Decoding murine cytomegalovirus

Lodha, Manivel; Muchsin, Ihsan; Jürges, Christopher; Juranić Lisnić, Vanda; L'Hernault, Anne; Rutkowski, Andrzej J.; Prusty, Bhupesh K.; Grothey, Arnhild; Milić, Andrea; Hennig, Thomas; ...

Source / Izvornik: **PLOS Pathogens, 2023, 19**

Journal article, Published version
Rad u časopisu, Objavljena verzija rada (izdavačev PDF)

https://doi.org/10.1371/journal.ppat.1010992

Permanent link / Trajna poveznica: https://urn.nsk.hr/urn:nbn:hr:184:762028

Rights / Prava: Attribution 4.0 International/Imenovanje 4.0 međunarodna

Download date / Datum preuzimanja: **2024-10-03**

Repository / Repozitorij:

Repository of the University of Rijeka, Faculty of Medicine - FMRI Repository

UNIRI DIGITALNA KNJIŽNICA

dabar
DIGITALNI AKADEMSKI ARHIVI I REPOZITORIJI

PLOS PATHOGENS

RESEARCH ARTICLE

Decoding murine cytomegalovirus

Manivel Lodha[1☯], Ihsan Muchsin[1☯], Christopher Jürges[1], Vanda Juranic Lisnic[2,3], Anne L'Hernault[4], Andrzej J. Rutkowski[4], Bhupesh K. Prusty[1], Arnhild Grothey[1], Andrea Milic[1], Thomas Hennig[1], Stipan Jonjic[2,3], Caroline C. Friedel[5], Florian Erhard[1]*, Lars Dölken[1,4,6]*

1 Institute for Virology and Immunobiology, Julius-Maximilians-Universität-Würzburg, Würzburg, Germany, **2** Department of Histology and Embryology, Faculty of Medicine, University of Rijeka, Rijeka, Croatia, **3** Center for Proteomics, Faculty of Medicine, University of Rijeka, Rijeka, Croatia, **4** Department of Medicine, University of Cambridge, Addenbrookes Hospital, Cambridge, United Kingdom, **5** Institute of Informatics, Ludwig-Maximilians-Universität München, Munich, Germany, **6** Helmholtz Institute for RNA-based Infection Research (HIRI), Würzburg, Germany

☯ These authors contributed equally to this work.
* florian.erhard@uni-wuerzburg.de (FE); lars.doelken@uni-wuerzburg.de (LD)

Check for updates

OPEN ACCESS

Citation: Lodha M, Muchsin I, Jürges C, Juranic Lisnic V, L'Hernault A, Rutkowski AJ, et al. (2023) Decoding murine cytomegalovirus. PLoS Pathog 19(5): e1010992. https://doi.org/10.1371/journal.ppat.1010992

Editor: Edward S. Mocarski, Emory University, UNITED STATES

Received: November 9, 2022

Accepted: March 17, 2023

Published: May 12, 2023

Copyright: © 2023 Lodha et al. This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data Availability Statement: The gedi toolkit, which was used for mapping and most of the analysis steps, is available on GitHub (https://github.com/erhard-lab/gedi). iTiSS, which is a module for gedi is available separately on GitHub (https://github.com/erhard-lab/iTiSS). The source code of all additional custom scripts generated for generating the figures, tables and analyzing the data in general can be found at Zenodo (https://doi.org/10.5281/zenodo.6861955). A genome browser including all data is available at https://doi.org/10.5281/zenodo.7105431. All sequencing data produced in this study are available at GEO

Abstract

The genomes of both human cytomegalovirus (HCMV) and murine cytomegalovirus (MCMV) were first sequenced over 20 years ago. Similar to HCMV, the MCMV genome had initially been proposed to harbor ≈170 open reading frames (ORFs). More recently, omics approaches revealed HCMV gene expression to be substantially more complex comprising several hundred viral ORFs. Here, we provide a state-of-the art reannotation of lytic MCMV gene expression based on integrative analysis of a large set of omics data. Our data reveal 365 viral transcription start sites (TiSS) that give rise to 380 and 454 viral transcripts and ORFs, respectively. The latter include >200 small ORFs, some of which represented the most highly expressed viral gene products. By combining TiSS profiling with metabolic RNA labelling and chemical nucleotide conversion sequencing (dSLAM-seq), we provide a detailed picture of the expression kinetics of viral transcription. This not only resulted in the identification of a novel MCMV immediate early transcript encoding the m166.5 ORF, which we termed *ie4*, but also revealed a group of well-expressed viral transcripts that are induced later than canonical true late genes and contain an initiator element (Inr) but no TATA- or TATT-box in their core promoters. We show that viral upstream ORFs (uORFs) tune gene expression of longer viral ORFs expressed in *cis* at translational level. Finally, we identify a truncated isoform of the viral NK-cell immune evasin m145 arising from a viral TiSS downstream of the canonical m145 mRNA. Despite being ≈5-fold more abundantly expressed than the canonical m145 protein it was not required for downregulating the NK cell ligand, MULT-I. In summary, our work will pave the way for future mechanistic studies on previously unknown cytomegalovirus gene products in an important virus animal model.

Author summary

We conducted a comprehensive characterization and reannotation of murine cytomegalovirus (MCMV) gene expression during lytic infection in murine fibroblasts using an

(accession number GSE212289). Third party annotations of our MCMV genome are available on NCBI under the accessions BK063393 and BK063394.

Funding: This work was supported by a grant from the Deutsche Forschungsgemeinschaft (FOR 2830, DO 1275/7-1 and ER 927/1-1 to LD and FE, respectively) and in the framework of the Research Unit FOR5200 DEEP-DV (443644894) project FR 2938/11-1 to CCF. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

integrative multi-omics approach. This unveiled hundreds of novel transcripts that explained the expression of close to 300 so far unknown viral open reading frames (ORFs). Interestingly, small viral ORFs (sORFs) were amongst the most highly expressed viral gene products and thus presumably encode for important viral microproteins of unknown function. We show that sORFs located upstream of larger ORFs tune the expression of the downstream ORFs by repressing their translation. We classified viral transcription start sites (TiSS) based on their expression kinetics obtained by the combination of metabolic RNA labelling with transcription start sites profiling. This not only identified a so far unknown viral immediate-early transcript (*ie4*, m166.5 RNA) but also revealed a novel class of viral late transcripts that are expressed later than canonical true late genes and lack TATA box-like motifs. We exemplify for the m145 locus how so far unknown viral TiSS give rise to abundantly expressed truncated viral proteins. In summary, we provide a state-of-the-art annotation of an important model virus, which will be instrumental for future studies on CMV biology, immunology and pathogenesis.

Introduction

Human cytomegalovirus (HCMV) is a ubiquitous pathogen that establishes a life-long infection upon primary infection [1]. While primary infection is mostly asymptomatic, HCMV is responsible for a significant morbidity and mortality in immunocompromised patients and neonates. There is currently no vaccine. The strict species specificity of HCMV poses a major challenge in understanding cytomegalovirus (CMV) pathogenesis [2]. Murine cytomegalovirus (MCMV) exhibits significant similarity to HCMV and represents a widely used model to study CMV pathogenesis [2,3]. Traditionally, CMV gene expression is temporally regulated and classified into immediate early (IE), early (E) and late (L) gene expression [4]. In contrast to viral *ie* gene expression (α genes), the expression of E genes (β genes) requires *de novo* expression of the major viral transcription factor IE3 and thus viral protein synthesis [5]. Viral L gene expression depends on viral DNA replication as well as expression of the viral late gene transcription factor (LTF) complex that binds to a TATA-like (TATT) motif in the proximal promoters of viral late genes [6–10]. L genes were further sub-divided into leaky late (γ1) and true late (γ2) genes based on their differential sensitivity to DNA synthesis inhibitors [11]. Moreover, recent temporal classification of HCMV [12] and CCMV [13] described at least 5 distinct kinetic clusters of viral protein expression further complicating the landscape of CMV gene expression.

In recent years, high-throughput sequencing technologies, including ribosome profiling (Ribo-seq) [14] and RNA-seq [15] reshaped our understanding of the coding capacity of herpesviruses including HCMV [16], HSV-1 [17], KSHV [18] and EBV [19]. Strikingly, these studies revealed the presence of hundreds of novel viral open reading frames (ORFs). These arise from promiscuous transcription initiation within the viral genome. Many of these novel viral ORFs are small ORFs (sORFs) of <100 amino acids (aa) in size. They may not contribute to the stable viral proteome but rather encode for cryptic unstable microproteins of unknown significance [17,20]. Depending on their genome location with respect to the larger viral ORFs, they are referred to as upstream ORFs (uORFs), upstream overlapping ORFs (uoORFs), internal ORFs (iORFs), or downstream ORFs (dORFs) [17,21].

The 230-kb genome of the MCMV Smith strain was initially predicted to encode 170 protein coding sequences (CDS), many of which share homology to HCMV [22]. To date, a state-of-the art reannotation of the MCMV genome including mRNAs, short ORFs and isoforms of

canonical ORFs as well as an overarching hierarchical nomenclature has been lacking. Nevertheless, additional viral gene products have been discovered through various genetic [23,24], *in silico* [25] and proteomic approaches [20,26]. This also includes the identification of different MCMV protein isoforms, which arise from alternative viral transcripts expressed with distinct kinetics [27]. A prominent example for the need for a comprehensive annotation of the MCMV genome was the identification of the 83-amino acid microprotein MATp1 [28]. MATp1 is expressed from the most abundant MCMV transcript (MAT) upstream of the coding sequence (CDS) of the spliced m169 gene [29]. Initially dismissed by *in silico* predictions due to its small size (≈83 aa), MATp1 acts in concert with the viral m04 protein and specific MHC-I allotypes in a trimeric complex to evade missing-self recognition by natural killer (NK) cells [28]. Furthermore, recognition of this trimeric complex by at least three activating NK-cell receptors explains intrinsic resistance of certain mouse strains to MCMV infection [28]. These findings highlight the importance of studying gene expression at single nucleotide resolution using unbiased, integrative multi-omics approaches to fully understand the coding potential of MCMV. Finally, the wealth of novel viral gene products requires a revised nomenclature.

We recently utilized a multi-omics approach coupled with integrative computational analysis to decipher the transcriptome and translatome of herpes simplex virus 1 (HSV-1) [17]. Here, we use a similar approach to comprehensively identify, characterize and hierarchically annotate MCMV gene products expressed during lytic infection of murine NIH-3T3 fibroblasts (**Fig 1**). Our new annotation comprises 365 viral transcription start sites (TiSS) that give rise to 380 and 454 viral transcripts and ORFs, respectively. TiSS profiling combined with metabolic RNA labelling and chemical nucleotide conversion sequencing (dSLAM-seq) resolved the kinetics of viral gene expression and their regulation by core promoter motifs. Abundant transcription initiation and alternative TiSS usage throughout lytic infection explained the expression of hundreds of novel viral ORFs and small ORFs, as well as N-terminal extensions (NTE) and truncations (NTT) thereof revealed by ribosome profiling. We employed the same nomenclature strategy as employed for HSV-1 to annotate novel MCMV transcripts and ORFs. In brief, transcription initiating ≥500 nt distant from another transcript was given a new identifier, starting with '.5' to provide room for future additional ORFs in case any TiSS or ORFs was missed. Transcripts arising from alternative TiSS located within <500 nt upstream or downstream of the main (canonical) transcript in a given locus were labelled as '*1', '*2',. . . and '#1', '#2',. . ., respectively. Previously identified protein coding sequences (CDS) of the Rawlinson reference annotation [22] were annotated as 'CDS', e.g., m04 CDS. All large novel ORFs were annotated as 'ORFs'. Small ORFs were annotated as 'uORF', 'uoORF', 'iORF', 'dORF' or 'sORF' depending on their relative location to their respective CDS or ORFs. N-terminal extensions ('NTEs') and truncations ('NTTs') of viral ORFs were annotated with '*1', '#2',. . . and '#1', '#2',. . ., respectively. Alternative spliced products were labelled as isoforms ('Iso1', 'Iso2'. . .). Transcripts and ORFs for which no corresponding ORF or TiSS, respectively, could be identified were labelled as 'orphan'. The fully reannotated MCMV genome was deposited to the NCBI GenBank Third Party Annotation database, with and without the BAC sequence under accessions BK063393 and BK063394 respectively. In summary, our work provides a state-of-the-art annotation of an important virus model.

Results

Characterization of the MCMV transcriptome

To identify the full complement of MCMV transcripts in lytic infection of fibroblasts, we profiled viral gene expression in MCMV-infected NIH-3T3 fibroblasts throughout the first three days of infection using multiple next-generation sequencing approaches (**Fig 1**). This included:

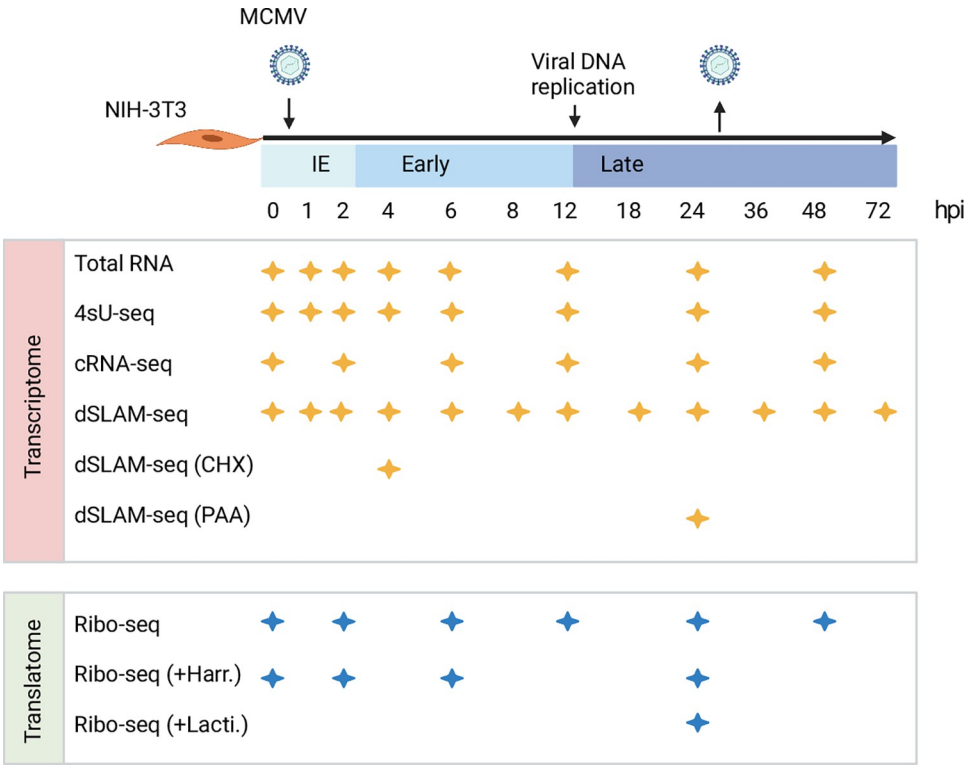

Fig 1. Overview of applied omics approaches. MCMV gene expression was analyzed in Swiss murine embryonic fibroblasts (NIH-3T3) infected with BAC-derived wild-type MCMV at an MOI of 10. Viral transcription start sites (TiSS) and splicing events were determined through total RNA-seq, 4sU-seq, cRNA-seq and dSLAM-seq (n = 2; including one biological replicate for dSLAM-seq with cycloheximide (CHX; 4 h) or phosphonoacetic acid (PAA; 24 h) treatment). To decipher the MCMV translatome, four biological replicates of ribosome profiling were performed. Enrichment of reads at translation start sites (TaSS) was improved by pre-treating cells with Harringtonine–Harr. (two biological replicates) or Lactimidomycin–Lacti. (one biological replicate) for 30 min. The available time points and conditions are indicated by stars for any given approach.

https://doi.org/10.1371/journal.ppat.1010992.g001

(***i***) RNA-seq of total RNA (Total RNA-seq) and (***ii***) newly transcribed RNA obtained by metabolic RNA labelling using 4-thiouridine (4sU-seq) [30] from the same samples. To analyze temporally resolved promoter usage, we performed transcription start sites (TiSS) profiling by (***iii***) cRNA-seq [16,17] as well as (***iv***) dSLAM-seq, a novel combination of differential RNA-seq (dRNA-seq) [31] with metabolic RNA labelling and thiol(SH)-linked alkylation of RNA (SLAM-seq) [32]. A representative example of the obtained data are shown for the M25 locus in **Fig 2A**. cRNA-seq is a modified total RNA sequencing protocol that is based on circularization of RNA fragments (hence termed cRNA-seq) [17]. It allows both TiSS identification based on a moderate enrichment (median: 8-fold) of reads starting at 5' RNA ends (**Fig 2B**) and quantification of total transcript levels. In contrast, dSLAM-seq provides a much greater enrichment of TiSS (median: 24-fold) by selectively enriching reads at 5' ends of cap-protected RNA fragments resistant to 5'-3' Xrn1 exonuclease digest (**Figs 2B and S1A**). Importantly, dSLAM-seq combines 1 h 4sU labelling immediately prior to cell lysis, followed by RNA isolation and chemical conversion of the introduced 4sU residues to a cytosine analogue (SLAM-seq). The latter facilitates computational identification of sequencing reads derived from newly transcribed RNA ('new RNA') based on the introduced U-to-C conversions [33]. Selective analysis of new RNA in dSLAM-seq data thus reveals the true temporal kinetics of gene expression for each viral TiSS. In addition, we also included dSLAM-seq samples pre-treated with

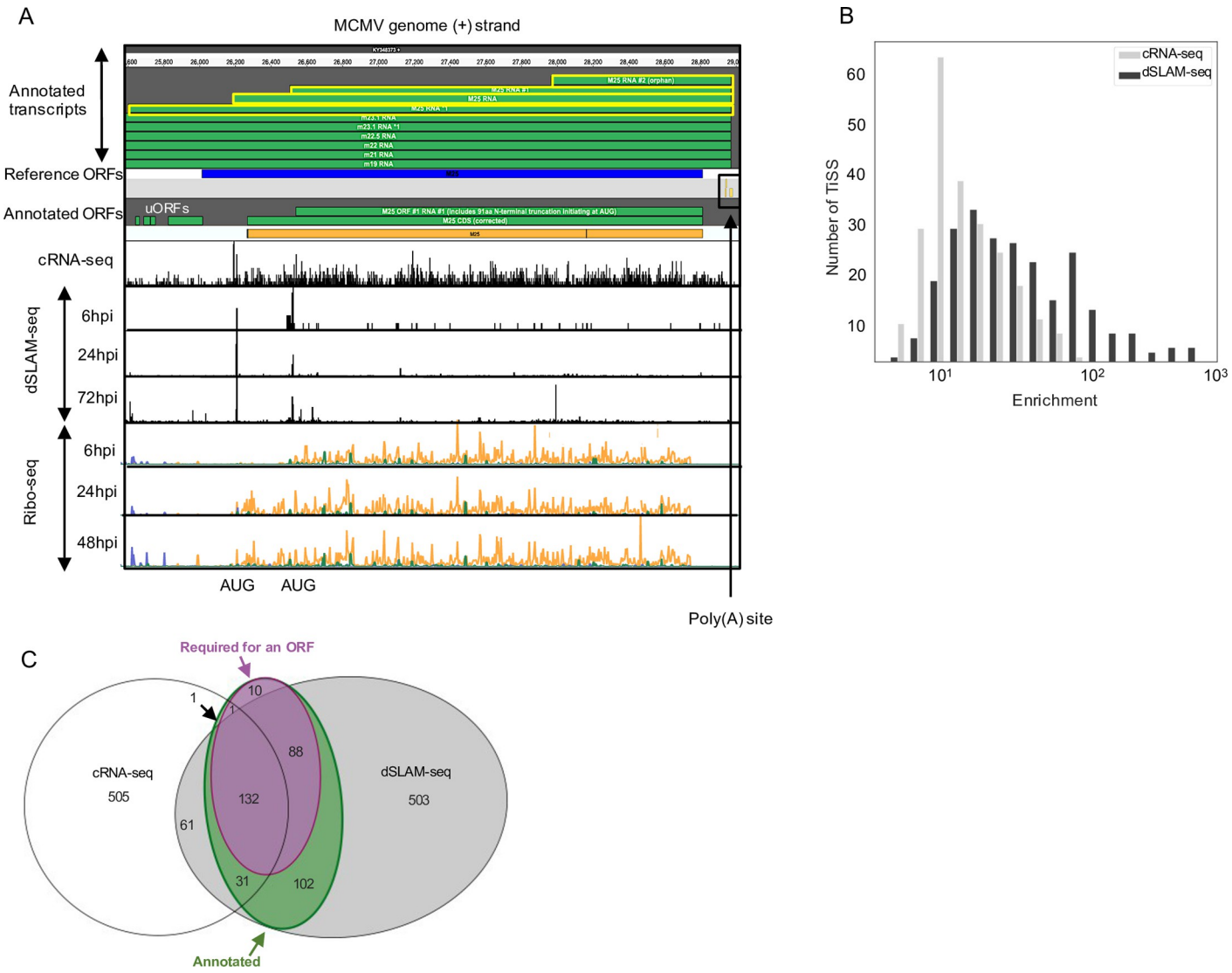

Fig 2. Characterization of the MCMV transcriptome. A. Screenshot of MCMV gene expression showing annotated transcripts and ORFs in the M25 locus with 5' read enrichment at TiSS as depicted by cRNA-seq and dSLAM-seq as well as Ribo-seq data, respectively. Four viral transcripts, which initiate within the M25 region of the MCMV genome, are highlighted in yellow (TiSS scores in S4 Table). The schematic depicts translation of the 130 and 105 kDa M25 protein isoforms validated in a recent study [27] and validated by our Ribo-seq data. The M25 RNA *1 also encodes four small ORFs (M25 uORFs 1–3 and M25 uoORF) of 6, 11, 8 and 63 aa, respectively whose expression levels and kinetics (6 hpi) correspond to their respective transcript (M25 RNA *1) and were hence annotated. dSLAM-seq data are depicted in linear scale, Ribo-seq data in logarithmic scale. **B.** Graphical representation of 5' read enrichment obtained by dSLAM-seq and cRNA-seq approaches. **C.** Venn diagram depicting the number of TiSS identified by both cRNA-seq and dSLAM-seq. TiSS included in the final annotation are depicted in the green circle as 'annotated'. TiSS labelled as 'Required for an ORF' represent TiSS that are required to explain the translation of a downstream ORF (no other TiSS within 500 nt upstream of the ORF).

https://doi.org/10.1371/journal.ppat.1010992.g002

chemical inhibitors of protein synthesis and viral DNA replication, namely cycloheximide (CHX; 4 hours post infection (hpi)) and phosphonoacetic acid (PAA; 24 hpi), respectively. A detailed overview of the analyzed time points and conditions is shown in **Fig 1.**

Reliable identification of viral TiSS requires integrative analysis of multiple data sets from different TiSS profiling approaches and kinetic studies [17]. We thus employed our recently published integrative TiSS analysis pipeline iTiSS [34], which identifies statistically significant peaks arising from TiSS profiling read accumulations across the genome. We furthermore scored these TiSS candidates according to a variety of additional criteria, including an increase

in upstream to downstream read coverage in cRNA-seq and 4sU-seq data, temporal changes in cRNA-seq and dSLAM-seq read counts and the presence of translated ORFs identified by Ribo-seq, for which no other transcript could otherwise be identified. This resulted in a maximum score of 7 for any given candidate TiSS. We then manually inspected all candidate TiSS using our in-house MCMV genome browser, which combines all data sets, time points and conditions (**Fig 2A**). In total, we identified and annotated 365 unique MCMV TiSS (**Fig 2C**), satisfying the given set of criteria (**S1B Fig**). Some TiSS were common for alternatively spliced products and differential poly(A) site usage resulting in a total of 380 MCMV transcripts. The complete list of all MCMV transcripts and the scores of their respective TiSS are included in **S1 Table**.

We next analyzed splicing events in the MCMV transcriptome based on our total RNA-seq and 4sU-seq data. We first identified all unique reads spanning exon-exon junctions by at least 10 nt (see **Methods** for details). We identified 366 splicing events, most of which only occurred at very low levels (**Fig 3**) with minimal exon-spanning translational activity. We thus decided not to include them into our new reference annotation and only retained 28 splicing events. Six of these splicing events had already been reported by Lisnic *et al*. [29] and several of these had been successfully validated using RT-PCR and 3' sequencing in the same study as well as other studies (**S2 Table**). To independently validate the identified splice sites and investigate the impact of the corresponding transcript isoforms on translation, we utilized our Ribo-seq data. We confirmed an alternative splice donor site in the m133 locus as suggested by Rawlinson *et al*; [22] leading to the expression of two protein isoforms from differentially spliced ORFs (**S2A Fig**). Splicing of both the most abundant transcript (MAT) within the m169 locus [28] and of a highly expressed transcript in the M116 locus were readily confirmed in our data (**S2B Fig**), the latter readily explained a recently validated spliced protein, M116.1p, which was found to be crucial for efficient infection of mononuclear phagocytes [35]. We also confirmed a previously reported spliced ORF in the m147.5 locus [36] (**S2C Fig**) along with a novel splicing event in the m124 locus, leading to a correction of the previously annotated m124 ORF (**S2D Fig**) [22]. While we readily observed the well-described MCMV 7.2 kb intron [37,38], we were unable to detect the overlapping 8 kb intron reported in the same study [37]. 4sU-seq data also revealed multiple alternative donor sites in the m60-m73.5 locus (**S3 Fig**), which expressed several weakly expressed isoforms of the m73.5 ORF, the most dominant being the M73-m73.5 spliced ORF. Our data demonstrate that splicing in the MCMV transcriptome is much more prevalent than previously thought but mostly comprises low level splicing events in addition to the previously described splicing events. A complete list of annotated splicing events, which we included into our new reference annotation of the MCMV genome, is included in **S2 Table** and a list of all 366 putative 4sU-seq based introns are included in **S3 Table**.

Temporal regulation of viral transcription

Metabolic RNA labelling and chemical nucleotide conversion combined with dSLAM-seq enabled us to analyze real-time transcriptional activity of each individual viral TiSS in 'new RNA' throughout the course of lytic infection. Utilizing maximal new RNA levels throughout infection for each TiSS obtained from our dSLAM-seq data, we grouped transcripts according to levels of gene expression (high, mid and low transcription). Many core promoters of eukaryotic genes contain TATA boxes [39], which are also prevalent in herpesvirus genomes [17]. T/A rich regions indicative of TATA box-like motifs (TBM) were much more prevalent in highly transcribed viral genes than in lowly transcribed viral genes (**Fig 4A**). In mammalian cells, TiSS are marked by an initiator element (Inr), characterized by a pyrimidine-purine dinucleotide [40]. As previously observed for HSV-1 [17], Inr elements were also prevalent for MCMV

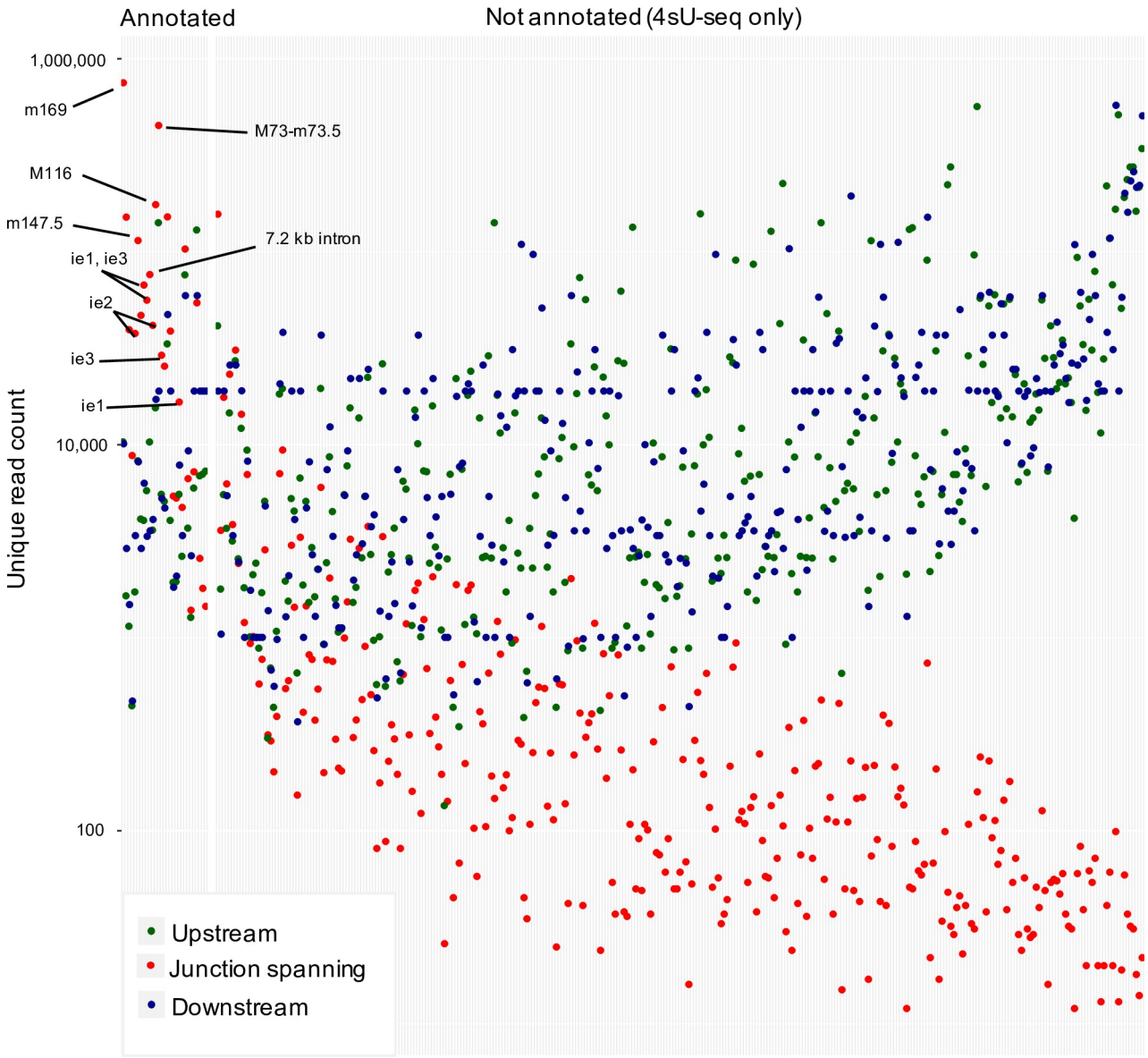

Fig 3. Identification of MCMV splicing events. Mapped reads from 4sU-seq and total RNA-seq identified 366 putative splicing events in the MCMV transcriptome. The y-axis displays the number of reads occurring at a spliced region, further categorized into reads spanning exon-exon junctions (red) by at least 10 nt as well as non-exon-spanning reads upstream (green) and downstream (blue). Putative splicing events were sorted based on the ratio of spliced (red) to unspliced (green + blue) reads. Only 28 of the 366 putative splicing events were included into our new reference annotation because they (***i***) had already been identified by others (16/28; S2 Table), (***ii***) were highly abundant, or (***iii***) affected the coding sequence of an MCMV ORF or sORF. To avoid unnecessary complexity in the revised annotation of the MCMV transcriptome, we excluded the other (putative) splicing events from our new reference annotation.

https://doi.org/10.1371/journal.ppat.1010992.g003

TiSS irrespective of their expression levels. This confirmed reliable identification of TiSS even for the most weakly utilized viral TiSS.

MCMV immediate early genes (*ie1*, *ie2* and *ie3*) are expressed within the first hour of infection and do not require viral protein synthesis and are thus resistant to inhibition of protein synthesis by cycloheximide (CHX). To identify novel MCMV immediate early genes, our dSLAM-seq experiment included a single replicate of 4 h of CHX treatment, which was initiated at the time of infection. Interestingly, CHX treatment not only confirmed the two

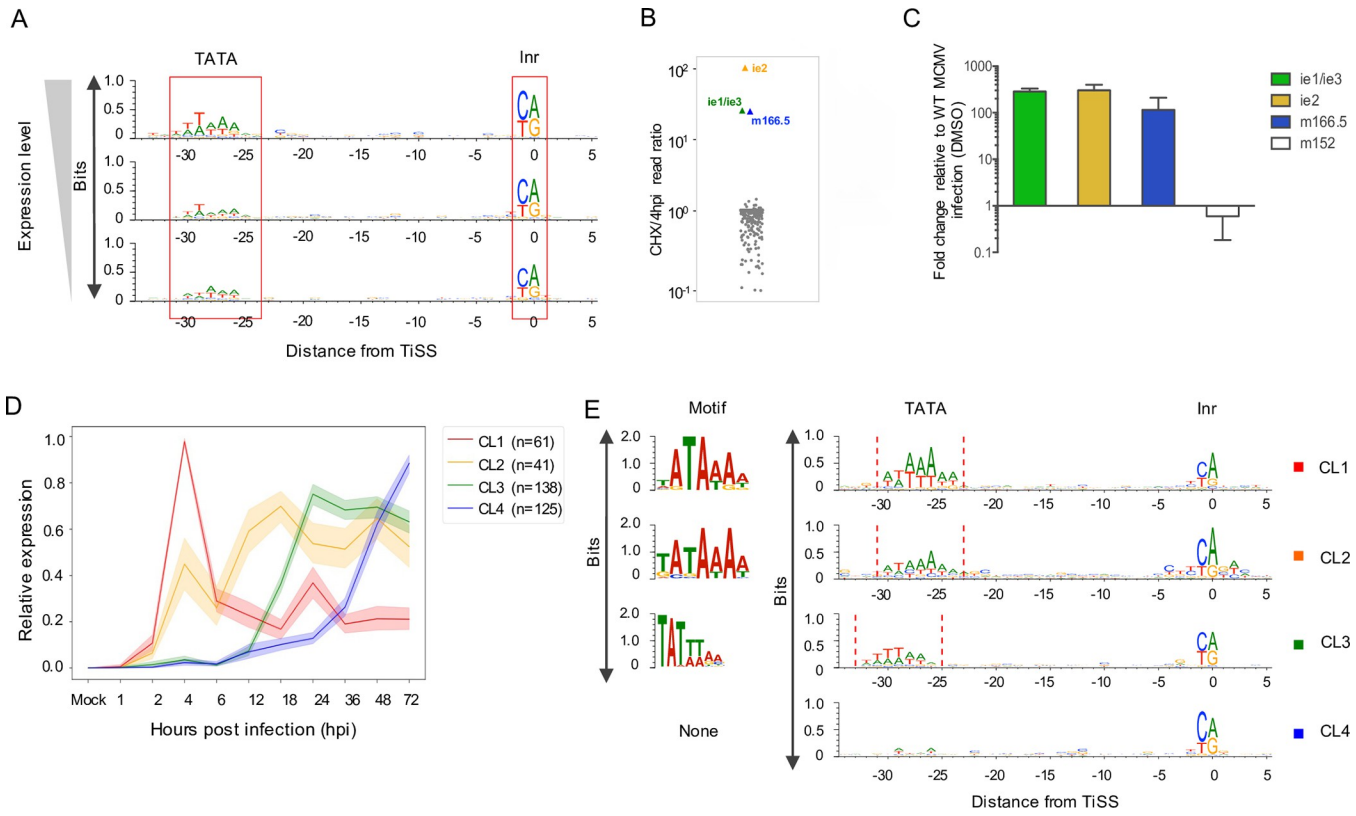

Fig 4. dSLAM-seq reveals distinct core promoter motifs associated with viral gene expression kinetics and a novel viral *ie* gene (*ie4*). A. Depiction of core promoter motifs of viral TiSS clustered according to their maximal transcription rates (new RNA derived from the dSLAM-seq data) in three equally sized bins (high, mid and low). The TATA box and initiator element (Inr) are shown. **B.** Ratio of new RNA levels at 4 hpi with and without cycloheximide (CHX) treatment (n = 1) are shown for the 365 MCMV TiSS (grey dots). This identified three immediate early TiSS, namely *ie1/ie3*, *ie2* and *ie4* (m166.5), highlighted in colored dots. **C.** Validation of the m166.5 RNA (*ie4)* as a so far unknown MCMV immediate early gene by qRT-PCR. qRT-PCR was performed on total RNA isolated from MCMV-infected NIH-3T3 cells harvested at 4 hpi with and without CHX treatment. GAPDH was used as a housekeeping gene and results were plotted as fold change relative to MCMV infection under DMSO treatment for three biological replicates. **D.** Graphical depiction of 4 clusters (CL1-4) of viral TiSS obtained by unsupervised clustering on new RNA derived from the dSLAM-seq data (relative expression levels shown). Relative levels were calculated for each TiSS on the basis of new RNA levels which were normalized (reads per million) and scaled such that the maximum across any given time point is 1. **E.** Graphical and sequence logo depiction of core promoter motifs identified for CL1-4 clusters through MEME motif analysis. Please note that the TATA-/TATT-box motif in cluster CL3 is shifted by 2 nt to the left compared to cluster CL1. TBM: TATA-box like motif, Inr: Initiator element.

https://doi.org/10.1371/journal.ppat.1010992.g004

immediate-early TiSS of *ie1/ie3* and *ie2* but revealed an additional immediate early TiSS, namely the m166.5 RNA encoding the m166.5 ORF of 446 aa (**Figs 4B and S4A)**. qRT-PCR analysis confirmed these findings revealing a ~100-fold increased TiSS usage upon CHX treatment by 4 hpi compared to the untreated control (Fig 4C). We thus termed m166.5 immediate early gene 4 (*ie4*). The respective m166.5 ORF has been shown to encode a nuclear protein [23], but lacks functional characterization. In contrast to the other MCMV *ie* genes, *ie4* does not contain any introns. Interestingly, all three immediate early TiSS (*ie1/ie3*, *ie2* and *ie4)* show identical transcription kinetics throughout infection (S4B Fig). This included an early peak at 2 hpi and a low at 6 hpi followed by a continuous, PAA-sensitive rise in transcription until late in infection (72 hpi).

To identify distinct temporal classes of lytic MCMV gene expression, we performed an unsupervised clustering of the 365 unique viral TiSS according to their temporal expression kinetics by 'new RNA'. We found four distinct clusters (CL1-4) to provide the most convincing clustering results. The respective clusters differed both in the onset of viral gene expression as well as subsequent rise or drop thereof (Fig 4D). CL1 expression peaked at 4 hpi followed by

strong downregulation in transcriptional activity despite viral DNA replication. While expression of CL2 genes was readily detectable by 4 hpi, thereby marking them as early genes, their expression increased only weakly at later times of infection. In contrast to the viral IE and E genes, viral L genes require viral DNA replication as well as the late viral transcription factor complex (LTF) [41]. The highly conserved CMV LTF is comprised of six viral proteins and binds to a modified TATA-box, i.e., a TATT motif [10]. Canonical TATA boxes were a hallmark of CL1 and CL2 transcripts (**Fig 4E**). In contrast, promoters of CL3 transcripts harbored TATT motifs. Interestingly, similar to our observations for HCMV [42], TATT motifs were shifted away from the Inr by 2 bp compared to the canonical TATA boxes in CL1 genes. The CL3 cluster comprises the canonical MCMV late genes that commonly encode for structural virion components. Expression of CL3 TiSS gradually increased over time, peaking at 24 hpi, i.e., slightly after the initiation of viral DNA replication, and commonly plateaued at late times of infection (>36 hpi). Interestingly, CL4 promoters did not harbor a TATT or TATA motif and expression of CL4 transcripts both rose significantly later than of CL3 and still continued to rise >36 hpi. However, CL4 transcripts were also expressed at lower levels than transcripts in CL3, possibly due to the absence of TATA or TATT motifs. This raised the question whether CL4 transcripts are indeed regulated differently than CL3 transcripts or whether their distinct kinetics are only observed due to lower transcriptional activity. To discern CL4 as an independent cluster, we segregated transcripts in the CL3 and CL4 clusters into four quartiles according to levels of TiSS expression using our dSLAM-seq data. CL3 and CL4 transcripts exhibited distinct kinetics for all quartiles (**S5A Fig**). Furthermore, even the least strongly expressed genes in CL3 were associated with a distinctly positioned TATT motif, while even the most highly expressed CL4 transcripts were not (**S5B Fig**). This indicated CL4 to represent a so far unknown class of viral transcripts that are expressed with delayed late kinetics and whose expression is dependent on viral DNA replication but not on the viral LTF. A list of all transcripts with their respective clusters and TBM are listed in **S4 Table**.

While our unsupervised clustering revealed major differences in the overall temporal expression profiles of viral TiSS, it did not consider when in infection the respective transcripts were first transcribed. Accordingly, the three immediate early TiSS were placed into cluster CL3 due to their steady increase in transcription after 6 hpi, which resembled the CL3 profile, although their first peak of expression already occurs at 1–2 hpi in contrast to 4 hpi for CL1 and CL2. Furthermore, we noticed that TiSS in cluster CL2, while showing an overall similar expression profile along the whole time course, could be subdivided into two distinct clusters that significantly differed in the onset of transcription (2–4 hpi vs. 6–12 hpi). We thus introduced manual criteria that specifically included information about the first time point when a TiSS showed expression (see **Methods**). This resulted in 6 classes of viral transcription kinetics (TR0-5) (**Figs 5A and S6**). The three immediate early TiSS were placed into Tr0. While transcription of TR1 genes (n = 34) peaked at 4 hpi and then dropped by at least 2-fold throughout infection, transcription of TR2 genes (n = 55) beyond 4 hpi did neither drop nor rise >2-fold and >4-fold, respectively. In contrast, transcription of TR3 TiSS (n = 13) was only weakly if at all detectable by 6 hpi. However, transcription then rapidly increased by 12 hpi but did not increase more than 2-fold thereafter. This is in stark contrast to the canonical viral late genes of TR4 (n = 97), which generally only started to be sufficiently transcribed by 18–24 hpi, i.e., well after the onset of viral DNA replication. Finally, TR5 TiSS (n = 131) were expressed with delayed kinetics and showed increasing transcriptional activity until very late in infection (36–72 hpi). In total, 9 and 23 TiSS could either not be unambiguously or not at all classified into one of the 6 TiSS clusters.

Attribution to TR1 was characteristic for the key viral immune evasins, e.g., m04, m06, m152, m154 and m155, which need to be rapidly expressed upon virus entry. In contrast, TR2

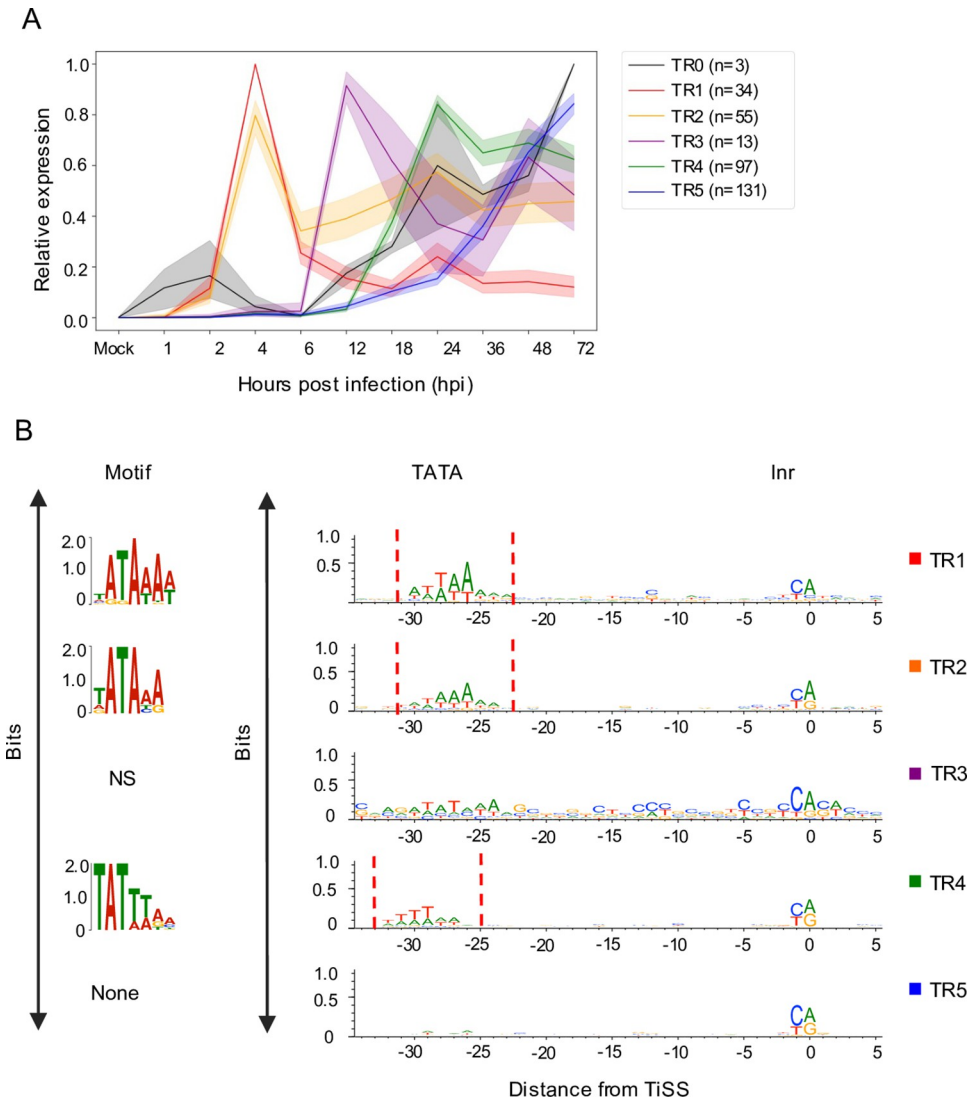

Fig 5. Transcription kinetics of viral TiSS. A. Graphical depiction of 6 clusters (TR0-5) of viral TiSS obtained by manual classification (for criteria see Methods) based on new RNA derived from the dSLAM-seq data (relative expression levels shown) computed similarly as in Fig 4D. **B.** Graphical and sequence logo depiction of core promoter motifs identified for TR0-5 clusters through MEME motif analysis computed as in Fig 4E. Please note that the TATA-/TATT-box motif in cluster TR4 is shifted by 2 nt to the left compared to cluster TR1. The TR3 cluster did not enrich for a specific motif, which is at least in parts due to the small number of TiSS (n = 13) in this cluster. TBM: TATA-box like motif, Inr: Initiator element, NS: Not significant.

https://doi.org/10.1371/journal.ppat.1010992.g005

kinetics were typically observed for viral proteins involved in viral DNA replication, e.g., M54, M57, M70, M105 and M114. While 2 of the 6 viral LTF components (M87 and M91) classified into TR3, three other components (M49, M79 and M95) did not quite fulfill the criteria for TR3 but showed extensive transcription by 12 hpi similar to TR3 kinetics. Only M92 was allocated into TR5 but nevertheless already showed weak expression at the onset of viral DNA replication (12 hpi). The presence of a TATT-motif shifted by 2 bp compared to the canonical TATA-motif of TR1 genes was characteristic for TR4 genes (**Fig 5B**). Here, our more stringent cut-offs removed many of the CL3 TiSS (39 of 75) that did not harbor an upstream TATT-motif. Our TR classification thus sharpened the respective TiSS annotations. Many of the TR5 genes remain poorly studied. However, TR5 also comprised conserved cytomegalovirus genes

including M48, M50, M51, M75, M92, and M104. While further subclustering of TR5 TiSS may be required, it is important to note that the 131 TR5 TiSS were largely devoid of a TATT-motif or TATA-motif in their respective core promoters. We thus hypothesize that transcription initiation of TR5 transcripts does not require the viral LTF but is merely driven by the excessive amounts of viral DNA at late stages of infection.

In summary, our data reveals six classes of lytic MCMV transcription kinetics. We decided to refer to TR0 as 'immediate early (α)', TR1 as 'early (β1)', TR2 as 'maintained early (β2)', TR3 as 'delayed early (β3)', TR4 as 'canonical late (γ1)' and TR5 as 'delayed late (γ2)' transcripts.

To assess the impact of TATA and TATT motifs on the kinetics and extent of viral gene expression, we utilized a dual color reporter virus (MCMV_Δm152-EGFP_SCP-IRES-mCherry). This virus expresses eGFP instead of the coding sequence of the m152 early gene (TR1) and mCherry expressed from an internal-ribosomal entry site (IRES) downstream of the late gene m48.2 CDS encoding for SCP (TR4). We mutated the TATA box of the m152 promoter to create a TATT motif. Upon infection of NIH-3T3 cells with the two viruses for 6 to 72 hours, we analyzed eGFP and mCherry fluorescence by microscopy (**Fig 6A**) and flow cytometry (**Fig 6B**). Consistent with our dSLAM-seq data, hardly any mCherry expression was observed within the first 12 h of infection and subsequent mCherry expression was sensitive to inhibition of viral DNA replication by PAA treatment. Interestingly, the TATA>TATT single point mutation was sufficient to render the m152 promoter PAA-sensitive and change the temporal expression profile towards late kinetics. However, the TATA>TATT mutation only altered the kinetics but not the maximum mean fluorescence intensity (MFI) of eGFP expression throughout infection. Furthermore, introduction of the TATT-motif did not abrogate m152 promoter activity early in infection and only reduced total eGFP expression levels during the first 12 h of infection by ≈2-fold indicating promiscuous binding of the cellular TATA binding protein to the artificially generated TATT sequence. Thus, while the TATT-motif defines sensitivity of a promoter to viral DNA replication, other promoter motifs or features define viral promoter activity during the early phase of infection.

Decoding the MCMV translatome

To decode the MCMV translatome, we employed ribosome profiling (Ribo-seq) along with translation start site (TaSS) profiling [43] across a time course of MCMV-infected NIH-3T3 cells (**Fig 1**). We identified and annotated a total of 454 MCMV ORFs including 232 small ORFs (**S5 Table**). Small ORFs included short novel ORFs (e.g., sORFs), some previously studied ORFs (e.g., m41.1 ORF, whose names were unaltered for consistency) as well as some N-terminal truncations <100 aa in length (annotated as '#1', '#2',. . .). Using the annotation described by Rawlinson *et al.* [22] as reference, we confirmed 150 out of the 170 predicted CDS (**Fig 7A**). Putative CDS with no signs of translation are included in **S6 Table**. Interestingly, most of the predicted CDS that we were unable to detect were low-scoring predictions as per previously described criteria and no corresponding TiSS could be identified. The absence of corresponding transcripts in MCMV infection of fibroblasts explains the absence of detectable levels of translation. As the respective transcripts may be expressed in other cell types or conditions, we nevertheless maintained these CDS in our new genome annotation but labelled them as 'orphan; not expressed'. Additionally, we detected 11 previously validated ORFs, annotated as ORFs (**S7 Table**). Overall, we identified 170 previously annotated ORFs (CDS), 68 ORFs comprising novel ORFs which included 11 ORFs validated in several studies (**S7 Table**), 108 short ORFs, 73 uORFs, 15 uoORFs, 19 uORFs and 1 dORF (**Fig 7B**), accounting for a total of 232 small ORFs (<100 aa in length) and 222 large ORFs (>100aa in length). N-

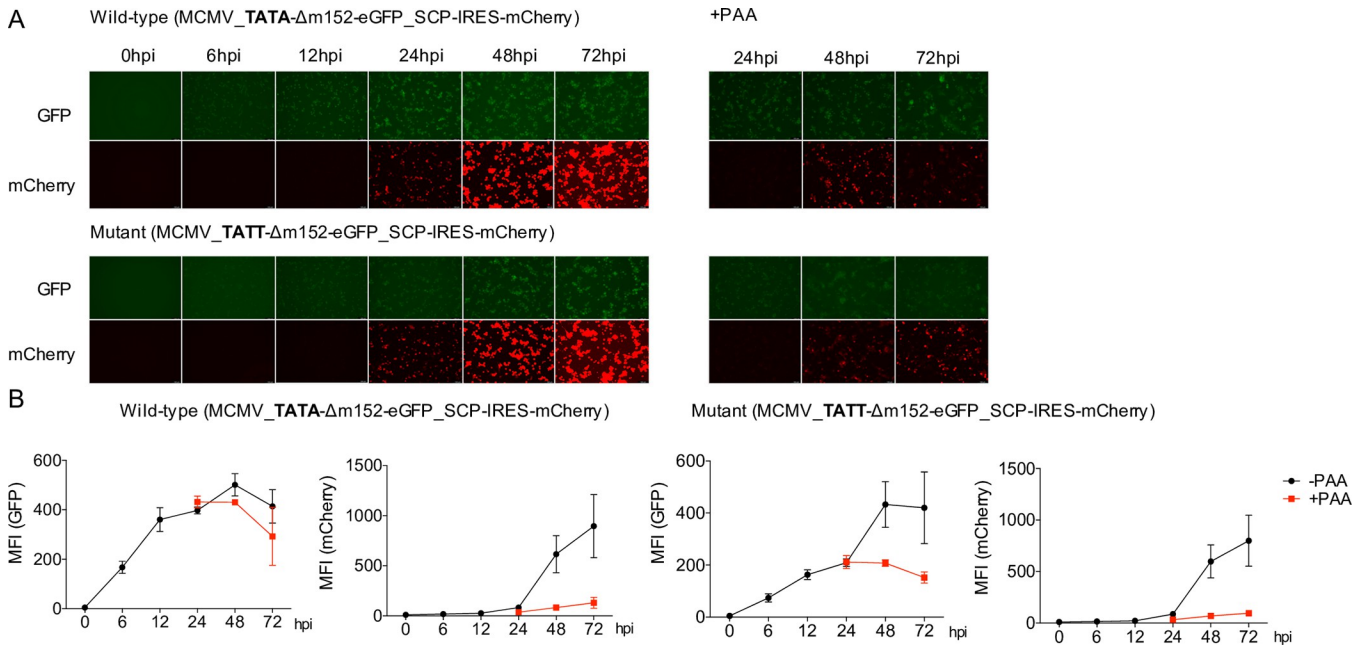

Fig 6. Converting a TATA box to a TATT box is sufficient to alter viral gene expression kinetics. A. NIH-3T3 cells were infected with a two-color MCMV reporter virus (MCMV_TATA-Δm152-eGFP_SCP-IRES-mCherry) and the TATA>TATT mutant thereof (MCMV_TATT-Δm152-eGFP_SCP-IRES-mCherry) at an MOI of 5 for the indicated time points with and without PAA treatment. mCherry and eGFP expression were analyzed through fluorescence microscopy. Representative images of three biological replicates (n = 3) are shown. **B.** Cells were fixed and eGFP and mCherry levels were analyzed quantitatively through flow cytometry and mean fluorescent intensity (MFI) values were plotted for three biological replicates (n = 3) along with standard deviation (S.D.) with and without PAA pre-treatment.

https://doi.org/10.1371/journal.ppat.1010992.g006

terminal truncated or extended products were observed for all of the above mentioned classes of viral open reading frames. It is important to note that some of the previously identified ORFs (S7 Table), e.g., m41.1, are <100 aa in size and thus represent sORFs, but their names were unaltered for consistency with previous studies.

Specific viral transcripts initiating less than 500 nt upstream of the respective ORFs explained translation of 366 of the 454 MCMV ORFs. Only for 88 viral ORFs (66 of 232 small ORFs) no TiSS could be identified within the upstream 500 nt. These were labeled as 'orphan' in our final annotation. The majority (50 of 57, 88%) of novel large viral ORFs initiated at canonical AUG start codons. Alternative start codons included ACG (1 ORFs/9 sORFs), GUG (1 ORFs/ 4 sORFs) and CUG (5 ORFs/ 10 small ORFs), with 12% of novel large ORFs (Fig 7C) and 12% of novel small ORFs (Fig 7D) initiating at non-canonical codons. Most of the 27 NTTs and 10 NTEs identified by our pipeline resulted from alternative TiSS usage. Consistent with the rules applied for CDS identification by Rawlinson *et al.* [22], all NTTs initiated from AUG start codons whereas NTEs predominantly initiated at non-canonical start codons, as these had previously not been considered (Fig 7E and 7F). As such, we identified an N-terminal truncation in the *ie2* locus, i.e., m128 CDS #1 RNA #1 expressed from a novel β1 transcript (m128 RNA #1), which confirms previous observations of a modified IE2 protein of 41 kDa [44]. (S7A and S7B Fig)

Characterization of a previously unknown N-terminally truncated ORF in the m145 locus

We identified a previously unknown NTT of the m145 CDS, which we termed m145 ORF #1. This ORF is expressed from a distinct transcript (m145 RNA #1) at 5-fold higher levels than

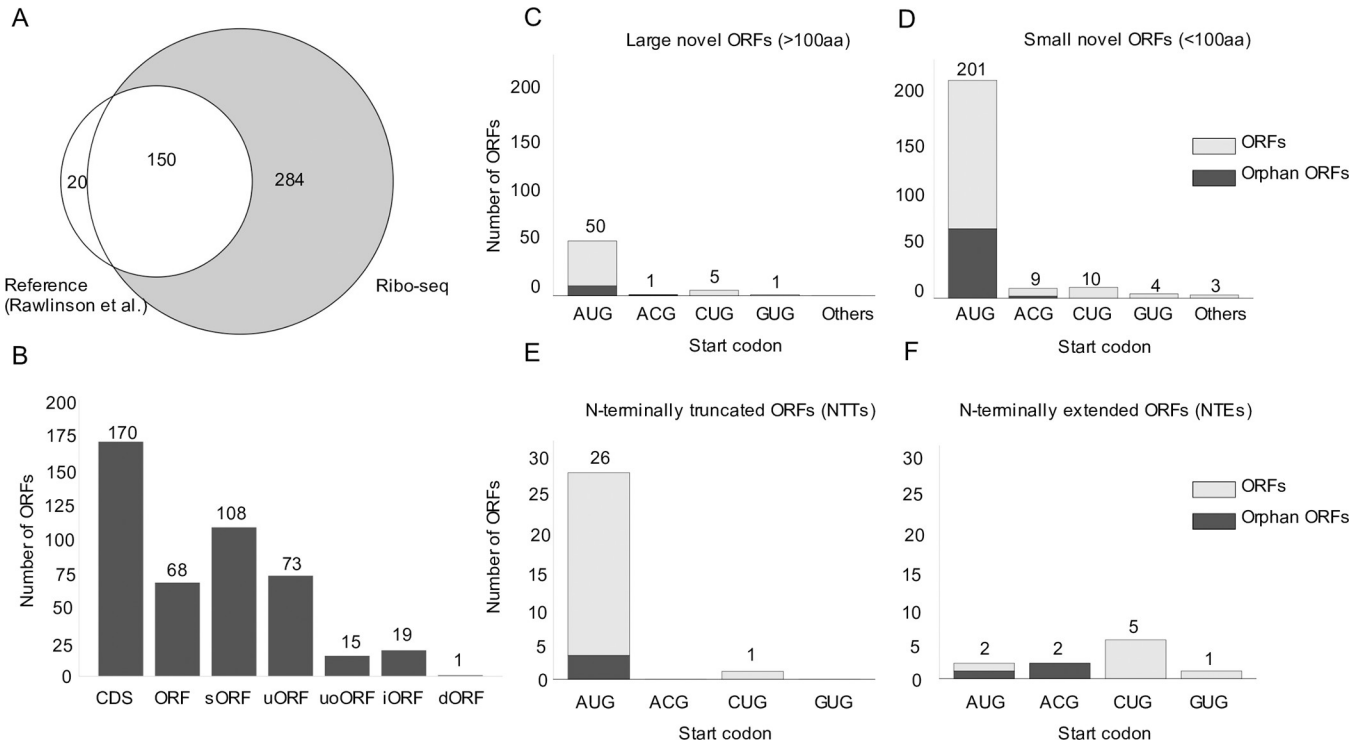

Fig 7. The MCMV translatome. A. Venn diagram depicting the number of MCMV ORFs in our revised MCMV genome annotation as detected by ribosome profiling compared to the Rawlinson *et al.* reference annotation [13]. **B.** Total number of viral ORFs annotated by ribosome profiling grouped into CDS (Rawlinson reference annotation), large ORFs, short ORFs (sORFs), upstream ORFs (uORFs), upstream overlapping ORFs (uoORFs), iORFs (internal ORFs) and downstream ORFs (dORFs). N-terminal extensions (NTEs) or truncations (NTTs) may span any of these defined groups of ORFs. **C-F.** Start codon usage for annotated novel large ORFs, small ORFs, NTEs and NTTs, respectively. ORFs in gray depict orphan ORFs, for which no TiSS could be identified. Each graph depicts the number of ORFs on the y-axis and the start codon usage on the x-axis.

https://doi.org/10.1371/journal.ppat.1010992.g007

the canonical m145 CDS, and lacks the first 340 aa of the 487 aa m145 CDS (**Fig 8A**). The glycoprotein encoded by the m145 CDS interferes with NK-cell activation by downregulating the stress-induced NK cell-activating ligand, MULT-I, predominantly in endothelial cells [45]. Considering the immunological significance of this locus, we sought to validate the N-terminally truncated ORF, m145 ORF #1 and assess its role in the regulation of MULT-I. After first validating the m145 proteins through plasmid expression systems (**S8A Fig**) using V5-tagged ORFs, we generated a C-terminally V5-tagged m145 CDS mutant virus (m145-V5) and analyzed expression in NIH-3T3 and SVEC 4–10 endothelial cells by Western blot (**S8B Fig**). This revealed expression of 4 different protein isoforms at ca. 70, 35, 20 and 13 kDa. While the 70 kDa isoform represents m145 CDS, the 20 kDa isoform constitutes the m145 ORF #1 as confirmed upon ectopic expression (**S8A Fig**). It is important to note that the canonical m145 CDS encodes a type I membrane protein (55 kDa), which contains a distinct signal peptide and is predicted to undergo N-linked glycosylation [45], thereby explaining the 70 kDa gene product. We created a panel of virus mutants (**Fig 8B**) to validate the expression of the four m145 gene products in SVEC 4–10 cells and characterize the respective isoforms. Mutation of the TATA box of the m145 RNA #1 promoter (Δm145 TATA RNA #1) adversely impacted the expression of all three small m145 isoforms (35 kDa, 20 kDa and 13 kDa), but not the 70 kDa isoform (**Fig 8C**). On the contrary, the 70 kDa isoform was selectively eliminated when a STOP codon was inserted 40 aa downstream of its AUG (Δm145 CDS) to terminate m145 CDS while avoiding reinitiation at alternative AUGs upstream (**Fig 8C**). Finally, mutating the m145 ORF #1 AUG start codon abrogated both the 35 and 20 kDa but not the 13 kDa isoform

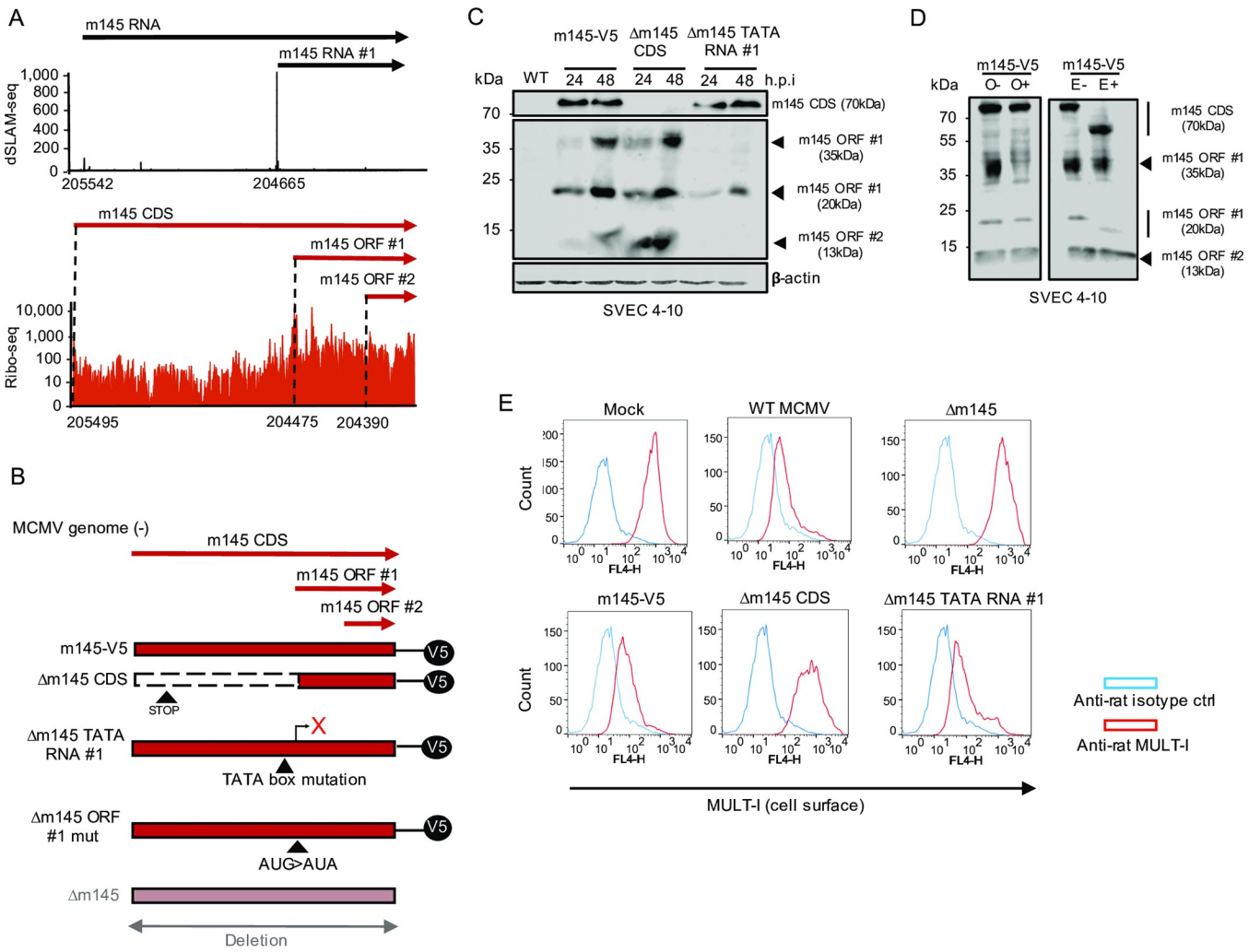

Fig 8. Characterization of N-terminally truncated ORFs in the m145 locus. A. ORFs and transcripts expressed from the m145 locus. This includes the so far unknown m145 ORF #1 and #2 expressed from m145 RNA #1. Coordinates for the TiSS and ORF start codon are shown for each transcript and ORF. dSLAM-seq data are shown in linear scale, Ribo-seq data in logarithmic scale. Aggregated reads across all time points mapping to the m145 locus are shown. **B.** Schematic representation of the MCMV mutants generated to characterize novel viral gene products encoded by the m145 locus. Mutant viruses were generated based on a reporter virus with a V5-tag inserted at the C-terminus of the canonical m145 CDS. The viruses were generated by *en passant* BAC mutagenesis as described in methods on this backbone. The ΔM145 CDS harbored a STOP codon at the 40th codon to skip additional AUGs downstream of the m145 CDS signal peptide which may have resulted in additional products, hindering accurate analysis of the locus. The ΔM145 TATA RNA #1 mutant included a mutation in the TATA box of the respective transcript to abrogate gene expression downstream while the ΔM145 ORF #1 mut mutant was created by mutating the start codon of m145 ORF #1. The ΔM145 virus is a previously created virus where the entire m145 locus (i.e. m145 CDS) was replaced by a kanamycin cassette. **C.** SVEC 4–10 murine endothelial cells were infected with the indicated viruses at an MOI of 1 for 24 and 48 h. V5-tagged m145 gene products were characterized by Western blot. Parental WT MCMV infection was used as negative control. **D.** SVEC 4–10 cells were infected for 48 h with the m145-V5 virus at an MOI of 1. Cells were harvested and treated with or without EndoH$_f$ (E) or O-glycosidase (O) to qualitatively analyze glycosylation patterns of m145 gene products via Western blot. The m145 CDS gene product of 70 kDa shifted to 55 kDa upon EndoH$_f$ treatment justifying its actual predicted weight. **E.** SVEC 4–10 cells were infected with m145 virus mutants at an MOI of 1 for 18 h and stained with rat anti-MULT-I and mouse anti-m04 antibodies following cell surface MULT-I analysis through flow cytometry by gating on infected cells (m04+). Anti-rat and anti-mouse isotype antibodies were utilized as negative controls. Western blots and flow cytometry histograms are a representative for two (n = 2) and three biological replicates (n = 3), respectively.

https://doi.org/10.1371/journal.ppat.1010992.g008

(S8C Fig). We conclude that the 35 kDa and 20 kDa represent post-translationally modified isoforms of m145 ORF #1, while the 13 kDa gene product results from inefficient ribosome scanning on the m145 RNA #1 and translation initiation at the next AUG start codon located 84 nt downstream. We thus annotated the 13 kDa isoform as an independent small ORF and named it m145 ORF #2 RNA #1 (= m145 ORF #1 translated from RNA #1).

Next, we asked whether the different isoforms translated from m145 RNA #1 represented glycosylated isoforms of m145 RNA #1 or novel gene products arising from the transcript. We first analyzed glycosylation patterns of the respective proteins through enzymatic treatment with $EndoH_f$ and O-glycosidase. This confirmed the glycosylated modification of the 20 kDa N- and 35 kDa O-linked isoforms. The former protein appeared at 16 kDa upon $EndoH_f$ treatment, justifying the predicted molecular weight of m145 ORF #1 while the latter band disappeared upon O-glycosidase treatment (**Fig 8D**). Interestingly, no O-linked glycosylated form of the larger protein encoded by m145 CDS was observed. We hypothesize that its signal peptide marks the protein to exclusively undergo N-linked glycosylation. In contrast, the 13 kDa gene product remained unaffected by glycosidase treatment. These findings confirm the existence of an additional truncated viral protein (m145 ORF #2) resulting from inefficient translation initiation at the m145 ORF #1 AUGs upstream. It is important to note that the expression of other truncated viral proteins may thus have been missed by our Ribo-seq data.

To clarify which of the m145 ORFs is responsible for downregulation of MULT-I, we analyzed cell surface expression of MULT-I through flow cytometry upon infection with the respective mutant viruses. The Δm145 ORF #1-V5 mutant downregulated MULT-I similar to WT MCMV, indicating that both m145 ORF #1 (despite being expressed at higher levels than m145 CDS) and m145 ORF #2 were not responsible for downregulating cell surface MULT-I and the phenotype was fully attributed to the longer isoform, namely the m145 CDS (**Fig 8E**). Our data also confirmed the importance of alternative TiSS usage in governing the expression of MCMV protein isoforms [27].

Viral uORFs tune viral gene expression

A substantial number of the novel viral ORFs that we identified represent uORFs, which are located completely upstream of a canonical ORF, and uoORFs, which start upstream and overlap with the canonical ORF. Since translation of u(o)ORFs impacts on translation of their downstream ORFs [46,47], we aimed to confirm this for selected MCMV u(o)ORFs using dual luciferase reporter assays. We cloned four candidate u(o)ORFs into the psiCheck-2 vector [48] upstream of firefly luciferase. We then mutated their AUG start codon(s) to abrogate translational regulation on the downstream out-of-frame firefly luciferase. This fully relieved translational repression on the downstream firefly *luc* gene confirming translation of the m169 uORF (MATp1) [28], m119.3 uORF along with uoORFs in the M35 and M48 locus (**Fig 9A-9D**). Interestingly, for both the m169 and m119.3 uORF, disruption of the first AUG was not sufficient to fully abrogate their inhibitory potential. However, subsequent mutation of downstream in- and out-of-frame AUG start codons consistently increased downstream luciferase expression. Only when all AUGs (up to 6 for m169 uORF) had been mutated, the observed rescue in luciferase activity matched the expression differences between the respective u(o)ORFs and their larger downstream counterparts observed by ribosome profiling. We confirm translation of several viral u(o)ORFs, which may serve to regulate downstream ORFs and/or express functional viral microproteins. Future studies should be performed to assess the regulatory and functional role of viral u(o)ORFs *in vitro* and *in vivo*.

Reannotation of the MCMV genome

The novel MCMV transcripts and ORFs identified by our approach generated the need for a revised annotation of the MCMV genome. We used the MCMV annotation provided by Rawlinson *et al.* [22] with its 170 viral ORFs as our reference annotation for the BAC-derived pSM3fr MCMV genome sequence [49] curated by our sequencing data. The sequence was corrected by eliminating the BAC sequence as well as corrections from Table 2 in [49] which were

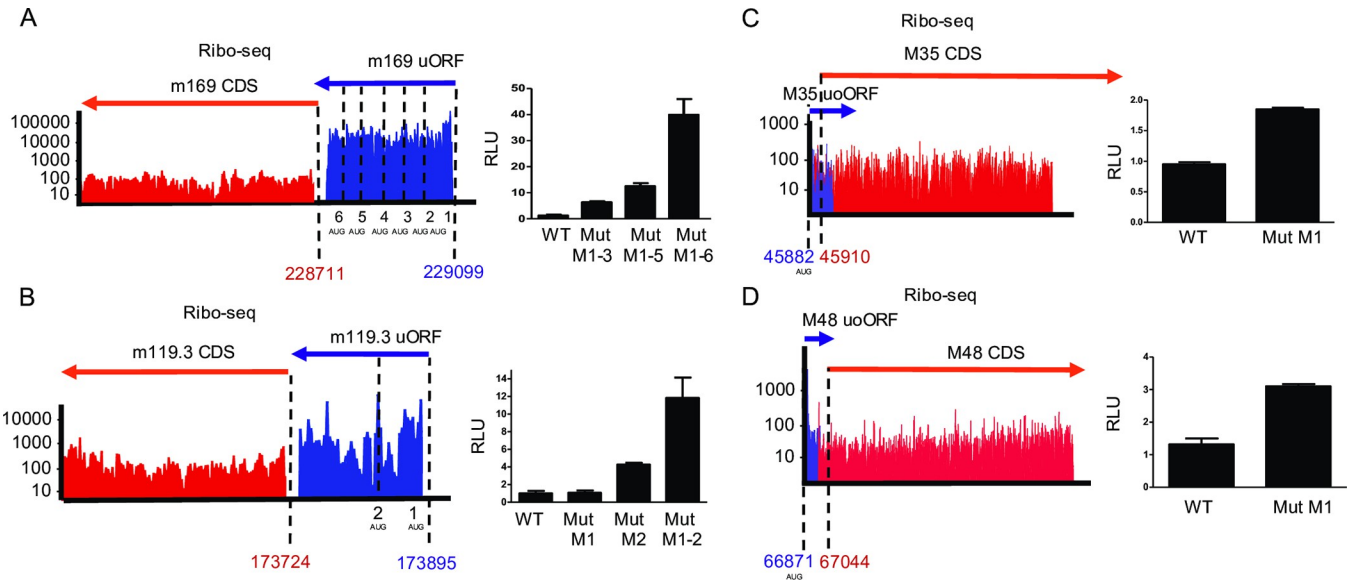

Fig 9. MCMV uORFs/uoORFs tune viral gene expression. Ribo-seq data (aggregated reads in logarithmic scale) of the respective viral genomic loci and their validation by dual luciferase assays are shown for m169 uORF (**A**), m119.3 uORF (**B**), M35 uoORF (**C**) and M48 uoORF (**D**). The number of AUG codons for the respective viral u(o)ORFs are indicated. Coordinates represent the start codons of the u(o)ORFs and ORFs. Both ORFs and u(o)ORFs were annotated based on translational start site profiling data, including Harringtonine and Lactimidomycin pre-treated samples and PRICE analysis. Manual curation assessed the presence of upstream TiSS, start codon usage and presence of STOP codons to identify relevant ORFs. psiCheck-2 reporter plasmids harbored the indicated MCMV u(o)ORFs (WT) and AUG start codon mutants thereof (Mut) upstream of *firefly-luc* reporter gene. Luciferase assay data at 48h post transfection are shown as mean RLU (Firefly/Renilla ratio) of three biological replicates (n = 3) plotted along with the standard error (S.E.M.).

https://doi.org/10.1371/journal.ppat.1010992.g009

not correctly incorporated previously in KY348373. The new annotated references (gb.) with and without the BAC sequence are uploaded as S2 and S3 Files respectively. All reference ORF names were maintained accordingly and named as 'CDS' (coding sequences) to distinguish these from novel viral 'ORFs'. Any viral ORFs that had previously been revised with minor changes were labelled as 'corrected' (**S8 Table**). We employed the same nomenclature strategy as for the HSV-1 annotation to annotate novel MCMV transcripts and ORFs without altering the existing nomenclature [17]. Briefly, transcription initiating ≥500 nt distant from another transcript was given a new identifier, starting with '.5' to provide room for future additional ORFs in case any TiSS or ORFs had been missed. Transcripts arising from alternative TiSS located within <500 nt upstream or downstream of the main (canonical) transcript in a given locus were labelled as '*1', '*2',. . . and '#1', '#2',. . ., respectively. All large novel ORFs were annotated as 'ORFs'. Small ORFs were annotated as 'uORF', 'uoORF', 'iORF', 'dORF' or 'sORF' depending on their relative location to their respective CDS or ORFs. NTEs and NTTs of ORFs were annotated with '*' and '#' respectively. An RNA identifier was used to explain ORFs that could be attributed to alternative TiSS. For example, M25 CDS #1 RNA #1 indicates a truncated ORF (NTT) in the M25 locus translated from an alternative TiSS, namely M25 RNA #1, which initiates downstream of the canonical M25 RNA (Fig 2A). Alternative spliced products were labelled as ORF isoforms ('Iso1', 'Iso2',. . .). ORFs for which no TiSS could be detected were labelled as 'orphan'. Similarly, transcripts for which no ORF was identified as expressed within the first 500 nt were labelled as 'orphan'. In total, our final reference annotation includes 66 weakly expressed 'orphan' viral RNAs and 88 'orphan' viral ORFs. Reference CDS, which were undetected in our data (and usually lacked a corresponding transcript), were labelled as 'orphan; not expressed' but were nevertheless included into the final annotation. The fully reannotated MCMV genome with and without the corresponding BAC were deposited to the NCBI GenBank Third Party Annotation database.

In summary, promiscuous transcription initiation within the MCMV genome, novel splice isoforms and translation of uORFs and uoORFs upstream of major viral CDS/ORFs explained the novel viral gene products identified by our integrative multi-omics approach.

Discussion

Our study provides a state-of-the-art annotation of the MCMV genome by integrative analyses of a variety of high-throughput sequencing approaches to reveal the hierarchical organization of the entire MCMV transcriptome and translatome at single-nucleotide resolution. While several studies have described novel ORFs and transcripts in previously unannotated regions, our integrative reannotation of the MCMV genome provides a unifying nomenclature for all MCMV gene products. As previously observed for HSV-1, simple peak calling based on our dSLAM-seq and cRNA-seq data would have resulted in the identification of hundreds of additional putative TiSS. While our annotation clearly represents a conservative approach, we restricted the final TiSS to 365 reproducible TiSS by integrative analysis of dSLAM-seq, cRNA-seq and 4sU-seq data. Careful manual inspection of all TiSS candidates in relation to the available Ribo-seq data further increased the reliability of the final TiSS that were included into the new reference annotation. The validity of this approach was confirmed by the strong overrepresentation of Inr elements at the viral TiSS even for the most weakly utilized TiSS. This is consistent with previous findings for HSV-1 and supports the accuracy of our annotation workflow [17]. The vast majority of TiSS were required to explain the expression of novel uORFs, uoORFs, iORFs and splice isoforms, and validated novel NTEs and NTTs revealed by ribosome profiling. Accordingly, only 66 TiSS (of 380, 17.37%) were labelled as orphan while 88 ORFs (of 454, 19.38%) could not be attributed to a viral transcript initiating within 500 nt upstream. Most of these TiSS (46 of 66; 70%) represented TR5 (γ2) TiSS indicating that translation of the corresponding ORFs or sORFs they encode might only have become detectable at >48 hpi by Ribo-seq and was missed by our Ribo-seq analysis.

We observed a striking number (n = 366) of putative splicing events in the MCMV transcriptome. However, the majority of these only occurred at low frequencies. We thus decided to include only a conservative 28 splicing events into our new reference annotation.

Interestingly, dSLAM-seq combined with 4 h of cycloheximide treatment revealed a novel unspliced *ie* gene, namely m166.5 RNA (*ie4*), which we subsequently confirmed by qRT-PCR. The function of ie4 remains unclear and deserves further studies. The expression of all three *ie* TiSS (ie1/ie3, ie2 and m166.5) was enhanced >300-fold upon inhibition of protein synthesis consistent with a lack of self-inhibition upon CHX treatment. After a first peak of transcription at 1–2 hpi, their expression already started to rise again at 6 hpi and then continued to rise until very late in infection. This increase was abolished upon PAA treatment.

Clustering transcripts by 'new RNA' through dSLAM-seq revealed four distinct clusters describing the kinetics of viral gene expression (CL1-CL4). Inspection of individual TiSS assigned to all clusters indicated that our unsupervised clustering was predominantly based on the overall temporal expression profiles while the first onset of expression only played a minor role for clustering. Thus, we defined manual classification criteria that were based on the CL clustering but more accurately defined the transcription kinetics of the viral TiSS, also taking into account the onset and first peak of expression. This resulted in 6 distinct transcription kinetics (TR0-5), which we refer to as immediate early (α = TR0), early (β1 = TR1), maintained early (β2 = TR2), delayed early (β3 = TR3), canonical late (γ1 = TR4) and delayed late (γ2 = TR5). However, we would like to point out that many viral genes comprise >1 viral TiSS. The expression kinetics of the respective proteins thus reflects the composite regulation of the 6 TR kinetics as exemplified by the *ie2* locus (**S7 Fig**).

In contrast to the TR2 (β2) TiSS, which are characteristic for many viral genes involved in viral DNA replication, 5 of the 6 LTF components showed expression kinetics either belonging to or consistent with TR3 (β3) kinetics. TR1 (β1) and TR4 (γ1) promoters were associated with distinct TATA- and TATT-box elements, respectively, thus explaining the expression of early and late genes as shown for various herpesviruses. Similar to HCMV [42], the TATT-motif in TR4 (γ1) promoters that is recognized by the viral LTF complex tended to be located by about 2 nt further upstream of the TiSS in comparison to the canonical TATA-box motif in the promoters of TR1 and cellular genes. By mutating the TATA box of an early (β1) gene, m152, to a TATT motif, we demonstrate that viral late kinetics and PAA-dependence are mediated by the TATT motif and thus the viral LTF complex. However, mutation of a TATA- to a TATT-motif had little impact on the absolute transcriptional output and did not qualitatively affect transcriptional activity of the m152 promoter early in infection. It is important to note that we only introduced a single A-to-T mutation but did not shift the TATT-box away from the m152 Inr element by 2 nt as typically observed for TR4 (γ1) genes. This may explain residual m152 early expression of the TATT mutant. However, other factors, which include cellular transcription factors activated early in infection, may also contribute to m152 early gene expression.

The delayed kinetics of cluster TR5 (γ2) and the absence of a TATT-box element were surprising. The respective transcripts came up significantly later in infection than cluster TR4 (γ1) and commonly continued to rise until 72 hpi. Their expression is thus unlikely to be dependent on the TATT-specific viral LTF. We hypothesize that transcription initiation of TR5 (γ2) transcripts is driven by weak transcription initiation mediated solely by the Inr element in the context of extensive amounts of viral DNA late in infection. While experimental proof will require studies using LTF-deficient MCMV mutants, our findings indicate that the viral LTF becomes rate limiting late in infection and that TR5 (γ2) TiSS represent LTF-independent transcription at very late stages of infection.

Recently, the Price lab reported on the identification of ≈7,500 transcription start site regions (TSRs) in the HCMV genome during lytic infection of fibroblasts, which corresponds, on average, to a TSR every 65 nt, using PRO-seq and PRO-cap [50]. These were corroborated by additional studies from the same lab attributing their expression kinetics at least in parts to the viral IE2 protein and LTF [10,51]. While our TiSS profiling data do not exclude the presence of a much larger set of TSRs for MCMV, the TiSS we identified (***i***) correspond to stable RNAs and (***ii***) are sufficient to explain the (near) complete MCMV translatome identified by ribosome profiling. Importantly, the presence of thousands of additional stable viral transcripts should have resulted in translation initiation at hundreds of additional AUGs and thus viral (s)ORFs observable by Ribo-seq. We conclude that the number of stable MCMV transcripts that are actively translated is unlikely to exceed our annotation by an order of magnitude. Importantly, the PROseq/PROcap approach not only detects stable transcripts but also transcription of highly unstable transcripts including promoter- and enhancer-derived RNAs. Interestingly, dSLAM-seq and STRIPE-seq analysis on HCMV-infected fibroblasts, which both only detect stable transcripts, only confirmed ≈1,700 of the >7,000 TSRs but indicated extensive non-productive (pervasive) transcription of the HCMV genome [42]. Our data for MCMV are consistent with our findings for HCMV showing that a large fraction of the >7,000 TSRs reported for HCMV presumably do not correspond to stable viral transcripts. It will be interesting to study whether transcription initiation is as promiscuous in lytic MCMV infection as observed for HCMV.

Our TiSS profiling data provide strong additional evidence for the newly identified ORFs and small ORFs detected by Ribo-seq. In the vast majority of cases, the respective novel ORFs initiate from the first AUG downstream of the respective TiSS. An excellent example of this is m145 ORF #1. It is translated from a so far unknown viral transcript (m145 RNA #1) that

initiates in the middle of the m145 CDS. However, as we demonstrated for the 13 kDa m145 ORF #2, inefficient ribosomal scanning of m145 RNA #1 also explains translation initiation at the next downstream AUG resulting in the expression of this truncated protein isoform. Although the less abundantly expressed m145 CDS was responsible for the published effects on MULT-I [45], our findings confirm expression of at least two additional viral proteins (m145 ORF #1 and #2) and implicate differentially glycosylated gene products expressed from the m145 locus. While we were a bit surprised to see that the less prominently expressed m145 CDS accounted for the reported regulation of MULT-I, high expression of m145 ORF #1 may well have confounded the interpretations of previous *in vivo* experiments [52]. Further studies are required to functionally characterize the role of the additional proteins expressed from the m145 locus.

Similar to HCMV [16], 227 of 284 novel MCMV ORFs (80%) were <100 aa in size, a substantial fraction of which represented uORFs or uoORFs. Their cellular counterparts have been implicated to control gene expression of their downstream ORFs at the translational level [46,47]. By identifying both the u(o)ORFs and their corresponding TiSS, our data will now enable functional studies pertaining to u(o)ORF-mediated gene regulation in CMV infection. However, small MCMV ORFs may nevertheless encode for abundant microproteins with important functions. The potential of such novel sORF-encoded viral microproteins for productive infection was recently demonstrated for the m169 uORF encoding an NK cell immune evasin [28] and the m41.1 gene product [53] that blocks mitochondrial apoptosis. Mass spectrometry and structural biology data should thus be reanalyzed to look for novel CMV microproteins in all 6 frames. For HCMV, such a 6-frame analysis of whole proteome mass spectrometry data has already been performed [20]. Finally, small ORFs have also been implicated to generate antigenic peptides, resembling rapidly generated DRiP-derived peptides [54]. Such peptides generated from microproteins may form a major component of the antigenic repertoire [43,54,55], playing a role in various diseases [21]. Our revised annotation of the MCMV genome now enables to assess their role in antigen presentation and immune evasion in the MCMV model.

Materials and methods

Cell culture, viruses and infection

NIH-3T3 (ATCC CRL-1658) Swiss mouse embryonic fibroblasts were grown in DMEM (Dulbecco's Modified Eagle's Medium) supplemented with 100 IU/mL penicillin (pen), 100 μg/mL streptomycin (strep) and 10% NCS (New-born calf serum). M2-10B4 (ATCC CRL-1972) fibroblasts were grown in RPMI-1640 (Roswell Park Memorial Institute Medium) supplemented with 100 IU/mL pen, 100 μg/mL strep and 10% FCS (Fetal calf serum). 293T (ATCC CRL-3216) human embryonic kidney (HEK) epithelial cells and SVEC 4–10 mouse endothelial cells (ATCC CRL-2181) were grown in DMEM supplemented with 100 IU/mL pen, 100 μg/mL strep and 10% FCS. All cells were grown in 5% CO_2 at 37˚C. All viruses were generated by infecting M2-10B4 cells after virus reconstitution. BAC-derived MCMV Smith strain was utilized for all sequencing experiments [49]. Infected cells and supernatants were harvested after >90% infection for virus purification and titration of virus stocks was conducted by standard plaque assays on NIH-3T3 cells [3]. The Δm145 virus has been published previously [52]. Infections were conducted using centrifugal enhancement at 800g for 30 min in 6-well plates followed by incubation at 37˚C in 5% CO_2 for 30 min. Media change following incubation marked the 0-hour time point of infection. An MOI of 10 was used for all high-throughput experiments.

Virus mutagenesis and reconstitution

The MCMV Smith strain bacterial artificial chromosome (BAC) in GS1783 *E. coli* [49] was used to construct MCMV virus mutants using *en passant* mutagenesis [56], as described previously. Selected clones were verified by restriction enzyme digestion and Sanger sequencing of the respective locus. BAC DNA was purified using the NucleoBond BAC 100 kit (Macherey-Nagel #740579) and were transfected into early passage NIH-3T3 cells in 6 well plates using TransIT-X2 dynamic delivery transfection system (Mirus). Viruses from cell culture supernatants were passaged on M2-10B4 cells followed by virus purification and titration [3]. All primers along with cloning strategies utilized are described in **S9 Table**. Briefly, m145 virus mutants were generated as follows. PCR products harboring mutations and homologies to adjacent MCMV sequences for each of the mutants were generated from their respective primers listed in **S9 Table** for BAC cloning. The Δm145 CDS mutant PCR product harbored a STOP codon mutation (CAC>UGA) at the 40th codon of m145 CDS. The Δm145 TATA RNA #1 mutant was developed by mutating the TATA box (TATATATAT>TATCTACAT) of m145 RNA #1 and the Δm145 ORF #1 mut was developed by mutating the start codon of m145 ORF #1 (AUG>AUA). The MCMV_TATA-Δm152-eGFP_SCP-IRES-mCherry virus was generated as described in **S9 Table** which was used as a backbone for generating the MCMV_TATT-Δm152-eGFP_SCP-IRES-mCherry virus where the PCR product for BAC mutagenesis harbored a TATAAAAA>TATTAAAA mutation.

RT-qPCR analysis

Wild-type MCMV infections were performed as described for dSLAM-seq in 12-well plates using centrifugal enhancement at 800g/30 minutes. Cycloheximide (50 μg/mL) treatment was performed at 0hpi. DMSO was used as mock treatment. Samples were harvested at 4hpi, followed by RNA extraction using the Zymo Quick Microprep kit including an additional gDNA digestion step using TURBO DNAse (Life technologies). 300–400 ng RNA was used to prepare cDNA utilizing the Bimake 5X qRT All-in-one- cDNA synthesis mix. A 1:5 dilution of the obtained cDNA was subject to 2-step qPCR using the SYBR green qPCR MasterMix (2X) by MedChemExpress as described by the manufacturer. qPCR was performed on the Roche LightCycler® 96. Each qPCR included two technical replicates per gene. The obtained data were analyzed by ddCt analysis for three biological replicates. Mean and SEM were plotted using Graphpad Prism. Primers used are listed in **S9 Table**.

Plasmids and transfection

The psiCheck-2 vector was utilized for validating uORFs/uoORFs by dual luciferase assays [48]. All uORF/uoORF constructs were purchased as gene block fragments from Integrated DNA Technologies (IDT) bearing homologies to psiCheck-2 BstBI and ApaI sites. Cloning was performed using the In-fusion HD Cloning Plus kit (Takara Bio) as per manufacturer's instructions, followed by transformation in Stellar competent cells (Takara Bio). uORF/uoORF start codon mutants were generated by double-fragment infusion cloning using two PCR products bearing homologous ends containing mutations. MCMV m145 ORFs were cloned into pCREL-IRES-Neon expression plasmids with a C-terminal V5-tag between Spe-I and Cla-I restriction sites using infusion cloning. All plasmids were sequenced and purified using the PureYield Promega Midiprep system. For luciferase assays, plasmids were transfected in NIH-3T3 cells in a 96-well plate using Lipofectamine 3000 (Invitrogen). Luciferase readings were measured 48 hours' post-transfection using the Dual-Glo Luciferase assay system (Promega), as per manufacturer's instructions using the Centro XS^3 LB960 system (Berthold Technologies). For Western blot, 6-well plates seeded with HEK293T cells were

transfected with the m145-expressing plasmids using TransIT-X2 dynamic delivery transfection system (Mirus) and cells were harvested at 48 hours' post transfection. All primers and synthetic constructs used are described in S9 Table. All restriction enzymes were purchased from NEB. Luciferase data mean values (Firefly/Renilla ratio) were plotted along with standard error (SEM) as relative light units (RLU) for three biological replicates using Graphpad Prism.

Western blot

Cells were lysed with 2X Laemmli sample buffer (Cold Spring Harbor protocols) with 20% β-Mercaptoethanol. Lysed samples were sonicated and heated at 95˚C/10 minutes. Tris-Glycine SDS-PAGE (12%) and wet transfer (Tris-Glycine-20% Methanol) on 0.2 μm Nitrocellulose membrane (Amersham Protran) were performed using the Mini Gel Tank (Life technologies). Membranes were subsequently subject to blocking in 5% (v/v) skimmed milk in 1X PBST (Phosphate buffered saline– 0.1% Tween 20) at room temperature for one hour. Samples were probed with rabbit anti-V5 antibody (Cell Signaling #13202S) at a 1:1000 dilution, overnight at 4˚C and then probed with a 1:1000 dilution of α anti-rabbit IgG-Horseradish peroxidase (HRP)–Sigma Aldrich A0545. All antibodies were diluted in 5% (v/v) milk in 1X PBST. Proteins were analyzed by visualizing the blots on LI-COR Odyssey FC Imaging System. For O-glycosidase (NEB P0733S) treatment, samples were lysed in 1X RIPA lysis buffer containing anti-protease cocktail (cOmplete, Mini Protease Inhibitor Cocktail, Roche) along with denaturing buffer supplied by NEB. Treatment with O-Glycosidase and Neuraminidase (NEB P0720S) was conducted as per manufacturer's instructions for one hour at 37˚C. A similar protocol was performed for $EndoH_f$ (NEB P0703S). ß-actin was used as a housekeeping control and immunoblotting was performed using mouse anti- ß-actin primary monoclonal antibody (C4- sc-47778 Santa Cruz Biotechnology, Inc.), and the fluorescent IRDye 680 RD goat anti-mouse IgG (Licor) was used as a secondary antibody. Both antibodies were diluted 1:1000 in 1X PBST. All western blot images were processed through ImageStudio Lite.

Flow cytometry

Uninfected and MCMV-infected SVEC 4–10 were washed with 1X PBS and detached using TrypLE Express (Gibco) 18 hpi followed by blocking in 10% FCS-PBS (1X) for 30 minutes. Cells were stained with rat anti-MULT-I and/or mouse anti-MCMV m04 at a dilution of 1:100 including isotype controls for MULT-I (eBioscience Rat IgG2a kappa control eBR2a) and m04 (eBioscience Mouse IgG2b kappa control eBMG2b) as well as only secondary antibody controls by incubating for 30 minutes on ice. Both anti-MULT-I and anti-m04 antibodies were provided by Stipan Jonjic. Followed by primary antibody staining, cells were stained by Invitrogen Goat anti-Rat IgG (H+L) Alexa Fluor 647 (MULT-I) and/or Abcam goat polyclonal anti-Mouse Alexa Fluor 488 (m04) at a dilution of 1:1000 for 30 minutes on ice. All antibodies were diluted in 10% FCS-PBS (1X). Cells were finally suspended in FACS buffer (1X PBS with 0.5% BSA, 0.02% sodium azide). Flow cytometry was performed using the BD Biosciences FACS Calibur Cell Quest Pro system. Gating and further analysis was performed using FlowJo 10. Briefly, live SVEC 4–10 cells were gated for anti-mouse Alexa Fluor 488 bound MCMV infected cells via the FL-1 channel (488 nm Argon ion laser and 530/30 filter) followed by histogram visualization of cell surface expression levels of MULT-I bound by anti-rat Alexa Fluor 647 using the FL-4 channel (635 nm Red diode laser and 661/16 filter). Flow cytometry analysis was similarly performed by analyzing GFP (FL-1) and mCherry expression (FL-3), post fixing in 4% formaldehyde and MFI values and SD for each time point/condition were plotted using Graphpad Prism for three biological replicates. Prior to fixing, the samples were analyzed qualitatively via microscopy at 10X resolution using the Leica DMi8 system.

Transcription start site (TiSS) profiling

Cycloheximide treatment at 50 μg/mL was conducted at the time of infection and phosphonoacetic acid (PAA) treatment was conducted at 300 μg/mL one-hour post infection. cRNA-seq and dSLAM-seq were performed as described [17] with minor modifications. For all dSLAM-seq samples, 4sU labelling was initiated by adding 400 μM for 60 minutes before harvest using TRI reagent (Sigma Aldrich) as described by manufacturer and purified by standard phenol-chloroform extraction. Total RNA was re-suspended in 1X PBS buffer. U-to-C conversion were initiated by iodoacetamide (IAA) treatment as described previously [32] and RNA was re-purified using RNeasy MinElute (Qiagen). Efficiency of IAA conversion was checked by converting 1mM 4sU and analyzing the change in absorption (loss of absorption maximum at 365 nm) upon IAA treatment [32]. Following this, library preparation using the dRNA-seq protocol and Xrn-I digestion was performed by the Core Unit Systems Medicine (Würzburg) as described previously for HSV-1 [17]. Sequencing was performed on NextSeq500 (Illumina). For cRNA-seq, the same protocol was utilized as for HSV-1 [17]. 5' read enrichment was obtained using chemical RNA fragmentation (50–80 nt fragments) and libraries were prepared using 3' adaptor ligation and circularization. Libraries were sequenced on a HiSeq 2000 at the Beijing Genomics Institute in Hong Kong. Total RNA-seq and 4sU-seq was conducted as described [30]. Briefly, 4sU labelling was conducted at 500μM for 60 minutes for the time points described in **Fig 1**. Cells were lysed in Trizol (Invitrogen) and total and 4sU-labelled (newly transcribed RNA) were isolated as per previous protocols. Libraries were prepared using the stranded TruSeq RNA-Seq protocol (Illumina, San Diego, USA) as described, and libraries were sequenced by synthesis sequencing at 2 × 101 nt on a HiSeq 2000 (Illumina).

Ribosome profiling

Ribosome profiling time-course (lysis in presence of cycloheximide) experiments were conducted as described [16] for time-points as shown in **Fig 1** for four biological replicates. Additionally, translation start site (TaSS) profiling was performed by culturing cells in medium containing either Harringtonine (2 μg/ml) or Lactimidomycin (50 μM) for 30 min prior to harvesting. Two biological replicates were generated for Harringtonine pre-treatment and one for Lactimidomycin. Libraries were generated as described for cRNA-seq [17], which introduces a 2 + 3 nt unique molecular identifier (UMI), facilitating the removal of PCR duplicates from sequencing libraries. All libraries were sequenced on a HiSeq 2000 at the Beijing Genomics Institute in Hong Kong.

Data analysis and statistics

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 1.2 against the mouse genome (mm10), the mouse transcriptome (Ensembl 90), and MCMV (KY348373, checked and corrected according to mutations listed in the previous publication [49]). 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 and sharing the same UMI were collapsed to a single copy. Two observed UMIs that differed by only a single base are likely due to a sequencing error and were therefore considered to be the same UMI. 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.

dSLAM-seq and 4sU-seq data 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 UMIs were used.

Our dSLAM-seq and cRNA-seq TiSS profiling data were analyzed with our iTiSS analysis pipeline (available at https://github.com/erhard-lab/iTiSS) [34], which identifies potential TiSS at single-nucleotide resolution. The SPARSE_PEAK module was used for dSLAM-seq. For cRNA-seq data, DENSE_PEAK, DENSITY, and KINETIC modules were used. For each replicate, reads were pooled from all time points. Subsequently, for each dataset, TiSSMerger2, a subprogram in iTiSS, was used to merge TiSS with a +/- 10 bp window. Correspondingly, all TiSS from all datasets were merged using TiSSMerger2 also with a +/- 10 bp window. iTiSS assigned a score ranged from 1 to 4 for each TiSS based on several criteria:

i. Significant accumulation of the 5′-end of reads in both replicates of the dSLAM-seq dataset at the TiSS (SPARSE_PEAK module).

ii. Significant accumulation of the 5′-end of reads in both replicates of the cRNA-seq dataset at the TiSS (DENSE_PEAK module).

iii. Stronger transcriptional activity downstream than upstream of the potential TiSS in both cRNA-seq replicates (DENSITY module).

iv. Significant temporal changes in TiSS read levels during the course of infection in both cRNA-seq replicates (KINETIC module).
We also included 3 additional criteria for scoring.

v. Stronger transcriptional activity downstream than upstream of the potential TiSS in both 4sU-seq replicates.

vi. Significant temporal changes in TiSS read levels during the course of infection in both 4sU-seq replicates.

vii. The presence of an ORF at most 250 bp downstream, which was not yet explained by another transcript.

Thus, in total, we assigned a score between 1 to 7 for each TiSS. We then manually inspected the final list of TiSS using our MCMV genome browser and selected TiSS with a prominent signal to be included in our annotation. A histogram was created showing the number of criteria fulfilled by all annotated TiSS. In addition, we also created a heat map and a bar plot to compare cRNA-seq and dSLAM-seq by calculating the enrichment of reads at TiSS compared to +/- 100bp region around the TiSS (**S1A Fig**). Both figures indicate that dSLAM-seq provides a better signal-to-noise ratio compared to cRNA-seq.

The total RNA count for each annotated TSS was calculated by counting the number of reads whose 5' end is within a +/- 5 bp window of a given TiSS. Subsequently, for each transcript, Uridine to cytosine (U-to-C) conversion rates, error rates, and new-to-total RNA ratios (NTRs) were estimated by analyzing dSLAM-seq data using GRAND-SLAM [33]. Only reads with 5' ends inside +/- 5 bp window of an annotated TiSS were considered. Newly synthesized RNA count of each TiSS was then calculated by multiplying NTR value with total RNA count obtained from dSLAM-seq data.

We grouped TiSS into three groups based on their expression level (**Fig 4A**). For each group, we generated sequence logos from the -34 to +5 bp window around TiSS using WebLogo [57].

All n = 365 TiSS were clustered using the k-means clustering algorithm [58] into four transcription classes (CL1-4) based on new RNA expression (**Fig 4D**). Clustering was repeated 10,000 times with different random initializations. Cluster centroids from each clustering replication were then clustered again one more time to obtain a consensus centroid. This consensus centroid was used for final clustering of TiSS. Classification into TR0-TR5 (**Fig 5A**) was performed via manual curation and application of the following cut-offs. Immediate-early genes were classified as TR0 based on enrichment upon CHX treatment. TR1 included TiSS with peak expression at 4 hpi followed by downregulation by at least 2-fold, such that expression at 4 hpi was more than twice the maximal expression for 6–72 hpi. TR2 included TiSS that did not classify as TR1 but had expression at 4 hpi that was at least 25% of maximal expression at later times of infection (6–72 hpi). TR3 expression uniquely initiated between 6 and 12 hpi. For these TiSS, expression during the first 6 h of infection was less than 25% of the expression thereafter (12–72 hpi), while expression at 12 hpi was at least >50% of maximal expression thereafter (18–72 hpi). TR4 expression did not rise to relevant levels until 18 hpi, i.e., maximal expression for 1–12 hpi < 20% of maximal expression for 18–72 hpi. Moreover, expression between 18–24 hpi was already >50% of the maximal expression thereafter (36–72 hpi). TR5 TiSS only reached maximal expression late in infection (36–72 hpi). Maximal expression at 36–72 hpi was at least >50% of maximal expression at 18–24 hpi (i.e., not TR4).

For each TiSS cluster, promoter motifs and their location were searched using MEME [59] with *-evt 0.01*, *-nmotifs 5*, *-minw 4*, and *-maxw 7* parameters. We searched the motifs inside the -34 to +5 bp window of a given TiSS. In addition, we also generated sequence logos from each cluster using the same window. CL3 and CL4 were analyzed further by grouping each of them into four groups (quantiles) based on the expression values (**S5A Fig**). Sequence logos were generated using the same procedure as mentioned before.

We used our in-house tool PRICE version 1.0.4 [43] to predict MCMV ORFs. A list of putative ORFs was then manually inspected by using the MCMV genome viewer to select *bona-fide* ORFs which were then included in the final annotation. We grouped these ORFs into CDS (ORFs which are included in previous annotation), ORF (ORFs with length ≧ 100 amino acids (aa) which are not in previous annotation), sORF (ORFs with length < 100 aa), uORF (ORFs located upstream of the canonical ORF, but inside the transcript region), uoORF (ORFs located upstream of the canonical ORF and also overlap the canonical ORF but in a different frame), iORF (ORFs located inside a canonical ORF but in a different frame), and dORF (ORFs located downstream of the canonical ORF, but inside the transcript region).

Identification of poly(A) sites and splicing events

4sU-seq reads were first filtered for rRNA reads by aligning reads against rRNA sequences using BWA [60] with a seed size (parameter -k) of 25. If both reads in a read pair aligned to rRNA without errors, they were removed from further analysis. Filtered 4sU-seq reads and all total RNA-seq reads were aligned against the MCMV genome using ContextMap version 2.7.9 [61] (using BWA as short read aligner and allowing at most 5 mismatches and a maximum indel size of 3). ContextMap also identifies reads containing part of the poly(A) tail and predicts poly(A) sites from these reads as previously described [61]. Default parameters were used for poly(A) site prediction. Candidate splice junctions were predicted if >10 reads were identified by ContextMap in at least one sample that overlapped at least 10 nt on both sides of junction. All viral introns are listed in **S3 Table.** All viral poly(A) sites are listed in **S10 Table.**

Supporting information

S1 File. Description of high-throughput sequencing datasets used in this study.
(DOCX)

S2 File. Fully annotated gb file of the BAC of the MCMV reference genome adapted from KY348373 [49] and corrected.
(GB)

S3 File. Fully annotated gb file of the MCMV reference genome adapted from KY348373 [49] and corrected (without the BAC sequence).
(GB)

S1 Table. List of all MCMV transcripts. List of all identified and annotated viral transcripts. TSS = transcription start site; TTS = transcription termination site. Transcription start sites with no evidence of downstream ORFs were annotated as 'orphans'. Alternative transcription start sites of a canonical transcript (RNA) were annotated with a * (TiSS upstream the canonical TiSS) or # (TiSS downstream the canonical TiSS). Scores obtained from iTiSS transcription start site calling are shown for all 380 transcripts. All coordinates are shown in the 0-based system.
(XLSX)

S2 Table. List of all splicing events annotated and identified through 4sU-seq analysis. List of all splicing events that were included into the new MCMV reference genome annotation. All coordinates are shown in the 0-based system.
(XLSX)

S3 Table. List of putative introns detected by 4sU-seq. List of all putative splicing events that were not included into the new MCMV genome reference annotation. All coordinates are shown in the 0-based system.
(XLSX)

S4 Table. Kinetic clusters of MCMV transcripts. The table depicts all TSS annotations, their genomic location and temporal kinetic classes (CL and TR). TATA-box motifs have been indicated for each TiSS. Scores derived from iTiSS for each TiSS are shown in column F. New RNA (I-AG) and total RNA (AH-BC) read counts are displayed for each replicate per time point of dSLAM-seq datasets. Ratios for 4 hpi samples (+/-CHX) computed using new RNA are displayed in column AG and ratios for 24 hpi samples (+/-PAA) computed using total RNA are displayed in column BF. All coordinates are shown in the 0-based system.
(XLSX)

S5 Table. List of all MCMV ORFs. List of all MCMV ORFs including their name, ORF type, coordinates, strand and length of predicted protein products. ORF types included previously annotated ORFs from Rawlinson *et al* [22]. (CDS), novel large ORFs (>100aa) and ORFs validated from previously studies (ORFs), ORFs <100 aa annotated as short ORFs (sORFs), upstream/upstream overlapping ORFs (uORFs/uoORFs), internal ORFs (iORFs), downstream ORFs (dORFs). N-terminally truncated or extended ORFs were annotated with a '#1','#2',. . . or '*1','*2',. . ., respectively. All coordinates are shown in the 0-based system.
(XLSX)

S6 Table. List of all unidentified CDS predicted by Rawlinson *et al* [22]. List of all CDS predicted by Rawlinson *et al*. that were not observed in our data. Their respective CDS were nevertheless included in our revised MCMV genome annotation as 'not expressed/orphan' CDS. *

CDS that are overlapping other MCMV genes by greater than 60% and are thus less likely to be protein coding and have no homologs in herpesviruses or cellular proteins as described by Rawlinson *et al*. All coordinates are shown in the 0-based system.
(XLSX)

S7 Table. List of previously unannotated ORFs confirmed by us and validated in several studies. Table of all MCMV ORFs that have been identified by other studies and that were confirmed by our data. All coordinates are shown in the 0-based system.
(XLSX)

S8 Table. Detected ORFs with minor corrections i.e. 'CDS (corrected)' as verified by previous studies. Table of all previously reported MCMV CDS that required minor corrections based on our data. * Initiation at a downstream AUG. All coordinates are shown in the 0-based system.
(XLSX)

S9 Table. List of primers and gene constructs used.
(XLSX)

S10 Table. List of MCMV poly(A) sites. List of all MCMV poly(A) sites that were annotated based on untemplated adenines on sequencing reads. All coordinates are shown in the 0-based system.
(XLSX)

S1 Fig. Characterization of the MCMV transcriptome. A. Heat maps comparing read enrichment at transcription start sites (TiSS) in the cRNA-seq and dSLAM-seq data. The x axis represents distance from TiSS (+/-100 bp) for 222 TiSS displayed along the y-axis detected by both dSLAM-seq and cRNA-seq. Colors represent maximal read count in $\log_2$ scale across all samples in cRNA-seq and dSLAM-seq, respectively. **B**. Histogram depiction of the number of MCMV TiSS satisfying the indicated number of criteria of the iTiSS algorithm. A detailed description of the employed criteria is included in methods. Scores for all TiSS assigned by iTiSS are shown in **S4 Table**.
(TIF)

S2 Fig. Examples of MCMV splicing events. Each schematic depicts viral gene expression and splicing in a given locus. Aggregated reads of Ribo-seq, cRNA-seq and dSLAM-seq data across all time points of infection are shown. Ribo-seq data are indicated in logarithmic scale, cRNA-seq and dSLAM-seq data in linear scale. The arrows at the top depict the annotated transcripts (black), poly(A) sites (PAS; in red) and ORFs (colored depending on the translated frame (yellow, purple and green)). The bold dotted line represents introns detected by 4sU-seq. **A**. In the m133 locus, splicing of two introns leads to the expression of both a known (Iso1) and a novel spliced ORF (Iso2), the latter is expressed by an alternative donor site, as predicted by Rawlinson *et al*. [22]. The m133 RNA #1 (orphan) transcript did not bear any evidence for downstream translational start sites and was annotated as a spliced transcript with delayed late kinetics, spliced similarly to m133 RNA Iso2 as evident from the overall read accumulation in cRNA-seq and dSLAM-seq (48 hpi) **B**. In the M116 locus, splicing explained a truncated M116 CDS Iso2 (M116.1p) revealed by ribosome profiling, whose transcript may terminate at an earlier poly(A) site (PAS). A second PAS downstream serves for the transcript encoding the unspliced M116 CDS. Here, transcription continues past the first PAS resulting in M116 RNA Iso1 which overlaps with the M115 and M114 transcripts whose expression levels correspond to their respective ORFs and were hence annotated. High levels of gene expression across the M116 locus resulted in a number of putative TiSS depicted in the dSLAM-seq

with no evidence of downstream translation, uniform cRNA-seq read distribution and overall lower gene expression as compared to the canonical TiSS. Hence, they were attributed to experimental noise and not annotated. **C**. In the m147.5 locus, splicing leads to the expression of a previously validated spliced ORF. **D**. In the m124 locus, splicing necessitates correction of the previously annotated m124 ORF. No TiSS were identified due to very low levels of transcriptional activity across this locus. Coordinates of the start codons and splicing acceptor and donor sites are displayed.
(TIF)

S3 Fig. Splicing events in the m60-m73.5 locus. Graphs represent TiSS profiling data (black) from cRNA-seq and dSLAM-seq as well as ORFs called by Ribo-Seq (different colors represent different frames of translation). Aggregated reads of Ribo-seq, cRNA-seq and dSLAM-seq data across all time points of infection are shown. Ribo-seq data are indicated in logarithmic scale, cRNA-seq and dSLAM-seq data in both linear and logarithmic scale. **A.** Spliced ORFs are depicted by exons connected with a dotted line representing introns at the bottom. Multiple splicing events were observed in the m60-73.5 locus, of which the m60 RNA and M73-m73.5 spliced transcripts have already been validated previously (see **S2 Table**). Of note, translation occurs in different fames upstream of splicing thereby explaining translation in different frames in the common downstream exon. For a given frame of translation at the second exon, the expression levels correlated well with the respective upstream exons. **B.** Shown is a zoomed schematic view of the dSLAM-seq data portraying every TiSS depicted in panel **A** in linear scale. Coordinates of the start codons/exon start sites and TiSS for all spliced events are indicated.
(TIF)

S4 Fig. The m166.5 RNA constitutes a novel *ie* gene (*ie4*). A. Cycloheximide (CHX) treatment combined with dSLAM-seq identified a so far unknown viral immediate early transcript in the m166.5 locus. Aggregated reads of Ribo-seq, cRNA-seq and dSLAM-seq data across all time points of infection in the m166-m167 locus are shown. Ribo-seq data are depicted in log scale, cRNA-seq and dSLAM-seq data (-/+CHX; 4 hpi) in linear scale. The m166.5 immediate-early transcript (termed *ie4*) and its corresponding m166.5 ORF overlap with the m167 CDS (orphan) and partially overlap with the N-terminal part of the m166 CDS. Unlike the m166 RNA, m166.5 RNA is expressed despite CHX pre-treatment implying immediate-early gene kinetics. **B.** Line graphs representing gene expression (new RNA levels) of the three *ie* genes (four when counting *ie1* and *ie3* as separate genes despite their use of the same TiSS) over time for two replicates per gene. New RNA expression levels of phosphonoacetic acid (PAA; 24 hpi)- or CHX (4 hpi)-treated samples (n = 1) are indicated as a star/triangle respectively for a given gene. Increase in new RNA levels for a given gene under CHX treatment defines them as immediate-early genes (TR0). Coordinates of the start codons and TiSS of all indicated gene products are displayed.
(TIF)

S5 Fig. Clusters CL3 and CL4 represent TiSS with distinct expression kinetics. A. Quantile groups segregated according to levels of expression for CL3 and CL4 transcripts on the basis of new RNA for the respective viral TiSS obtained from dSLAM-seq data. The x-axis displays hours post infection. Relative expression is shown on the y-axis. CL3 and CL4 transcripts are indicated by green and blue lines, respectively. **B.** Motif analysis (MEME) for all four quantiles for CL3 and CL4 transcripts. NS: Not significant.
(TIF)

S6 Fig. Kinetics of all TiSS in the clusters TR0-TR5. Line graphs representing all TiSS within the six TR clusters (TR0-5). The y-axis represents relative new RNA levels across the infection time-course (x-axis).
(TIF)

S7 Fig. Gene expression in the MCMV *ie2* locus. A. Schematic of the *ie2* locus. The canonical TiSS is represented by the dominant spliced *ie2* transcript (m126-m128 RNA) comprising 2 introns and only one *ie2* coding exon initiating at the first AUG (186087) shown i.e., m128 CDS (*ie2* Exon 3). A second AUG represents a truncated isoform (m128 CDS #1 RNA #1). Aggregated reads of Ribo-seq, cRNA-seq and dSLAM-seq data across all time points of infection are shown. Ribo-seq data are indicated in log scale, cRNA-seq and dSLAM-seq data in linear scale. **B.** Graphs represent TiSS profiling data (black) from dSLAM-seq including kinetics for 6 and 24 hpi and ORFs called by Ribo-seq (Colored) for the same time points for a given replicate. The arrows above depict manual annotations of transcripts and ORFs. Alternative transcription initiation at the *ie2* (m128) locus led to the expression of an N-terminally truncated ORF expressed from an early TiSS (m128 RNA #1) whose expression was not influenced by PAA treatment. cRNA-seq and dSLAM-seq data are represented in linear scale, Ribo-seq in logarithmic scale. Coordinates of all TiSS, start codons and spliced junctions in the m126-m128 locus are displayed.
(TIF)

S8 Fig. Validation of m145 ORFs. Both the m145 CDS and m145 ORF #1 were cloned into expression plasmids (pCREL-IRES-Neon) with their expression driven by a CMV promoter. **A.** Expression of the two viral ORFs was validated via transfection of the respective plasmids into HEK293T cells. Western blots were performed at 48 h post transfection. Both the 55 kDa (non-glycosylated) and 70 kDa isoforms of m145 CDS were detected whereas only a single 20 kDa isoform of m145 ORF #1 was detected. **B.** The m145-V5 virus described in Fig 7B was used to infect both NIH-3T3 and SVEC 4–10 cells at an MOI of 1 for the respective time points to validate the m145 gene products, whose expression was similar in both cell lines. **C**. A start codon mutant of m145 ORF #1 (△m145 ORF #1 mut) was utilized to infect SVEC 4–10 cells for 48 h. Western blot analysis revealed expression of m145 ORF #2 to remain unaffected. WT indicates wild-type MCMV. β-actin was used as a housekeeping control. Images are a single representative of 2 biological replicates (n = 2) for each experiment. Additional gene products (35 kDa and 13 kDa) of m145 RNA #1 were not detected in the expression plasmid system in S7A. Absence of the 13 kDa isoform (m145 ORF #2) is likely to be due to the optimized Kozak sequence of the employed expression vector, which prevents ribosomes from bypassing the main AUG start codon. The reason for the absence of the 35 kDa isoform remains unclear. We hypothesize that this may be due to differences in the employed cell system or requirements for other viral gene products.
(TIF)

Author Contributions

Conceptualization: Manivel Lodha, Vanda Juranic Lisnic, Anne L'Hernault, Andrzej J. Rutkowski, Thomas Hennig, Caroline C. Friedel, Florian Erhard, Lars Dölken.

Data curation: Manivel Lodha, Ihsan Muchsin, Christopher Jürges, Vanda Juranic Lisnic, Anne L'Hernault, Thomas Hennig, Caroline C. Friedel, Florian Erhard, Lars Dölken.

Formal analysis: Manivel Lodha, Ihsan Muchsin, Christopher Jürges, Anne L'Hernault, Thomas Hennig, Caroline C. Friedel, Florian Erhard, Lars Dölken.

Funding acquisition: Florian Erhard, Lars Dölken.

Investigation: Manivel Lodha, Anne L'Hernault, Andrzej J. Rutkowski, Bhupesh K. Prusty, Arnhild Grothey, Andrea Milic, Thomas Hennig, Lars Dölken.

Methodology: Manivel Lodha, Christopher Jürges, Vanda Juranic Lisnic, Anne L'Hernault, Andrzej J. Rutkowski, Thomas Hennig, Caroline C. Friedel, Florian Erhard, Lars Dölken.

Project administration: Stipan Jonjic, Florian Erhard.

Resources: Manivel Lodha, Bhupesh K. Prusty, Stipan Jonjic, Caroline C. Friedel, Florian Erhard.

Software: Manivel Lodha, Ihsan Muchsin, Christopher Jürges, Caroline C. Friedel, Florian Erhard.

Supervision: Florian Erhard, Lars Dölken.

Validation: Manivel Lodha, Ihsan Muchsin, Vanda Juranic Lisnic, Caroline C. Friedel, Florian Erhard.

Visualization: Manivel Lodha, Ihsan Muchsin, Christopher Jürges, Florian Erhard, Lars Dölken.

Writing – original draft: Manivel Lodha, Ihsan Muchsin, Thomas Hennig, Florian Erhard.

Writing – review & editing: Manivel Lodha, Florian Erhard.

References

1. Griffiths P, Reeves M. Pathogenesis of human cytomegalovirus in the immunocompromised host. Nat Rev Microbiol. 2021; 19(12):759–73. Epub 2021/06/26. https://doi.org/10.1038/s41579-021-00582-z PMID: 34168328; PubMed Central PMCID: PMC8223196 number 2020135.6 assigned to University College London (UCL), entitled 'hCMV antibody and vaccine target', that deals with a novel antigenic domain on HCMV glycoprotein B (gB). UCL received funds from Takeda pharmaceuticals to compensate for the time P.G. spent as a member of the end-point committee for a randomized clinical trial (RCT) of maribavir. The authors declare no other competing interests.
2. Tang Q, Maul GG. Mouse cytomegalovirus crosses the species barrier with help from a few human cytomegalovirus proteins. J Virol. 2006; 80(15):7510–21. Epub 2006/07/15. https://doi.org/10.1128/JVI.00684-06 PMID: 16840331; PubMed Central PMCID: PMC1563706.
3. Brune W, Hengel H, Koszinowski UH. A mouse model for cytomegalovirus infection. Curr Protoc Immunol. 2001;Chapter 19:Unit 19.7. Epub 2008/04/25. https://doi.org/10.1002/0471142735.im1907s43 PMID: 18432758.
4. Forte E, Zhang Z, Thorp EB, Hummel M. Cytomegalovirus Latency and Reactivation: An Intricate Interplay With the Host Immune Response. Front Cell Infect Microbiol. 2020; 10:130. Epub 2020/04/17. https://doi.org/10.3389/fcimb.2020.00130 PMID: 32296651; PubMed Central PMCID: PMC7136410.
5. Angulo A, Ghazal P, Messerle M. The major immediate-early gene ie3 of mouse cytomegalovirus is essential for viral growth. J Virol. 2000; 74(23):11129–36. Epub 2000/11/09. https://doi.org/10.1128/jvi.74.23.11129-11136.2000 PMID: 11070009; PubMed Central PMCID: PMC113196.
6. Chapa TJ, Johnson LS, Affolter C, Valentine MC, Fehr AR, Yokoyama WM, et al. Murine cytomegalovirus protein pM79 is a key regulator for viral late transcription. J Virol. 2013; 87(16):9135–47. Epub 2013/06/14. https://doi.org/10.1128/JVI.00688-13 PMID: 23760242; PubMed Central PMCID: PMC3754071.
7. Chapa TJ, Perng YC, French AR, Yu D. Murine cytomegalovirus protein pM92 is a conserved regulator of viral late gene expression. J Virol. 2014; 88(1):131–42. Epub 2013/10/18. https://doi.org/10.1128/JVI.02684-13 PMID: 24131717; PubMed Central PMCID: PMC3911726.
8. Pan D, Han T, Tang S, Xu W, Bao Q, Sun Y, et al. Murine Cytomegalovirus Protein pM91 Interacts with pM79 and Is Critical for Viral Late Gene Expression. J Virol. 2018;92(18). Epub 2018/07/13. https://doi.org/10.1128/JVI.00675-18 PMID: 29997217; PubMed Central PMCID: PMC6146718.
9. Han T, Hao H, Sleman SS, Xuan B, Tang S, Yue N, et al. Murine Cytomegalovirus Protein pM49 Interacts with pM95 and Is Critical for Viral Late Gene Expression. J Virol. 2020;94(6). Epub 2020/01/04. https://doi.org/10.1128/JVI.01956-19 PMID: 31896598; PubMed Central PMCID: PMC7158740.

10. Li M, Hu Q, Collins G, Parida M, Ball CB, Price DH, et al. Cytomegalovirus late transcription factor target sequence diversity orchestrates viral early to late transcription. PLoS Pathog. 2021; 17(8):e1009796. Epub 2021/08/03. https://doi.org/10.1371/journal.ppat.1009796 PMID: 34339482; PubMed Central PMCID: PMC8360532.
11. In: Arvin A, Campadelli-Fiume G, Mocarski E, Moore PS, Roizman B, Whitley R, et al., editors. Human Herpesviruses: Biology, Therapy, and Immunoprophylaxis. Cambridge: Cambridge University Press; 2007.
12. Weekes MP, Tomasec P, Huttlin EL, Fielding CA, Nusinow D, Stanton RJ, et al. Quantitative temporal viromics: an approach to investigate host-pathogen interaction. Cell. 2014; 157(6):1460–72. Epub 2014/06/07. https://doi.org/10.1016/j.cell.2014.04.028 PMID: 24906157; PubMed Central PMCID: PMC4048463.
13. Phan QV, Bogdanow B, Wyler E, Landthaler M, Liu F, Hagemeier C, et al. Engineering, decoding and systems-level characterization of chimpanzee cytomegalovirus. PLoS Pathog. 2022; 18(1):e1010193. Epub 2022/01/05. https://doi.org/10.1371/journal.ppat.1010193 PMID: 34982803; PubMed Central PMCID: PMC8759705.
14. Ingolia NT, Brar GA, Rouskin S, McGeachy AM, Weissman JS. The ribosome profiling strategy for monitoring translation in vivo by deep sequencing of ribosome-protected mRNA fragments. Nat Protoc. 2012; 7(8):1534–50. Epub 2012/07/28. https://doi.org/10.1038/nprot.2012.086 PMID: 22836135; PubMed Central PMCID: PMC3535016.
15. Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet. 2009; 10(1):57–63. Epub 2008/11/19. https://doi.org/10.1038/nrg2484 PMID: 19015660; PubMed Central PMCID: PMC2949280.
16. Stern-Ginossar N, Weisburd B, Michalski A, Le VT, Hein MY, Huang SX, et al. Decoding human cytomegalovirus. Science. 2012; 338(6110):1088–93. Epub 2012/11/28. https://doi.org/10.1126/science.1227919 PMID: 23180859; PubMed Central PMCID: PMC3817102.
17. Whisnant AW, Jürges CS, Hennig T, Wyler E, Prusty B, Rutkowski AJ, et al. Integrative functional genomics decodes herpes simplex virus 1. Nat Commun. 2020; 11(1):2038. Epub 2020/04/29. https://doi.org/10.1038/s41467-020-15992-5 PMID: 32341360; PubMed Central PMCID: PMC7184758.
18. Arias C, Weisburd B, Stern-Ginossar N, Mercier A, Madrid AS, Bellare P, et al. 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. 2014; 10(1):e1003847. Epub 2014/01/24. https://doi.org/10.1371/journal.ppat.1003847 PMID: 24453964; PubMed Central PMCID: PMC3894221.
19. Bencun M, Klinke O, Hotz-Wagenblatt A, Klaus S, Tsai MH, Poirey R, et al. Translational profiling of B cells infected with the Epstein-Barr virus reveals 5' leader ribosome recruitment through upstream open reading frames. Nucleic Acids Res. 2018; 46(6):2802–19. Epub 2018/03/13. https://doi.org/10.1093/nar/gky129 PMID: 29529302; PubMed Central PMCID: PMC5887285.
20. Nightingale K, Lin KM, Ravenhill BJ, Davies C, Nobre L, Fielding CA, et al. High-Definition Analysis of Host Protein Stability during Human Cytomegalovirus Infection Reveals Antiviral Factors and Viral Evasion Mechanisms. Cell Host Microbe. 2018; 24(3):447–60.e11. Epub 2018/08/21. https://doi.org/10.1016/j.chom.2018.07.011 PMID: 30122656; PubMed Central PMCID: PMC6146656.
21. Lodha M, Erhard F, Dölken L, Prusty BK. The Hidden Enemy Within: Non-canonical Peptides in Virus-Induced Autoimmunity. Front Microbiol. 2022; 13:840911. Epub 2022/03/01. https://doi.org/10.3389/fmicb.2022.840911 PMID: 35222346; PubMed Central PMCID: PMC8866975.
22. Rawlinson WD, Farrell HE, Barrell BG. Analysis of the complete DNA sequence of murine cytomegalovirus. J Virol. 1996; 70(12):8833–49. Epub 1996/12/01. https://doi.org/10.1128/JVI.70.12.8833-8849.1996 PMID: 8971012; PubMed Central PMCID: PMC190980.
23. Tang Q, Murphy EA, Maul GG. Experimental confirmation of global murine cytomegalovirus open reading frames by transcriptional detection and partial characterization of newly described gene products. J Virol. 2006; 80(14):6873–82. Epub 2006/07/01. https://doi.org/10.1128/JVI.00275-06 PMID: 16809293; PubMed Central PMCID: PMC1489029.
24. Lacaze P, Forster T, Ross A, Kerr LE, Salvo-Chirnside E, Lisnic VJ, et al. Temporal profiling of the coding and noncoding murine cytomegalovirus transcriptomes. J Virol. 2011; 85(12):6065–76. Epub 2011/04/08. https://doi.org/10.1128/JVI.02341-10 PMID: 21471238; PubMed Central PMCID: PMC3126304.
25. Brocchieri L, Kledal TN, Karlin S, Mocarski ES. Predicting coding potential from genome sequence: application to betaherpesviruses infecting rats and mice. J Virol. 2005; 79(12):7570–96. Epub 2005/05/28. https://doi.org/10.1128/JVI.79.12.7570-7596.2005 PMID: 15919911; PubMed Central PMCID: PMC1143683.
26. Kattenhorn LM, Mills R, Wagner M, Lomsadze A, Makeev V, Borodovsky M, et al. Identification of proteins associated with murine cytomegalovirus virions. J Virol. 2004; 78(20):11187–97. Epub 2004/09/

29. https://doi.org/10.1128/JVI.78.20.11187-11197.2004 PMID: 15452238; PubMed Central PMCID: PMC521832.

27. Kutle I, Sengstake S, Templin C, Glaß M, Kubsch T, Keyser KA, et al. The M25 gene products are critical for the cytopathic effect of mouse cytomegalovirus. Sci Rep. 2017; 7(1):15588. Epub 2017/11/16. https://doi.org/10.1038/s41598-017-15783-x PMID: 29138436; PubMed Central PMCID: PMC5686157.

28. Železnjak J, Lisnić VJ, Popović B, Lisnić B, Babić M, Halenius A, et al. The complex of MCMV proteins and MHC class I evades NK cell control and drives the evolution of virus-specific activating Ly49 receptors. J Exp Med. 2019; 216(8):1809–27. Epub 2019/05/31. https://doi.org/10.1084/jem.20182213 PMID: 31142589; PubMed Central PMCID: PMC6683999.

29. Juranic Lisnic V, Babic Cac M, Lisnic B, Trsan T, Mefferd A, Das Mukhopadhyay C, et al. Dual analysis of the murine cytomegalovirus and host cell transcriptomes reveal new aspects of the virus-host cell interface. PLoS Pathog. 2013; 9(9):e1003611. Epub 2013/10/03. https://doi.org/10.1371/journal.ppat.1003611 PMID: 24086132; PubMed Central PMCID: PMC3784481.

30. Rutkowski AJ, Erhard F, L'Hernault A, Bonfert T, Schilhabel M, Crump C, et al. Widespread disruption of host transcription termination in HSV-1 infection. Nat Commun. 2015; 6:7126. Epub 2015/05/21. https://doi.org/10.1038/ncomms8126 PMID: 25989971; PubMed Central PMCID: PMC4441252.

31. Sharma CM, Vogel J. Differential RNA-seq: the approach behind and the biological insight gained. Curr Opin Microbiol. 2014; 19:97–105. Epub 2014/07/16. https://doi.org/10.1016/j.mib.2014.06.010 PMID: 25024085.

32. Herzog VA, Reichholf B, Neumann T, Rescheneder P, Bhat P, Burkard TR, et al. Thiol-linked alkylation of RNA to assess expression dynamics. Nat Methods. 2017; 14(12):1198–204. Epub 2017/09/26. https://doi.org/10.1038/nmeth.4435 PMID: 28945705; PubMed Central PMCID: PMC5712218.

33. Jürges C, Dölken L, Erhard F. Dissecting newly transcribed and old RNA using GRAND-SLAM. Bioinformatics. 2018; 34(13):i218–i26. Epub 2018/06/29. https://doi.org/10.1093/bioinformatics/bty256 PMID: 29949974; PubMed Central PMCID: PMC6037110.

34. Jürges CS, Dölken L, Erhard F. Integrarative transcription start site identification with iTiSS. Bioinformatics. 2021. Epub 2021/03/16. https://doi.org/10.1093/bioinformatics/btab170 PMID: 33720332.

35. Ružić T, Juranić Lisnić V, Mahmutefendić Lučin H, Lenac Roviš T, Železnjak J, Cokarić Brdovčak M, et al. Characterization of M116.1p, a Murine Cytomegalovirus Protein Required for Efficient Infection of Mononuclear Phagocytes. J Virol. 2022; 96(2):e0087621. Epub 2021/10/28. https://doi.org/10.1128/JVI.00876-21 PMID: 34705561; PubMed Central PMCID: PMC8791281.

36. Loewendorf A, Krüger C, Borst EM, Wagner M, Just U, Messerle M. Identification of a mouse cytomegalovirus gene selectively targeting CD86 expression on antigen-presenting cells. J Virol. 2004; 78(23):13062–71. Epub 2004/11/16. https://doi.org/10.1128/JVI.78.23.13062-13071.2004 PMID: 15542658; PubMed Central PMCID: PMC524971.

37. Kulesza CA, Shenk T. Murine cytomegalovirus encodes a stable intron that facilitates persistent replication in the mouse. Proc Natl Acad Sci U S A. 2006; 103(48):18302–7. Epub 2006/11/16. https://doi.org/10.1073/pnas.0608718103 PMID: 17105807; PubMed Central PMCID: PMC1838746.

38. Schwarz TM, Volpe LA, Abraham CG, Kulesza CA. Molecular investigation of the 7.2 kb RNA of murine cytomegalovirus. Virol J. 2013; 10:348. Epub 2013/12/04. https://doi.org/10.1186/1743-422x-10-348 PMID: 24295514; PubMed Central PMCID: PMC4220806.

39. Smale ST, Kadonaga JT. The RNA polymerase II core promoter. Annu Rev Biochem. 2003; 72:449–79. Epub 2003/03/26. https://doi.org/10.1146/annurev.biochem.72.121801.161520 PMID: 12651739.

40. Carninci P, Sandelin A, Lenhard B, Katayama S, Shimokawa K, Ponjavic J, et al. Genome-wide analysis of mammalian promoter architecture and evolution. Nat Genet. 2006; 38(6):626–35. Epub 2006/04/29. https://doi.org/10.1038/ng1789 PMID: 16645617.

41. Gruffat H, Marchione R, Manet E. Herpesvirus Late Gene Expression: A Viral-Specific Pre-initiation Complex Is Key. Front Microbiol. 2016; 7:869. Epub 2016/07/05. https://doi.org/10.3389/fmicb.2016.00869 PMID: 27375590; PubMed Central PMCID: PMC4893493.

42. Jürges CS, Lodha M, Khanh Le-Trilling VT, Bhandare P, Wolf E, Zimmermann A, et al. Multi-omics reveals principles of gene regulation and pervasive non-productive transcription in the human cytomegalovirus genome. bioRxiv. 2022:2022.01.07.472583. https://doi.org/10.1101/2022.01.07.472583

43. Erhard F, Halenius A, Zimmermann C, L'Hernault A, Kowalewski DJ, Weekes MP, et al. Improved Ribo-seq enables identification of cryptic translation events. Nat Methods. 2018; 15(5):363–6. Epub 2018/03/13. https://doi.org/10.1038/nmeth.4631 PMID: 29529017; PubMed Central PMCID: PMC6152898.

44. Messerle M, Keil GM, Koszinowski UH. Structure and expression of murine cytomegalovirus immediate-early gene 2. J Virol. 1991; 65(3):1638–43. Epub 1991/03/01. https://doi.org/10.1128/JVI.65.3.1638-1643.1991 PMID: 1847480; PubMed Central PMCID: PMC239953.

45. Krmpotic A, Hasan M, Loewendorf A, Saulig T, Halenius A, Lenac T, et al. NK cell activation through the NKG2D ligand MULT-1 is selectively prevented by the glycoprotein encoded by mouse cytomegalovirus gene m145. J Exp Med. 2005; 201(2):211–20. Epub 2005/01/12. https://doi.org/10.1084/jem.20041617 PMID: 15642742; PubMed Central PMCID: PMC2212792.

46. Vattem KM, Wek RC. Reinitiation involving upstream ORFs regulates ATF4 mRNA translation in mammalian cells. Proc Natl Acad Sci U S A. 2004; 101(31):11269–74. Epub 2004/07/28. https://doi.org/10.1073/pnas.0400541101 PMID: 15277680; PubMed Central PMCID: PMC509193.

47. Young SK, Wek RC. Upstream Open Reading Frames Differentially Regulate Gene-specific Translation in the Integrated Stress Response. J Biol Chem. 2016; 291(33):16927–35. Epub 2016/07/01. https://doi.org/10.1074/jbc.R116.733899 PMID: 27358398; PubMed Central PMCID: PMC5016099.

48. Dölken L, Malterer G, Erhard F, Kothe S, Friedel CC, Suffert G, et al. Systematic analysis of viral and cellular microRNA targets in cells latently infected with human gamma-herpesviruses by RISC immunoprecipitation assay. Cell Host Microbe. 2010; 7(4):324–34. Epub 2010/04/24. https://doi.org/10.1016/j.chom.2010.03.008 PMID: 20413099.

49. Jordan S, Krause J, Prager A, Mitrovic M, Jonjic S, Koszinowski UH, et al. Virus progeny of murine cytomegalovirus bacterial artificial chromosome pSM3fr show reduced growth in salivary Glands due to a fixed mutation of MCK-2. J Virol. 2011; 85(19):10346–53. Epub 2011/08/05. https://doi.org/10.1128/JVI.00545-11 PMID: 21813614; PubMed Central PMCID: PMC3196435.

50. Parida M, Nilson KA, Li M, Ball CB, Fuchs HA, Lawson CK, et al. 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. 2019; 10(1). Epub 2019/02/14. https://doi.org/10.1128/mBio.02047-18 PMID: 30755505; PubMed Central PMCID: PMC6372792.

51. Isomura H, Stinski MF, Murata T, Yamashita Y, Kanda T, Toyokuni S, et al. The human cytomegalovirus gene products essential for late viral gene expression assemble into prereplication complexes before viral DNA replication. J Virol. 2011; 85(13):6629–44. Epub 2011/04/22. https://doi.org/10.1128/JVI.00384-11 PMID: 21507978; PubMed Central PMCID: PMC3126524.

52. Hiršl L, Brizić I, Jenuš T, Juranić Lisnić V, Reichel JJ, Jurković S, et al. Murine CMV Expressing the High Affinity NKG2D Ligand MULT-1: A Model for the Development of Cytomegalovirus-Based Vaccines. Front Immunol. 2018; 9:991. Epub 2018/06/06. https://doi.org/10.3389/fimmu.2018.00991 PMID: 29867968; PubMed Central PMCID: PMC5949336.

53. Crosby LN, McCormick AL, Mocarski ES. Gene products of the embedded m41/m41.1 locus of murine cytomegalovirus differentially influence replication and pathogenesis. Virology. 2013; 436(2):274–83. Epub 2013/01/09. https://doi.org/10.1016/j.virol.2012.12.002 PMID: 23295021; PubMed Central PMCID: PMC3557549.

54. Yewdell JW, Antón LC, Bennink JR. Defective ribosomal products (DRiPs): a major source of antigenic peptides for MHC class I molecules? J Immunol. 1996; 157(5):1823–6. Epub 1996/09/01. PMID: 8757297.

55. Yewdell JW. Immunology. Hide and seek in the peptidome. Science. 2003; 301(5638):1334–5. Epub 2003/09/06. https://doi.org/10.1126/science.1089553 PMID: 12958347.

56. Tischer BK, Smith GA, Osterrieder N. En passant mutagenesis: a two step markerless red recombination system. Methods Mol Biol. 2010; 634:421–30. Epub 2010/08/03. https://doi.org/10.1007/978-1-60761-652-8_30 PMID: 20677001.

57. Crooks GE, Hon G, Chandonia JM, Brenner SE. WebLogo: a sequence logo generator. Genome Res. 2004; 14(6):1188–90. Epub 2004/06/03. https://doi.org/10.1101/gr.849004 PMID: 15173120; PubMed Central PMCID: PMC419797.

58. Fabian Pedregosa GV, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, Édouard Duchesnay. Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research. 2011; 12(85):2825–30.

59. Bailey TL, Johnson J, Grant CE, Noble WS. The MEME Suite. Nucleic Acids Res. 2015; 43(W1):W39–49. Epub 2015/05/09. https://doi.org/10.1093/nar/gkv416 PMID: 25953851; PubMed Central PMCID: PMC4489269.

60. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009; 25(14):1754–60. Epub 2009/05/20. https://doi.org/10.1093/bioinformatics/btp324 PMID: 19451168; PubMed Central PMCID: PMC2705234.

61. Bonfert T, Kirner E, Csaba G, Zimmer R, Friedel CC. ContextMap 2: fast and accurate context-based RNA-seq mapping. BMC Bioinformatics. 2015; 16:122. Epub 2015/05/01. https://doi.org/10.1186/s12859-015-0557-5 PMID: 25928589; PubMed Central PMCID: PMC4411664.