotein J)		0	0	0	0
US5 dORF	x	0	0	0	0
US6 CDS (Glycoprotein D)		8	0	4	0
US6 uORF	x	0	0	0	0
US7 CDS (Glycoprotein I)		5	0	0	0
US8.5 CDS (US8A)		3	0	0	0
US8 CDS *1 (includes 7aa N-terminal extension initiating from AUC)	x	14	0	0	0
US8 CDS (Glycoprotein E)		14	0	0	0
US9 CDS (Membrane protein)		1	0	0	0
US9 uoORF	x	0	0	0	0

Table B.6: List of ORFs for which no obvious corresponding transcript initiating within 500 nt upstream was identified.

Name	ID	Type	Gene	Transcript	Strand	Length (aa)	Start-codon	Stop-codon	TiSS	Stop
IRL2 dORF RNA iso1 (orphan, located within intron 1)	IRL2_dORF_RNA_iso1	dORF	IRL2	IRL2_RNA_iso1	-	42	AUG	UGA	123731	123603
ORF-O (TaSS probably 76nt upstream of ORF-P initiating from ACG start codon; no evidence of frame-shift; orphan)	ORF-O	ORF	ORF-O	-	-	271	ACG	UAG	1261	446
ORF-P (orphan)	ORF-P	ORF	ORF-P	-	-	233	AUG	UAG	1185	484
UL15 iORF 1 (located in intron; orphan)	UL15_iORF_1	iORF	UL15	-	+	26	CUG	UAG	33181	33261
UL15 iORF 2 (located in intron; orphan)	UL15_iORF_2	iORF	UL15	-	+	73	AUG	UAG	33303	33524
UL15 iORF 3 (located in exon 2; orphan)	UL15_iORF_3	iORF	UL15	UL15_RNA	+	46	AUG	UGA	33696	33836
UL15.4 sORF 1 (orphan; possible alternative TaSS at 30021, AUG)	UL15.4_sORF_1	sORF	UL15.4	-	-	65	AUG	UAG	30071	29874
UL15.4 sORF 2 (orphan)	UL15.4_sORF_2	sORF	UL15.4	-	-	58	AUG	UAG	29834	29658
UL15.5 CDS (orphan; Capside associated protein)	UL15.5_CDS	CDS	UL15.5	-	+	293	AUG	UGA	33932	34813
UL17.4 sORF 1 (orphan)	UL17.4_sORF_1	sORF	UL17.4	-	-	59	AUG	UAA	34380	34201
UL17.4 sORF 2 (orphan, contains 36aa N-terminal extension of UL17 uORF 1 RNA *1 initiating from AUG)	UL17.4_sORF_2	sORF	UL17.4	-	-	67	AUG	UAA	34104	33901
UL17.5 ORF (orphan)	UL17.5_ORF	ORF	UL17.5	-	-	96	AUG	UAG	34752	34462
UL17.5 uoORF (orphan)	UL17.5_uoORF	uoORF	UL17.5	-	-	14	AUG	UAA	34787	34743
UL19.4 sORF (orphan)	UL19.4_sORF	sORF	UL19.4	-	+	6	AUG	UAG	39336	39356
UL20.5 CDS (orphan)	UL20.5_CDS	CDS	UL20.5	-	-	160	AUG	UAA	42016	41534
UL23.6 ORF (orphan)	UL23.6_ORF	ORF	UL23.6	-	-	154	AUG	UGA	48308	47844
UL25.4 ORF (orphan)	UL25.4_ORF	ORF	UL25.4	-	-	241	AUG	UGA	49438	48713
UL28.6 ORF (orphan)	UL28.6_ORF	ORF	UL28.6	-	+	348	AUG	UAA	60902	61948
UL34.5 sORF 1 (orphan)	UL34.5_sORF_1	sORF	UL34.5	-	-	28	AUG	UAG	70540	70454
UL34.5 sORF 2 (orphan)	UL34.5_sORF_2	sORF	UL34.5	-	-	68	AUG	UGA	70137	69931
UL37 uoORF *1 (orphan, N-terminal 23aa extension of UL37 uoORF initiating from AUG)	UL37_uoORF_*1	uoORF	UL37	-	-	93	AUG	UAG	84265	83984
UL37.3 ORF (orphan; possible 7aa N-terminal extension initiating from AUG at 81763)	UL37.3_ORF	ORF	UL37.3	-	+	104	AUG	UAG	81784	82098
UL37.4 ORF (orphan, TaSS unclear: most likely at 83320 from AUG)	UL37.4_ORF	ORF	UL37.4	-	+	265	AUG	UAG	83320	84117
UL37.4 uORF (orphan, probable uORF of the UL37.4 ORF)	UL37.4_uORF	uORF	UL37.4	-	+	80	CUG	UGA	82960	83202
UL43.5 iORF (orphan)	UL43.5_iORF	iORF	UL43.5	-	-	25	AUG	UAG	93317	93240
UL43.5 ORF (orphan; TaSS unclear)	UL43.5_ORF	ORF	UL43.5	-	-	124	UUG	UGA	93454	93080
UL43.5 uORF (orphan)	UL43.5_uORF	uORF	UL43.5	-	-	16	CUG	UGA	93571	93521
UL43.6 ORF (orphan)	UL43.6_ORF	ORF	UL43.6	-	-	311	AUG	UGA	95715	94780
UL43.6 uORF (orphan; may have alternative TaSS further upstream)	UL43.6_uORF	uORF	UL43.6	-	-	6	AUG	UGA	95807	95787
UL44.5 ORF (orphan)	UL44.5_ORF	ORF	UL44.5	-	-	162	AUG	UAG	96765	96277
UL5.6 dORF (orphan)	UL5.6_dORF	dORF	UL5.6	UL5.6_RNA	+	63	AUG	UAG	14389	14580
UL51.5 sORF 1 (orphan)	UL51.5_sORF_1	sORF	UL51.5	-	+	11	AUG	UAG	108309	108344
UL51.5 sORF 2 (orphan)	UL51.5_sORF_2	sORF	UL51.5	-	+	51	AUG	UGA	108545	108700
UL6.7 sORF 1 (orphan)	UL6.7_sORF_1	sORF	UL6.7	-	-	6	AUG	UAG	17919	17899
UL6.7 sORF 2 (orphan)	UL6.7_sORF_2	sORF	UL6.7	-	-	43	AUG	UGA	17858	17727
UL8.3 ORF (orphan)	UL8.3_ORF	ORF	UL8.3	-	+	121	AUG	UAG	19923	20288
UL9.4 dORF (orphan)	UL9.4_dORF	dORF	UL9.4	-	+	22	AUG	UGA	22786	22854
UL9.4 ORF (orphan)	UL9.4_ORF	ORF	UL9.4	-	+	118	AUG	UAG	22355	22711
US11.5 ORF (orphan)	US11.5_ORF	ORF	US11.5	-	+	126	AUG	UGA	145298	145678
US2.6 ORF (orphan; probable alternative TaSS from first downstream AUG)	US2.6_ORF	ORF	US2.6	-	-	315	AUG	UAG	135970	135023
US4.5 ORF (orphan, TaSS unclear)	US4.5_ORF	ORF	US4.5	-	-	222	AUG	UGA	136894	136226

Table B.7: Primers and gene synthesis constructs used.

Name	Use	Template	Sequence
UL19for	Forward primer for UL19 lentiviral vector cloning	BAC HSV1(17+)Lox-CheVP26	CTGGAGAATTGGCTAGCATGGCCGCTCCCAACCGCGA
Ul19rev	Reverse primer for UL19 lentiviral vector cloning	BAC HSV1(17+)Lox-CheVP26	CGGTTACGTAGTCGACTTACAGAGCCAGTCCCTTG
pEP-Kan-pr1	Forward primer for all PCR 1 reactions	pEP-Kan	GTATATCTGGCCCGTACATCGATCT
H5-UL19nte-pepkan	Reverse primer for UL19 NTE PCR 1	pEP-Kan	CACGGGTGGCCGGGCGTTGGGGTGGATCTTAGCCTCCCCGGGCCAATATCGACTACAAAGACCATGACGGTGATTATAAAGATCATGACATCGATTACAAGGATGACGATGACAAAGATATCAGGATGACGACGATAAGTAGGG
UL19nte-pepkan-for	Forward primer for UL19 NTE PCR 2	UL19 NTE PCR 1	CACGGGTGGCCGGGCGTTGG
H3-UL19nte-pepkan	Reverse primer for UL19 NTE PCR 2	UL19 NTE PCR 1	TCGCGGTTGGGAGCGGCCATGGGGTCGCGTGGAGATCGGCTGTCTCTAGCTTTGTCATCGTCATCCTTGTAATCGATGTCATGATCTTTATAATCACCGTCATGGTCTTTGTAGTCGTATATCTGGCCCGTACATCGATCT
H5-UL54nte-pepkan	Reverse primer for UL54 NTE PCR 1	pEP-Kan	GGACATCCACCACCCGGCCGGTGGTGGTGTGCAGCCGTGTTCCAACCACGGACTACAAAGACCATGACGGTGATTATAAAGATCATGACATCGATTACAAGGATGACGATGACAAAAGGATGACGACGATAAGTAGGG
UL54nte-pepkan-for	Forward primer for UL54 NTE PCR 2	UL54 NTE PCR 1	GGACATCCACCACCCGGCC
H3-UL54nte-pepkan	Reverse primer for UL54 NTE PCR 2	UL54 NTE PCR 1	CCGGTAGCGAGCGACCGGGCCCGAATCGGGGAGAGGCACCGAAGCGTGACTTTGTCATCGTCATCCTTGTAATCGATGTCATGATCTTTATAATCACCGTCATGGTCTTTGTAGTCGTATATCTGGCCCGTACATCGATCT
H5-US5nte-pepkan	Reverse primer for US5 NTE PCR 1	pEP-Kan	GGGCCTCAGAACCCACCCGAAACGGCCAACGGACGTCTGAGCCAGGCCTGGACTACAAAGACCATGACGGTGATTATAAAGATCATGACATCGATTACAAGGATGACGATGACAAAGAATTCAGGATGACGACGATAAGTAGGG
US5nte-pepkan-for	Forward primer for US5 NTE PCR 2	US5 NTE PCR 1	GGGCCTCAGAACCCACCCGA
H3-US5nte-pepkan	Reverse primer for US5 NTE PCR 2	US5 NTE PCR 1	GACATCGCGACACACAGAACGCCAAGTCGTGTGCTGTTTCTCCGGATAGCTTTGTCATCGTCATCCTTGTAATCGATGTCATGATCTTTATAATCACCGTCATGGTCTTTGTAGTCGTATATCTGGCCCGTACATCGATCT
H5-US3nte-pepkan	Reverse primer for US3 NTE PCR 1	pEP-Kan	GGGGAGGGTGATTTACGGGGGGGTAGGTCAGGGGGTGGGTCGTCAAACTGGACTACAAAGACCATGACGGTGATTATAAAGATCATGACATCGATTACAAGGATGACGATGACAAAGAATTCAGGATGACGACGATAAGTAGGG
US3nte-pepkan-for	Forward primer for US3 NTE PCR 2	US3 NTE PCR 1	GGGGAGGGTGATTTACGGGG
H3-US3nte-pepkan	Reverse primer for US3 NTE PCR 2	US3 NTE PCR 1	GCCAACCAACGAGCACCCCGAACGACGGGCCCCGGGGTTTTAAGGAGCGGTTTGTCATCGTCATCCTTGTAATCGATGTCATGATCTTTATAATCACCGTCATGGTCTTTGTAGTCGTATATCTGGCCCGTACATCGATCT
H5-ATG-UL19-pepkan	Reverse primer for UL19 ATG PCR 1	pEP-Kan	CTCCCCGGGCCAATATCGCTAGAGACAGCCGATCTCCACGCGACCCCATGGACTACAAAGACCATGACGGTGATTATAAAGATCATGACATCGATTACAAGGATGACGATGACAAAGATATCAGGATGACGACGATAAGTAGGG
UL19atg-pepkan-for	Forward primer for UL19 ATG PCR 2	UL19 ATG PCR 1	CTCCCCGGGCCAATATCGCT
H3-ATG-UL19-pepkan	Reverse primer for UL19 ATG PCR 2	UL19 ATG PCR 1	ACCATGGCCGCGGCATACCGGTATCCCGGAGGGTCGCGGTTGGGAGCGGCTTTGTCATCGTCATCCTTGTAATCGATGTCATGATCTTTATAATCACCGTCATGGTCTTTGTAGTCGTATATCTGGCCCGTACATCGATCT
H5-ATG-UL54-pepkan	Reverse primer for UL54 ATG PCR 1	pEP-Kan	AGGCCATATCCGACACCCCAGCCCCGACGGCAGCCGACAGCCCGGTCATGGACTACAAAGACCATGACGGTGATTATAAAGATCATGACATCGATTACAAGGATGACGATGACAAAAGGATGACGACGATAAGTAGGG
UL54atg-pepkan-for	Forward primer for UL54 ATG PCR 2	UL54 ATG PCR 1	AGGCCATATCCGACACCCCA
H3-ATG-UL54-pepkan	Reverse primer for UL54 ATG PCR 2	UL54 ATG PCR 1	CTGTCGGAGAGGTCCAGGCCGAGGTCAATTAGCATATCAATGTCAGTCGCTTTGTCATCGTCATCCTTGTAATCGATGTCATGATCTTTATAATCACCGTCATGGTCTTTGTAGTCGTATATCTGGCCCGTACATCGATCT
H5-ATG-US5-pepkan	Reverse primer for US5 ATG PCR 1	pEP-Kan	TGGCTATCCGGAGAAACAGCACACGACTTGGCGTTCTGTGTGTCGCGATGGACTACAAAGACCATGACGGTGATTATAAAGATCATGACATCGATTACAAGGATGACGATGACAAAGAATTCAGGATGACGACGATAAGTAGGG
US5atg-pepkan-for	Forward primer for US5 ATG PCR 2	US5 ATG PCR 1	TGGCTATCCGGAGAAACAGC
H3-ATG-US5-pepkan	Reverse primer for US5 ATG PCR 2	US5 ATG PCR 1	GCCCCCACGAGGCTTCCCAAAAGCCCCAGATGCCAGACTGCGCGCAGAGATTTGTCATCGTCATCCTTGTAATCGATGTCATGATCTTTATAATCACCGTCATGGTCTTTGTAGTCGTATATCTGGCCCGTACATCGATCT
H5-ATG-US3-pepkan	Reverse primer for US3 ATG PCR 1	pEP-Kan	GGCCCGTCGTTCGGGGTGCTCGTTGGTTGGCACTCACGGTGCGGCGAATGGACTACAAAGACCATGACGGTGATTATAAAGATCATGACATCGATTACAAGGATGACGATGACAAAGAATTCAGGATGACGACGATAAGTAGGG
US3atg-pepkan-for	Forward primer for US3 ATG PCR 2	US3 ATG PCR 1	GGCCCGTCGTTCGGGGTGCT
H3-ATG-US3-pepkan	Reverse primer for US3 ATG PCR 2	US3 ATG PCR 1	TCCTTCCTCCTGCCCTGTCCCCCGTAAACGCGACAAAACTTACGACAGGCTTTGTCATCGTCATCCTTGTAATCGATGTCATGATCTTTATAATCACCGTCATGGTCTTTGTAGTC GTATATCTGGCCCGTACATCGATCT
UL50-NTE-Zeo	Direct electroporation for UL50 NTE	Synthesis	ACGAGCAAGACAAACAACAGGCCCGCCCGACAGACCCTTCTGGGGGGGCCCATCGTCCCTAACGACTACAAAGACCATGACGGTGATTATAAAGATCATGACATCGATTACAAGGATGACGATGACAAAAGGAAGCATGATAAGTAGGGATAACAGGGTAATCGATTTATTGGATCCCGGACCGTGTTGACAATTAATCATCGGCATAGTATATCGGCATAGTATAATACGACAAGGTGAGGAACTAAACCATGGCAAAACTGACCAGCGCAGTTCCGGTTCTGACCGCACGTGATGTTGCCGGTGCCGTTGAATTTTGGACCGATCGTCTGGGTTTTAGCCGTGATTTTGTGGAAGATGATTTTGCCGGTGTTGTTCGTGATGATGTTACCCTGTTTATTAGCGCAGTTCAGGATCAGGTTGTTCCGGATAATACCCTGGCATGGGTTTGGGTTCGTGGTCTGGATGAACTGTATGCAGAATGGTCAGAAGTTGTGAGCACCAATTTTCGTGATGCAAGCGGTCCGGCAATGACCGAAATTGGTGAACAGCCGTGGGGTCGTGAATTTGCACTGCGTGATCCGGCAGGTAATTGTGTTCATTTTGTTGCAGAAGAACAGGATTAACACTCGACCTCGATTAATTAATTGTAACACTGGCATTACCCTGTTATCCCTAGATCTGCTGCCGACTACAAAGACCATGACGGTGATTATAAAGATCATGACATCGATTACAAGGATGACGATGACAAAAGGAAGATGAGTCAGTGGGGATCCGGGGCGATCCTTGTCCAGCCGGACAGCTTGGGTCGGGGGTACGATGGCGACTGGC
UL50-ATG-Zeo	Direct electroporation for UL50 ATG	Synthesis	CAGGCCCGCCCGACAGACCCTTCTGGGGGGGCCCATCGTCCCTAACAGGAAGATGGACTACAAAGACCATGACGGTGATTACATGATAAGTAGGGATAACAGGGTAATCGATTTATTGGATCCCGGACCGTGTTGACAATTAATCATCGGCATAGTATATCGGCATAGTATAATACGACAAGGTGAGGAACTAAACCATGGCAAAACTGACCAGCGCAGTTCCGGTTCTGACCGCACGTGATGTTGCCGGTGCCGTTGAATTTTGGACCGATCGTCTGGGTTTTAGCCGTGATTTTGTGGAAGATGATTTTGCCGGTGTTGTTCGTGATGATGTTACCCTGTTTATTAGCGCAGTTCAGGATCAGGTTGTTCCGGATAATACCCTGGCATGGGTTTGGGTTCGTGGTCTGGATGAACTGTATGCAGAATGGTCAGAAGTTGTGAGCACCAATTTTCGTGATGCAAGCGGTCCGGCAATGACCGAAATTGGTGAACAGCCGTGGGGTCGTGAATTTGCACTGCGTGATCCGGCAGGTAATTGTGTTCATTTTGTTGCAGAAGAACAGGATTAACACTCGACCTCGATTAATTAATTGTAACACTGGCATTACCCTGTTATCCCTAGATCTGCTGCGACTACAAAGACCATGACGGTGATTATAAAGATCATGACATCGATTACAAGGATGACGATGACAAAAGTCAGTGGGGATCCGGGGCGATCCTTGTCCAGCCGGACAGCTTGGGTC
H5-TRL2A-pepkan	Reverse primer for RL2A PCR 1	pEP-Kan	TCTGGTCGCATTTGCACCTCGGCACTCGGAGCGAGACGCAGCAGCCAGGCGACTACAAAGACCATGACGGTGATTATAAAGATCATGACATCGATTACAAGGATGACGATGACAAAAGGATGACGACGATAAGTAGGG
TRL2A-pepkan-for	Reverse primer for RL2A 3xFLAG PCR 2	RL2A FLAG PCR 1	TCTGGTCGCATTTGCACCTC
H3-TRL2A-pepkan	Forward primer for RL2A 3xFLAG PCR 2	RL2A FLAG PCR 1	AACGTAGGCGGGGCTTCTGTGGTGATGCGGAGAGGGGGCGGCCCGAGTCTTTTGTCATCGTCATCCTTGTAATCGATGTCATGATCTTTATAATCACCGTCATGGTCTTTGTAGTCGTATATCTGGCCCGTACATCGATCT
RL-AMP_H5	Forward primer for replacing part of repeat long with AmpR	pcDNA3	CTAGCGAGTTAGACAGGCAAGCACTACTCGCCTCTGCACGCACTGGCACTTTTCGGGG
RL-AMP_H3	Reverse primer for replacing part of repeat long with AmpR	pcDNA3	GTCGTCCAGGTCGTCGTCATCCTCGTCCGTGGTGGGCTCCGGGTGGGTGTGGTCTGACAGTTACCAATG

Appendix C

Integrative multi-omics reveals principles of gene regulation and pervasive transcription of transient RNAs in the human cytomegalovirus genome

Table C.1: Estimated T→C conversion rates and error rates for dRNA-seq and STRIPE-seq data

Rate	dSLAM.R1.mock	dSLAM.R1.2hpi	dSLAM.R1.6hpi	dSLAM.R1.12hpi	dSLAM.R1.24hpi	
single_old	0.002551	0.002441	0.002594	0.002384	0.002543	
single_new	0.07576	0.07325	0.09407	0.08262	0.1115	
double_old	0.000532	0.0005356	0.0005318	0.0005308	0.0005387	
double_new	0.05833	0.05696	0.07978	0.0672	0.08854	
	dSLAM.R1.72hpi	**dSLAM.R2.mock1**	**dSLAM.R2.mock2**	**dSLAM.R2.2hpi**	**dSLAM.R2.6hpi**	
single_old	0.00258	0.002217	0.002076	0.002483	0.002162	
single_new	0.131	0.04022	0.04012	0.05114	0.06627	
double_old	0.0005354	0.0006169	0.0005649	0.0006958	0.0004816	
double_new	0.09056	0.03819	0.03597	0.0272	0.05225	
	dSLAM.R2.12hpi	**dSLAM.R2.24hpi**	**dSLAM.R2.48hpi**	**dSLAM.R2.72hpi**	**dSLAM.R2.96hpi**	**dSLAM.R2.GCV.72hpi**
single_old	0.002185	0.002225	0.00218	0.00206	0.00221	0.002227
single_new	0.05311	0.08586	0.09116	0.08805	0.06579	0.09052
double_old	0.0005052	0.0004795	0.0005107	0.0004516	0.0005397	0.000478
double_new	0.04538	0.06798	0.07899	0.08075	0.06071	0.06975

Rate	STRIPE.Mock.minus.A	STRIPE.Mock.minus.B	STRIPE.Mock.plus.A	STRIPE.Mock.plus.B	STRIPE.TB40.24hpi.A
single_old	0.0003819	0.0002928	0.0002963	0.0004481	0.0004373
single_new	0.05805	0.04011	0.05024	0.05322	0.05733
double_old	0.0004921	0.0005016	0.0007833	0.0004824	0.0009217
double_new	0.0004033	0.07383	0.05125	0.0495	0.06302
	STRIPE.TB40.24hpi.B	**STRIPE.TB40.72hpi.A**	**STRIPE.TB40.72hpi.B**	**STRIPE.TB40.72hpi.GCV.A**	**STRIPE.TB40.72hpi.GCV.B**
single_old	0.0003055	0.0003986	0.000523	0.000363	0.0003839
single_new	0.05538	0.0638	0.07183	0.05592	0.06192
double_old	0.0007681	0.0006368	0.0006949	0.001334	0.0005978
double_new	0.05953	0.05825	0.0598	0.06542	0.0551
	STRIPE.TB40.72hpi.PAA.A	**STRIPE.TB40.72hpi.PAA.B**	**STRIPE.BAC2.24hpi.A**	**STRIPE.BAC2.24hpi.B**	**STRIPE.BAC2.72hpi.A**
single_old	0.0003969	0.0003127	0.0006939	0.0006412	0.001037
single_new	0.0353	0.03095	0.07453	0.07421	0.06099
double_old	0.0009096	0.000623	0.001108	0.0008361	0.00133
double_new	0.05413	0.05699	0.07764	0.07315	0.06089
	STRIPE.BAC2.72hpi.B	**STRIPE.BAC2.72hpi.GCV.A**	**STRIPE.BAC2.72hpi.GCV.B**	**STRIPE.BAC2.72hpi.PAA.A**	**STRIPE.BAC2.72hpi.PAA.B**
single_old	0.000922	0.001098	0.001184	0.001185	0.0007703
single_new	0.05761	0.1816	0.1549	0.1202	0.13
double_old	0.001028	0.00104	0.001083	0.001226	0.0008165
double_new	0.06565	0.2462	0.232	0.1291	0.1371

Table C.2: Identified TSS by iTiSS [Jürges et al., 2018] in human and HCMV, respectively. Due to its size this table is omitted from the manuscript. The table is available on BioRxiv.

Table C.3: All promoter motifs identified for each TSS in human and HCMV, respectively. The numbers inside each column depict the relative size from the TSS. TSS has position 0. A negative number means in upstream direction, a positive number in downstream direction. Due to is size this table is omitted from the manuscript. The table is available on BioRxiv.

Table C.4: Predicted transcription factors in HCMV. Transcription factors were predicted using TFM-Explorer [Tonon et al., 2010] in the top 500 TSS per timepoint

2hpi

Rank	Factor	Matrix ID	Location	Sequences	P-Value
1	HIF1A::ARNT	MA0259.1	+[-0089:+0002]	24	2.04E-08
2	Pax6	MA0069.1	+[-0071:-0037]	11	1.66E-06
3	CREB1	MA0018.2	+[-0082:-0047]	13	2.10E-05
4	Pax2	MA0067.1	+[-0061:-0010]	14	2.48E-05
5	NFIL3	MA0025.1	+[-0060:-0030]	9	9.03E-05
6	BRCA1	MA0133.1	+[-0094:-0057]	14	9.50E-05

6hpi

Rank	Factor	Matrix ID	Location	Sequences	P-Value
1	HIF1A::ARNT	MA0259.1	+[-0089:-0010]	25	2.65E-10
2	CREB1	MA0018.2	+[-0084:-0049]	16	1.77E-07
3	MYC::MAX	MA0059.1	+[-0088:-0044]	13	4.80E-06
4	SRF	MA0083.1	+[-0062:-0033]	10	5.02E-06
5	MIZF	MA0131.1	+[-0087:-0013]	13	1.21E-05
6	Mycn	MA0104.2	+[-0087:-0044]	15	1.72E-05
7	FOXD1	MA0031.1	+[-0077:-0022]	14	2.47E-05
8	Ar	MA0007.1	+[-0100:-0046]	12	3.82E-05

12hpi

Rank	Factor	Matrix ID	Location	Sequences	P-Value
1	HIF1A::ARNT	MA0259.1	+[-0089:-0019]	22	1.82E-09
2	MYC::MAX	MA0059.1	+[-0093:-0048]	15	2.13E-07
3	SRF	MA0083.1	+[-0062:-0033]	11	6.92E-07
4	HLF	MA0043.1	+[-0068:-0032]	10	3.04E-05
5	Myc	MA0147.1	+[-0092:-0047]	13	5.23E-05
6	FOXO3	MA0157.1	+[-0079:-0029]	17	7.39E-05

24hpi

Rank	Factor	Matrix ID	Location	Sequences	P-Value
1	TBP	MA0108.2	+[-0100:-0023]	41	1.05E-25
2	HIF1A::ARNT	MA0259.1	+[-0089:-0010]	24	2.65E-10
3	MYC::MAX	MA0059.1	+[-0088:-0018]	20	3.70E-09
4	CREB1	MA0018.2	+[-0082:-0047]	15	1.74E-07
5	Myc	MA0147.1	+[-0087:-0047]	15	7.50E-07
6	Mycn	MA0104.2	+[-0087:-0044]	17	9.54E-07
7	Arnt	MA0004.1	+[-0077:-0042]	8	5.63E-06
8	FOXO3	MA0157.1	+[-0079:-0029]	18	7.56E-06
9	BRCA1	MA0133.1	+[-0093:-0044]	17	1.82E-05
10	USF1	MA0093.1	+[-0078:-0043]	10	2.65E-05
11	HLF	MA0043.1	+[-0068:-0032]	10	3.04E-05

48hpi

Rank	Factor	Matrix ID	Location	Sequences	P-Value
1	TBP	MA0108.2	+[-0055:-0023]	44	1.81E-41
2	HIF1A::ARNT	MA0259.1	+[-0097:+0002]	29	1.63E-10
3	MYC::MAX	MA0059.1	+[-0091:-0019]	19	2.97E-08
4	BRCA1	MA0133.1	+[-0097:-0046]	19	2.06E-07
5	FOXO3	MA0157.1	+[-0097:-0029]	22	3.21E-06
6	Myc	MA0147.1	+[-0091:-0047]	14	1.03E-05
7	FOXD1	MA0031.1	+[-0098:-0027]	17	1.10E-05

72hpi

Rank	Factor	Matrix ID	Location	Sequences	P-Value
1	TBP	MA0108.2	+[-0100:-0030]	36	8.09E-22
2	HIF1A::ARNT	MA0259.1	+[-0097:+0002]	34	1.18E-13
3	MYC::MAX	MA0059.1	+[-0094:+0000]	18	4.04E-07
4	CREB1	MA0018.2	+[-0082:-0037]	14	7.99E-07
5	FOXD1	MA0031.1	+[-0043:-0006]	13	6.20E-06
6	USF1	MA0093.1	+[-0078:-0043]	10	2.65E-05
7	Arnt::Ahr	MA0006.1	+[-0099:-0066]	12	5.48E-05
8	Ar	MA0007.1	+[-0097:-0046]	12	8.66E-05
9	NFIL3	MA0025.1	+[-0067:-0012]	12	9.92E-05

96hpi

Rank	Factor	Matrix ID	Location	Sequences	P-Value
1	TBP	MA0108.2	+[-0100:-0030]	38	1.05E-22
2	HIF1A::ARNT	MA0259.1	+[-0097:+0002]	32	2.31E-12
3	NFIL3	MA0025.1	+[-0067:-0012]	14	1.03E-06
4	USF1	MA0093.1	+[-0078:-0043]	10	2.65E-05
5	FOXD1	MA0031.1	+[-0043:-0006]	12	3.00E-05
6	CREB1	MA0018.2	+[-0082:-0037]	12	5.81E-05
7	MYC::MAX	MA0059.1	+[-0091:-0007]	15	6.22E-05

Appendix D

Statements of individual author contributions to the papers and individual Figures

"Dissertation Based on Several Published Manuscripts"

Statement of individual author contributions and of legal second publication rights

(If required please use more than one sheet)

Publication (complete reference): Jürges, C., Dölken, L. & Erhard, F. Dissecting newly transcribed and old RNA using GRAND-SLAM. *Bioinformatics* **34**, i218--i226 (2018)

Participated in	**Author Initials,** Responsibility decreasing from left to right				
Study Design Methods Development	FE FE	 CSJ			
Data Collection	FE/CSJ				
Data Analysis and Interpretation	FE/CSJ	LD			
Manuscript Writing Writing of Introduction Writing of Materials & Methods Writing of Discussion Writing of First Draft	 FE FE FE FE	 CSJ CSJ CSJ CSJ			

Explanations (if applicable):

Publication (complete reference): Whisnant, A. W. & Jürges, C. S. *et al.* Integrative functional genomics decodes herpes simplex virus 1. *Nat. Commun.* **11**, 2038 (2020)

Participated in	**Author Initials,** Responsibility decreasing from left to right				
Study Design Methods Development	LD/FE CSJ/AWW				
Data Collection	CSJ/AWW	TH	TH/EW/BP/AJR/ AL/LaDj/MG/KD/ JM/RA/NJM/FWHK/ GM/CB/SK/CL/TD/ RZ/ML/FG/PJL/CCF		
Data Analysis and Interpretation	CSJ	FE/LD	CCF		
Manuscript Writing Writing of Introduction Writing of Materials & Methods Writing of Discussion Writing of First Draft	 AWW CSJ CSJ AWW/CSJ	 CSJ AWW AWW	 LD/FE LD/FE LD/FE LD/FE		

Explanations (if applicable): Co-First author publication with AWW. AWW is a virologist and therefore took the responsibility for the biological part of the paper, whereas I (CSJ), as a bioinformatician, were solely responsible for the analysis, methods and evaluation of the data.

Publication (complete reference): Jürges, C. S., Dölken, L. & Erhard, F. Integrative transcription start site identification with iTiSS. *Bioinformatics* 37(18): 3056-3057 (2021).					
Participated in	**Author Initials,** Responsibility decreasing from left to right				
Study Design Methods Development	CSJ CSJ				
Data Collection	CSJ				
Data Analysis and Interpretation	CSJ	FE			
Manuscript Writing Writing of Introduction Writing of Materials & Methods Writing of Discussion Writing of First Draft	 CSJ CSJ CSJ CSJ	 FE FE FE FE	 LD LD		

Explanations (if applicable):

Publication (complete reference): Jürges, C.S., Lodha, M., Trilling, M., Bhandare, P., Wold, E., Zimmermann, A., Le-Trilling, K., Dölken, L. & Erhard, F.					
Participated in	**Author Initials,** Responsibility decreasing from left to right				
Study Design Methods Development	FE/LD CSJ				
Data Collection	CSJ	ML/PB	MT/EW/AZ/KL		
Data Analysis and Interpretation	CSJ	FE/LD			
Manuscript Writing Writing of Introduction Writing of Materials & Methods Writing of Discussion Writing of First Draft	 CSJ CSJ CSJ CSJ	 FE/LD FE/LD FE/LD FE/LD			

Explanations (if applicable): A shorter version of this manuscript will be submitted to Nature Microbiology. This version is available on BioRxiv.

The doctoral researcher confirms that she/he has obtained permission from both the publishers and the co-authors for legal second publication.

The doctoral researcher and the primary supervisor confirm the correctness of the above mentioned assessment.

Christopher Sebastian Jürges	16.12.2021	Würzburg	
Doctoral Researcher's Name	Date	Place	Signature
Florian Erhard	16.12.2021	Würzburg	
Primary Supervisor's Name	Date	Place	Signature

"Dissertation Based on Several Published Manuscripts"

Statement of individual author contributions to figures/tables/chapters included in the manuscripts

(If required please use more than one sheet)

Publication (complete reference): Jürges, C., Dölken, L. & Erhard, F. Dissecting newly transcribed and old RNA using GRAND-SLAM. *Bioinformatics* **34**, i218--i226 (2018)

Figure	**Author Initials,** Responsibility decreasing from left to right				
1	FE	CSJ			
2	CSJ	FE			
3	CSJ/FE				
4	FE	CSJ			
5	FE	CSJ			
6	FE	CSJ			
7	FE	CSJ			
Chapters	**Author Initials,** Responsibility decreasing from left to right				
1	FE	CSJ			
2	FE	CSJ			
3	CSJ/FE				
4	CSJ/FE				
5	FE	CSJ			

Explanations (if applicable):

Publication (complete reference): Whisnant, A. W. & Jürges, C. S. *et al.* Integrative functional genomics decodes herpes simplex virus 1. *Nat. Commun.* **11**, 2038 (2020)

Figure	**Author Initials,** Responsibility decreasing from left to right				
1	CSJ	LD			
2	CSJ				
3	CSJ				
4	CSJ	FE			
5	AWW				
6	CSJ	AWW			
Suppl. Figures	**Author Initials,** Responsibility decreasing from left to right				
1	CSJ				
2	CSJ				
3	CSJ				
4	CSJ				
5	AW				
6	AW				

7	CSJ				
8	CSJ				
9	FE				
Suppl. Tables	**Author Initials,** Responsibility decreasing from left to right				
1	CSJ				
2	CSJ				
3	CSJ				
4	CSJ				
5	CSJ/TD				
6	CSJ				
7	CSJ				
8	CSJ				
9	CSJ				
10	AW				
Chapters	**Author Initials,** Responsibility decreasing from left to right				
1	CSJ/AW	LD/FE			
2	CSJ/AW	LD/FE			
3	CSJ/AW	LD/FE			

Explanations (if applicable): Co-First author publication with AWW. AWW is a virologist and therefore took the responsibility for the biological part of the paper, whereas I (CSJ), as a bioinformatician, were solely responsible for the analysis, methods and evaluation of the data.

Publication (complete reference): Jürges, C. S., Dölken, L. & Erhard, F. Integrative transcription start site identification with iTiSS. *Bioinformatics* 37(18): 3056-3057 (2021).

Figure	**Author Initials,** Responsibility decreasing from left to right				
1	CSJ	FE			
Suppl. Figures	**Author Initials,** Responsibility decreasing from left to right				
1	CSJ				
2	CSJ				
3	CSJ				
4	CSJ				
5	CSJ				
Suppl. Tables	**Author Initials,** Responsibility decreasing from left to right				
1	CSJ				
2	CSJ				
Chapters	**Author Initials,** Responsibility decreasing from left to right				
1	CSJ	FE			
2	CSJ	FE			
3	CSJ	FE			
4	CSJ	FE			

Explanations (if applicable):

Publication (complete reference): Jürges, C.S., Lodha, M., Trilling, M., Bhandare, P., Wold, E., Zimmermann, A., Le-Trilling, K., Dölken, L. & Erhard, F.					
Figure	**Author Initials,** Responsibility decreasing from left to right				
1	CSJ	FE			
2	CSJ	FE			
3	CSJ	FE			
4	CSJ	FE			
5	CSJ/FE				
6	CSJ/FE				
7	CSJ	FE			
Suppl. Figures	**Author Initials,** Responsibility decreasing from left to right				
1	CSJ	FE			
2	CSJ	FE			
3	CSJ	FE			
4	CSJ	FE			
5	CSJ	FE			
6	CSJ	FE			
7	CSJ	FE			
8	CSJ	FE			
Suppl. Tables	**Author Initials,** Responsibility decreasing from left to right				
1	CSJ				
2	CSJ				
3	CSJ				
4	CSJ				
Chapters	**Author Initials,** Responsibility decreasing from left to right				
1	CSJ	FE	LD		
2	CSJ	FE	LD		
3	CSJ	FE	LD		

Explanations (if applicable): A shorter version of this manuscript will be submitted to Nature Microbiology. This version is available on BioRxiv.

I also confirm my primary supervisor's acceptance.

Christopher Sebastian Jürges 16.12.2021 Würzburg

_____

Doctoral Researcher's Name Date Place Signature

Bibliography

[Adams et al., 1991] Adams, M. D., Kelley, J. M., Gocayne, J. D., Dubnick, M., Polymeropoulos, M. H., Xiao, H., Merril, C. R., Wu, A., Olde, B., Moreno, R. F., and al. Et (1991). Complementary DNA sequencing: expressed sequence tags and human genome project. *Science*, 252(5013):1651 LP – 1656.

[Amati et al., 2001] Amati, B., Frank, S. R., Donjerkovic, D., and Taubert, S. (2001). Function of the c-Myc oncoprotein in chromatin remodeling and transcription. *Biochimica et Biophysica Acta (BBA) - Reviews on Cancer*, 1471(3):M135–M145.

[Arias et al., 2014] Arias, C., Weisburd, B., Stern-Ginossar, N., Mercier, A., Madrid, A. S., Bellare, P., Holdorf, M., Weissman, J. S., and Ganem, D. (2014). 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.

[Arthur and Vassilvitskii, 2007] Arthur, D. and Vassilvitskii, S. (2007). K-Means++: The Advantages of Careful Seeding. In *Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '07, pages 1027–1035, USA. Society for Industrial and Applied Mathematics.

[Bailey and Elkan, 1994] Bailey, T. L. and Elkan, C. (1994). Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proceedings. International Conference on Intelligent Systems for Molecular Biology*, 2:28–36.

[Balázs et al., 2018] Balázs, Z., Tombácz, D., Szűcs, A., Snyder, M., and Boldogkői, Z. (2018). Dual Platform Long-Read RNA-Sequencing Dataset of the Human Cytomegalovirus Lytic Transcriptome .

[Baradaran et al., 1994] Baradaran, K., Dabrowski, C. E., and Schaffer, P. A. (1994). Transcriptional analysis of the region of the herpes simplex virus type 1 genome containing the UL8, UL9, and UL10 genes and identification of a novel delayed-early gene product, OBPC. *Journal of virology*, 68(7):4251–4261.

[Barbosa et al., 2013] Barbosa, C., Peixeiro, I., and Romão, L. (2013). Gene Expression Regulation by Upstream Open Reading Frames and Human Disease. *PLOS Genetics*, 9(8):e1003529.

[Bartel, 2009] Bartel, D. P. (2009). MicroRNAs: target recognition and regulatory functions. *Cell*, 136(2):215–233.

[Bashir et al., 2012] Bashir, A., Klammer, A. A., Robins, W. P., Chin, C.-S., Webster, D., Paxinos, E., Hsu, D., Ashby, M., Wang, S., Peluso, P., Sebra, R., Sorenson, J., Bullard, J., Yen, J., Valdovino, M., Mollova, E., Luong, K., Lin, S., LaMay, B., Joshi, A., Rowe, L., Frace, M., Tarr, C. L., Turnsek, M., Davis, B. M., Kasarskis, A., Mekalanos, J. J., Waldor, M. K., and Schadt, E. E. (2012). A hybrid approach for the automated finishing of bacterial genomes. *Nature Biotechnology*, 30(7):701–707.

[Batista et al., 2014] Batista, P. J., Molinie, B., Wang, J., Qu, K., Zhang, J., Li, L., Bouley, D. M., Lujan, E., Haddad, B., Daneshvar, K., Carter, A. C., Flynn, R. A., Zhou, C., Lim, K.-S., Dedon, P., Wernig, M., Mullen, A. C., Xing, Y., Giallourakis, C. C., and Chang, H. Y. (2014). m(6)A RNA modification controls cell fate transition in mammalian embryonic stem cells. *Cell stem cell*, 15(6):707–719.

[Beaudoing et al., 2000] Beaudoing, E., Freier, S., Wyatt, J. R., Claverie, J. M., and Gautheret, D. (2000). Patterns of variant polyadenylation signal usage in human genes. *Genome Res*, 10(7):1001–1010.

[Bell and Tora, 1999] Bell, B. and Tora, L. (1999). Regulation of Gene Expression by Multiple Forms of TFIID and Other Novel TAFII-Containing Complexes. *Experimental Cell Research*, 246(1):11–19.

[Bell et al., 2011] Bell, J. T., Pai, A. A., Pickrell, J. K., Gaffney, D. J., Pique-Regi, R., Degner, J. F., Gilad, Y., and Pritchard, J. K. (2011). DNA methylation patterns associate with genetic and gene expression variation in HapMap cell lines. *Genome biology*, 12(1):R10–R10.

[Bencun et al., 2018] Bencun, M., Klinke, O., Hotz-Wagenblatt, A., Klaus, S., Tsai, M.-H., Poirey, R., and Delecluse, H.-J. (2018). Translational profiling of B cells infected with the Epstein-Barr virus reveals 5' leader ribosome recruitment through upstream open reading frames. *Nucleic Acids Research*, 46(6):2802–2819.

[Bencurova et al., 2018] Bencurova, E., Gupta, S., Sarukhanyan, E., and Dandekar, T. (2018). Identification of Antifungal Targets Based on Computer Modeling. *Journal of Fungi*, 4(3):81.

[Blake et al., 1990] Blake, M. C., Jambou, R. C., Swick, A. G., Kahn, J. W., and Azizkhan, J. C. (1990). Transcriptional initiation is controlled by upstream GC-box interactions in a TATAA-less promoter. *Molecular and cellular biology*, 10(12):6632–6641.

[Bleidorn, 2016] Bleidorn, C. (2016). Third generation sequencing: technology and its potential impact on evolutionary biodiversity research. *Systematics and Biodiversity*, 14(1):1–8.

[Bolger et al., 2014] Bolger, A. M., Lohse, M., and Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, 30(15):2114–2120.

[Boppana et al., 1992] Boppana, S. B., Pass, R. F., Britt, W. J., Stagno, S., and Alford, C. A. (1992). Symptomatic congenital cytomegalovirus infection: neonatal morbidity and mortality. *The Pediatric infectious disease journal*, 11(2):93–99.

[Bray et al., 2016] Bray, N. L., Pimentel, H., Melsted, P., and Pachter, L. (2016). Near-optimal probabilistic RNA-seq quantification. *Nature Biotechnology*, 34(5):525–527.

[Bresnahan and Shenk, 2000] Bresnahan, W. A. and Shenk, T. (2000). A subset of viral transcripts packaged within human cytomegalovirus particles. *Science*, 288(5475):2373–2376.

[Cabrera-Quio et al., 2016] Cabrera-Quio, L. E., Herberg, S., and Pauli, A. (2016). Decoding sORF translation – from small proteins to gene regulation. *RNA Biology*, 13(11):1051–1059.

[Calton et al., 2004] Calton, C. M., Randall, J. A., Adkins, M. W., and Banfield, B. W. (2004). The Pseudorabies Virus Serine/Threonine Kinase Us3 Contains Mitochondrial, Nuclear and Membrane Localization Signals. *Virus Genes*, 29(1):131–145.

[Camacho et al., 2009] Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., and Madden, T. L. (2009). BLAST+: architecture and applications. *BMC Bioinformatics*, 10(1):421.

[Carninci et al., 2006] Carninci, P., Sandelin, A., Lenhard, B., Katayama, S., Shimokawa, K., Ponjavic, J., Semple, C. A. M., Taylor, M. S., Engström, P. G., Frith, M. C., Forrest, A. R. R., Alkema, W. B., Tan, S. L., Plessy, C., Kodzius, R., Ravasi, T., Kasukawa, T., Fukuda, S., Kanamori-Katayama, M., Kitazume, Y., Kawaji, H., Kai, C., Nakamura, M., Konno, H., Nakano, K., Mottagui-Tabar, S., Arner, P., Chesi, A., Gustincich, S., Persichetti, F., Suzuki, H., Grimmond, S. M., Wells, C. A., Orlando, V., Wahlestedt, C., Liu, E. T., Harbers, M., Kawai, J., Bajic, V. B., Hume, D. A., and Hayashizaki, Y. (2006). Genome-wide analysis of mammalian promoter architecture and evolution. *Nature Genetics*, 38(6):626–635.

[Chao and Price, 2001] Chao, S. H. and Price, D. H. (2001). Flavopiridol Inactivates P-TEFb and Blocks Most RNA Polymerase II Transcription in Vivo. *Journal of Biological Chemistry*, 276(34):31793–31799.

[Cherrington and Mocarski, 1989] Cherrington, J. M. and Mocarski, E. S. (1989). Human cytomegalovirus ie1 transactivates the alpha promoter-enhancer via an 18-base-pair repeat element. *Journal of virology*, 63(3):1435–1440.

[Chu et al., 2015] Chu, Q., Ma, J., and Saghatelian, A. (2015). Identification and characterization of sORF-encoded polypeptides. *Critical Reviews in Biochemistry and Molecular Biology*, 50(2):134–141.

[Chu and Corey, 2012] Chu, Y. and Corey, D. R. (2012). RNA sequencing: platform selection, experimental design, and data interpretation. *Nucleic acid therapeutics*, 22(4):271–274.

[Core and Adelman, 2019] Core, L. and Adelman, K. (2019). Promoter-proximal pausing of RNA polymerase II: a nexus of gene regulation. *Genes & Development*, 33(15-16):960–982.

[Cox and Mann, 2008] Cox, J. and Mann, M. (2008). MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nature Biotechnology*, 26(12):1367–1372.

[Cramer, 2019] Cramer, P. (2019). Organization and regulation of gene transcription. *Nature*, 573(7772):45–54.

[Crick, 1958] Crick, F. (1958). On protein synthesis. *Symposia of the Society for Experimental Biology*, 12:138–163.

[Dandekar et al., 2000] Dandekar, T., Huynen, M., Regula, J. T., Ueberle, B., Zimmermann, C. U., Andrade, M. A., Doerks, T., Sánchez-Pulido, L., Snel, B., Suyama, M., Yuan, Y. P., Herrmann, R., and Bork, P. (2000). Re-annotating the Mycoplasma pneumoniae genome sequence: adding value, function and reading frames. *Nucleic acids research*, 28(17):3278–3288.

[Davis et al., 2013] Davis, M. P. A., van Dongen, S., Abreu-Goodger, C., Bartonicek, N., and Enright, A. J. (2013). Kraken: a set of tools for quality control and analysis of high-throughput sequence data. *Methods (San Diego, Calif.)*, 63(1):41–49.

[Davis et al., 2015] Davis, Z., Verschueren, E., Jang, G., Kleffman, K., Johnson, J., Park, J., Von Dollen, J., Maher, M., Johnson, T., Newton, W., Jäger, S., Shales, M., Horner, J., Hernandez, R., Krogan, N., and Glaunsinger, B. (2015). Global Mapping of Herpesvirus-Host Protein Complexes Reveals a Transcription Strategy for Late Genes. *Molecular Cell*, 57(2):349–360.

[Davison, 2010] Davison, A. J. (2010). Herpesvirus systematics. *Veterinary Microbiology*, 143(1):52–69.

[Davison et al., 2003] Davison, A. J., Dolan, A., Akter, P., Addison, C., Dargan, D. J., Alcendor, D. J., McGeoch, D. J., and Hayward, G. S. (2003). The human cytomegalovirus genome revisited: comparison with the chimpanzee cytomegalovirus genome. *J Gen Virol*, 84(Pt 1):17–28.

[Demmler, 1991] Demmler, G. J. (1991). Infectious Diseases Society of America and Centers for Disease Control. Summary of a workshop on surveillance for congenital cytomegalovirus disease.

[Depledge et al., 2019] Depledge, D. P., Srinivas, K. P., Sadaoka, T., Bready, D., Mori, Y., Placantonakis, D. G., Mohr, I., and Wilson, A. C. (2019). Direct RNA sequencing on nanopore arrays redefines the transcriptional complexity of a viral pathogen. *Nature Communications*, 10(1):754.

[Dobin et al., 2013] Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T. R. (2013). STAR: ultrafast universal RNA-seq aligner. *Bioinformatics (Oxford, England)*, 29(1):15–21.

[Dölken et al., 2008] Dölken, L., Ruzsics, Z., Radle, B., Friedel, C. C., Zimmer, R., Mages, J., Hoffmann, R., Dickinson, P., Forster, T., Ghazal, P., and Koszinowski, U. H. (2008). High-resolution gene expression profiling for simultaneous kinetic parameter analysis of RNA synthesis and decay. *RNA.*, 14(9):1959–1972.

[Domingo et al., 2008] Domingo, E., Parrish, C. R., and Holland, J. J. (2008). Origin and Evolution of Viruses. In *Origin and Evolution of Viruses*, pages 447–476. Academic Press.

[Draper et al., 1986] Draper, K. G., Devi-Rao, G., Costa, R. H., Blair, E. D., Thompson, R. L., and Wagner, E. K. (1986). Characterization of the genes encoding herpes simplex virus type 1 and type 2 alkaline exonucleases and overlapping proteins. *Journal of virology*, 57(3):1023–1036.

[Dridi et al., 2018] Dridi, S., Richerioux, N., Gonzalez Suarez, C. E., Vanharen, M., Sanabria-Solano, C., and Pearson, A. (2018). A Mutation in the UL24 Gene Abolishes Expression of the Newly Identified UL24.5 Protein of Herpes Simplex Virus 1 and Leads to an Increase in Pathogenicity in Mice. *Journal of virology*, 92(20):e00671—-18.

[Duffy et al., 2015] Duffy, E. E., Rutenberg-Schoenberg, M., Stark, C. D., Kitchen, R. R., Gerstein, M. B., and Simon, M. D. (2015). Tracking Distinct RNA Populations Using Efficient and Reversible Covalent Chemistry. *Mol Cell*, 59(5):858–866.

[E. et al., 2000] E., G. A., J., D. C. A., and M., M. J. (2000). Human Cytomegalovirus Virions Differentially Incorporate Viral and Host Cell RNA during the Assembly Process. *Journal of Virology*, 74(19):9078–9082.

[Eid et al., 2009] Eid, J., Fehr, A., Gray, J., Luong, K., Lyle, J., Otto, G., Peluso, P., Rank, D., Baybayan, P., Bettman, B., Bibillo, A., Bjornson, K., Chaudhuri, B., Christians, F., Cicero, R., Clark, S., Dalal, R., Dewinter, A., Dixon, J., Foquet, M., Gaertner, A., Hardenbol, P., Heiner, C., Hester, K., Holden, D., Kearns, G., Kong, X., Kuse, R., Lacroix, Y., Lin, S., Lundquist, P., Ma, C., Marks, P., Maxham, M., Murphy, D., Park, I., Pham, T., Phillips, M., Roy, J., Sebra, R., Shen, G., Sorenson, J., Tomaney, A., Travers, K., Trulson, M., Vieceli, J., Wegener, J., Wu, D., Yang, A., Zaccarin, D., Zhao, P., Zhong, F., Korlach, J., and Turner, S. (2009). Real-time DNA sequencing from single polymerase molecules. *Science (New York, N.Y.)*, 323(5910):133–138.

[Erhard, 2018] Erhard, F. (2018). Estimating pseudocounts and fold changes for digital expression measurements. *Bioinformatics*.

[Erhard et al., 2019] Erhard, F., Baptista, M. A. P., Krammer, T., Hennig, T., Lange, M., Arampatzi, P., Jürges, C. S., Theis, F. J., Saliba, A.-E., and Dölken, L. (2019). scSLAM-seq reveals core features of transcription dynamics in single cells. *Nature*, 571(7765):419–423.

[Erhard et al., 2018] Erhard, F., Halenius, A., Zimmermann, C., L'Hernault, A., Kowalewski, D. J., Weekes, M. P., Stevanovic, S., Zimmer, R., and Dölken, L. (2018). Improved Ribo-seq enables identification of cryptic translation events. *Nature Methods*, 15(5):363–366.

[Erhard and Zimmer, 2015] Erhard, F. and Zimmer, R. (2015). Count ratio model reveals bias affecting NGS fold changes. *Nucleic Acids Res*, 43(20):e136.

[Fields et al., 2013] Fields, B. N., Knipe, D. M. D. M., and Howley, P. M. (2013). *Fields virology*. Wolters Kluwer Health/Lippincott Williams & Wilkins.

[Finnen et al., 2010] Finnen, R. L., Roy, B. B., Zhang, H., and Banfield, B. W. (2010). Analysis of filamentous process induction and nuclear localization properties of the HSV-2 serine/threonine kinase Us3. *Virology*, 397(1):23–33.

[Garalde et al., 2018] Garalde, D. R., Snell, E. A., Jachimowicz, D., Sipos, B., Lloyd, J. H., Bruce, M., Pantic, N., Admassu, T., James, P., Warland, A., Jordan, M., Ciccone, J., Serra, S., Keenan, J., Martin, S., McNeill, L., Wallace, E. J., Jayasinghe, L., Wright, C., Blasco, J., Young, S., Brocklebank, D., Juul, S., Clarke, J., Heron, A. J., and Turner, D. J. (2018). Highly parallel direct RNA sequencing on an array of nanopores. *Nature Methods*, 15(3):201–206.

[Garg et al., 2016] Garg, R., Shankar, R., Thakkar, B., Kudapa, H., Krishnamurthy, L., Mantri, N., Varshney, R. K., Bhatia, S., and Jain, M. (2016). Transcriptome analyses reveal genotype- and developmental stage-specific molecular responses to drought and salinity stresses in chickpea. *Scientific reports*, 6:19228.

[Gatherer et al., 2011] Gatherer, D., Seirafian, S., Cunningham, C., Holton, M., Dargan, D. J., Baluchova, K., Hector, R. D., Galbraith, J., Herzyk, P., Wilkinson, G. W., and Davison, A. J. (2011). High-resolution human cytomegalovirus transcriptome. *Proc Natl Acad Sci U S A*, 108(49):19755–19760.

[Gaudermann et al., 2006] Gaudermann, P., Vogl, I., Zientz, E., Silva, F. J., Moya, A., Gross, R., and Dandekar, T. (2006). Analysis of and function predictions for previously conserved hypothetical or putative proteins in Blochmannia floridanus. *BMC microbiology*, 6:1.

[Georgakilas et al., 2020] Georgakilas, G. K., Perdikopanis, N., and Hatzigeorgiou, A. (2020). Solving the transcription start site identification problem with ADAPT-CAGE: a Machine Learning algorithm for the analysis of CAGE data. *Scientific Reports*, 10(1):877.

[Govind et al., 2009] Govind, G., Harshavardhan, V. T., Patricia, J. K., Dhanalakshmi, R., Senthil Kumar, M., Sreenivasulu, N., and Udayakumar, M. (2009). Identification and functional validation of a unique set of drought induced genes preferentially expressed in response to gradual water stress in peanut. *Molecular genetics and genomics : MGG*, 281(6):591–605.

[Griffiths and Reeves, 2021] Griffiths, P. and Reeves, M. (2021). Pathogenesis of human cytomegalovirus in the immunocompromised host. *Nature Reviews Microbiology*.

[Gruffat et al., 2016] Gruffat, H., Marchione, R., and Manet, E. (2016). Herpesvirus Late Gene Expression: A Viral-Specific Pre-initiation Complex Is Key. *Frontiers in microbiology*, 7:869.

[Gu et al., 2016] Gu, Z., Eils, R., and Schlesner, M. (2016). Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics*.

[Gu et al., 2014] Gu, Z., Gu, L., Eils, R., Schlesner, M., and Brors, B. (2014). circlize implements and enhances circular visualization in R. *Bioinformatics*, 30(19):2811–2812.

[Gupte et al., 1991] Gupte, S. S., Olson, J. W., and Ruyechan, W. T. (1991). The major herpes simplex virus type-1 DNA-binding protein is a zinc metalloprotein. *The Journal of biological chemistry*, 266(18):11413–11416.

[Haberle et al., 2015] Haberle, V., Forrest, A. R. R., Hayashizaki, Y., Carninci, P., and Lenhard, B. (2015). CAGEr: precise TSS data retrieval and high-resolution promoterome mining for integrative analyses. *Nucleic acids research*, 43(8):e51–e51.

[Haberle and Stark, 2018] Haberle, V. and Stark, A. (2018). Eukaryotic core promoters and the functional basis of transcription initiation. *Nature Reviews Molecular Cell Biology*, 19(10):621–637.

[Hagemeier et al., 1992] Hagemeier, C., Walker, S., Caswell, R., Kouzarides, T., and Sinclair, J. (1992). The human cytomegalovirus 80-kilodalton but not the 72-kilodalton immediate-early protein transactivates heterologous promoters in a TATA box-dependent mechanism and interacts directly with TFIID. *Journal of Virology*, 66(7):4452–4456.

[Hann et al., 1998] Hann, L. E., Cook, W. J., Uprichard, S. L., Knipe, D. M., and Coen, D. M. (1998). The role of herpes simplex virus ICP27 in the regulation of UL24 gene expression by differential polyadenylation. *J Virol*, 72(10):7709–7714.

[Hennig et al., 2018] Hennig, T., Michalski, M., Rutkowski, A. J., Djakovic, L., Whisnant, A. W., Friedl, M.-S., Jha, B. A., Baptista, M. A. P., L'Hernault, A., Erhard, F., Dölken, L., and Friedel, C. C. (2018). HSV-1-induced disruption of transcription termination resembles a cellular stress response but selectively increases chromatin accessibility downstream of genes. *PLOS Pathogens*, 14(3):e1006954.

[Hernández et al., 2015] Hernández, S., Franco, L., Calvo, A., Ferragut, G., Hermoso, A., Amela, I., Gómez, A., Querol, E., and Cedano, J. (2015). Bioinformatics and Moonlighting Proteins. *Frontiers in bioengineering and biotechnology*, 3:90.

[Herzog et al., 2017] Herzog, V. A., Reichholf, B., Neumann, T., Rescheneder, P., Bhat, P., Burkard, T. R., Wlotzka, W., von Haeseler, A., Zuber, J., and Ameres, S. L. (2017). Thiol-linked alkylation of RNA to assess expression dynamics. *Nature Methods*, 14(12):1198–1204.

[Hiller et al., 2004] Hiller, M., Huse, K., Szafranski, K., Jahn, N., Hampe, J., Schreiber, S., Backofen, R., and Platzer, M. (2004). Widespread occurrence of alternative splicing at NAGNAG acceptors contributes to proteome plasticity. *Nature Genetics*, 36(12):1255–1257.

[Hinnebusch et al., 2016] Hinnebusch, A. G., Ivanov, I. P., and Sonenberg, N. (2016). Translational control by 5'-untranslated regions of eukaryotic mRNAs. *Science (New York, N.Y.)*, 352(6292):1413–1416.

[Hinnebusch and Lorsch, 2012] Hinnebusch, A. G. and Lorsch, J. R. (2012). The mechanism of eukaryotic translation initiation: new insights and challenges. *Cold Spring Harbor perspectives in biology*, 4(10):a011544.

[Hiroki et al., 2011] Hiroki, I., F., S. M., Takayuki, M., Yoriko, Y., Teru, K., Shinya, T., and Tatsuya, T. (2011). The Human Cytomegalovirus Gene Products Essential for Late Viral Gene Expression Assemble into Prereplication Complexes before Viral DNA Replication. *Journal of Virology*, 85(13):6629–6644.

[Hobbs et al., 2014] Hobbs, M., Pavasovic, A., King, A. G., Prentis, P. J., Eldridge, M. D. B., Chen, Z., Colgan, D. J., Polkinghorne, A., Wilkins, M. R., Flanagan, C., Gillett, A., Hanger, J., Johnson, R. N., and Timms, P. (2014). A transcriptome resource for the koala (Phascolarctos cinereus): insights into koala retrovirus transcription and sequence diversity. *BMC genomics*, 15(1):786.

[Hough et al., 2000] Hough, C. D., Sherman-Baust, C. A., Pizer, E. S., Montz, F. J., Im, D. D., Rosenshein, N. B., Cho, K. R., Riggins, G. J., and Morin, P. J. (2000). Large-Scale Serial Analysis of Gene Expression Reveals Genes Differentially Expressed in Ovarian Cancer. *Cancer Research*, 60(22):6281 LP – 6287.

[Huerta-Cepas et al., 2017] Huerta-Cepas, J., Forslund, K., Coelho, L. P., Szklarczyk, D., Jensen, L. J., von Mering, C., and Bork, P. (2017). Fast Genome-Wide Functional

Annotation through Orthology Assignment by eggNOG-Mapper. *Molecular biology and evolution*, 34(8):2115–2122.

[Hunter et al., 2009] Hunter, S., Apweiler, R., Attwood, T. K., Bairoch, A., Bateman, A., Binns, D., Bork, P., Das, U., Daugherty, L., Duquenne, L., Finn, R. D., Gough, J., Haft, D., Hulo, N., Kahn, D., Kelly, E., Laugraud, A., Letunic, I., Lonsdale, D., Lopez, R., Madera, M., Maslen, J., McAnulla, C., McDowall, J., Mistry, J., Mitchell, A., Mulder, N., Natale, D., Orengo, C., Quinn, A. F., Selengut, J. D., Sigrist, C. J. A., Thimma, M., Thomas, P. D., Valentin, F., Wilson, D., Wu, C. H., and Yeats, C. (2009). InterPro: the integrative protein signature database. *Nucleic acids research*, 37(Database issue):D211–D215.

[Ingolia et al., 2009] Ingolia, N. T., Ghaemmaghami, S., Newman, J. R., and Weissman, J. S. (2009). Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. *Science*, 324(5924):218–223.

[Jiang et al., 2011] Jiang, L., Schlesinger, F., Davis, C. A., Zhang, Y., Li, R., Salit, M., Gingeras, T. R., and Oliver, B. (2011). Synthetic spike-in standards for RNA-seq experiments. *Genome research*, 21(9):1543–1551.

[Johnstone et al., 2016] Johnstone, T. G., Bazzini, A. A., and Giraldez, A. J. (2016). Upstream ORFs are prevalent translational repressors in vertebrates. *The EMBO journal*, 35(7):706–723.

[Jonas and Izaurralde, 2015] Jonas, S. and Izaurralde, E. (2015). Towards a molecular understanding of microRNA-mediated gene silencing. *Nature reviews. Genetics*, 16(7):421–433.

[Jovasevic and Roizman, 2010] Jovasevic, V. and Roizman, B. (2010). The novel HSV-1 US5-1 RNA is transcribed off a domain encoding US5, US4, US3, US2 and alpha22. *Virology journal*, 7(1):103.

[Jürges et al., 2018] Jürges, C., Dölken, L., and Erhard, F. (2018). Dissecting newly transcribed and old RNA using GRAND-SLAM. *Bioinformatics*, 34(13):i218—-i226.

[Jürges et al., 2021] Jürges, C. S., Dölken, L., and Erhard, F. (2021). Integrative transcription start site identification with iTiSS. *Bioinformatics*.

[Kenzelmann et al., 2007] Kenzelmann, M., Maertens, S., Hergenhahn, M., Kueffer, S., Hotz-Wagenblatt, A., Li, L., Wang, S., Ittrich, C., Lemberger, T., Arribas, R., Jonnakuty, S., Hollstein, M. C., Schmid, W., Gretz, N., Grone, H. J., and Schutz, G. (2007). Microarray analysis of newly synthesized RNA in cells and animals. *Proc Natl Acad Sci U S A*, 104(15):6164–6169.

[Kimberlin et al., 2001] Kimberlin, D. W., Lin, C. Y., Jacobs, R. F., Powell, D. A., Frenkel, L. M., Gruber, W. C., Rathore, M., Bradley, J. S., Diaz, P. S., Kumar, M., Arvin, A. M.,

Gutierrez, K., Shelton, M., Weiner, L. B., Sleasman, J. W., de Sierra, T. M., Soong, S. J., Kiell, J., Lakeman, F. D., and Whitley, R. J. (2001). Natural history of neonatal herpes simplex virus infections in the acyclovir era. *Pediatrics*, 108(2):223–229.

[King et al., 2012] King, A., Lefkowitz, E., Adams, M. J., and Carstens, E. B. (2012). Order - Herpesvirales. In King, A. M. Q., Adams, M. J., Carstens, E. B., and Lefkowitz, E. J. B. T. V. T., editors, *Virus Taxonomy*, pages 99–107. Elsevier, San Diego.

[Kober et al., 2013] Kober, L., Zehe, C., and Bode, J. (2013). Optimized signal peptides for the development of high expressing CHO cell lines. *Biotechnology and bioengineering*, 110(4):1164–1173.

[Kolodziejczyk et al., 2015] Kolodziejczyk, A., Kim, J. K., Svensson, V., Marioni, J., and Teichmann, S. (2015). The Technology and Biology of Single-Cell RNA Sequencing. *Molecular Cell*, 58(4):610–620.

[Koren et al., 2012] Koren, S., Schatz, M. C., Walenz, B. P., Martin, J., Howard, J. T., Ganapathy, G., Wang, Z., Rasko, D. A., McCombie, W. R., Jarvis, E. D., and Phillippy, A. M. (2012). Hybrid error correction and de novo assembly of single-molecule sequencing reads. *Nature Biotechnology*, 30(7):693–700.

[Kukurba and Montgomery, 2015] Kukurba, K. R. and Montgomery, S. B. (2015). RNA Sequencing and Analysis. *Cold Spring Harbor protocols*, 2015(11):951–969.

[Kwak et al., 2013] Kwak, H., Fuda, N. J., Core, L. J., and Lis, J. T. (2013). Precise Maps of RNA Polymerase Reveal How Promoters Direct Initiation and Pausing. *Science*, 339(6122):950–953.

[La et al., 2003] La, S. B., Stéphane, A., Catherine, R., Liang, J., Xavier, d. L., Michel, D., Richard, B., Jean-Michel, C., and Didier, R. (2003). A Giant Virus in Amoebae. *Science*, 299(5615):2033.

[Langmead and Salzberg, 2012] Langmead, B. and Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature Methods*, 9(4):357–359.

[Langmead et al., 2009] Langmead, B., Trapnell, C., Pop, M., and Salzberg, S. L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.*, 10(3):R25.

[Lee et al., 2016] Lee, S. H., Caviness, K., Albright, E. R., Lee, J.-H., Gelbmann, C. B., Rak, M., Goodrum, F., and Kalejta, R. F. (2016). Long and Short Isoforms of the Human Cytomegalovirus UL138 Protein Silence IE Transcription and Promote Latency. *Journal of virology*, 90(20):9483–9494.

[Letunic et al., 2015] Letunic, I., Doerks, T., and Bork, P. (2015). SMART: recent updates, new developments and status in 2015. *Nucleic acids research*, 43(Database issue):D257–D260.

[Li et al., 2009] Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., and 1000 Genome Project Data Processing Subgroup (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics (Oxford, England)*, 25(16):2078–2079.

[Li et al., 2020] Li, M., Ball, C. B., Collins, G., Hu, Q., Luse, D. S., Price, D. H., and Meier, J. L. (2020). Human cytomegalovirus IE2 drives transcription initiation from a select subset of late infection viral promoters by host RNA polymerase II. *PLOS Pathogens*, 16(4):e1008402.

[Li et al., 2021] Li, M., Hu, Q., Collins, G., Parida, M., Ball, C. B., Price, D. H., and Meier, J. L. (2021). Cytomegalovirus late transcription factor target sequence diversity orchestrates viral early to late transcription. *PLOS Pathogens*, 17(8):e1009796.

[Li et al., 2011] Li, Z., Zhang, Z., Yan, P., Huang, S., Fei, Z., and Lin, K. (2011). RNA-Seq improves annotation of protein-coding genes in the cucumber genome. *BMC genomics*, 12:540.

[Lifton et al., 1978] Lifton, R. P., Goldberg, M. L., Karp, R. W., and Hogness, D. S. (1978). The Organization of the Histone Genes in Drosophila melanogaster: Functional and Evolutionary Implications. *Cold Spring Harbor Symposia on Quantitative Biology*, 42:1047–1051.

[Liu and Roizman, 1991] Liu, F. Y. and Roizman, B. (1991). The promoter, transcriptional unit, and coding sequence of herpes simplex virus 1 family 35 proteins are contained within and in frame with the UL26 open reading frame. *Journal of virology*, 65(1):206–212.

[Lu et al., 2020] Lu, S., Wang, J., Chitsaz, F., Derbyshire, M. K., Geer, R. C., Gonzales, N. R., Gwadz, M., Hurwitz, D. I., Marchler, G. H., Song, J. S., Thanki, N., Yamashita, R. A., Yang, M., Zhang, D., Zheng, C., Lanczycki, C. J., and Marchler-Bauer, A. (2020). CDD/SPARCLE: the conserved domain database in 2020. *Nucleic acids research*, 48(D1):D265–D268.

[Lurain et al., 2006] Lurain, N. S., Fox, A. M., Lichy, H. M., Bhorade, S. M., Ware, C. F., Huang, D. D., Kwan, S.-P., Garrity, E. R., and Chou, S. (2006). Analysis of the human cytomegalovirus genomic region from UL146 through UL147A reveals sequence hypervariability, genotypic stability, and overlapping transcripts. *Virology Journal*, 3(1):4.

[Mahat et al., 2016] Mahat, D. B., Kwak, H., Booth, G. T., Jonkers, I. H., Danko, C. G., Patel, R. K., Waters, C. T., Munson, K., Core, L. J., and Lis, J. T. (2016). Base-pair-resolution genome-wide mapping of active RNA polymerases using precision nuclear run-on (PRO-seq). *Nature protocols*, 11(8):1455–1476.

[Malani, 2010] Malani, P. N. (2010). Mandell, Douglas, and Bennett's Principles and Practice of Infectious Diseases. *JAMA*, 304(18):2067–2071.

[Malone et al., 1990] Malone, C. L., Vesole, D. H., and Stinski, M. F. (1990). Transactivation of a human cytomegalovirus early promoter by gene products from the immediate-early gene IE2 and augmentation by IE1: mutational analysis of the viral proteins. *Journal of virology*, 64(4):1498–1506.

[Mann and Jensen, 2003] Mann, M. and Jensen, O. N. (2003). Proteomic analysis of post-translational modifications. *Nature Biotechnology*, 21(3):255–261.

[Mapelli et al., 2005] Mapelli, M., Panjikar, S., and Tucker, P. A. (2005). The crystal structure of the herpes simplex virus 1 ssDNA-binding protein suggests the structural basis for flexible, cooperative single-stranded DNA binding. *The Journal of biological chemistry*, 280(4):2990–2997.

[Marcinowski et al., 2012] Marcinowski, L., Lidschreiber, M., Windhager, L., Rieder, M., Bosse, J. B., Radle, B., Bonfert, T., Gyory, I., de Graaf, M., Prazeres da Costa, O., Rosenstiel, P., Friedel, C. C., Zimmer, R., Ruzsics, Z., and Dölken, L. (2012). Real-time Transcriptional Profiling of Cellular and Viral Gene Expression during Lytic Cytomegalovirus Infection. *PLoS Pathog*, 8(9):e1002908.

[Margulies et al., 2005] Margulies, M., Egholm, M., Altman, W. E., Attiya, S., Bader, J. S., Bemben, L. A., Berka, J., Braverman, M. S., Chen, Y.-J., Chen, Z., Dewell, S. B., Du, L., Fierro, J. M., Gomes, X. V., Godwin, B. C., He, W., Helgesen, S., Ho, C. H., Irzyk, G. P., Jando, S. C., Alenquer, M. L. I., Jarvie, T. P., Jirage, K. B., Kim, J.-B., Knight, J. R., Lanza, J. R., Leamon, J. H., Lefkowitz, S. M., Lei, M., Li, J., Lohman, K. L., Lu, H., Makhijani, V. B., McDade, K. E., McKenna, M. P., Myers, E. W., Nickerson, E., Nobile, J. R., Plant, R., Puc, B. P., Ronan, M. T., Roth, G. T., Sarkis, G. J., Simons, J. F., Simpson, J. W., Srinivasan, M., Tartaro, K. R., Tomasz, A., Vogt, K. A., Volkmer, G. A., Wang, S. H., Wang, Y., Weiner, M. P., Yu, P., Begley, R. F., and Rothberg, J. M. (2005). Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, 437(7057):376–380.

[Martí-Carreras and Maes, 2019] Martí-Carreras, J. and Maes, P. (2019). Human cytomegalovirus genomics and transcriptomics through the lens of next-generation sequencing: revision and future challenges. *Virus Genes*, 55(2):138–164.

[Mayer et al., 2015] Mayer, A., di Iulio, J., Maleri, S., Eser, U., Vierstra, J., Reynolds, A., Sandstrom, R., Stamatoyannopoulos, J. A., and Churchman, L. S. (2015). Native elongating transcript sequencing reveals human transcriptional activity at nucleotide resolution. *Cell*, 161(3):541–554.

[McFarlane et al., 2011] McFarlane, S., Nicholl, M. J., Sutherland, J. S., and Preston, C. M. (2011). Interaction of the human cytomegalovirus particle with the host cell induces hypoxia-inducible factor 1 alpha. *Virology*, 414(1):83–90.

[McGeoch et al., 1995] McGeoch, D. J., Cook, S., Dolan, A., Jamieson, F. E., and Telford, E. A. R. (1995). Molecular Phylogeny and Evolutionary Timescale for the Family of Mammalian Herpesviruses. *Journal of Molecular Biology*, 247(3):443–458.

[McGregor et al., 1996] McGregor, F., Phelan, A., Dunlop, J., and Clements, J. B. (1996). Regulation of herpes simplex virus poly (A) site usage and the action of immediate-early protein IE63 in the early-late switch. *Journal of virology*, 70(3):1931–1940.

[McLauchlan et al., 1992] McLauchlan, J., Phelan, A., Loney, C., Sandri-Goldin, R. M., and Clements, J. B. (1992). Herpes simplex virus IE63 acts at the posttranscriptional level to stimulate viral mRNA 3' processing. *J Virol*, 66(12):6939–6945.

[McLauchlan et al., 1989] McLauchlan, J., Simpson, S., and Clements, J. B. (1989). Herpes simplex virus induces a processing factor that stimulates poly(A) site usage. *Cell*, 59(6):1093–1105.

[Melé et al., 2015] Melé, M., Ferreira, P. G., Reverter, F., DeLuca, D. S., Monlong, J., Sammeth, M., Young, T. R., Goldmann, J. M., Pervouchine, D. D., Sullivan, T. J., Johnson, R., Segrè, A. V., Djebali, S., Niarchou, A., Consortium, T. G., Wright, F. A., Lappalainen, T., Calvo, M., Getz, G., Dermitzakis, E. T., Ardlie, K. G., and Guigó, R. (2015). The human transcriptome across tissues and individuals. *Science*, 348(6235):660 LP – 665.

[Melvin et al., 1978] Melvin, W. T., Milne, H. B., Slater, A. A., Allen, H. J., and Keir, H. M. (1978). Incorporation of 6-thioguanosine and 4-thiouridine into RNA. Application to isolation of newly synthesised RNA by affinity chromatography. *Eur.J Biochem.*, 92(2):373–379.

[Meyer and Jaffrey, 2014] Meyer, K. D. and Jaffrey, S. R. (2014). The dynamic epitranscriptome: N6-methyladenosine and gene expression control. *Nature reviews. Molecular cell biology*, 15(5):313–326.

[Mischo and Proudfoot, 2013] Mischo, H. E. and Proudfoot, N. J. (2013). Disengaging polymerase: Terminating RNA polymerase II transcription in budding yeast. *Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms*, 1829(1):174–185.

[Moldován et al., 2018] Moldován, N., Tombácz, D., Szűcs, A., Csabai, Z., Snyder, M., and Boldogkői, Z. (2018). Multi-Platform Sequencing Approach Reveals a Novel Transcriptome Profile in Pseudorabies Virus. *Frontiers in Microbiology*, 8:2708.

[Morris et al., 2008] Morris, S. R., Bauer, H. M., Samuel, M. C., Gallagher, D., and Bolan, G. (2008). Neonatal herpes morbidity and mortality in California, 1995-2003. *Sexually transmitted diseases*, 35(1):14–18.

[Murphy et al., 2003] Murphy, E., Rigoutsos, I., Shibuya, T., and Shenk, T. E. (2003). Reevaluation of human cytomegalovirus coding potential. *Proc Natl Acad Sci U S A*, 100(23):13585–13590.

[Ns and Mettenleiter, 1996] Ns, A. J. and Mettenleiter, T. C. (1996). Identification and Characterization of Pseudorabies Virus dUTPase. *JOURNAL OF VIROLOGY*, 70(2):1242–1245.

[Ogle and Roizman, 1999] Ogle, W. O. and Roizman, B. (1999). Functional anatomy of herpes simplex virus 1 overlapping genes encoding infected-cell protein 22 and US1.5 protein. *Journal of virology*, 73(5):4305–4315.

[Ogorodnikov et al., 2016] Ogorodnikov, A., Kargapolova, Y., and Danckwardt, S. (2016). Processing and transcriptome expansion at the mRNA 3' end in health and disease: finding the right end. *Pflugers Archiv : European journal of physiology*, 468(6):993–1012.

[Ohshima et al., 1981] Ohshima, Y., Okada, N., Tani, T., Itoh, Y., and Itoh, M. (1981). Nucleotide sequences of mouse genomic loci including a gene or pseudogene for U6 (4.8S) nuclear RNA. *Nucleic acids research*, 9(19):5145–5158.

[Pandya-Jones and Black, 2009] Pandya-Jones, A. and Black, D. L. (2009). Co-transcriptional splicing of constitutive and alternative exons. *RNA*, 15(10):1896–1908.

[Parida et al., 2019] Parida, M., Nilson, K. A., Li, M., Ball, C. B., Fuchs, H. A., Lawson, C. K., Luse, D. S., Meier, J. L., and Price, D. H. (2019). Nucleotide Resolution Comparison of Transcription of Human Cytomegalovirus and Host Genomes Reveals Universal Use of RNA Polymerase II Elongation Control Driven by Dissimilar Core Promoter Elements. *mBio*, 10(1).

[Parry et al., 2010] Parry, T. J., Theisen, J. W. M., Hsu, J.-Y., Wang, Y.-L., Corcoran, D. L., Eustice, M., Ohler, U., and Kadonaga, J. T. (2010). The TCT motif, a key component of an RNA polymerase II transcription system for the translational machinery. *Genes & development*, 24(18):2013–2018.

[Perera, 2000] Perera, L. P. (2000). The TATA Motif Specifies the Differential Activation of Minimal Promoters by Varicella Zoster Virus Immediate-early Regulatory Protein IE62*. *Journal of Biological Chemistry*, 275(1):487–496.

[Perng et al., 2002] Perng, G.-C., Maguen, B., Jin, L., Mott, K. R., Kurylo, J., BenMohamed, L., Yukht, A., Osorio, N., Nesburn, A. B., Henderson, G., Inman, M., Jones, C., and Wechsler, S. L. (2002). A novel herpes simplex virus type 1 transcript (AL-RNA) antisense to the 5' end of the latency-associated transcript produces a protein in infected rabbits. *Journal of virology*, 76(16):8003–8010.

[Philippe et al., 2020] Philippe, L., van den Elzen, A. M. G., Watson, M. J., and Thoreen, C. C. (2020). Global analysis of LARP1 translation targets reveals tunable and dynamic features of 5' TOP motifs. *Proceedings of the National Academy of Sciences*, 117(10):5319 LP – 5328.

[Pol et al., 2016] Pol, J., Kroemer, G., and Galluzzi, L. (2016). First oncolytic virus approved for melanoma immunotherapy. *OncoImmunology*, 5(1):e1115641.

[Policastro et al., 2020] Policastro, R. A., Raborn, R. T., Brendel, V. P., and Zentner, G. E. (2020). Simple and efficient profiling of transcription initiation and transcript levels with STRIPE-seq. *Genome Research*.

[Poon et al., 2006] Poon, A. P. W., Benetti, L., and Roizman, B. (2006). U(S)3 and U(S)3.5 protein kinases of herpes simplex virus 1 differ with respect to their functions in blocking apoptosis and in virion maturation and egress. *Journal of virology*, 80(8):3752–3764.

[Porrua et al., 2016] Porrua, O., Boudvillain, M., and Libri, D. (2016). Transcription Termination: Variations on Common Themes. *Trends in Genetics*, 32(8):508–522.

[Porrua and Libri, 2015] Porrua, O. and Libri, D. (2015). Transcription termination and the control of the transcriptome: why, where and how to stop. *Nat Rev Mol Cell Biol*, 16(3):190–202.

[Proudfoot, 2016] Proudfoot, N. J. (2016). Transcriptional termination in mammals: Stopping the RNA polymerase II juggernaut. *Science (New York, N.Y.)*, 352(6291):aad9926.

[Pultoo et al., 2000] Pultoo, A., Jankee, H., Meetoo, G., Pyndiah, M. N., and Khittoo, G. (2000). Detection of cytomegalovirus in urine of hearing-impaired and mentally retarded children by PCR and cell culture. *The Journal of communicable diseases*, 32(2):101–108.

[Quick et al., 2014] Quick, J., Quinlan, A. R., and Loman, N. J. (2014). A reference bacterial genome dataset generated on the MinION™ portable single-molecule nanopore sequencer. *GigaScience*, 3:22.

[Rabani et al., 2014] Rabani, M., Raychowdhury, R., Jovanovic, M., Rooney, M., Stumpo, D. J., Pauli, A., Hacohen, N., Schier, A. F., Blackshear, P. J., Friedman, N., Amit, I., and Regev, A. (2014). High-Resolution Sequencing and Modeling Identifies Distinct Dynamic RNA Regulatory Strategies. *Cell*, 159(7):1698–1710.

[Rajcani et al., 2004] Rajcani, J., Andrea, V., and Ingeborg, R. (2004). Peculiarities of herpes simplex virus (HSV) transcription: an overview. *Virus Genes*, 28(3):293–310.

[Randall et al., 1997] Randall, G., Lagunoff, M., and Roizman, B. (1997). The product of ORF O located within the domain of herpes simplex virus 1 genome transcribed during latent infection binds to and inhibits in vitro binding of infected cell protein 4 to its cognate DNA site. *Proc Natl Acad Sci U S A*, 94(19):10379–10384.

[Riml et al., 2017] Riml, C., Amort, T., Rieder, D., Gasser, C., Lusser, A., and Micura, R. (2017). Osmium-Mediated Transformation of 4-Thiouridine to Cytidine as Key To Study RNA Dynamics by Sequencing. *Angewandte Chemie International Edition*, 56(43):13479–13483.

[Ritchie et al., 2009] Ritchie, W., Flamant, S., and Rasko, J. E. J. (2009). Predicting microRNA targets and functions: traps for the unwary.

[Rosner et al., 2013] Rosner, M., Schipany, K., and Hengstschlager, M. (2013). Merging high-quality biochemical fractionation with a refined flow cytometry approach to monitor nucleocytoplasmic protein expression throughout the unperturbed mammalian cell cycle. *Nat Protoc*, 8(3):602–626.

[Rutkowski et al., 2015] Rutkowski, A. J., Erhard, F., L'Hernault, A., Bonfert, T., Schilhabel, M., Crump, C., Rosenstiel, P., Efstathiou, S., Zimmer, R., Friedel, C. C., and Dölken, L. (2015). Widespread disruption of host transcription termination in HSV-1 infection. *Nat Commun*, 6:7126.

[Sambucetti et al., 1989] Sambucetti, L. C., Cherrington, J. M., Wilkinson, G. W., and Mocarski, E. S. (1989). NF-kappa B activation of the cytomegalovirus enhancer is mediated by a viral transactivator and by T cell stimulation. *The EMBO journal*, 8(13):4251–4258.

[Sandbaumhüter et al., 2013] Sandbaumhüter, M., Döhner, K., Schipke, J., Binz, A., Pohlmann, A., Sodeik, B., and Bauerfeind, R. (2013). Cytosolic herpes simplex virus capsids not only require binding inner tegument protein pUL36 but also pUL37 for active transport prior to secondary envelopment. *Cellular Microbiology*, 15(2):248–269.

[Sandberg, 2014] Sandberg, R. (2014). Entering the era of single-cell transcriptomics in biology and medicine. *Nature Methods*, 11(1):22–24.

[Sandelin et al., 2007] Sandelin, A., Carninci, P., Lenhard, B., Ponjavic, J., Hayashizaki, Y., and Hume, D. A. (2007). Mammalian RNA polymerase II core promoters: insights from genome-wide studies. *Nature reviews. Genetics*, 8(6):424–436.

[Sandri-Goldin, 2007] Sandri-Goldin, R. M. (2007). *Initiation of transcription and RNA synthesis, processing and transport in HSV and VZV infected cells.*

[Sandri-Goldin, 2011] Sandri-Goldin, R. M. (2011). The many roles of the highly interactive HSV protein ICP27, a key regulator of infection. *Future Microbiol*, 6(11):1261–1277.

[Sanger et al., 1977] Sanger, F., Nicklen, S., and Coulson, A. R. (1977). DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences*, 74(12):5463 LP – 5467.

[Sarisky and Hayward, 1996] Sarisky, R. T. and Hayward, G. S. (1996). Evidence that the UL84 gene product of human cytomegalovirus is essential for promoting oriLyt-dependent DNA replication and formation of replication compartments in cotransfection assays. *Journal of virology*, 70(11):7398–7413.

[Saxonov et al., 2006] Saxonov, S., Berg, P., and Brutlag, D. L. (2006). A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters. *Proceedings of the National Academy of Sciences of the United States of America*, 103(5):1412–1417.

[Schmidt et al., 2002] Schmidt, S., Bork, P., and Dandekar, T. (2002). A Versatile Structural Domain Analysis Server Using Profile Weight Matrices. *Journal of Chemical Information and Computer Sciences*, 42(2):405–407.

[Schofield et al., 2018] Schofield, J. A., Duffy, E. E., Kiefer, L., Sullivan, M. C., and Simon, M. D. (2018). TimeLapse-seq: adding a temporal dimension to RNA sequencing through nucleoside recoding. *Nature methods*, 15(3):221–225.

[Schwalb et al., 2016] Schwalb, B., Michel, M., Zacher, B., Frühauf, K., Demel, C., Tresch, A., Gagneur, J., and Cramer, P. (2016). TT-seq maps the human transient transcriptome. *Science (New York, N.Y.)*, 352(6290):1225–1228.

[Schwanhausser et al., 2011] Schwanhausser, B., Busse, D., Li, N., Dittmar, G., Schuchhardt, J., Wolf, J., Chen, W., and Selbach, M. (2011). Global quantification of mammalian gene expression control. *Nature*, 473(7347):337–342.

[Sedlazeck et al., 2013] Sedlazeck, F. J., Rescheneder, P., and von Haeseler, A. (2013). NextGenMap: fast and accurate read mapping in highly polymorphic genomes. *Bioinformatics*, 29(21):2790–2791.

[Sekulovich et al., 1988] Sekulovich, R. E., Leary, K., and Sandri-Goldin, R. M. (1988). The herpes simplex virus type 1 alpha protein ICP27 can act as a trans-repressor or a trans-activator in combination with ICP4 and ICP0. *J Virol*, 62(12):4510–4522.

[Sessegolo et al., 2019] Sessegolo, C., Cruaud, C., Da Silva, C., Cologne, A., Dubarry, M., Derrien, T., Lacroix, V., and Aury, J.-M. (2019). Transcriptome profiling of mouse samples using nanopore sequencing of cDNA and RNA molecules. *Scientific Reports*, 9(1):14908.

[Sharma et al., 2010] Sharma, C. M., Hoffmann, S., Darfeuille, F., Reignier, J., Findeiß, S., Sittka, A., Chabas, S., Reiche, K., Hackermüller, J., Reinhardt, R., Stadler, P. F., and Vogel, J. (2010). The primary transcriptome of the major human pathogen Helicobacter pylori. *Nature*, 464(7286):250–255.

[Sharma and Vogel, 2014] Sharma, C. M. and Vogel, J. (2014). Differential RNA-seq: the approach behind and the biological insight gained. *Current Opinion in Microbiology*, 19:97–105.

[Shiraki et al., 2003] Shiraki, T., Kondo, S., Katayama, S., Waki, K., Kasukawa, T., Kawaji, H., Kodzius, R., Watahiki, A., Nakamura, M., Arakawa, T., Fukuda, S., Sasaki, D., Podhajska, A., Harbers, M., Kawai, J., Carninci, P., and Hayashizaki, Y. (2003). Cap

analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage. *Proceedings of the National Academy of Sciences*, 100(26):15776 LP – 15781.

[Sinclair, 2010] Sinclair, J. (2010). Chromatin structure regulates human cytomegalovirus gene expression during latency, reactivation and lytic infection. *Biochimica et biophysica acta*, 1799(3-4):286–295.

[Smale and Kadonaga, 2003] Smale, S. T. and Kadonaga, J. T. (2003). The RNA Polymerase II Core Promoter. *Annual Review of Biochemistry*, 72(1):449–479.

[Smith et al., 1992] Smith, I. L., Hardwicke, M. A., and Sandri-Goldin, R. M. (1992). Evidence that the herpes simplex virus immediate early protein ICP27 acts post-transcriptionally during infection to regulate gene expression. *Virology*, 186(1):74–86.

[Smith et al., 2017] Smith, T. S., Heger, A., and Sudbery, I. (2017). UMI-tools: Modelling sequencing errors in Unique Molecular Identifiers to improve quantification accuracy. *Genome Research*.

[Stagno et al., 1986] Stagno, S., Pass, R. F., Cloud, G., Britt, W. J., Henderson, R. E., Walton, P. D., Veren, D. A., Page, F., and Alford, C. A. (1986). Primary Cytomegalovirus Infection in Pregnancy: Incidence, Transmission to Fetus, and Clinical Outcome. *JAMA*, 256(14):1904–1908.

[Starck et al., 2016] Starck, S. R., Tsai, J. C., Chen, K., Shodiya, M., Wang, L., Yahiro, K., Martins-Green, M., Shastri, N., and Walter, P. (2016). Translation from the 5' untranslated region shapes the integrated stress response. *Science (New York, N.Y.)*, 351(6272):aad3867.

[Stern-Ginossar et al., 2012] Stern-Ginossar, N., Weisburd, B., Michalski, A., Vu, T. K. L., Hein, M. Y., Huang, S. X., Ma, M., Shen, B., Qian, S. B., Hengel, H., Mann, M., Ingolia, N. T., and Weissman, J. S. (2012). Decoding Human Cytomegalovirus. *Science (New York, N.Y.)*, 338(6110):1088–1093.

[Stevens et al., 1987] Stevens, J. G., Wagner, E. K., Devi-Rao, G. B., Cook, M. L., and Feldman, L. T. (1987). RNA complementary to a herpesvirus alpha gene mRNA is prominent in latently infected neurons. *Science (New York, N.Y.)*, 235(4792):1056–1059.

[Szklarczyk et al., 2011] Szklarczyk, D., Franceschini, A., Kuhn, M., Simonovic, M., Roth, A., Minguez, P., Doerks, T., Stark, M., Muller, J., Bork, P., Jensen, L. J., and von Mering, C. (2011). The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored. *Nucleic acids research*, 39(Database issue):D561–D568.

[Tang et al., 2016] Tang, S., Patel, A., and Krause, P. R. (2016). Herpes simplex virus ICP27 regulates alternative pre-mRNA polyadenylation and splicing in a sequence-dependent manner. *Proc Natl Acad Sci U S A*, 113(43):12256–12261.

[Tang et al., 2019] Tang, S., Patel, A., and Krause, P. R. (2019). Hidden regulation of herpes simplex virus 1 pre-mRNA splicing and polyadenylation by virally encoded immediate early gene ICP27. *PLOS Pathogens*, 15(6):e1007884.

[Terhune et al., 2004] Terhune, S. S., Schroer, J., and Shenk, T. (2004). RNAs Are Packaged into Human Cytomegalovirus Virions in Proportion to Their Intracellular Concentration. *Journal of Virology*, 78(19):10390–10398.

[Thodberg et al., 2019] Thodberg, M., Thieffry, A., Vitting-Seerup, K., Andersson, R., and Sandelin, A. (2019). CAGEfightR: analysis of 5'-end data using R/Bioconductor. *BMC Bioinformatics*, 20(1):487.

[Tischer et al., 2010] Tischer, B. K., Smith, G. A., Osterrieder, N., and Passant, E. (2010). Mutagenesis: A Two Step Markerless Red Recombination System. In *Methods in molecular biology (Clifton, N.J.)*, volume 634, pages 421–430.

[Tomasec et al., 2005] Tomasec, P., Wang, E. C. Y., Davison, A. J., Vojtesek, B., Armstrong, M., Griffin, C., McSharry, B. P., Morris, R. J., Llewellyn-Lacey, S., Rickards, C., Nomoto, A., Sinzger, C., and Wilkinson, G. W. G. (2005). Downregulation of natural killer cell–activating ligand CD155 by human cytomegalovirus UL141. *Nature Immunology*, 6(2):181–188.

[Tombacz et al., 2017] Tombacz, D., Csabai, Z., Szucs, A., Balazs, Z., Moldovan, N., Sharon, D., Snyder, M., and Boldogkoi, Z. (2017). Long-Read Isoform Sequencing Reveals a Hidden Complexity of the Transcriptional Landscape of Herpes Simplex Virus Type 1. *Front Microbiol*, 8:1079.

[Tonon et al., 2010] Tonon, L., Touzet, H., and Varré, J.-S. (2010). TFM-Explorer: mining cis-regulatory regions in genomes. *Nucleic acids research*, 38(Web Server issue):W286–W292.

[Tripp et al., 2011] Tripp, H. J., Hewson, I., Boyarsky, S., Stuart, J. M., and Zehr, J. P. (2011). Misannotations of rRNA can now generate 90% false positive protein matches in metatranscriptomic studies. *Nucleic acids research*, 39(20):8792–8802.

[Ussuf et al., 1995] Ussuf, K. K., Anikumar, G., and Nair, P. M. (1995). Newly synthesised mRNA as a probe for identification of wound responsive genes from potatoes. *Indian J Biochem.Biophys.*, 32(2):78–83.

[Uvarovskii and Dieterich, 2017] Uvarovskii, A. and Dieterich, C. (2017). pulseR: Versatile computational analysis of RNA turnover from metabolic labeling experiments. *Bioinformatics (Oxford, England)*, 33(20):3305–3307.

[Vattem and Wek, 2004] Vattem, K. M. and Wek, R. C. (2004). Reinitiation involving upstream ORFs regulates ATF4 mRNA translation in mammalian cells. *Proceedings of the National Academy of Sciences of the United States of America*, 101(31):11269–11274.

[Verbruggen et al., 2009] Verbruggen, N., Hermans, C., and Schat, H. (2009). Molecular mechanisms of metal hyperaccumulation in plants. *New Phytologist*, 181(4):759–776.

[Vo ngoc et al., 2017] Vo ngoc, L., Wang, Y.-L., Kassavetis, G. A., and Kadonaga, J. T. (2017). The punctilious RNA polymerase II core promoter. *Genes & Development*, 31(13):1289–1301.

[Wagih, 2017] Wagih, O. (2017). *ggseqlogo: A 'ggplot2' Extension for Drawing Publication-Ready Sequence Logos*.

[Wang et al., 2014] Wang, X., Lu, Z., Gomez, A., Hon, G. C., Yue, Y., Han, D., Fu, Y., Parisien, M., Dai, Q., Jia, G., Ren, B., Pan, T., and He, C. (2014). N6-methyladenosine-dependent regulation of messenger RNA stability. *Nature*, 505(7481):117–120.

[Wang et al., 2009] Wang, Z., Gerstein, M., and Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet*, 10(1):57–63.

[Waterhouse et al., 2018] Waterhouse, A., Bertoni, M., Bienert, S., Studer, G., Tauriello, G., Gumienny, R., Heer, F. T., de Beer, T. A. P., Rempfer, C., Bordoli, L., Lepore, R., and Schwede, T. (2018). SWISS-MODEL: homology modelling of protein structures and complexes. *Nucleic acids research*, 46(W1):W296–W303.

[Weekes et al., 2014] Weekes, M. P., Tomasec, P., Huttlin, E. L., Fielding, C. A., Nusinow, D., Stanton, R. J., Wang, E. C., Aicheler, R., Murrell, I., Wilkinson, G. W., Lehner, P. J., and Gygi, S. P. (2014). Quantitative temporal viromics: an approach to investigate host-pathogen interaction. *Cell*, 157(6):1460–1472.

[Wenger et al., 2019] Wenger, A. M., Peluso, P., Rowell, W. J., Chang, P.-C., Hall, R. J., Concepcion, G. T., Ebler, J., Fungtammasan, A., Kolesnikov, A., Olson, N. D., Töpfer, A., Alonge, M., Mahmoud, M., Qian, Y., Chin, C.-S., Phillippy, A. M., Schatz, M. C., Myers, G., DePristo, M. A., Ruan, J., Marschall, T., Sedlazeck, F. J., Zook, J. M., Li, H., Koren, S., Carroll, A., Rank, D. R., and Hunkapiller, M. W. (2019). Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. *Nature Biotechnology*, 37(10):1155–1162.

[Whisnant et al., 2020] Whisnant, A. W., Jürges, C. S., Hennig, T., Wyler, E., Prusty, B., Rutkowski, A. J., L'hernault, A., Djakovic, L., Göbel, M., Döring, K., Menegatti, J., Antrobus, R., Matheson, N. J., Künzig, F. W. H., Mastrobuoni, G., Bielow, C., Kempa, S., Liang, C., Dandekar, T., Zimmer, R., Landthaler, M., Grässer, F., Lehner, P. J., Friedel, C. C., Erhard, F., and Dölken, L. (2020). Integrative functional genomics decodes herpes simplex virus 1. *Nature Communications*, 11(1):2038.

[Williamson et al., 1990] Williamson, W. D., Percy, A. K., Yow, M. D., Gerson, P., Catlin, F. I., Koppelman, M. L., and Thurber, S. (1990). Asymptomatic congenital cytomegalovirus infection. Audiologic, neuroradiologic, and neurodevelopmental abnormalities during the first year. *American journal of diseases of children (1960)*, 144(12):1365–1368.

[Windhager et al., 2012] Windhager, L., Bonfert, T., Burger, K., Ruzsics, Z., Krebs, S., Kaufmann, S., Malterer, G., L'Hernault, A., Schilhabel, M., Schreiber, S., Rosenstiel, P., Zimmer, R., Eick, D., Friedel, C. C., and Dölken, L. (2012). Ultrashort and progressive 4sU-tagging reveals key characteristics of RNA processing at nucleotide resolution. *Genome Res*, 22(10):2031–2042.

[Wise et al., 2021] Wise, L. M., Xi, Y., and Purdy, J. G. (2021). Hypoxia-Inducible Factor 1α (HIF1α) Suppresses Virus Replication in Human Cytomegalovirus Infection by Limiting Kynurenine Synthesis. *mBio*, 12(2).

[Wohlrab et al., 1982] Wohlrab, F., Garrett, B. K., and Francke, B. (1982). Control of Expression of the Herpes Simplex Virus-Induced Deoxypyrimidine Triphosphatase in Cells Infected with Mutants of Herpes Simplex Virus Types 1 and 2 and Intertypic Recombinants. *JOURNAL OF VIROLOGY*, 43(3):935–942.

[Woodford et al., 1988] Woodford, T. A., Schlegel, R., and Pardee, A. B. (1988). Selective isolation of newly synthesized mammalian mRNA after in vivo labeling with 4-thiouridine or 6-thioguanosine. *Anal.Biochem.*, 171(1):166–172.

[Wu et al., 2008] Wu, H.-J., Wang, A. H.-J., and Jennings, M. P. (2008). Discovery of virulence factors of pathogenic bacteria. *Current Opinion in Chemical Biology*, 12(1):93–101.

[Wu and Watanabe, 2005] Wu, T. D. and Watanabe, C. K. (2005). GMAP: a genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics*, 21(9):1859–1875.

[Yates et al., 2015] Yates, A., Akanni, W., Amode, M. R., Barrell, D., Billis, K., Carvalho-Silva, D., Cummins, C., Clapham, P., Fitzgerald, S., Gil, L., Girón, C. G., Gordon, L., Hourlier, T., Hunt, S. E., Janacek, S. H., Johnson, N., Juettemann, T., Keenan, S., Lavidas, I., Martin, F. J., Maurel, T., McLaren, W., Murphy, D. N., Nag, R., Nuhn, M., Parker, A., Patricio, M., Pignatelli, M., Rahtz, M., Riat, H. S., Sheppard, D., Taylor, K., Thormann, A., Vullo, A., Wilder, S. P., Zadissa, A., Birney, E., Harrow, J., Muffato, M., Perry, E., Ruffier, M., Spudich, G., Trevanion, S. J., Cunningham, F., Aken, B. L., Zerbino, D. R., and Flicek, P. (2015). Ensembl 2016. *Nucleic Acids Research*, 44(D1):D710–D716.

[Yinon et al., 2010] Yinon, Y., Farine, D., Yudin, M. H., Gagnon, R., Hudon, L., Basso, M., Bos, H., Delisle, M. F., Farine, D., Menticoglou, S., Mundle, W., Ouellet, A.,

Pressey, T., Roggensack, A., Yudin, M. H., Boucher, M., Castillo, E., Gruslin, A., Money, D. M., Murphy, K., Ogilvie, G., Paquet, C., Van Eyk, N., and van Schalkwyk, J. (2010). Cytomegalovirus Infection in Pregnancy. *Journal of Obstetrics and Gynaecology Canada*, 32(4):348–354.

[Young and Wek, 2016] Young, S. K. and Wek, R. C. (2016). Upstream Open Reading Frames Differentially Regulate Gene-specific Translation in the Integrated Stress Response. *J Biol Chem*, 291(33):16927–16935.

[Yue et al., 2015] Yue, Y., Liu, J., and He, C. (2015). RNA N6-methyladenosine methylation in post-transcriptional gene expression regulation. *Genes & development*, 29(13):1343–1355.

[Železnjak et al., 2019] Železnjak, J., Lisnić, V. J., Popović, B., Lisnić, B., Babić, M., Halenius, A., L'Hernault, A., Roviš, T. L., Hengel, H., Erhard, F., Redwood, A. J., Vidal, S. M., Dölken, L., Krmpotić, A., and Jonjić, S. (2019). The complex of MCMV proteins and MHC class I evades NK cell control and drives the evolution of virus-specific activating Ly49 receptors. *The Journal of experimental medicine*, 216(8):1809–1827.