UNIVERSITY OF RIJEKA
SCHOOL OF MEDICINE

Vanda Juranić Lisnić

ANALYSIS OF MURINE CYTOMEGALOVIRUS TRANSCRIPTOME

DOCTORAL THESIS

Mentors:
Prof.Dr.Sc. Joanne Trgovcich
Prof.Dr.Sc. Astrid Krmpotić

Rijeka, 2013

Mentors:
Prof.Dr.Sc. Joanne Trgovcich, associate prof.
Prof.Dr.Sc. Astrid Krmpotić, full prof.

The doctoral thesis was defended on 8th November 2013. at School of Medicine, University of Rijeka, Croatia in front of the following committee:

1. Prof.Dr.Sc. Bojan Polić
2. Dr.Sc. Magdalena Grce
3. Prof.Dr.Sc. Ivica Pavić
4. Prof.Dr.Sc. Joanne Trgovcich
5. Prof.Dr.Sc. Astrid Krmpotić

Thesis contains 154 pages, 31 figures and 14 tables.

The research described in this thesis was performed at the Department for histology and embryology and Center for proteomics, School of Medicine, University of Rijeka under the supervision of Prof.Dr.Sc. Stipan Jonjić and Prof.Dr.Sc. Astrid Krmpotić and at the Ohio State University under the supervision of Prof.Dr.Sc. Joanne Trgovcich.

This research was financed by the Ministry of Science, Education and Sports, grant "The role of immune-subversive cytomegaloviral genes in latency" (grant no. 062-0621261-1268; project leader Astrid Krmpotić), and Unity Through Knowledge Fund grant "Transcriptomic approach to viral disease research" (project leaders Joanne Trgovcich and Stipan Jonjic, grant no. 08/07)

"*It was the best of times, it was the worst of times, it was the age of wisdom, it was the age of foolishness, it was the epoch of belief, it was the epoch of incredulity, it was the season of Light, it was the season of Darkness, it was the spring of hope, it was the winter of despair"… (C. Dickens, A Tale of Two Cities)*

I OWE MY THANKS TO:

The Boss Stipan Jonjić for a wonderful opportunity to do Science in his group, for his encouragement and trust, for his relentless drive and for not shying away from the hard questions or obstacles of any kind;

My mentors Joanne Trgovcich and Astrid Krmpotić who supported me in more ways than I can count (or thank);

Members of the "Dobrovoljno vatrogasno društvo" of Proteomika and Hista for picking over 100,000 bacterial colonies with me, sorting and FACSing throughout the night, producing hectoliters of various proteins and antibodies and filling endless forms and still managing to make it all fun;

And finally thanks to my family for keeping me sane and accepting to ride the PhD rollercoaster with me.

ANALIZA TRANSKRIPTOMA MIŠJEG CITOMEGALOVIRUSA

PROŠIRENI SAŽETAK

Svrha istraživanja

Humani citomegalovirus široko je rasprostranjen patogen, a posebno je opasan za trudnice, novorođenčad i imunosuprimirane pacijente. Nažalost, efikasnog cjepivo nema, a postoji potreba i za efikasnijim i manje toksičnim antiviralnim lijekovima. Glavne prepreke razvoju novih antiviralnih ljekova i cjepiva jesu: (1) specifičnost za vrstu i (2) praznine u našem znanju i razumjevanju virusnih gena, interakcijama virusnih gena i domaćina te kako te interakcije izazivaju bolest. Prva prepreka uspješno se nadvladava korištenjem animalnih virusa, posebice mišjeg citomegalovirusa (MCMV). S ciljem nadvladavanja druge prepreke u ovom je radu provedena detaljna analiza transkriptoma mišjeg citomegalovirusa (MCMV) te analiza transkriptoma stanica domaćina tijekom litičke infekcije citomegalovirusom.

Materijali i metode

Transkriptom MCMV analiziran je na dva načina: klasičnom analizom cDNK knjižnice i sekvencioniranjem dobivenih klonova te analizom transkriptoma uz pomoć sekvencioniranja nove generacije (odnosno RNK-sekvencioniranjem, eng. *RNASeq*) koja omogućava paralelno praćenje i transkriptoma domaćina uz transkriptom virusa. Analiza transkriptoma domaćina rezultira vrlo dugačkim listama diferencijalno reguliranih gena iz kojih je teško izvući neko biološko značenje. Stoga su liste diferencijalno reguliranih gena domaćina podvrgnute analizi termina genske ontologije (eng. *gene ontology analysis* odnosno GO analiza) i analizom de-reguliranih bioloških puteva. Transkripcijski kompleksne regije genoma MCMV dodatno su analizirane Northern hibridizacijom i metodom RT-PCR dok je korelacija između količine transkripata i proteina odabranih gena analizirana metodom Western blot. Na kraju, funkcija novog, prekrojenog transkripta MAT (most abundant transcript; najzastupljeniji transkript) analizirana je uz pomoć reporterskih stanica koje nose aktivacijske Ly49 receptore.

Rezultati

Ovaj rad predstavlja prvu detaljnu analizu transkriptoma MCMV-a korištenjem komplementarnih metoda analize cDNK knjižnice i RNASeq analizom, a rezultirala je identifikacijom brojnih novih transkripata MCMV, uključujući i nove prekrojene transkripte kao i transkripte koji se prepisuju sa intergenskih regija. Ustanovljeno je da najjače izraženi

virusni transkripti imaju nepoznatu funkciju i često su pogrešno anotirani. Najjače izražen transkript (tzv. MAT transkript) prvi je virusni transkript koji ima i kodirajuću i nekodirajuću funkciju. Naime, nedavno je pokazano da se na njegovom 3' netranslatiranom kraju (3' UTR, od eng. *3' untranslated region*) nalazi vezno mjesto za staničnu mikro-RNK [18, 84], dok je u ovom radu pokazano da on također kodira i barem još dva proteina. Uz ove dvije navedene funkcije, u ovom je radu otkriveno da je 5' netranslatirani kraj (5'UTR) MAT transkripta bitan virusni faktor kojeg na inificranim stanicama prepoznaju stanice prirodne ubojice pomoću aktivacijskih receptora Ly49.

Analiza transkriptoma stanica domaćina pokazala je da litička infekcija virusom MCMV izaziva izrazite promjene u ekspresijskom profilu gena domaćina: ekspresija gotovo trećine gena miša promijenila se uslijed infekcije virusom MCMV. Geni čija se ekspresija pojačava tijekom infekcije uglavnom su geni uključeni u upalne i imunološke procese, međutim neki pripadaju i skupini transkripcijskih faktora te genima povezanim s razvojem i diferencijacijom. Ovi rezultati u skladu su s dosadašnjim saznanjima o CMV-u kao virusu koji izaziva upalu te uzrokuje razvojne poremećaje tijekom kongenitalnih infekcija. Brojni geni čija se ekspresija smanjila tijekom infekcije povezani su sa funkcijama čija je uloga u infekciji za sada nepoznata poput dugačkih intergenskih nekodirajućih RNK, *antisense* RNK ili malih nukleolarnih RNK. GO analiza rezultirala je detekcijom disreguliranih bioloških puteva koji još do sada nisu bili povezani sa citomegalovirusnom infekcijom, a koji imaju potencijal rasvjetljavanja nekih nepoznanica u patogenezi citomegalovirusne infekcije.

Zaključci

Jedno od najznačajnijih otkrića u ovome radu jest dokaz izuzetne kompleksnosti transkriptoma MCMV-a koji dosad nije bio ovako sustavno istraživan niti su postojale transkriptomske mape. Ova analiza transkriptoma MCMV-a predstavlja važan prvi korak ka razvoju boljih genomskih mapa i reanotaciji genoma MCMV-a. Analiza odgovora stanica domaćina na infekciju dala je novi pogled na molekularne interakcije između virusa i domaćina i otvorila brojna nova područja istraživanja koja imaju potencijal da p pronađu nove mogućnosti liječenja bolesti izazvanih CMV-om.

Ključne riječi

mišji citomegalovirus, MCMV, transkriptom, ekspresija gena, izbjegavanje imunološkog odgovora, aktivacijski receptori Ly49

TRANSCRIPTOMIC ANALYSIS OF MURINE CYTOMEGALOVIRUS

SUMMARY

Human cytomegalovirus (HCMV) is a ubiquitous human pathogen responsible for devastating congenital disease and life-threatening complications in immune-suppressed patients. Available treatments have many shortcomings and effective vaccine is still lacking. Major obstacles to progress in vaccine and antiviral drug development are (1) species specificity of HCMV, and (2) gaps in our understanding of viral genes and their interaction with host genes. First limitation is circumvented by the use of model animal viruses, especially murine cytomegalovirus (MCMV). We sought to alleviate the second problem by studying MCMV transcriptome using two approaches: classical cDNA library analysis and next generation sequencing (RNASeq). This dual analysis revealed incredible complexity of MCMV transcriptome, detected numerous novel viral spliced and unspliced transcripts as well as transcription from intergenic regions, and showed that expression levels of viral transcripts vary by several orders of magnitude. Unexpectedly, most top expressed genes were of unknown functions and were improperly annotated. Therefore, this analysis provides the first comprehensive overview of MCMV transcriptome, underscores the necessity of transcriptomic analyses in providing evidence-based genome annotation and could serve as the first step towards re-annotation of MCMV genome. The most abundant viral transcript, recently identified as a noncoding RNA regulating cellular microRNAs [18, 84], was shown to also code for a novel protein(s). This is the first viral transcript that functions both as a noncoding RNA and an mRNA. In this work it is also shown that this transcript's 5' UTR plays a role in NK cell recognition of infected cells via activating Ly49 receptors.

Analysis of host transcriptome showed that lytic infection revealed that many unexpected gene groups are disregulated in response to the infection. Such systematic analysis may shed new light on cytomegalovirus pathogenesis and suggests new areas of research.

Key words

murine cytomegalovirus, MCMV, transcriptome, gene expression, NK cell evasion, activating Ly49 receptors

TABLE OF CONTENTS

1. INTRODUCTION

1.2 HERPESVIRUSES

Herpesviruses (*Herpesviridae*) are a family of large DNA viruses infecting a wide range of vertebrate and invertebrate hosts: mammals, birds, reptiles, amphibians, fish and oyster. Most of the human population is infected with at least one herpesvirus and the infection lasts for life. In fact, herpesviral infections are one of the leading causes of viral disease in humans and the infection with one herpesvirus does not preclude the infection with another species. Herpesviruses derive their name from the Greek word *herpein,* which means to creep and reflects the spreading of the skin lesions and the propensity of these viruses to cause recurrent infections.

Herpesviruses are classified into 3 subfamilies. Alpha-herpesviruses (*Alphaherpesvirinae*) target mucosal epithelium during lytic infection, have short replicative cycle and may infect a wide variety of host tissues. Beta-herpesviruses (*Betaherpesvirinae*) are strictly species specific and have longer reproductive cycle but have a broad cell tropism. Gamma-herpesviruses (*Gammaherpesvirinae*) are lymphotropic and the two both gamma-herpesviruses that infect humans have been associated with malignancies [37]. Of over 25 viruses in the herpesvirus family, eight are human pathogens: herpes simplex viruses type 1 and 2 (HSV1 and HSV2) and Varicella-zoster virus (VZV) belong to *Alphaherpesvirinae* subfamily, human cytomegalovirus (HCMV, also called human herpesvirus 5 (HHV-5)) and human herpesviruses 6 and 7 to *Betaherpesvirinae,* and Kaposi's sarcoma-associated virus (KSHV) and Epstein-Barr virus (EBV) to *Gammaherpesvirinae* subfamily [75] (Table 1).

Table 1. Human herpes viruses and their characteristics.

Family	Common name (abbreviation)	ICTV name (abbreviation)	Target cell type	Site of latency	Disease association
α	Herpes simplex virus 1 (HSV-1)	Human herpes virus 1 (HHV-1)	epithelial and keratinocyte	neuron	oral and/or genital herpes
	Herpes simplex virus 2 (HSV-2)	Human herpes virus 2 (HHV-2)	epithelial and keratinocyte	neuron	oral and/or genital herpes
	Varicella-zoster virus (VZV)	Human herpes virus 3 (HHV-3)	epithelial, keratinocyte, T cell, sebocyte, monocyte, endothelial, Langerhans and PBMC	neuron	chickenpox and shingles
β	Cytomegalovirus (CMV)	Human herpes virus 5 (HHV-5)	macrophage, dendritic, endothelial, smooth muscle, epithelial and fibroblast	CD34+ HSC*, monocyte	infectious mononucleosis-like syndrome, retinitis, congenital disease
	human B-lymphotrophic virus (HBLV)	Human herpes virus 6 variant A (HHV-6A)	T cell	bone marrow progenitor	sixth disease
	human B-lymphotrophic virus (HBLV)	Human herpes virus 6 variant B (HHV-6B)	T cell	bone marrow progenitor	sixth disease
	Human herpes virus 7 (HHV-7)	Human herpes virus 7 (HHV-7)	T cell	T cell	Pityriasis rosea
γ	Epstein-Barr virus (EBV), lymphocryptovirus	Human herpesvirus 4 (HHV-4)	B cell, epithelial	B cell	infectious mononucleosis, various B cell lymphomas, nasopharyngeal carcinoma
	Kaposi's sarcoma-associated herpes virus (KSHV)	Human herpes virus 8 (HHV-8)	lymphocytes	B cell	Kaposi's sarcoma, primary effusion lymphoma, multicentric Castleman's disease

*HSC=hematopoietic stem cell

1.2.1 Virus and genome structure

An infectious herpesviral particle or virion has a multilayered architecture and consists of the following layers: lipid envelope, amorphous protein coat called tegument and icosahedral capsid (also called nucleocapsid) that encapsulates herpesviral genome. Herpesvirus virions vary in size from 120 up to 260 nm in diameter and the source of this variation seems to be the thickness of tegument and the state of the envelope [111].

The envelope is a lipid bilayer derived from cellular membranes obtained during virion maturation and egress, and is spiked with virally encoded glycoproteins. In fact, the vast majority, if not all, virion glycoproteins are located on the capsid. These proteins are required for entry of the virus particle into target cells and are also targeted by the host's immune response [90, 111].

The tegument is the most structurally diverse part of the herpesviral virions (reviewed in [13]). This proteinaceous layer links the capsid with the envelope and contains nearly half of the whole protein mass of the virion. The tegument of HSV-1 contains more than 20 virally encoded proteins, and at least 30 have been found in the tegument of HCMV. Although functional homologs exist between most tegument proteins of herpesviruses, only a small number exhibit structural homologies. Virally encoded tegument proteins enter the cell with the virus particle and are then able to quickly modify cellular environment to suit viral needs (e.g. by managing host protein synthesis shut-off and/or by mediating evasion of cellular antiviral responses) and to regulate the expression of viral genes. Some viral tegument proteins play roles in maintaining the structural stability of the capsids and directing the acquisition of the virus envelope. In addition to virally encoded proteins, the tegument also contains proteins of host origin as well as some viral RNAs. Cellular proteins found in the tegument are mostly proteins that are present in the cell in high abundance (components of the cytoskeleton, some heat shock proteins, annexin) so it is still unclear whether these proteins are actively or passively incorporated into the tegument [75, 89, 90].

The nucleocapsid consists of five conserved proteins (the major capsid protein pUL191, pUL18, pUL38, pUL35 and pUL6) and has a highly ordered icosahedral shape of approximately 130 nm in diameter. While there is little genetic similarity between mammalian herpesviruses and their more distant relatives, herpesviruses of reptiles, birds, fish, amphibians and the oyster, the capsid structure is conserved. Also, although genomes of different herpesviruses range in size from 125 to 240 kb (with beta-herpesviruses having the

largest genome), their capsids are roughly of the same size. Inside the capsid, herpesviral genome is packaged without any histones or analogous proteins [38, 75, 89].

While different members of the herpesvirus family have little sequence homology, their genome organization and structure are quite similar. Most contain two unique regions – unique long (UL) and unique short (US) bounded by direct or inverted repeats (Figure 1). Terminal repeats are thought to play a role during virus replication by enabling genome circularization. Due to the high number of repetitive sequences, the genome size of individual strains can vary [36].

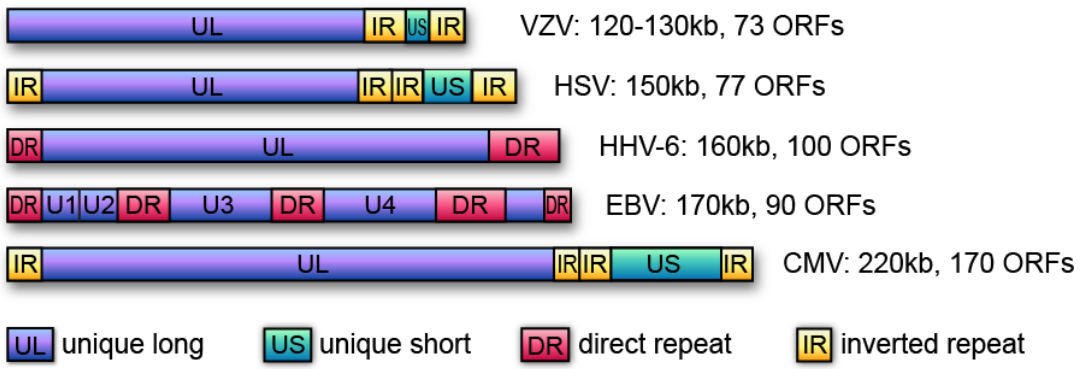


Figure 1. Genome organization of several herpesviruses, their size and number of open reading frames (ORFs).

Available whole genome sequences of herpesviruses are expanding every day, with over 60 available today. However, the ease of whole genome sequencing is not followed by the ease of describing gene content. As will be discussed later in the introduction, there currently exists no set of criteria for efficient detection of all genes, especially in dense viral genomes. Currently the most commonly applied method in herpesviral genome annotation is the use of comparative genomics that compares putative open reading frames (ORFs) to related, well defined ORFs of other herpesviruses [36].

In general, typical herpesviral gene is adapted for efficient expression inside its host cell and has the following structure: promoter or regulatory sequence located 50-200 bp upstream from TATA box, transcription initiation site 20 to 25 bp downstream from the TATA box, 5' untranslated region (UTR) of 30 to 300 bp and 10 to 30 bp long 3'UTR followed by polyadenylation signal. Most genes are transcribed by RNA polymerase II, although some (mostly non-coding RNAs) use polymerase III. Genes often overlap, so promoter/regulatory

sequences of one gene can be located in the coding region of another gene. Transcription from an internal methionine yielding N-terminally truncated protein is also described. Although these proteins share parts of the reading frame and thus have similar aminoacid composition, they may play completely different roles. Such proteins may be translated from a single, shared transcript or from another 3' co-terminal transcript with variable 5' start. In addition to protein-coding genes, herpesviruses encode non-coding RNAs (ncRNAs) whose function is largely enigmatic [111].

Approximately 40 genes, called herpesvirus core genes, are conserved among all three mammalian herpesviral subfamilies and are usually located in the central part of the genome. These genes are of crucial importance for herpesviral growth and are involved in genome replication, packaging of viral DNA and capsid structure and formation. Core genes are divided into seven core gene blocks where each block contains 2 to up to 12 genes. Arrangement of the core gene blocks is conserved at the level of subfamily. In addition to protein-coding genes, some regulatory sequences are also conserved (i.e. lytic DNA replication) as well as sequences located at the genome termini [36, 111].

Herpesviral genes can be broadly separated into two categories based on their dispensability for viral growth in cell culture: essential and non-essential. Essential genes govern transcription of other viral genes, replication and virion assembly, while non-essential genes regulate cellular and/or host immune responses. However, since the classification of genes to "essential" or "non-essential" gene is based only on the gene's dispensability for *in vitro* growth, the designation can be a bit misleading. A major part of the success of herpeviruses as persistent pathogens is due to the fact that large portions of their genomes are dedicated towards immune evasion genes which are essential for viral growth and spread *in vivo*. So, while deletion of an immune evasion gene will not impair viral growth in cell culture, such viruses are severely attenuated *in vivo*. Further complicating the division into essential and non-essential is the fact that many herpesviral genes have multiple functions. Herpesviral genes may also be classified based on the timing of their transcriptional activation (see chapter 1.2.2) [111].

Some herpesviral genes are of host origin and a consequence of long virus-host co-evolution. The acquisition of the host genes seems to involve reverse transcription step (probably as a consequence of co-infection with a retrovirus or) due to the lack of introns in the viral version of these genes. Genes involved in immune regulation are usually targets of herpesviral

molecular piracy, although conserved genes encoding viral DNA polymerase and dUTPase also seem to be appropriated from the host [36, 111].

1.2.2 Herpesvirus life cycle

At the cellular level, infection with a herpesvirus may result in lytic or latent infection. Lytic infection is characterized by activation of a specific cascade of viral genes resulting in intensive genome replication, generation of new viral particles and consequent lysis of the infected cell. In certain cell types, however, the viral genome remains dormant after reaching the nucleus, only a fraction of viral genes are expressed and no new viral progeny is made. *In vivo* such virus is invisible to the immune system. This ability to enter latency has made herpesviruses such successful life-long persistent pathogens. Yet, despite its importance, it seems that different herpesviruses do not share any latency associated strategy or genes. The virus may reactivate from latency at any time during life after primary infection, usually following immunosuppression or similar stress. The purpose of such reactivation events is further dissemination of the virus inside its host or between hosts [111].

Lytic life cycle of herpesviruses begins with the recognition of the target cell via glycoprotein complexes present on the virus envelope. Conserved herpesviral glycoprotein B (gB) and heterodimers consisting of glycoproteins H and L (gH/gL complex) are responsible for the entry of the virus into the host cell. In HCMV, the gH/gL complex associates with additional proteins (UL74 gene product gO or complex of glycoproteins encoded by UL128 and UL130) that regulate the entry into different cell types [93]. The envelope then fuses with the cell membrane, the capsid surrounded by tegument enters cytoplasm and travels to the nucleus using cellular microtubule motor system. This process is regulated by proteins of the tegument and capsid. Only the genome enters the nucleus through the nuclear pore via concerted action of capsid and tegument proteins, where viral genes begin their transcription in temporally ordered manner [90, 93]. First genes that are transcribed are called immediate early (IE) or alpha genes and are involved in the regulation of transcription and translation of early proteins in order to "optimize" the host cell for viral gene expression and genome replication (reviewed in [128]). They do not require *de novo* protein synthesis. Only one regulatory gene is conserved among different herpesviruses - multifunctional regulator of expression (MRE). Its best known function is the prevention of splicing at the early times post infection (PI), which favors the export of mostly unspliced viral transcripts from the nucleus and their subsequent translation. This inhibition is released at late times post infection. MRE

is also implicated in the shut-off of host gene transcription through yet unexplained mechanisms [93]. After IE genes early (E) or beta genes are expressed. E genes do require *de novo* protein synthesis after virus entry into the cell and are mostly involved in virus DNA replication (reviewed in [141]). Finally, last expressed are the late (L) or gamma genes which encode proteins needed for virus assembly and egress, and their transcription is often dependent on DNA replication (reviewed in [2]).

Viral DNA is replicated in a rolling circle manner from circularized genome utilizing six virally encoded genes that belong to the core gene group. Replication begins at defined and conserved origins of replication (*ori*) and different herpesviruses can have from one (beta-herpesviruses) to up to three (HSV-1 and HSV-2) *ori* sites. In the nucleus, viral DNA replication starts near nuclear domain 10 (ND10) structures, which become disrupted as the infection progresses. The viral DNA replication is so efficient that at late times post infection it can completely overtake the nucleus with viral DNA levels equaling that of cellular DNA [93].

After replication, the viral DNA is packaged into the capsid inside the nucleus. Capsids are also made in the nucleus from capsid proteins generated in the cytoplasm. Packaging of the genome into the capsids is regulated by conserved cleavage/packaging (*pac*) site in the genome and several conserved viral proteins. *Pac* sites serve as recognition sequences for viral packaging machinery to cleave the genome from the multi-genome concatamer that results from the rolling-circle type of replication. Finally, port capping protein (PCP) seals the filled nucleocapsid. Some parts of the tegument are also added to nuclear virions [93, 132].

Newly made nucleocapsids are now too big to pass through the nuclear pore and require two-stage pass through the inner and then outer nuclear membrane in a process termed nuclear egress. While this process is not fully understood, it is known that both viral and cellular proteins are required for its successful completion. In this process, the capsids first bud through inner nuclear membrane resulting in the formation of primary enveloped virions in the perinuclear space (the space between inner and outer nuclear membrane). Heterodimeric complex of two conserved viral proteins called nuclear egress complex (NEC) is required for this process. In human cytomegalovirus these proteins are called UL50 and UL53. The primary envelope then fuses with the outer nuclear membrane and releases the virions into the cytoplasm. During this, the virion loses its primary envelope (de-envelopment) [90]. Inside the cytoplasm, the majority of tegument proteins are added (including some host proteins) to

the capsids in a special cytoplasmic compartment called cytoplasmic virus assembly compartment (cVAC) made of rearranged or modified host organelles. Finally, capsids associate with microtubular system in order to traverse the cytoplasm and reach the final envelopment place (the Golgi apparatus) where they receive their final envelope by poorly understood mechanisms. Following the final envelopment, the virions are released from the cell by exocitosis [93].

The complexity of new virus assembly results in generation of a multitude of different aberrant virus particles (e.g. dense bodies, non-infectious enveloped particles) in addition to infectious viral particles. Non-infectious aberrant particles can be found both in infected cell and extracellular compartment, sometimes in a much greater amount than infectious virus, and it has been proposed that they function as decoys to overwhelm and saturate the immune response [38].

1.3 CYTOMEGALOVIRUS

Betaherpesvirinae subfamily of the herpesvirus family contains 4 genera: cytomegaloviruses (CMVs), muromegalovirus, roseolovirus and proboscivirus. Human cytomegalovirus (HCMV) is the most frequent viral cause of congenital infections, often causing devastating congenital disease with life-long sensorineural sequelae [19] and with annual prevalence of 0.1-2% among newborns [15]. Deaths and permanent disabilities associated with congenital CMV infection affect more newborns than Down's syndrome, fetal alcohol syndrome, or neural-tube defects [112]. In immunocompetent adults, HCMV infection usually passes with little or no symptoms; however, in immunosuppressed patients like AIDS, cancer or transplant patients, it can cause a multitude of life-threatening conditions affecting multiple tissues and organs and is the leading cause of complications and graft loss in transplant patients [12]. Recently, HCMV has been linked to lung injury in trauma patients [31], and recognized as a possible co-factor in some cancers and atherosclerosis [125, 126]. In contrast to HCMV diseases arising from impaired immune response, involvement of HCMV infection in atherosclerosis and cancerogenesis is not associated with high level of unchecked viral replication. Like other herpesviruses, HCMV infects the majority of world's population with infection rates ranging from 20% to almost 100% among adults, depending on socioeconomic status [142].

HCMV displays a broad cellular tropism – it can infect fibroblasts, endothelial cells, epithelial cells, monocytes/macrophages, smooth muscle cells, stromal cells, neuronal cells, neutrophils, trophoblasts and hepatocytes. As a consequence, HCMV can spread to multiple organs and tissues to cause the disease. Entry into target cells is mediated by herpesviral envelope glycoproteins gB, gH/gL and gM/hN complexes, which interact with a number of different cellular receptors. The initial contact is made by the interaction of envelope glycoprotein with cellular proteoglycan heparan sulfate [92]. Infected cells acquire typical enlarged (or cytomegalic) shape. The entry of the virus into the host cell is followed by intense cellular antiviral response: activation of interferon-stimulated genes, production of interferon β and other inflammatory cytokines whose function is to alert the immune system of the viral intrusion [30]. To counter this, HCMV has developed a plethora of modulators directed against infected cell response and nearly every aspect of immune response. Therefore, fully functional immune response is needed for the efficient control of HCMV. The virus is, however, never completely eradicated, even in immunocompetent hosts, as it can enter latency during which it is invisible to the immune system.

Infectious virus particles are shed through all bodily fluids of the infected person: urine, saliva, breast milk, semen and tears, even if the patient does not have any clinical symptoms.

Despite its importance in human health and decades of research, treatment options are scarce and burdened with high toxicities as well as drug-resistant virus strains [3]. Progress in antiviral drug and vaccine development depends on good understanding of viral genes, their products and their interactions with host genes. A major obstacle to HCMV research is its strict species specificity that precludes the use of experimental animals in HCMV research. Immune evasion is a major cause of HCMV's success as ineradicable pathogen and complex immune interactions are very hard to investigate using only *in vitro* analyses. Thus, to properly understand HCMV, other CMV viruses must be used.

1.3.1 Murine cytomegalovirus – the model virus

There are several CMVs that infect animals suited for *in vivo* work: rat CMV (RCMV), guinea pig CMV (GPCMV), murine CMV (MCMV), porcine CMV (PCMV), rhesus macaque CMV (RhCMV) and chimpanzee CMV (CCMV). Of these, murine CMV is the most widely used since it shares many biological, genetic and pathological properties with HCMV. Mice

are most widely used experimental animals in medical research due to their size and cost-effectiveness and availability of numerous mutant strains.

MCMV has now been successfully used to investigate many pathological facets of CMV infection; from humoral and cellular immunity and host responses, to congenital CMV infection and infection in immunosuppressed hosts. The cloning and application of BAC mutagenesis [88] opened a possibility of constructing various gene deletion mutants that have been instrumental in the identification and characterization of a multitude of immune evasion and other genes.

The construction of deletion mutants depends on the accuracy of genomic maps. The first sequence of HCMV was published in 1990 [8, 22] and of MCMV in 1996 [106]. However, despite decades of research many questions remain: a definitive genomic map, a catalogue of gene products and information regarding how these gene products interact with the host and ultimately cause the disease.

1.3.2 Analysis of MCMV genome

The first sequence of MCMV identified a double-stranded DNA genome of 230,278 bp in size and with a GC content of 58.7% [106]. Unlike HCMV, MCMV does not contain large internal repeats but is arranged as a single unique sequence bounded by short (31 bp) terminal direct repeats not represented anywhere else in the genome. MCMV was also found to contain a few short direct and inverse repeats.

Rawlinson *et al.* sequenced Smith strain of MCMV [106] and identified 170 ORFs using homology searches, comparison with other herpes viruses and *in silico* prediction software with the following criteria: minimal length of 300 bp and less than 60% overlap with adjacent ORFs. The genomes of MCMV and HCMV were shown to be very similar at genetic and nucleotide compositional levels, although overall arrangement of the genomes differ. The central 180-bp regions of HCMV and MCMV are co-linear and contain conserved herpesvirus genes interspersed with genes unique to MCMV (Figure 2). All known enzyme homologs encoded by HCMV were also found in MCMV, as well as numerous structural and tegument proteins. Of nine families of homologous proteins described in HCMV, six have their sequence homologous gene families in MCMV (US22, UL25, UL82 and the GCRs).

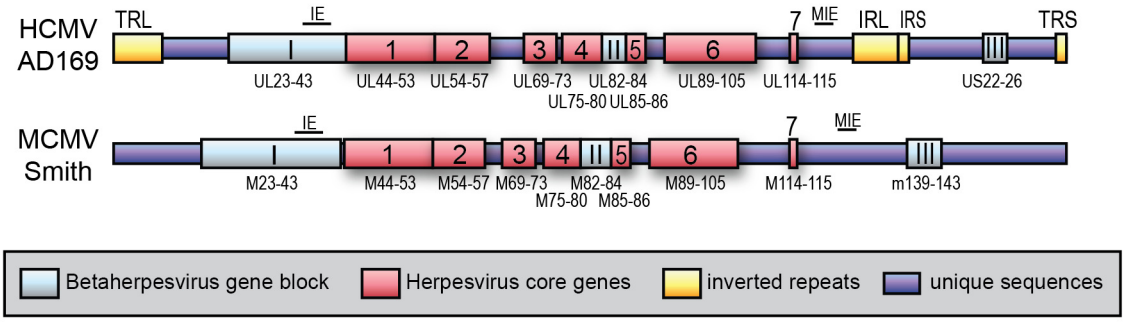


Figure 2. Comparison of HCMV and MCMV genome structures.

In 2004., Kattenhorn *et al.* [58] analyzed proteins associated with MCMV virions using liquid chromatography-tandem mass spectrometry (MS) and comparing the obtained peptide sequences with a database of putative MCMV ORFs containing Rawlinson's genes as well as 12 putative novel genes the authors predicted using GeneMark software. This analysis confirmed some of the previously predicted ORFs but also identified peptides coming from two previously unannotated regions (*m166.5* and and $ORF_{105932–106072}$ which was later confirmed by Scalzo *et al.* [114]) and detected a few sequencing errors which led to the re-annotation of the *m20* and *M31* ORFs.

Brocchieri *et al.* [16] used a purely computational approach to reanalyze the MCMV genome in 2005. These authors argued that commonly employed parameters for ORF discovery (ORF length >100 bp, less than 60% of overlap between adjacent genes, only ATG as ORF start codon) were not well suited for gene discovery in herpesviruses and thus led to the exclusion of known CMV gene products, especially small ORFs or ORFs with multiple splicing events. Due to genome size restriction imposed by viral capsid, they argued, a greater degree of overlap between adjacent genes is highly likely in order to preserve precious genomic space. Finally, sequencing errors as well as posttranscriptional modifications, splicing, alternative translation initiation and stop codon suppression may all confound ORF prediction software. They have therefore used a less restrictive approach to the prediction of protein-coding genes based on translational frame analysis taking into account frame-specific G+C content. No assumptions on the ORF size have been made and no restrictions on the degree of overlapping with neighboring genes. Based on their analysis, the authors suggested a substantial revision of the MCMV and RCMV genome annotations including 14 new putative ORFs for MCMV as well as new translation start sites and stop sites for 18 and 4 previously annotated genes, respectively. While their approach successfully predicted frameshift extensions to *m20* and

M31 genes previously reported by Kattenhorn *et al.* [58], this analysis lacks further experimental confirmation of other predictions.

Tang *et al*. [133] used two new ORF prediction tools (MacVector and GenePicker) and predicted 14 new ORFs, in addition to the ones previously predicted by Rawlinson *et al.* [106]. They then constructed DNA microarray assay based on their and Rawlinson's predicted ORFs and were able to confirm the expression of 172 predicted genes, 7 of which were newly predicted by their analysis. Expression of 10 previously predicted ORFs was not detectable in fibroblasts either using microarray or RT-PCR, whereas 2 of these were detected in macrophages indicating that some MCMV genes might show tissue/cell-type specificity.

Analysis of genome stability after *in vivo* and *in vitro* passage [24] demonstrated high genome stability of MCMV in the absence of selective pressure. In total 452 differences between their and Rawlinson's sequence were identified, of which 50 were insertion/deletions (indels) and 402 single-base pair substitutions. While most changes were detected in the central coding region, ORF containing the most sequence differences was immunoevasin *m04*.

Despite years of research, the definitive genomic map of MCMV or indeed the definitive number of genes is still lacking. Currently there exist two genomic annotations of MCMV: one based on Rawlinson's annotation with 170 ORFs (can be found under GenBank accession number GU305914.1) and NCBI reference sequence annotation (GenBank accession number NC_004065.1) that identifies 160 ORFs. Aside from different ORF number, the differences between the two annotations are mostly minor, as can be seen in Figure 10, chapter 4.2.1. The most significant change is the change of ORF names introduced by newer NC_004065.1 annotation. However, since the majority of publications in the MCMV field use nomenclature from modified Rawlinson's annotation, this is the nomenclature that is preferentially used throughout the text of this thesis.

1.3.3 Immune responses to MCMV infection and immune evasion

MCMV infection of laboratory mice is the most commonly used model to investigate immune responses elicited by CMV infection and viral immune evasion genes. CMVs are expert immune modulators: large portion of their genomes have been dedicated to immune evasion genes. While many of these genes have been dubbed as "non-essential" genes due to their

dispensability for *in vitro* growth, the deletion of the majority leads to significant attenuation *in vivo*.

MCMV actively prevents recognition by both arms of the immune system: innate and adaptive immunity, as well as cellular antiviral mechanisms: from repression of host gene expression and translation, interferon responses and apoptosis to evasion of antibody responses and NK and T lymphocytes.

For efficient control of MCMV, natural killer cells are one of the most important cells of the innate arm of the immune system. Natural killer (NK) cells are a subset of bone-marrow-derived lymphocytes that can kill suspicious cells without prior sensitization. This ability makes NK cells an important factor in organisms' defense against transformed cells as well as in early control of various pathogens, especially viruses. In order to exert their function, NK cells survey their surroundings through a panel of inhibitory and activating receptors. The decision on whether to kill or spare particular cell depends on the balance of signals coming from these receptors and MCMV very effectively manipulates both (reviewed in [74]). In order to evade activating NK cell receptors like NKG2D, MCMV encodes multiple proteins dedicated to downregulation of ligands for these receptors; *m145* affects surface portion of MULT-1, *m152* targets RAE-1 family and H60 is targeted by *m155* [51, 64, 65, 76, 77] while *m138*/fcr-1 downregulates surface portions of H60, MULT-1 and RAE-1ε [5, 71] (Figure 3).

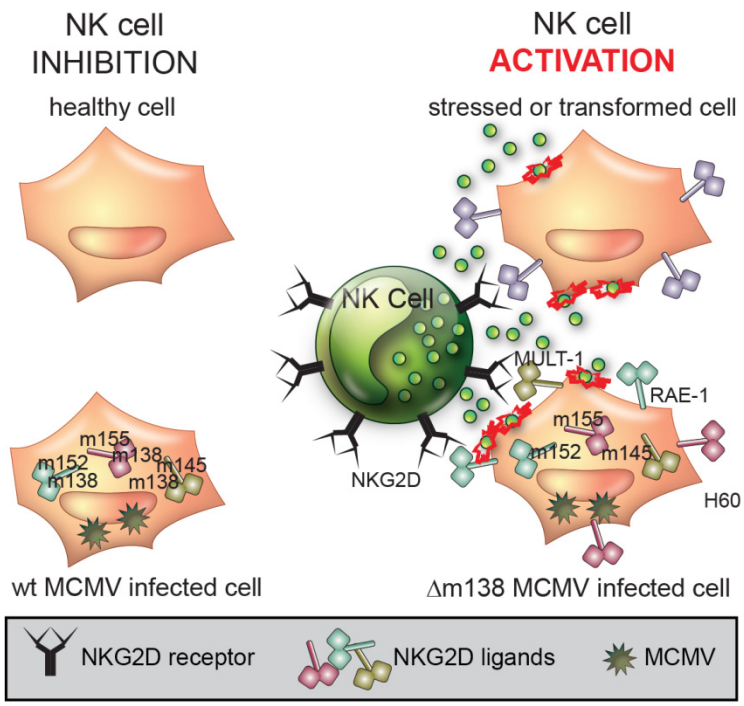


Figure 3. MCMV evasion of NKG2D receptors on NK cells. NKG2D is an important activating NK cell capable of overriding signals coming from inhibitory receptor and has thus been extensively targeted by MCMV. Ligands for NKG2D (H60, Mult-1 and Rae proteins) are expressed only when the cell is stressed, undergoing transformation or infection. To counter recognition by NK cells, MCMV encodes four proteins that downregulate cell surface expression of NKG2D ligands. Deletion of either one of these immunoevasive genes results in engagement of NKG2D, activation of NK cells and NK cell mediated lysis of the infected cell.

In contrast to the evasion of activating NK receptors, MCMV is actively trying to engage inhibitory receptors to keep NK cells in inhibited state. The host responded by duplicating some inhibitory receptors and turning them into activating versions [1, 81]. An excellent example of such evolutionary arms race between the virus and its host is recognition of virally encoded m157 protein by Ly49 receptors. In 129J mouse strain m157 is recognized by an inhibitory Ly49I receptor resulting in the inhibition of NK cells and subsequent high viral titers [6] (Figure 4). In C57Bl/6 mouse strain, the same protein is recognized by activating Ly49H receptor, making this mouse strain highly resistant to MCMV infection [6, 122]. It is important to note that these studies analyzed only the Smith strain of MCMV. Different interactions of Smith MCMV encoded m157 protein in different mouse strains is shown in Figure 4. Different wild isolates of MCMV exhibit variations in their *m157* gene and most are not recognized by Ly49H [32, 123, 134]. In addition, serial passage of Ly49H-sensitive MCMV through Ly49H^{+} mice leads to loss-of-function mutations in *m157* gene [44, 134].

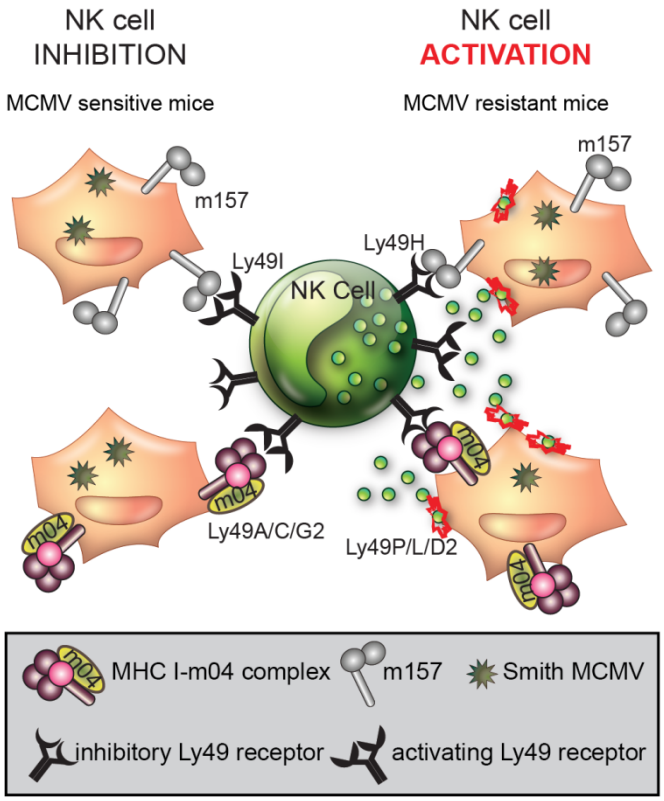


Figure 4. Modulation of NK cell responses through Ly49 receptors. To avoid recognition by NK cells, MCMV avoids recognition of its proteins by activating NK cell receptors while actively trying to engage inhibitory NK cell receptors. However, in evolutionary arms battle against this pervasive pathogen, different mouse strains have developed activating receptors capable of recognizing viral proteins that have originally served as ligands for inhibitory receptors. The best known example is virally encoded m157 protein which is recognized by the inhibitory NK cell receptor Ly49I in MCMV-sensitive mice, while in MCMV-resistant mice the same protein is recognized by activating NK cell receptor Ly49H. A similar example is m04-MHC I complex, which is recognized by inhibitory NK cell receptors in MCMV-sensitive mice and by activating receptors in MCMV-resistant mice.

Evasion of CD8 T cells is connected with the evasion of NK cells. In order to evade CD8 T cells, MCMV must remove MHC I from the cell surface to prevent the presentation of virally encoded peptides. For this purpose, MCMV encodes two proteins: m152 that arrests the maturation of MHC I molecules in ERGIC compartment [149], and m06 that redirects MHC I to lysosomes for degradation [109]. Although the absence of MHC I from the cell surface protects the virus from CD8 T cells, it also renders such cells sensitive to NK cell recognition through the absence of engagement of inhibitory receptors, a phenomenon also known as "missing-self" recognition (reviewed in [57]). To prevent this, MCMV encodes a third protein – *m04*/gp34, which forms complexes with some MHC I molecules and allows them to reach the cell surface [59, 61]. Viral proteins regulating cell surface expression of MHC I molecules

and their modes of action are shown in Figure 5. *m04*/MHC I complexes on the cell surface serve as ligands for inhibitory Ly49 receptors preventing NK cells from "missing-self" reaction [7]. Similar to *m157*-Ly49H/I axis (Figure 4), the mice responded by developing activating Ly49 receptors capable of recognizing the same complexes [60, 103] and conferring resistance to MCMV in these mouse strains.

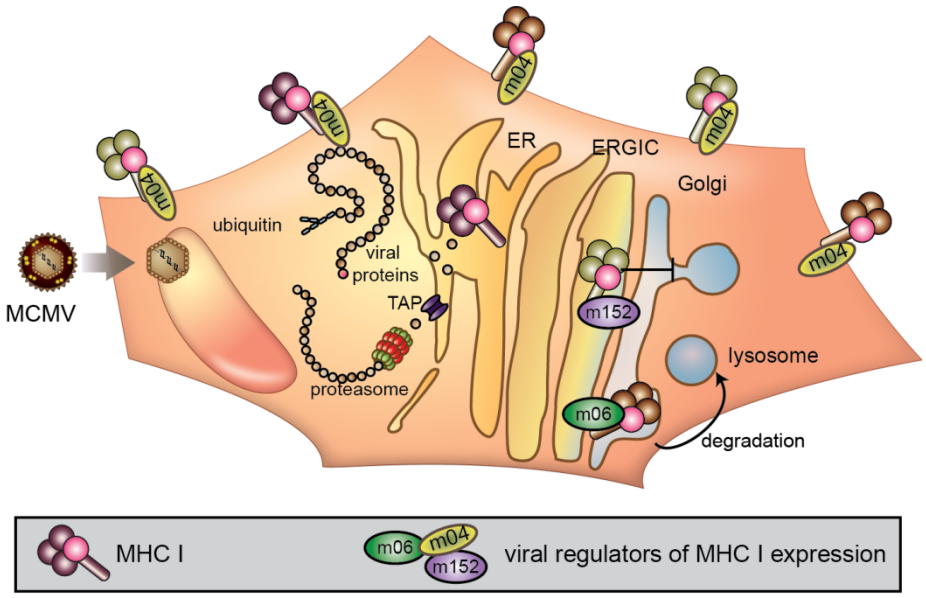


Figure 5. Viral proteins regulating cell surface expression of MHC I molecules. Normal, healthy cells display MHC I molecules on their cell surface loaded with cellular peptides. These peptides are generated from all cellular peptides by the proteasome and are then loaded into the MHC I molecules. CD8 T cells are educated during their development not to react to peptides produced from proteins of healthy cells; however, viral or aberrant proteins result in CD8 T cell recognition and activation. In order to avoid recognition by CD8 T cells, MCMV downregulates MHC I from the cell surface; however, downregulation of all MHC I would render it sensitive to NK-cell-mediated "missing-self" recognition and lysis. Therefore, MCMV encodes *m04*/gp34 that makes complexes with some MHC I, brings them to the cell surface and serves as a ligand for inhibitory NK cell receptors.

1.4 TRANSCRIPTOMICS

Transcriptomics is the study of transcriptomes: the complete sets of RNAs (transcripts) transcribed from the genome of a specific cell at a specific time and under specific conditions.

1.4.1 Transcriptome is more complex than genome

According to the central dogma of molecular biology, genetic information encoded in a gene is relayed to a protein via messenger RNA (mRNA). Until very recently, mRNA was viewed as a mere bridge between the DNA and the protein; an expendable copy of the valuable

genetic material needed to make the main workforce of the cell – the protein. A genome was considered to consist of coding DNA – a DNA that gives rise to proteins or functional RNA molecules (transport or ribosomal RNAs) – and non-coding DNA, often called junk DNA, with no discernible function. This view was largely based on bacterial genomes where 90% of the genome encodes a protein. With the development of faster and better sequencing tools, sequencing of larger genomes became possible and it became evident that increasing complexity of an organism is not followed by proportional increase in gene numbers (G-value paradox). On the contrary, the increasing complexity of an organism seems to be followed by a decrease in the fraction of protein-coding DNA in the genome; so nematode *Caenorhabditis elegans* and human have approximately the same number of genes, but 24% of nematode's genome contains protein-coding sequences, while in mammals protein-coding genes make only 2% of the whole genome [119].

In contrast to the genome size, proteome size is related to the complexity of an organism. Genes of higher eukaryotes consist of protein-coding parts (exons) interspersed with non-coding DNA (introns). After transcription, introns in pre-mRNA are excised by a tightly regulated process called splicing to produce mature mRNA. In addition to introns, occasionally an exon can be spliced out, thus giving rise to a different mRNA and different protein (reviewed in [82]). This process is called alternative splicing and it is considered to be the most important source of protein diversity in vertebrates [11, 47]. The vast majority of all multiexon protein-coding genes in higher mammals are alternatively spliced [139]. In addition to increasing the number of possible proteins, alternative splicing exhibits tissue specificity and inducibility [82].

The number of proteins therefore exceeds the number of protein-coding genes. It is now becoming more and more apparent that the same is true for RNA. For a very long time the transcriptome was thought to consist of mostly ribosomal RNA, transfer RNA and of messenger RNA, which constitutes only 1-5% of all RNAs. In the past 10 years, it has become increasingly evident that mammalian transcriptome is far more complex than previously anticipated. In addition to alternative splice isoforms, the list of non-coding RNAs keeps growing (Figure 6). With the advent of tiling microarrays, and later, next generation sequencing (discussed in chapter 1.4.2.3), it was found that transcription in mammals is highly pervasive (over 90% of mammalian genome is transcribed (reviewed in [27]) and is not confined to protein-coding genes. It is, therefore, now considered that the complexity of

the transcriptome, rather than the number of encoded genes, is a distinguishing factor between simpler and more complex life forms [119].

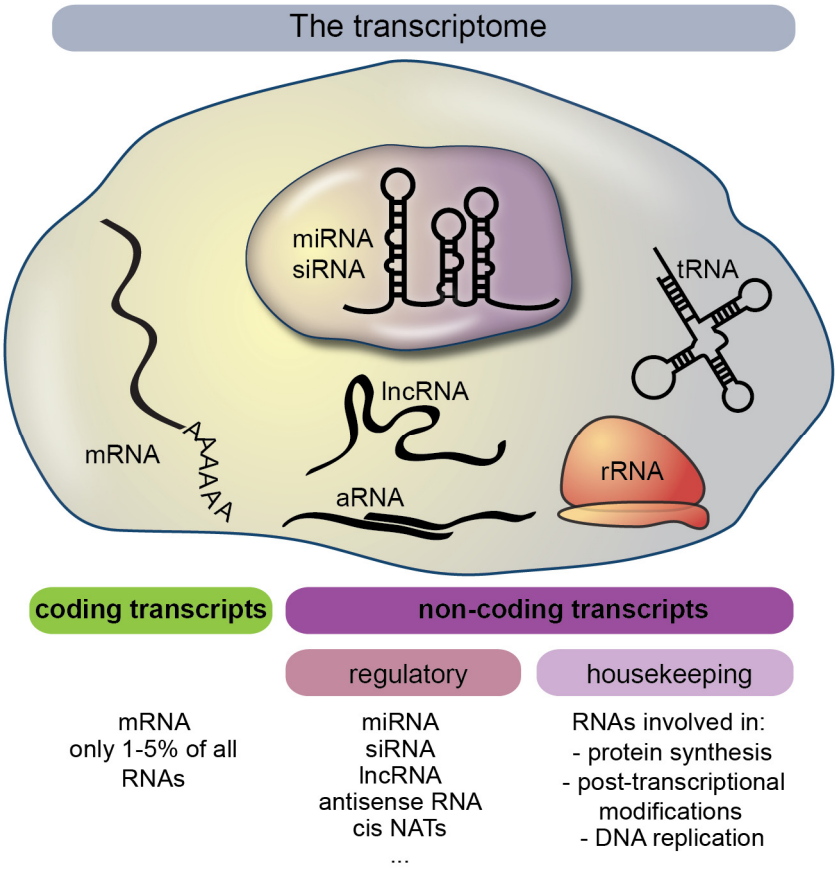


Figure 6. The transcriptome. The transcriptome is a full set of transcripts that accumulate in a specific cell at a specific time and under specific conditions. The transcripts can roughly be divided into coding transcripts (those that are translated and therefore encode a protein) and non-coding, which are not translated but function as catalytic, structural or regulatory RNAs.

The list of non-coding RNAs (ncRNA) has grown substantially in the past few years: in addition to rRNAs and tRNAs, we now recognize a wide variety of non-coding RNAs that can range in size from several nucleotides to several kilobases: microRNAs (miRNAs), small nucleolar RNAs (snoRNAs), antisense RNAs (asRNAs or aRNAs), long non-coding RNAs (lncRNAs), etc. [33]. Many ncRNAs display tissue- and condition-specific expression regulation, specific sub-cellular localization and are associated with human disease, indicating that they are an important part of our transcriptomes (reviewed in [33, 144]) rather than a consequence of transcriptional "noise". The functions of many ncRNAs are still unknown but increasing number of evidence suggests they regulate expression of other genes, especially during development and through different mechanisms ranging from degradation of mRNAs

(small interfering RNAs), inhibition of translation (miRNAs) to chromatin modifications, methylation of regulatory sequences (long ncRNAs) and dosage compensation (long ncRNAs) (reviewed in [73].

The implications of these findings for biological and biomedical sciences are profound. As sequencing technology has advanced and sequences of more and more organisms have become available, the researchers have become faced with a growing number of sequences and genes with unknown functions. Functions of many new genes were determined by reverse genetic approach, where the gene of interest is mutated or removed from the genome and its function determined by the resulting phenotype. However, if the disrupted gene is overlapped by regulatory non-coding RNA or an antisense RNA is transcribed from the non-coding strand, the observed phenotype could also be a consequence of disrupted regulatory RNAs rather than disrupted protein-coding gene.

1.4.2 Transcriptome analysis – why and how

As our knowledge about the complexity of a transcriptional profile of a cell has increased, the need for better genomic maps and potentially transcriptomic maps has become more urgent. Unlike the genome, the transcriptome differs among different cell types in a single organism, is highly dynamic and may be subject to fast changes in response to different stimuli (stress, cell cycle phase, infection, presence of different drugs, etc). Therefore, the main goal of transcriptomics is qualitative and quantitative characterization of all transcripts that accumulate in a particular cell or tissue at specific time and under specific conditions. Understanding transcriptional profiles of different cell types and under different conditions is essential for our understanding of the differences between various cell populations, their development and pathogenesis of diseases. The biomedical field, especially cancer and drug research, has been very fast to embrace gene expression profiling in order to elucidate mechanisms of various diseases, find potential drug targets or understand a drug's effect (reviewed in [140]).

While analysis of transcriptional profile of a single gene or locus can be analyzed by numerous traditional molecular biology methods that have been available for decades (Northern blot, RACE, transcript mapping), whole transcriptome analysis has become possible only recently. Because the transcriptome size greatly exceeds the genome site, whole transcriptome analysis requires high-throughput methods.

Transcriptome analysis methods can roughly be divided into sequence-based and hybridization-based approaches. Sanger-based sequencing of cDNA clones (described in chapter 1.4.2.1) was one of the first methods employed to analyze transcriptomes but it is very low-throughput, labor intensive, expensive and not reliably quantitative. Hybridization-based microarray analysis (chapter 1.4.2.2) provided first high-throughput method to analyze gene expression patterns and has dominated the field for over a decade. Finally, with the advent of new, cheaper, faster and even more high-throughput methods of sequencing, so-called next-generation sequencing (NGS; another commonly used term is massively parallel sequencing; discussed in chapter 1.4.2.3), sequencing-based transcriptome analysis quickly became the new standard in transcriptome analysis and made transcriptomics one of the fastest developing fields today. These three methods are discussed further below.

1.4.2.1 cDNA library analysis

Analysis of reverse-transcribed DNA (or coding DNA) started in the 1970s and has since become one of the fundamental tools of molecular biology. The technique is therefore well established and described in the literature [113], with commercially available premade cDNA libraries or easy-to-use kits. RNAs of interest are reverse transcribed to first strand cDNA by use of specific adaptors and primers or, in the case of polyadenylated RNA analysis, by use of polydT primers. Since a great portion of cellular RNAs are ribosomal RNAs, often depletion of ribosomal RNAs or isolation of specific RNA subset is performed before reverse transcription. Most often selection of polyadenylated transcripts is performed as this selection procedure not only effectively removes rRNAs, but also removes tRNAs and degraded transcripts. An important caveat is the fact that although most mRNAs are polyadenylated, some are not. In addition, most regulatory RNAs are not polyadenylated and thus in polyA libraries these transcripts are excluded.

After RNA selection and reverse transcription, single-stranded cDNA is then converted to double-stranded cDNA, cloned into appropriate vector, propagated in *E. coli* (or any other prokaryotic or eukaryotic organism, although *E. coli* is most frequently used) and sequenced. Based on the preparation steps, cDNA library can be classified as directional (strand specific) or random. cDNA clones in directional library retain strand information – their orientation in the vector reflects the original transcriptional polarity of the RNAs. In random cDNA library, the orientation of the cDNA clones is random and strand information is lost. With the discovery of antisense transcription, directional cDNA libraries became the preferred choice.

There are several drawbacks to cDNA library analysis. The first and most obvious is the fact that due to multiple cloning, selecting and sequencing steps, this technique is labor intensive and thus unsuitable for the analysis of big and complex transcriptomes that require high-throughput approaches. Second is the difficulty to obtain cDNA clones that represent full-length RNAs. Not only are RNases highly ubiquitous and resilient enzymes, RNA is sensitive due to its single-strandedness. Polyadenylated RNAs are most sensitive at their 5' ends where only 5' cap structure protects the mRNA from degradation and this sensitivity often results in generation of cDNA libraries that contain transcripts with truncations on their 5' ends, especially in the case of larger transcripts. This can make it hard to determine exact ends, especially exact 5' ends, of transcripts unless larger number of the same transcript is cloned. Additionally, smaller transcripts (either truncations or genuine smaller RNAs) are preferentially ligated into vector, enriching the library with smaller transcripts. This problem can partially be alleviated by size fractionation of cDNAs before ligation into vector. Finally, some rare transcripts may be difficult to clone and require large libraries to ensure deep coverage of the transcriptome.

There are also several benefits of cDNA library construction. Unlike microarray or RNASeq analysis, cDNA library analysis results in a physical library of cDNA clones which can then be further used in a multitude of assays. For instance, an interesting, full-length cDNA clone can be transfected into appropriate cell in order to determine its function and/or coding potential. Various truncated cDNA clones can be further used to map important regions of the transcript. As can be seen in the Materials and methods section, cDNA clones generated in this work were successfully used as Northern blot probes as well as positive-control templates for PCR for verification of splicing. Another benefit of cDNA analysis is that this transcriptome analysis is annotation independent: it does not rely on currently used genomic maps and annotations, and is thus especially well suited for discovery of novel transcripts, especially novel spliced transcripts, as well as antisense transcripts and transcripts coming from un-annotated regions, as was shown in the analysis of HCMV transcriptome by Zhang *et al.* [147].

1.4.2.2 Microarray analysis of transcriptome

DNA microarray consists of single-stranded DNA fragments (probes) attached to a solid substrate (glass, silicon chip or nylon) and is mostly used to determine gene expression levels. RNAs from the investigated cell or organism are isolated, converted to cDNA, fluorescently

or radioactively labeled and hybridized to the chip. After washing to remove non-specifically hybridized molecules, only sequences with a high degree of complementarity to the probes remain bound and the amount of signal is proportional to the abundance of transcripts in the sample. A laser scanner or charge-coupled device is used to record the fluorescent or radioactive signals respectively [48].

Since the first attempts at using arrays to monitor gene expression in 1995 [115], DNA microarrays have become a central part of a multitude of hybridization-based assays investigating not just transcriptomes but also DNA-protein interaction profiling, characterization of genetic variations like copy numbers or SNPs (single-nucleotide polymorphisms), genome-wide association studies, etc. DNA microarrays were the first practical technique for measuring gene expression at whole genome levels. Unlike cDNA analysis, microarray technology was well suited for high-throughput analyses as chips became more dense and allowed interrogation of whole genomes of simpler organisms. First DNA microarrays used cDNA libraries for probe designing. These microarrays, while remaining annotation independent, suffer from the same shortcomings and biases as cDNA library analysis. With better annotations came annotation-based gene or exome microarrays capable of monitoring not just gene expression but the use of alternative splice isoforms. Genome/exome arrays contain probes (usually around 50 bp long) corresponding to known or predicted genes/exomes. Some genome/exome arrays contain several different probes corresponding to different parts of the same gene/exome to increase the sensitivity of the array. Gene/exome DNA arrays have been widely used for over a decade, resulting in the development of a wide variety of commercially available, affordable and automatable platforms. Numerous computational analysis tools available as well as the development of nanoscale sample assays made microarrays even more popular and widespread, resulting in the application of DNA microarrays in almost every field of biology and biomedicine but especially in molecular profiling of various diseases by comparison of diseased and healthy cells and tissues, analysis of molecular pathways, toxico- and farmacogenomics, stem-cell research and others [54, 138]. As the analysis of non-coding RNAs gained momentum and importance, numerous non-coding transcripts have found their way into commercially available arrays. However, none of these arrays can detect novel transcripts.

DNA tiling arrays were developed to overcome the limitation of annotation and arrays based on previous knowledge. These arrays contained probes that spanned the whole genome in both sense and antisense orientations with various degrees of overlapping and could be used

for novel transcript detection and transcriptome mapping. Depending on the degree of overlapping between the probes from neighboring DNA regions, 5' and 3' ends of transcripts as well as novel splice junctions could be determined with resolution of several base pairs. The drawback of this approach was the need for a significantly increased number of different probes that increased the cost of the arrays and rendered them impractical for investigation of larger genomes.

Other drawbacks of microarray-based transcriptome analysis include: problematic analysis of highly related sequences due to cross-hybridization, variability in hybridization efficiency between the probes, signals from longer transcripts may obscure signals coming from overlapping shorter transcripts, low signal-to-noise ratio and detection range which may result in poor detection of rare transcripts [21, 120, 138]. The widespread use of microarrays resulted in the development of public databases, defined reporting standards and guidelines for good experimental design, nomenclature and file formats. Despite that, there have been several reports of inconsistent results from different research groups using the same platforms, variations between different platforms and differences in the interpretation of the same data [138]. Therefore, validation of selected genes using other techniques (Northern or Western blot, PCR or qPCR) is necessary before any conclusions are drawn.

1.4.2.3 RNASeq

Microarray analysis with its ability to simultaneously interrogate thousands of genes in multiple different samples has revolutionized biology and biomedical sciences, and led to the rapid development of systems biology. However, despite significant advances it has always been limited with the necessity of *a priori* decision on what part of the transcriptome will be analyzed (just mRNA, single chromosome with sense and antisense tilling arrays), especially when analyzing transcriptomes of complex organisms. The annotation-independent technique of cDNA analysis, on the other hand, was ill suited for high-throughput demands of systems biology. The development of next generation sequencing (NGS) enabled simultaneous fast sequencing of multiple targets without the need for cloning and thus effectively removed the two major obstacles of cDNA library analysis. RNASeq combined high-throughput ability of microarrays with the annotation independence of cDNA library analysis and with greater resolution.

In 1991 EMBL filed a patent application for large-scale DNA sequencing-by-synthesis technique that did not involve the use of gel electrophoresis (reviewed in [4]). Instead, DNA is sequenced by the addition of fluorescently labeled reversible terminator nucleotide to the growing DNA chain and measuring the resulting fluorescence by a sensitive CCD camera. A reaction mix contains all four nucleotides, each labeled with a different fluorophore. After the first nucleotide is incorporated and fluorescence measured, the dye is removed enabling the addition of next fluorophore-labeled nucleotide. This could be performed on a large number of different DNA molecules in parallel by miniaturizing the reaction. Each of these miniaturized reactions should be seeded by single DNA molecule which is attached to a solid surface (glass or silicone slides or beads) and then amplified in order for the fluorescent signal to be more visible. One of the names used for this new technology – massively parallel sequencing – underscores the main advantage of next-generation sequencing: the ability to interrogate numerous different sequences at once. In addition to speed, NGS has made sequencing of nucleic acids significantly cheaper than Sanger's method.

With the power to quickly sequence millions of molecules, NGS is the first technology powerful enough to allow direct sequencing of RNA. In next generation RNA sequencing (RNASeq in short), RNA from the cell is randomly fragmented, reverse transcribed to cDNA, attached to solid surface and then sequenced (for details on RNASeq procedure used in this thesis see chapter 3.2.9 and Figure 9). The obtained sequences are filtered for quality and aligned to target genome. Prior to sequencing, specific RNA subsets may be selected or removed from the RNA pool much in the same way as can be done for cDNA or microarray analysis (removal of rRNA or selection of polyA mRNAs). Quantification of transcripts is expressed as RPKM value (reads per kilobase per million mapped reads) that normalizes the number of reads mapped to a particular gene with the gene size and thus allows the comparison of expression levels of different genes within the same sample [96].

First RNASeq protocols resulted in 36-50 bp long sequences (called sequencing reads) and did not preserve strand information. In the past year numerous protocols that preserve strand information were developed as well as much longer sequencing reads. Strand information is very useful when interrogating transcriptomes of high-density genomes such as microbial or viral genomes. Longer sequencing reads and paired-end sequencing (sequencing protocol that sequences from both ends of DNA or cDNA molecule) have allowed *de novo* transcriptome assembly – analysis and reconstruction of transcriptome without reference genome. Thus

NGS has turned the tables and allowed transcriptomics to answer questions about genomes as well as enabled genomic and transcriptomic studies of non-model organisms [46].

In addition to annotation-independent transcriptome analysis, RNASeq allows precise quantification of transcripts and transcript isoforms, which is especially important when analyzing eukaryotic transcriptomes where on average a single gene produces 5 different transcript isoforms [140]. Unlike microarrays, RNASeq does not suffer from background noise and has a much wider dynamic range, allows single-nucleotide resolution of exact transcript ends, exhibits strong concordance between the platforms and lower technical variation (no need for technical replicates), is available for all species and can detect novel transcripts and transcripts that are a result of gene-fusion events as well as yield information about untranslated regions of transcripts [85, 143]. However, the unprecedented wealth of information and level of sensitivity bring in challenges of extracting useful biological meaning and necessitating the development of bioinformatic tools for aligning, visualizing and interpreting RNASeq data. In the beginning of RNASeq, most available tools were not user friendly and required extensive programming knowledge. Luckily, as RNASeq gains momentum, this is rapidly changing. In the past years, dozens of new RNASeq analyses and interpretation software have been developed, with user-friendly interface more adapted for biology and biomedical professionals. Advancement of NGS technology has led to longer and longer read lengths which open even more possibilities: from more accurate mapping of novel spliced events to *de novo* transcriptome-based genome sequence assembly. Downstream bioinformatic analysis of these results needs to follow these advancements. In short, NGS produces a wealth of information but in order to extract useful information good bioinformatic platform is needed. As will be discussed later, although bioinformatic analysis of NGS data is rapidly expanding, there are still outstanding issues. One especially important issue is transcript reconstruction for organisms with very dense genomes with multiple overlapping genes.

Other drawbacks include reduced efficiency of sequencing GC-rich transcripts or genomic regions, poor alignment of repetitive or highly related sequences where mapping software usually discards reads that can be aligned to multiple targets in the genome and potential overrepresentation of certain sequences that are more readily cleaved during fragmentation [117].

1.4.3 Transcriptomics of CMV

As was discussed in chapter 1.3.2, although the first sequence of MCMV has been available for 17 years and its genome annotation has seen several changes, the exact number of MCMV genes is still unknown. Additionally, the discovery of pervasiveness of antisense and non-coding transcription, and the increasing list of regulatory roles for non-coding transcripts in mammalian transcriptomes (reviewed in [119]) strongly argues in favor of the possibility that CMVs too possess non-coding transcripts.

One of the first transcriptomic studies of cytomegaloviruses employed cDNA library analysis of fibroblasts infected with AD169 strain of HCMV [147]. Although this analysis included only polyadenylated RNAs, it still detected abundant antisense transcription in the HCMV genome where 55% of all analyzed cDNA clones were antisense to known or predicted HCMV genes. Additionally, 45% of all analyzed cDNA clones came from the regions of the HCMV genome that were previously considered non-coding. This was the first report that showed abundant antisense and non-coding transcription in HCMV. Nowadays, in the light of active non-coding and antisense transcription in the transcriptome of virus' host, this finding is perhaps not so surprising. Herpesviruses have co-evolved with their hosts for millions of years and are known molecular pirates collecting useful genes from their hosts. Moreover, with genome size restrictions imposed by capsid size, the viruses could greatly benefit from antisense transcription.

In 2011 RNASeq was applied to HCMV transcriptome analysis confirming antisense transcription of most genomic regions but also showing that the majority of AS transcripts were expressed at much lower levels than their sense counterparts and accounted for only 8.7% of transcription from those regions [45]. In addition, it also showed that 65.1% of all polyadenylated transcripts are four non-coding RNAs (RNA2.7, RNA1.2, RNA4.9, and RNA5.0) that show minor overlapping with neighboring coding regions. Finally, this analysis resulted in the addition of four previously undetected new protein-coding regions to HCMV annotation (RL8A, RL9A, UL150A, andUS33A). Since only one time point (72 hours post infection) was used, there is still a possibility that HCMV transcriptome encodes even more alternatively spliced and novel coding and non-coding transcripts. Although both strand-specific and strand-nonspecific Illumina RNASeq protocols were used to sequence viral polyadenylated transcripts and long sequence reads were sequenced, HCMV genome proved to be too condensed and complex to enable genomic map reconstruction from transcriptomic

data. Data from the studies of Zhang *at al*. and Gatherer *et al.* [45, 147] pointed to an important problem in HCMV field: our annotations of HCMV are flawed and do not reflect the real transcriptional complexity of these viruses. The role of AS transcripts in context of viral infection is poorly understood, despite their pervasiveness. Many gene deletion mutants that have helped elucidate roles of viral genes have probably caused the deletion of AS transcript as well. Was the observed phenotype of the deletion virus the result of the deleted coding transcript or non-coding AS transcript? Clearly, better maps are needed and greater care needs to be taken when constructing new mutant viruses.

Analysis of HCMV proteome by applying ribosomal footprinting technique coupled with RNASeq revealed another layer of complexity of HCMV genome. Analysis by Stern-Ginossar *et al.* [127] identified 751 translated ORFs. The original number of predicted HCMV genes was up to 252 ORFs, so where do the additional 500 come from? The previously unidentified ORFs revealed by this analysis are internal ORFs lying within longer, previously described ORFs, either in frame (thus resulting in truncated proteins) or out of the frame (resulting in completely new proteins), short upstream ORFs lying upstream of the canonical ORFs, ORFs within AS transcripts and undetected short ORFs coming from distinct transcripts. Many of the new ORFs were very short (less than 20 codons), which may explain why prediction-based annotations have missed them. They were, however, confirmed by mass spectrometry or tagging approaches. The 24 previously annotated ORFs were not found to be translated. Most of the viral genes, newly detected ORFs included, showed tight temporal regulation of translation. Widespread use of alternative 5' start sites was also confirmed and was shown to follow tight temporal regulation as well.

The research of MCMV transcriptomics is much poorer than that of HCMV. Lacaze *et al.* [67] designed microarrays that contained 55-mer oligonucleotide probes in sense and antisense orientations to each 170 ORF predicted in MCMV genome according to the updated Rawlinson's annotation in order to test whether antisense transcription in MCMV is as pervasive as in HCMV. The authors analyzed NIH 3T3 cells infected with Smith strain MCMV at 0.5, 6.5, 24 and 48 hours post infection. Using stringent statistical testing, these authors detected 119 ORFs out of 170 tested ORFs as being expressed at all time points. It is important to note that in the effort to exclude false positive findings, many genes whose expression in the course of MCMV expression was previously validated or are homologs of HCMV genes known to be expressed, have not passed the threshold to be considered expressed in microarray experiment. These are M44, M70, M75, m135, m143, m144, m153,

and m157. The authors argue that the failure to detect majority of these genes is a consequence of very strict exclusion parameters, and this argument is underscored by the inability of the microarrays to detect the expression of *ie1*, and *ie2* was only detectable at 24 hours PI, while more sensitive techniques like RT-PCR could easily detect *ie1* and *ie2* already at 0.5 hours PI. This analysis detected antisense transcription from 35 loci throughout the MCMV genome, 4 of which could be confirmed by clones in the cDNA library described in this thesis. Based on these findings, the authors conclude that antisense transcription occurs relatively frequently in MCMV. One important caveat to this study is the use of just one probe per predicted gene. Since many of the predicted MCMV ORFs have not been experimentally validated, the correctness of the annotation is questionable. A signal from a probe does not prove the existence of that ORF; only that this particular region of the genome is transcribed. The real transcript could be completely different from the predicted ORFs, as is the case and is shown in this thesis.

Recently, RNASeq has also been applied in the analysis of MCMV transcriptome [83]. In this work, Marcinowski *et al.* analyzed newly transcribed RNA using 4-thiouridine labeling to dissect transcriptional activity of viral genes. This approach revealed a peak of viral gene expression very early after infection (1-2 hours PI). Interestingly, all genes were transcribed at 1-2 hours PI, not just immediate-early genes and viral reads accounted for 15% of all sequencing reads at that time point. At 5-6 hours PI the levels of viral gene transcription dropped to only 5% of all reads. The mechanism behind this peak and subsequent repression of viral gene expression is not fully understood. The authors of this work did not attempt to re-annotate MCMV genome based on RNASeq analysis.

All transcriptomic analyses conducted so far were based on the existing annotation provided by Rawlinson *et al.* [106], even the RNASeq analysis by Marcinowski *et al.* [83]. Rawlinson's annotation was based on the comparison of putative ORFs with those of other herpesviruses. As was discussed earlier in this chapter, first annotation-independent analysis of HCMV transcriptome [147] indicated numerous inconsistencies between the genomic map of HCMV and real transcriptional complexity. It is likely that the same will apply to MCMV. Therefore, in-depth analysis of MCMV transcriptome is the next logical and much needed step.

2. RESEARCH GOALS

The definitive genomic map of mouse cytomegalovirus is still lacking despite nearly two decades of research dedicated to MCMV genome. Currently used maps show only ORFs although a growing body of evidence suggests that non-coding transcription as well as antisense transcription are also present and play important roles in the transcriptomes of cytomegaloviruses [45, 67, 127, 147]. Moreover, a significant portion of annotated ORFs lacks experimental confirmation.

Murine cytomegalovirus is an important model virus for the study of human CMV disease. Thanks to the use of MCMV, numerous viral evasion mechanisms employed by these viruses have been elucidated, mostly by use of viral gene deletion mutants. Good gene deletion mutant targets a single gene of interest, while leaving the rest intact. In dense genomes such are those of viruses this task requires precise genomic and transcriptomic maps, especially in the light of recent findings of intense antisense transcription in CMV genomes [45, 147]. Although RNASeq analyses can now be used for *de novo* assembly of transcripts, dense viral genomes still pose problems for currently available bioinformatic tools.

The goal of this work was to perform annotation-independent and in-depth analysis of murine cytomegalovirus. To that aim we analyzed all polyadenylated MCMV transcripts that accumulate in the infected cells using two different, annotation-independent approaches in order to complement the shortcomings of each method method: cDNA library analysis and RNASeq analysis. In order to detect as many transcripts as possible, mouse embryonal fibroblasts (MEF) were used as this primary cell line is highly permissive for infection and has already been used in other genomic and transcriptomic analyses of MCMV. Although it has previously been shown that MCMV genome is highly stable in *in vitro* passage [24], the least passaged MCMV Smith strain was used to minimize the impact of random mutations on the structure and abundance of MCMV transcripts. Using the information obtained from this dual analysis, a map of MCMV transcriptomic profile was constructed showing multiple novel spliced and unspliced transcripts. Such a map will be a useful tool for CMV virologists and present an important first step in re-annotation of MCMV genome.

While only viral transcripts were analyzed in the course of cDNA analysis, RNASeq technology allowed the analysis of not only the virus transcriptome but also that of its host. Such high-throughput analyses usually yield extensive list of genes. To gain meaning,

multiple bioinformatics tools have been used to analyze the list of perturbed genes based on their gene ontology associations, function, involvement in certain signaling or canonical pathways, or localization to certain subcellular compartment or involvement in a particular disease. These analyses can identify deregulated pathways and potential diseases and conditions associated with deregulation of a certain pathway. As life-long persistent pathogen, HCMV is now being implicated in numerous autoimmune conditions and cancers – this analysis has the potential to confirm some of the speculations and point out some overlooked ones.

In short, this work presents the first in-depth and annotation-independent analysis of MCMV and its host's transcriptome.

3. MATERIALS AND METHODS

3.1 MATERIALS

3.1.1 Plasmids

The cDNA library was constructed by cloning cDNA fragments into pFIN2 plasmid (Figure 7). The pFIN2 is a modified pcDNA3.1(+) plasmid (Invitrogen) in which second PmeI restriction site inside a multiple-cloning site (MCS) was modified into PacI restriction site. This plasmid was chosen for the construction of cDNA library since it allows high-level expression in most mammalian cells. This feature enables us to transfect individual interesting cDNA clones into various mammalian cell lines and use different *in vitro* assays to test their function.

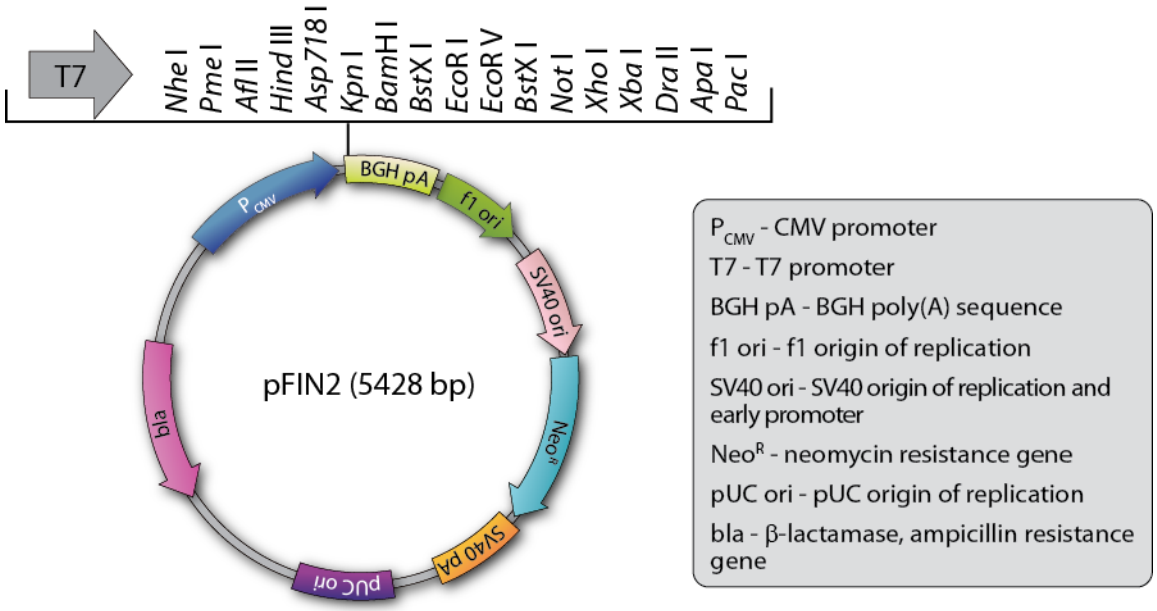


Figure 7. pFIN2 plasmid used for generation of cDNA library. cDNA fragments were directionally cloned into EcoRI and PacI restriction sites inside the MCS (shown at the top) under the control of T7 promoter (gray arrow).

For bacterial expression, pQE-30 plasmid (Qiagen) was used. pQE-30 is low-copy expression plasmid for the production of N-terminally His-tagged proteins (6×His).

3.1.2 Bacterial strains

For cDNA library construction, XL-1 Blue Supercompetent *E. coli* (Stratagene) was used.

The m169 viral protein was produced in *E. coli* Bl21 pREP4 strain. This bacterial strain permits high-level expression of proteins under the IPTG-inducible *lacZ* promoter and contains an additional plasmid with lac repressor protein (pREP4), which prevents constitutive expression of cloned protein.

3.1.3 Cell lines

Primary mouse embryonic fibroblasts (MEFs) from BALB/c or Balb.K mice were prepared and maintained in DMEM supplemented with 3% FCS as described [17] and used between passages 3-5.

Immortalized MEF cell lines used in this thesis are MEF.K, SVEC4-10, and B12. All immortalized cell lines are adherent and were maintained in Dulbecco's Modified Eagle Medium (DMEM) with 10% FCS.

Immortalized murine BALB.K MEFs, (MEF.K) and SVEC4-10 were used for Jag2 studies.

MEF.K are SV40-immortalized MEF cells originating from Balb.K mice [60].

SVEC4-10 (ATCC CRL-2181) are SV40-transformed endothelial cells from C3H/J mouse.

B12 are SV40-immortalized MEF cells from Balb/c mice [39].

Raw264.7 are SV40 transformed peritoneal macrophages from Balb/c mice, grown in 10% RPMI.

SP2/0-Ag14 cells are murine myeloma cells originating from Balb/c mice and most commonly used as fusion partners, grown in 10% RPMI. They do not secrete immunoglobulin, are resistant to 8-azaguanine at 20 mg/mL and are HAT (hypoxanthine-aminopterin-thymidine) sensitive.

Ly49 2B4-NFAT reporter cell line with activating Ly49 receptors contains GFP reporter gene under the control of nuclear factor of activated T-cell (NFAT) promoter and was constructed

as previously described in [40]. The cells were grown in 10% RPMI as semi-adherent cell line.

3.1.4 Viruses

All viruses used in this work are listed in Table 2.

Table 2. Viruses used in this thesis

Virus	Description (deleted genome regions)	References
Smith (ATCC VR-1399)	wild type	[106]
MW97.01	wild type, BAC-derived Smith	[88]
K181	wild type, field isolate, GenBank acc no: AM886412.1	[108]
C4A	wild type, field isolate	gift from A. Redwood
C4D	wild type, field isolate	gift from A. Redwood
G4	wild type, field isolate	gift from A. Redwood
K6	wild type, field isolate	gift from A. Redwood
WP15B	wild type, field isolate, GenBank acc no: EU579860.1	gift from A. Redwood
Δm04	MW97.01 missing m04 gene (5300-6334)	[135]
Δ7S3-GFP	MW97.01 missing m167-m170, expresses GFP	[84]
Δm168- Δm169	deletion of m168 and m169	
Δm169- Δm170	deletion of m169 and m170	
Δm168	deletion of m168 (227 920-228 462)	[84]
Δm169	deletion of m169 (C; 228 313–228 708)	[84]
Δm170	deletion of m170 (C; 229 342–230 046)	[84]
m168-mut	mutated binding site for miR-27b in MAT region	[84]
C3X m169 Δ5'UTR	deletion of 5'UTR of MAT transcript	gift from L. Dölken
C3X m169 Δintron	deletion of intron in MAT region	gift from L. Dölken

Δm04, Δ7S3, Δm167, Δm168, Δm169, Δm170, Δm168-Δm169 and Δm169-Δm170 mutant viruses were generated by ET-cloning [137] using the full-length MCMV BAC pSM3fr [136].

Primers for the construction of the double deletion mutants have also been described [84] using the forward primer for the first gene and the reverse primer for the second gene.

3.1.5 Growth media for *E. coli*

Bacteria *E. coli* were grown in LB broth medium (composition shown in table below). All reagents were dissolved in double distilled water and autoclaved for 20 minutes at 121 °C.

LB broth	bacto-tryptone	10 g/L
	yeast extract	5 g/L
	NaCl	10 g/L
	(agar) (for agar plates)	15 g/L
	(ampicillin) (used as needed)	50-100 μg/mL

3.1.6 Animal cell media

All media used in this work are based on the commercially available DMEM or RPMI-1640 media (Gibco) whose composition is shown in the table below. Supplements (antibiotics, DMSO and/or FCS) are also shown below.

Dulbecco´s Modified Eagle Medium (DMEM)	fetal calf serum (FCS)	3-10%
	HEPES (pH 7.2)	10 mM
	L-glutamine	2 mM
	penicillin	10^5 U/L
	streptomycin	0.1 g/L
RPMI-1640	fetal calf serum (FCS)	3-10%
	HEPES (pH 7.2)	10 mM
	L-glutamine	2 mM
	penicillin	10^5 U/L
	streptomycin	0.1 g/L
	2-mercaptoethanol	$5 \cdot 10^{-5}$ g/mol

Freezing medium	RPMI-1640	70%
	fetal calf serum	20%
	dimethyl-sulfoxide (DMSO)	10%
Freezing medium for reporter cells	fetal calf serum	90%
	dimethyl-sulfoxide (DMSO)	10%
FACS medium	PBS, 1X	
	bovine serum albumin (BSA)	1%
	sodium azide	0.1%

3.1.7 Solutions and buffers

All buffers used in this work and their compositions are shown below. All buffers were made by dissolving the ingredients in double distilled water and were then sterilized by filtration if needed.

Phosphate-buffered saline (PBS) 10×	NaCl	140 mM
	KCl	2.7 mM
	Na_2HPO_4	6.5 mM
	KH_2PO_4	1.5 mM
	$CaCl_2$	0.7 mM
	$MgCl_2 \cdot 6H_2O$	0.7 mM
VBS buffer (pH adjusted to 7.8 with HCl)	Tris-HCl	50 mM
	KCl	12 mM
	Na_2EDTA	5 mM
TEN buffer for annealing oligonucleotides	Tris-HCM, pH 7.5	10 mM
	EDTA	0.1 M
	NaCl	25 mM

3.1.7.1 Buffers for purification of nucleic acids

GTE buffer (alkaline lysis solution 1)	glucose	50 mM
	EDTA	10 mM
	Tris-HCl (pH 8.0)	25 mM
NaOH/SDS (alkaline lysis solution 2)	NaOH	0.2 M
	SDS	1%
Neutralization solution (alkaline lysis solution 3); pH 4.8 adjusted with acetic acid	Sodium acetate	3 M
TE buffer	Tris-HCl (pH 9.0 ili 7.4)	10 mM
	EDTA (pH 8.0)	1 mM
	RNase (optional)	10 μg/mL

3.1.7.2 Buffers for gel electrophoresis of nucleic acids

Tris-acetate (TAE) buffer, 50X	Tris-base	242 g
	glacial acetic acid	57.1 mL
	500 mM EDTA (pH 8.0)	100 mL
	water	to 1 L
DNA loading buffer	bromophenol-blue	2.5 g/L
	xylene cyanol	2.5 g/L
	glycerol	1 mL
agarose gel	agarose	0.8-1.2%
	1× TAE buffer	
MOPS running buffer 10×	MOPS (pH 7)	0.2 M
	sodium acetate	20 mM
	EDTA (pH 8)	10 mM

formaldehyde agarose gel	agarose	1.5%
	1× MOPS buffer	
	formaldehyde, 37%	2.2 M
5× RNA loading buffer for formaldehyde agarose gels (10 mL)	bromophenol blue	pinch
	xylene cyanol	pinch
	EDTA, 500 mM	80 µL
	formaldehyde, 37%	720 µL
	glycerol, 100%	2 mL
	formamide	3084 µL
	10× MOPS	4 mL
	RNase-free water	to 10 mL

3.1.7.3 Buffers for transfer of nucleic acids to positively charged membranes

Washing buffer pH 7.5	maleic acid	0.1 M
	NaCl	0.15 M
	Tween 20	0.3%
Maleic acid buffer pH 7.5	maleic acid	0.1 M
	NaCl	0.15 M
Detection buffer pH 9.5	Tris-HCl	0.1 M
	NaCl	0.1 M
Saline-sodium citrate (SSC), 20×	NaCl	3 M
	sodium citrate	0.3 M

3.1.7.4 Buffers for isolation and separation of proteins by SDS-PAGE

RIPA (Radio Immunoprecipitation Assay) cell lysis buffer	Tris	25 mM
	NaCl	150 mM
	Na deoxycholate	1%
	SDS	0.1%
	NP40	1%
	EDTA	1 mM
	NaF	10 mM
	$NaVO_4$	1 mM
	PMSF	2 mM
NP40 cell lysis buffer	Tris	25 mM
	NaCl	150 mM
	NP40	1%
	NaF	10 mM
	$NaVO_4$	1 mM
	PMSF	2 mM
Laemmli running buffer, 10× pH 8.3	Tris	0.25 M
	glycine	1.92 M
	SDS	1%
2x sample loading buffer	Tris, 0.5 M, pH 6.8	1.2 mL
	glycerol, 100%	1.9 mL
	SDS; 10%	2 mL
	2-mercaptoethanol	0.5 mL
	bromophenol-blue, 1%	1 mL
	distilled water	to 16 mL

3.1.7.5 Buffer for transfer of proteins to PVDF membrane

Transfer buffer	Tris	20 mM
	glycine	150 mM

	methanol	10-20%
	SDS	0.005%

3.1.7.6 Buffers for Western blot

Transfer buffer	Tris	20 mM
	glycine	150 mM
	methanol	10-20%
	SDS	0.005%
TBST, 1×	Tris-HCl, pH 7.4	20 mM
	NaCl	150 mM
	Tween 20	0.05%
blocking buffer, 3% BSA	BSA	3%
	NaN_2	0.02%
	Tween 20	0.1%
	1× PBS	
blocking buffer, milk	non-fat milk	5%
	1× PBS	

3.1.8 Antibodies

All antibodies used in this work, their host, specificity, isotype and source are listed in the table below. They were used according to the manufacturers' instructions.

Antibody	Host	Specificity	Isotype	Source
m169	mouse	mouse	IgG2	hybridoma supernatant produced in house
m04	mouse	mouse		hybridoma supernatant produced in house
actin (clone C4)	mouse	mouse	IgG1, κ	Millipore
β-integrin (CD29; clone 9EG7)	rat	mouse	IgG2a, κ	BD
Jag2 (clone N19)	goat	mouse, rat, human	IgG	Santa Cruz

Antibody	Host	Specificity	Isotype	Source
Insm1	rabbit	human, mouse	polyclonal serum	LSBio
GAPDH	mouse	mouse, human	IgG1	Millipore
En2	rabbit	human, mouse	polyclonal	Thermo Scientific
Agtr2	rabbit	human, mouse	polyclonal	Sigma-Aldrich
Delta	rabbit	human, mouse	polyclonal	Santa Cruz
Trim71	goat	mouse	polyclonal	Thermo Scientific
anti mouse-POD	goat	mouse	IgG	Jackson ImmunoResearch
anti rat-POD	goat	rat	IgG	Jackson ImmunoResearch
anti goat IgG-HRP	rabbit	goat	polyclonal serum	Abcam

3.1.9 Oligonucleotides

Oligonucleotides used for cloning and sequencing in this study are listed in Table 3, while oligonucleotides used to generate probes for Northern blot are listed in Table 4.

Table 3. Oligonucleotides

Oligonucleotide	Sequence	Use
PacI primer-adapter	5’-GCGGCCGCTTAATTAACC(T)^{15}-3’	cDNA library generation
EcoRI-PmeI adapter	5’-AATTCCCGCGGGTTTAAACG-3’	cDNA library generation
	5’-Pho-CGTTTAAACCCGCGGG-3’	
m169 PCR primers	F: 5'-TTTTTGGATCCATGAGCAACGCGGTCCCGTTC-3'	bacterial expression of putative m169 protein
	R: 5'-TTTTTCTGCAGTCATCACGGGGGGCACCTACC-3'	
3’ sequencing primer	5’-GCACCTTCCAGGGTCAAGGAAG-3’	sequencing of cDNA clones from 3’ end

Table 4. cDNA clones and oligonucleotides used to generate probes for Northern blot

	antisense probe			sense probe
region	**clone name**	**genomic location**	**genomic strand**	**Oligonucleotides***
m15-m16	E119	14027-15700	+	F:AATTAACCCTCACTAAAGGGAAAAGTATTGCGTATAAGACACT
				R: TCAAGAAGATGTACCGTCAC
m20-19	IE205	21144-20434	-	F:AATTAACCCTCACTAAAGGGAGAAAAGATTCTTTATTGCGTCGAG
				R: AGCGCGATGCTGTTACG
	L57	21371-20436	-	F: NA
				R:NA
m72	L69	103534-104161	+	F:AATTAACCCTCACTAAAGGGGCTCCGGTCCGCCCGAAT
				R: GGCAGCTCCAGCGGACCC
m74	L147	104825-105449	+	F:AATTAACCCTCACTAAAGGGACAGAGGTGGCGAGCATCAAA
				R: GAAAAATTGTATCGGGTGCATGTTTTC
M75	L42	105878-106095	+	F:AATTAACCCTCACTAAAGGGAGAAAAGATTCTTTATTGCGTCGAG
				R: AGCGCGATGCTGTTACG
M100	E126	145353-144169	-	F:AATTAACCCTCACTAAAGGGCGCGTATCTCTTCGTTGTCCA
				R:ATTACCCGCGCATCATCGAC
M102	E14	147457-148161	+	F:AATTAACCCTCACTAAAGGGTCGTCTTTTGCAGTGTGTCT
				R: CATCCGCTTCATGGCCAC
M103	L51	148772-148169	-	F:AATTAACCCTCACTAAAGGGTTTTATTGTTCGAGGCGCTTT
				R: ACCTTCCTGACCGGCACCA
M116	E140	169140-168095	-	F:AATTAACCCTCACTAAAGGGCCTGCTGAGGAGTAGTCTTGG
				R: TGTCGGCGCGCTGCTCT

*Underlined sections are T3 promoter sequences

3.1.10 Other chemicals, enzymes, kits and membranes

restriction endonucleases	New England Biolabs
Amersham Hybond N$^+$	GE Healthcare

PVDF membrane	Roche
TRIzol	Invitrogen
vanadyl ribonucleosite complexes	Sigma
RNase OUT	Invitrogen
Freund's adjuvant	Sigma
BCA Protein Assay Kit	Pierce
Enhanced Chemiluminescence (ECL) Detection System	GE Healthcare
BSA (bovine serum antigen)	Roth

3.2 METHODS

3.2.1 Plasmid DNA purification

Plasmid DNA was purified either using QIAPrep Spin Miniprep kit (Qiagen) following the manufacturer's instructions or using standard alkaline lysis protocol [113].

3.2.2 General techniques for handling animal cells

Animal cells were cultured in Petri dishes or flasks at 37 °C in 5% CO_2 in DMEM or RPMI media supplemented with FCS as needed (see section 3.1.6). Adherent cell lines were detached by incubation with pre-warmed trypsin or 2 mM EDTA (when assessing surface proteins that are sensitive to trypsin). Semi-adherent cell lines (like Ly49 reporter cells) were detached by repeated washing with media using an automatic pipettor.

Cell number was determined by staining the cells with trypan blue stain and counting in Neubauer hemocytometer. A volume of 25 µL of cell suspension was mixed with 200 µL of trypan blue and loaded into Neubauer hemocytometer. Only unstained (live) cells were counted. The concentration of the cells was calculated using the following equation:

$$c\ (cells\ in\ mixture) = \frac{N(counted\ cells)*}{proportion\ of\ chamber*V(chamber)} \times \frac{V(sample\ dilution)}{V(original\ cell\ suspension)}$$

where V(chamber) is 0.001 mm^3 and V(sample dilution) is 25 µL.

3.2.3 Production of primary mouse embryonic fibroblasts

Primary mouse embryonic fibroblasts were derived from aseptically removed day 18 old embryo. After the removal of placenta and visible organs, the tissue was homogenized first by cutting with scissors and then by treatment with trypsin for 90 minutes with agitation. This homogenate was finally passed through wire mesh, washed with 3% DMEM and incubated in 3% DMEM for 24 hours. After 24 hours, new 3% DMEM medium was added and the cells were left to grow for 3-4 days until they were confluent. The cells were stored at -80 °C in a freezing medium at a concentration of $5 \cdot 10^6$ cells/aliquot.

3.2.4 Cryopreservation of animal cell lines

Cells were kept in a freezing medium at -80 °C for short-term storage and in liquid nitrogen for long-term storage. Cells grown in culture were detached (if adherent) by treatment with trypsin, washed with the medium, resuspended in freezing medium and slowly frozen in cryovials placed in isopropanol bath at -80 °C to achieve a cooling rate of 1-3 °C/minute. Unlike freezing, thawing of the cells was done quickly by placing the cryovials in a water bath at 37 °C. After they were thawed, the cells were washed with the medium and then cultured at 37 °C and 5% CO_2.

3.2.5 Production of tissue-derived virus and preparation of virus stocks

All viruses were produced on Balb/c primary MEFs by infecting the cells with 0.01 PFU/cell without centrifugal enhancement. After 4-5 days, when all cells displayed cytopathic effects, the cells and supernatant were collected using a cell scraper and cellular debris was pelleted by centrifugation at 6400×*g* for 20 minutes. The supernatant containing virus particles was then centrifuged at 26,000×*g* at 4 °C for 90 minutes. The supernatant was discarded, leaving 200 µL overlaying the pellet. The virus pellet was stored overnight on ice. The next day the pellet was resuspended in the medium and overlaid with 18 mL of 15% sucrose/VSB buffer in a new set of sterile ultracentrifuge tubes. The virus was gradient purified at 72,000×*g* at 4 °C for 2 h. Purified virus pellet was resuspended in 25% sucrose/VSB, then aliquoted and stored at -80 °C. The virus titer was determined by standard plaque assay on Balb/c MEF without centrifugal enhancement (see below, section 3.2.6) as described in [55].

3.2.6 Infection of adherent cells

Primary and transformed MEF cell lines, SVEC4-10 and Raw 264.7 cells were infected with MCMV in the course of this study. Since all of these cells are adherent cell lines, infection procedure was the same. Primary MEFs were infected with 0.1-0.5 PFU, while transformed cell lines were infected with 1-5 PFU/cell to ensure complete infection and to prevent uninfected fast-diving cells from overgrowing the infected ones. The cells were counted and plated at least 4 hours before infection in 10-mm Petri dishes (medium sized Petri-dishes). The virus was resuspended in 3 mL of medium/Petri dish. The medium overlaying the cells was removed by aspiration and the cells were overlaid by 3 mL of virus suspension and incubated at 37 °C in 5% CO_2 for 30 minutes. To facilitate synchronous entry of viruses into the cells, after 30-minute incubation the cells were centrifuged at 800×*g* for 30 minutes. This procedure, termed centrifugal enhancement, enhances the infection by 10- to 20-fold.

3.2.7 Isolation of MCMV genomic DNA

Smith MCMV-infected cBalb/c MEFs were harvested 72 h post infection, and viral DNA was isolated as described previously [147]. Briefly, infected cells were collected by centrifugation, the cell pellets were resuspended in 5 mL of 150 mM NaCl, 10 mM Tris (pH 7.4), and 1.5 mM $MgCl_2$. After incubation on ice, NP-40 was added to a final concentration of 0.1%. The lysate was centrifuged at 3,700 rpm for 20 min using a Beckman GS-6R centrifuge. The supernatant was collected and brought to a final concentration of 0.2% sodium dodecyl sulfate (SDS), 0.5 mM EDTA, and 50 mM β-mercaptoethanol. After incubation on ice and extraction with phenol-chloroform, the genomic DNA was precipitated with ethanol, resuspended in 1 mL of Tris-EDTA buffer, and treated with RNase (Sigma-Aldrich). The genomic DNA was further purified by centrifugation in a linear 5 to 20% (wt/vol) potassium acetate gradient at 40,000 rpm for 3.5 h at 20°C in a Beckman L7 Ultracentrifuge SW60 rotor. Following centrifugation, the DNA was collected, precipitated with ethanol, and resuspended in 50 µL of distilled water. The purified genomic DNA was digested with MseI, followed by phenol-chloroform extraction and ethanol precipitation. The digested genomic DNA was finally resuspended in 50 µL of sterile water.

3.2.8 Construction of MCMV cDNA library, positive selection of clones and sequencing

Total RNA was extracted from Smith MCMV-infected Balb/c MEF at 4, 8 and 12 h after infection (IE library); 16, 24 and 32 hrs after infection (E library); and 40, 60 and 80 hrs after infection (L library). No drugs were used to select for different temporal classes of transcripts, and equal amounts of RNA from each time point were pooled prior to library construction. cDNA libraries were generated as described previously for HCMV [147] by following the instruction manual for the SuperScript Plasmid System with Gateway Technology for cDNA Synthesis and Cloning (Invitrogen), with some minor modifications and is shown schematically in Figure 8. Briefly, total RNA was isolated using the TRIzol Reagent. A poly(T)-tailed PacI primer-adapter was used for first-strand cDNA synthesis (Table 3). After second-strand synthesis, an EcoRI-PmeI adapter was added to the 5' end and cDNAs were cleaved with PacI.

The EcoRI-PmeI adapter was generated by annealing 2 oligonucleotides listed in Table 3. Equimolar amounts of both oligonucleotides were mixed in TEN buffer, the sealed Eppendorf tube was placed in 1 L of boiling water, boiled for 5 minutes and left to slowly cool overnight.

cDNA fragments were size fractionated to remove excess adapters and prevent small cDNA clones from dominating the library. Different size fractions of cDNA clones were then inserted into pFIN2 (Figure 7) previously digested with EcoRI and PacI and transformed into XL1-Blue Supercompetent *E. coli* cells.

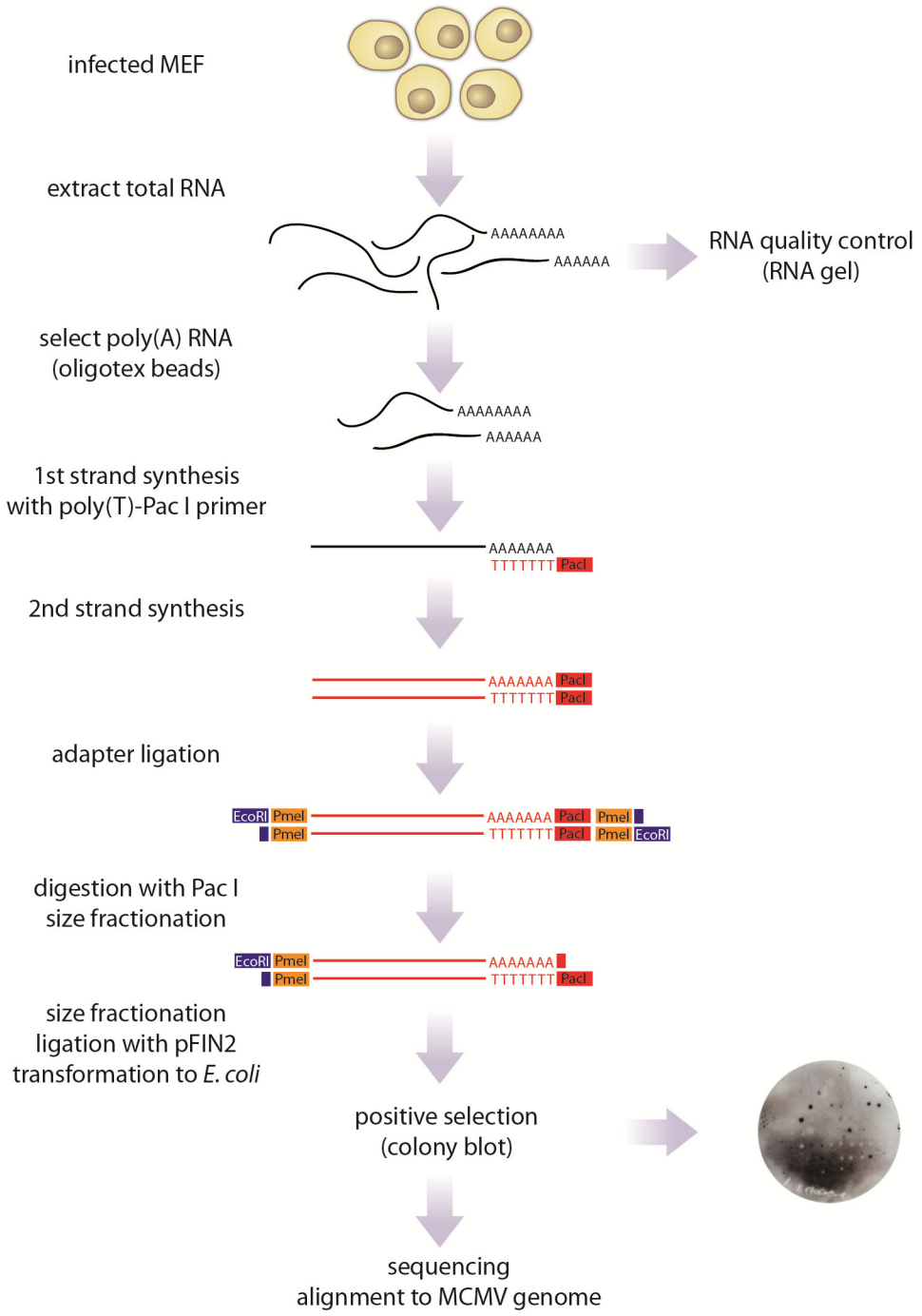


Figure 8. Schematic overview of cDNA library construction. Total RNA was isolated from the infected cells at various time points post infection to ensure transcripts of all temporal classes were included in the library. Majority of RNAs in a cell were ribosomal RNAs with 1-5% of messenger RNAs (mRNAs). Most mRNAs have polyadenylated [poly(A)] tails and can be selected by passage through poly(T) Oligotex column. Poly(A) tails were utilized again as primer binding site for first strand synthesis. PacI adapter was added on 5' end of poly(T) primer to allow directional cloning of cDNA fragments into pFIN2 plasmid. Second strand was synthesized by nick translation. Finally, EcoRI-PmeI primer adapter was added. Fragments were subsequently cut with PacI, thus producing different ends on 5' and 3' ends of cDNA clones, size fractionated and ligated into PacI and EcoRI

digested pFIN2 plasmid. Plasmids bearing viral cDNA fragments were detected by colony blot using viral DNA as a probe, sequenced from 5' and 3' ends and aligned to MCMV genome using megaBLAST.

Positive selection of viral cDNA clones was performed as described previously [147]. Briefly, random transformed bacterial colonies were picked on agarose plates (approximately 100 colonies per plate), transferred to nylon membrane and lysed as described in [53]. MseI-digested and DIG-labeled viral DNA was used as a probe and generated using DIG High Prime DNA Labeling and Detection Starter Kit II (Roche) according to the manufacturer's instructions. Plasmids harboring cDNA clones that reacted with the probe were isolated and sequenced from the 5' end using T7 primer for pcDNA3.1(+) or the 3' ends (primer listed in Table 3) or standard poly(T) primers at the OSU Plant-Microbe Genomics Facility. Sequences were compared to the MCMV Smith strain genome [GenBank accession no. NC_004065] using mega BLAST.

3.2.9 Next generation sequencing – library preparation, alignment and analysis

Total RNA was extracted from Balb/c MEF cells cultured in 100-mm^2 Petri dishes and exposed to 0.3 PFU/cell of the MW 97.01 strain of murine cytomegalovirus or mock-infected. At 4, 8, 12, 16, 24, 32, 40, 60 and 80 hours after infection, RNA was isolated using TRIzol Reagent. RNA integrity was assessed on Agilent Bioanalyzer and only samples with RNA index values of at least 9 were used. Equal amounts of RNA from each time point were pooled (0.3 µg of RNA per time point) and treated with DNaseI. Libraries were prepared with Illumina TruSeq RNA kit according to the manufacturer's instructions and sequenced on Illumina Genome Analyzer IIx as single-end 36-bp reads. Schematic overview of RNASeq library generation, alignment and analysis is shown in Figure 9.

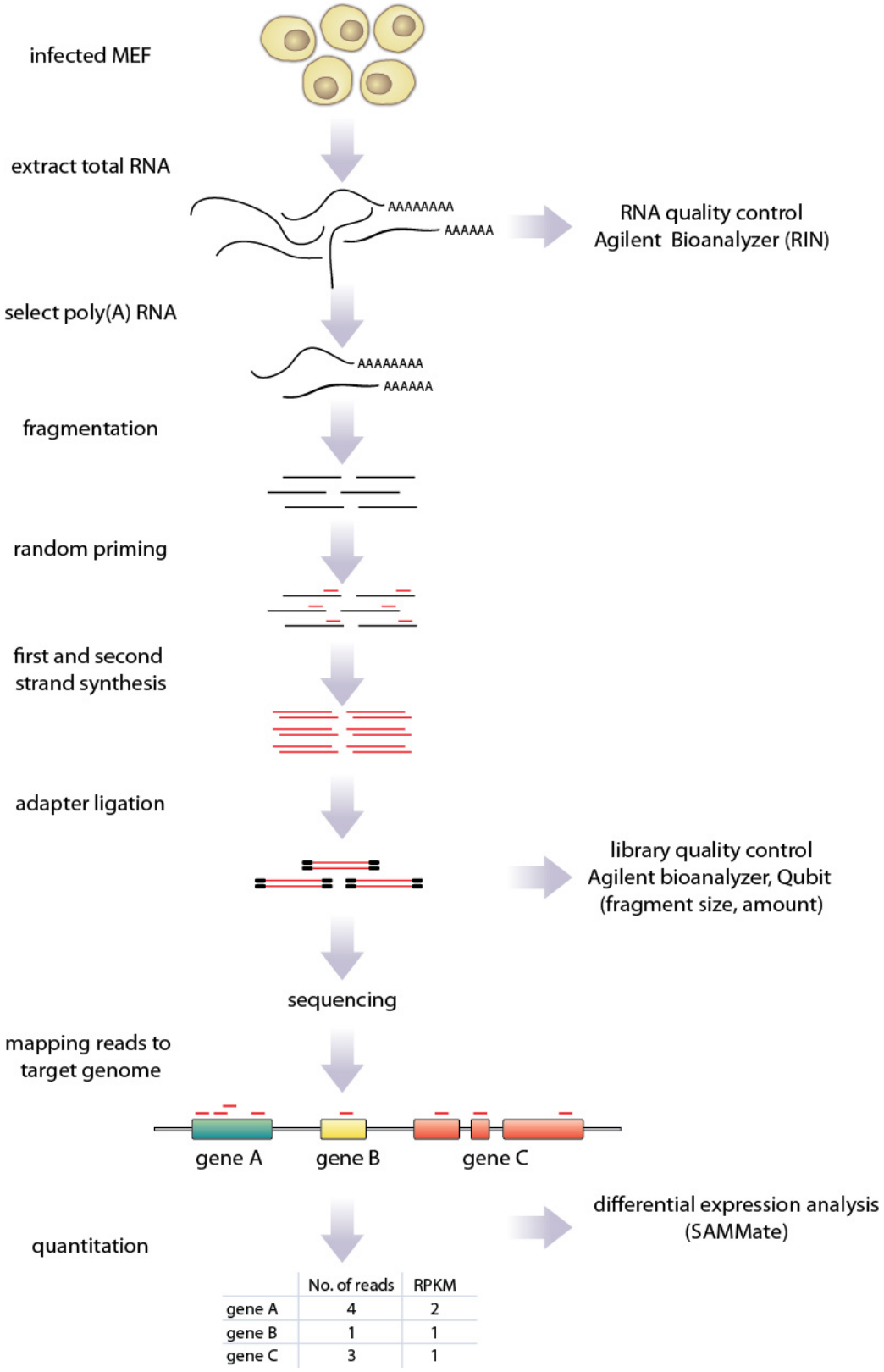


Figure 9. Schematic overview of RNASeq library generation, alignment and analysis. Total RNA was isolated from the infected MEF cells at the same time points used for generation of cDNA library. Equal amounts of RNA from each time point were pooled (0.3 µg for each time point), treated with DNase I and RNA quality assessed on Agilent bioanalyzer. Poly(A) RNA was selected on poly-T oligo-attached magnetic beads, fragmented, randomly primed and reverse transcribed to double

stranded cDNA fragments. Staggered cDNA ends of cDNA left after second strand synthesis were repaired by filling 5' overhangs and cleaving off 3' overhangs. Finally, 3' ends were adenylated to prevent cDNA fragments from ligating to one another during primer-adapter ligation and then primer-adaptors are added. cDNA library was checked for fragment size and concentration on Agilent bioanalyzer and then sequenced on Illumina GA IIx. Illumina's ELAND aligner was used to align reads to target genomes (mouse and MCMV) and the alignment was visualized in Integrated Genomics Viewer. Differential expression was calculated from ELAND alignments using SAMMate.

Reads were aligned to mouse (NCBI37/mm9 assembly) and MCMV genome (GenBank acc.no. NC_004065.1) using ELAND aligner or Bowtie aligner (for Dölken group data). Alignments were visualized using Integrative Genomics Viewer (http://www.broadinstitute.org/igv/) [110]. Differential gene expression was assessed by calculating RPKM (reads per kilobase of million mapped reads (RPKM) using SAMMate 2.6.1. release with EdgeR (http://sammate.sourceforge.net/) [145]. Gene ontology (GO) enrichment analysis was performed on filtered lists of differentially expressed genes ($p<0.05$) using GOrilla ranked list analysis [41, 42]. Ingenuity Core Analysis (Ingenuity® Systems, www.ingenuity.com) was used for gene interaction network and canonical pathway analysis. Gene lists were filtered for statistically significant differentially expressed genes ($p<0.05$) and a fold change cut-off of 2 was set to identify molecules whose expression was significantly differentially regulated. For network generation, these molecules (Network Eligible molecules) were overlaid onto a global molecular network developed from the information in the Ingenuity Knowledge Base based on their connectivity. The Functional Analysis of a network identified the biological functions and/or diseases that were most significant to the molecules in the network. Right-tailed Fisher's exact test was used to confirm that biological functions and/or a disease assigned to data sets were not due to chance.

3.2.10 Northern blot analysis

For MAT Northern blot analysis, RNA was isolated from mock- or MCMV-infected Balb/c MEFs collected 24 h after infection. RNA (10 µg/lane) was separated by formaldehyde agarose gel electrophoresis and transferred to a nylon membrane by capillary transfer over-night as described in [113]. After transfer, membranes were rinsed in water, 50mM NaOH, and 10× SSC, and cross-linked by baking and exposure to UV light at 800 J/cm^2.

Fragments corresponding to the MAT gene sequences derived from cDNA library clones E1, E125 and E134 (the longest MAT clones in our library) were used as a probe. Plasmid DNA

was isolated from bacterial glycerol stocks, cut with EcoRI and PacI (releases cloned cDNA fragment), followed by isolation of MAT DNA fragment from the gel using QIAquick gel extraction kit (Qiagen) according to manufacturer's instructions. Equal amounts of DNA from each cDNA clone were mixed and DIG-labeled using DIG High Prime DNA Labeling and Detection Starter Kit II (Roche) according to the manufacturer's instructions. To neutralize RNases, the probe was additionally treated with RNase OUT (40 μL/mL; Invitrogen) and vanadyl ribonucleoside complexes (500 μL per probe; Sigma). Membranes were hybridized to the probe overnight at 65 °C and detected according to the manufacturer's instructions.

For all other Northern blot analyses: RNA was isolated using TRIzol reagent from mock- or MCMV-infected Balb/c MEF at 10, 30 and 60 h after infection. RNA (1 μg/lane) was separated by formaldehyde agarose gel electrophoresis, transferred to positively charged nylon membrane and crosslinked by UV irradiation. Membranes were reacted to DIG-labeled probes overnight at 67 °C. Single-stranded DIG-labeled RNA probes were generated using Roche's DIG Northern Starter Kit. Antisense probes were generated by *in vitro* transcription from T7 present in pcDNA3.1 plasmids containing cDNA clones that harbor the desired gene fragments (Table 4). Therefore, antisense probes were identical to transcripts cloned in cDNA library and could detect transcripts antisense to cDNA clones. To generate sense probes, T3 promoter was added to 5' end of complimentary strand of the gene fragments used for antisense probes by PCR (Table 4). The PCR fragments were then transcribed *in vitro* and DIG labeled using T3 RNA polymerase. Care was taken to generate sense probes of length comparable to the corresponding sense probes.

3.2.11 Generation of the antibody against m169

The putative m169 gene sequence was amplified by PCR using viral DNA isolated from MCMV BAC pSM3fr using oligonucleotides listed in Table 2. Amplified PCR fragments were cut with BamHI and HindIII restriction endonucleases and inserted into pQE30 expression vector previously digested with the same endonucleases and then introduced to *E. coli* Bl21 pREP4 strain by heat-shock transformation. The protein was induced according to the manufacturer's instructions (QIAExpressionist, Qiagen) by IPTG and purified on a His-tag column.

Balb/c mice were immunized with 50 μg of purified protein and complete Freund's adjuvant according to the standard protocol for generation of monoclonal antibodies [99, 146] with

minor modifications. Second immunization with an equal amount of protein and incomplete Freund's adjuvant was performed 15 days after the first. Antibody titer in blood was followed by ELISA and when it reached adequate levels, animals were boosted with 50 μg of purified protein in PBS (1/3 of total volume intraperitoneally, 2/3 of total volume subcutaneously).

Mice were sacrificed by CO_2 asphyxiation 3 days after the boost and their spleen was aseptically removed. Single-cell suspension of splenocytes was obtained by passing the spleen through wire mesh. Erythrocytes were lysed by incubating the cell suspension for 5 minutes in erythrocyte lysing buffer, followed by washing the cells with RPMI. Fusion was performed by mixing equal amounts of SP2/0 and splenocytes, pelleting the cells at 800×*g* for 5 minutes, removing the supernatant and then adding 1 mL of warm PEG slowly drop by drop. The cells were gently mixed after the addition of each drop and then for another minute after the last drop. Finally, 9 mL of RPMI without HEPES were added slowly (in the course of 3-5 minutes) until macroscopic cell clumps appeared. Cells were then again pelleted at 100×*g* for 5 minutes and vigorously resuspended in prewarmed HAT-supplemented 20% RPMI to achieve cell density of 10^5 cells/mL. The cells were transferred to 96-well plate and incubated at 37 °C in 5% CO_2 for 8 days. The old medium was replaced with 200 μL of fresh HAT-supplemented 20% RPMI. After 3 additional days, supernatants of successfully fused cells (that survived in HAT medium) were tested using ELISA on purified m169 protein.

Supernatants from motherwells positive in ELISA were tested by immunoblot on purified MAT protein, and positive wells were rescreened by immunoblot using lysates from MEFs infected with WT, Δ7S3-GFP, Δm168-Δm169, Δm169-Δm170, Δm168, Δm169 and Δm170 mutants as described in sections 3.2.12 and 3.2.13.

3.2.12 SDS-PAGE gel electrophoresis

For SDS-PAGE analysis of proteins, primary MEF or transformed MEFs grown as adherent cells in monolayer were used. After the removal of the medium, the cells were lysed by direct addition of RIPA or NP40 cell lysis buffer and by scraping the cells. Cell lysate was sonicated to break genomic DNA and prevent clumping, and the remaining debris was pelleted by centrifugation in table-top centrifuge at max speed and 4 °C for 30 minutes. Protein concentration was determined using BCA protein assay kit according to the manufacturer's instructions. A mass of 30-100 μg of total protein was mixed with loading buffer, denatured at 95 °C/5 minutes and separated on 8–12% SDS-PAGE gels (depending on target protein size).

3.2.13 Western blot

The proteins were transferred to PVDF membranes electrophoretically. After the transfer, the membranes were quickly rinsed in distilled water 3 times and then blocked in 3% BSA blocking solution for half an hour at room temperature with constant shaking. Primary antibody was added in 3% BSA blocking solution and incubated overnight at 4 °C with constant shaking. The next day, primary antibody solution was removed (and stored at 4 °C for further use) and the membrane was washed 3 times for 5 minutes in TBST. Before the addition of secondary antibody, the membrane was again blocked by incubating in 5% non fat milk in TBST for 15 minutes with constant agitation at room temperature. Secondary antibody was also diluted in 5% non-fat milk in TBST and incubated with constant agitation at room temperature for 1 hour. After secondary antibody solution, the membrane was washed 3× for 5 minutes in TBST and developed with Enhanced chemiluminiscence detection system kit (GE Healthcare) according to the manufacturer's instructions and detected on Uvitec Alliance 4.7 (Uvitec Cambridge).

3.2.14 Ly49 reporter cell assay

Primary MEF cells (stimulator cells) were infected (MOI 1) with WT MCMV, field isolates of various MCMV mutants. After 12 hours PI, reporter cells were then added in RPMI with 10% FCS at 3:1 effector-to-target ratio in 24-well plates in duplicate samples and incubated for 24 hours. Engagement of Ly49 receptor was measured by flow cytometry (BD FACSAriaI or FACSVerse).

4. RESULTS

The main goal of this thesis was to provide an in-depth analysis of the transcriptome of MCMor this purpose, we employed two experimental approaches: classical cDNA cloning and sequencing of viral transcripts, and next-generation sequencing of cDNA generated from total cellular RNA (RNASeq). Using such dual analysis we were able to comprehensively analyze the transcriptome of MCMV and its host.

4.2 THE TRANSCRIPTOME OF MURINE CYTOMEGALOVIRUS

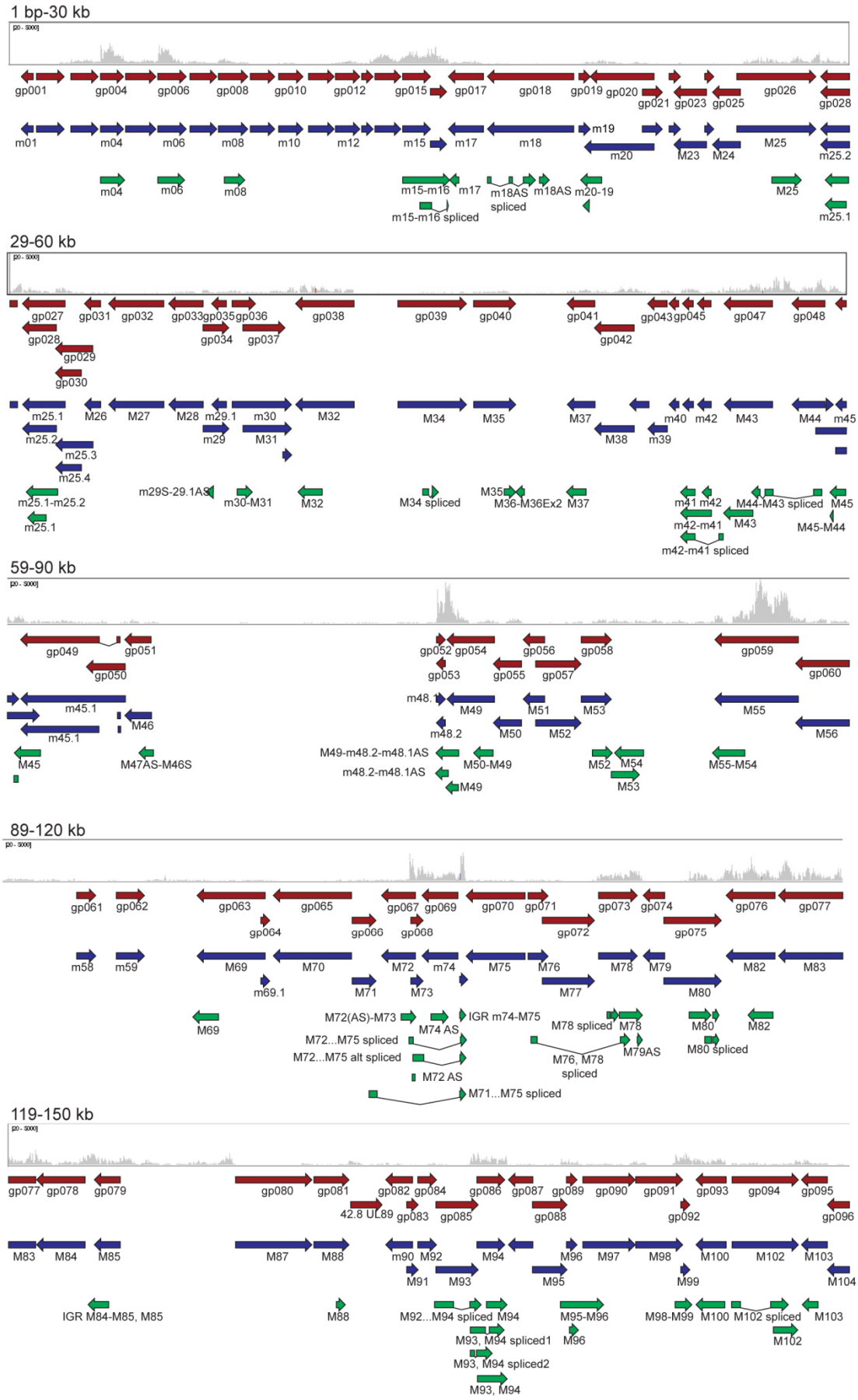
cDNA libraries representing the major temporal classes of viral gene expression [immediate early (IE), early (E) and late (L)] and were generated by collecting RNA from infected MEFs at 9 time points after infection (3 time points per temporal sub-library). A total of 448 cDNA clones were included in the final analyses (84 from the IE library, 163 from the E library, and 201 from the L library). The summary of all isolated clones as well as their temporal distribution compared to previously published temporal analyses is shown in Supplemental table 1.

For RNASeq analysis, RNA was collected at the same time points as in the cDNA library, pooled, converted to RNASeq library and sequenced on Illumina Genome Analyzer IIx. Of the 33,995,400 reads that passed the filter from infected cells, 11% aligned to MCMV genome indicating a 585-fold coverage of the viral genome. In order to allow comparison between the subsequent analyses of transcriptomes of various deletion mutants, BAC-derived MCMV strain MW97.01 was used instead of Smith MCMV. Genome structure of MW97.01 (defined as EcoRI digestion patern) and *in vivo* growth and virulence have previously been shown to be identical to Smith MCMV [136].

Transcriptomic data generated using these two experimental approaches were compared to currently available genome annotations (the NCBI reference sequence, GenBank Accession no. NC_004065.1, and a more recent sequence analysis of the Smith strain based on Rawlinson's annotation, GenBank Accession No. GU305914.1), as is shown in Supplemental table 1 at the end of the thesis. Supplemental table 2 shows the comparison between two annotations, quantification of RNASeq data in relation to the two annotations and comparison to the quantification obtained from the cDNA library.

4.2.1 Comparison of MCMV transcriptome to currently used annotations

A detailed comparison showing exact genomic locations of individual genes in annotations and transcripts detected in the cDNA library, as well as their temporal expression, is shown in Supplemental table 1, while a schematic overview comparing cDNA library and RNASeq analysis to two currently used MCMV annotations is shown in Figure 10.


1 bp-30 kb
gp001
gp004
gp006
gp008
gp010
gp012
gp015
gp017
gp018
gp019
gp020
gp021
gp023
gp025
gp026
gp028
m01
m04
m06
m08
m10
m12
m15
m17
m18
m19
m20
M23
M24
M25
m25.2
m04
m06
m08
m15-m16
m17
m18AS
m18AS spliced
m20-19
m15-m16 spliced
M25
m25.1
29-60 kb
gp027
gp031
gp032
gp033
gp035
gp036
gp038
gp039
gp040
gp041
gp043
gp045
gp047
gp048
gp028
gp034
gp037
gp042
gp029
gp030
m25.1
M26
M27
M28
m29.1
m30
M32
M34
M35
M37
m40
m42
M43
M44
m45
m25.2
m29
M31
M38
m39
m25.3
m25.4
m25.1-m25.2
m29S-29.1AS
m30-M31
M32
M34 spliced
M35
M36-M36Ex2
M37
m41 m42
M44-M43 spliced
M45
m42-m41
M43
M45-M44
m25.1
m42-m41 spliced
59-90 kb
gp049
gp051
gp052
gp054
gp056
gp058
gp059
gp050
gp053
gp055
gp057
gp060
m45.1
M46
m48.1
M49
M51
M53
M55
m45.1
m48.2
M50
M52
M56
M45
M47AS-M46S
M49-m48.2-m48.1AS
M50-M49
M52
M54
M55-M54
m48.2-m48.1AS
M53
M49
89-120 kb
gp061
gp062
gp063
gp065
gp067
gp069
gp070
gp071
gp073
gp074
gp076
gp077
gp064
gp066
gp068
gp072
gp075
m58
m59
M69
M70
M72
m74
M75
M76
M78
M79
M82
M83
m69.1
M71
M73
M77
M80
M69
M72(AS)-M73
M74 AS
IGR m74-M75
M78 spliced
M78
M80
M82
M72...M75 spliced
M79AS
M80 spliced
M72...M75 alt spliced
M76, M78 spliced
M72 AS
M71...M75 spliced
119-150 kb
gp077
gp078
gp079
gp080
gp081
gp082
gp084
gp086
gp087
gp089
gp090
gp091
gp093
gp094
gp095
42.8 UL89
gp083
gp085
gp088
gp092
gp096
M83
M84
M85
M87
M88
m90
M92
M94
M96
M97
M98
M100
M102
M103
M91
M93
M95
M99
M104
IGR M84-M85, M85
M88
M92...M94 spliced
M94
M95-M96
M98-M99
M100
M102 spliced
M103
M93, M94 spliced1
M96
M102
M93, M94 spliced2
M93, M94

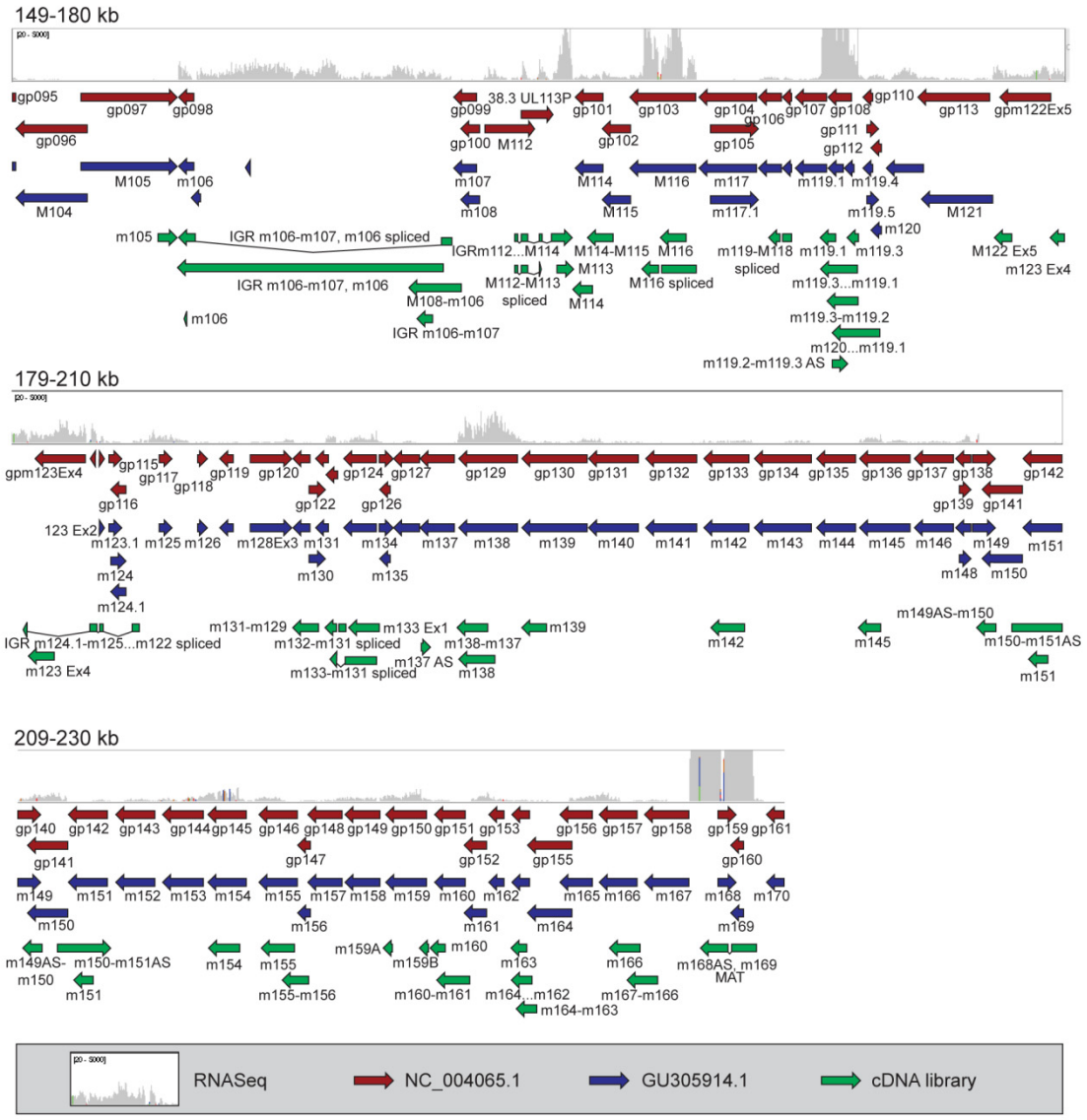


Figure 10. Comparison of cDNA cloning and RNASeq data in relation to current genome annotations. Comparison of cDNA library (green arrows) and RNASeq analysis of murine cytomegalovirus (gray histograms). The longest clone from each group of clones in the cDNA library is shown. ELAND alignments of RNASeq reads were loaded in Integrative Genomics Viewer and compared to NC_004065.1 (red arrows) and GU305914.1 (blue arrows). The data range for RNASeq data was set to 20-5000 (minimum of 20 and maximum of 5000 reads at each nucleotide is shown). Data is shown in 30-kb ranges with 1-kb overlap.

Figure 10 shows that the two current annotations (red and blue arrows) mostly agree. In contrast, MCMV transcripts identified through our classical cDNA cloning and sequencing (green arrows) diverged dramatically from both annotations. Although current bioinformatic tools cannot reconstruct dense genomes from transcriptomic data, the alignment of RNASeq

data can provide some insight into structure of certain well-expressed transcripts. As can be seen in Figure 10 our RNASeq data underscore significant differences between the currently available MCMV annotations and the real transcripts that accumulate in the infected cells.

Novel and divergent transcripts observed in this study can be divided into 4 groups:

1) transcripts overlapping more than 1 currently annotated gene;

2) transcripts shorter than currently annotated genes;

3) antisense transcripts;

4) novel spliced transcripts.

In addition to novel and divergent transcripts, we have also observed a lack of transcripts in the currently annotated regions. Several regions annotated as coding were not represented in cDNA library. For example, no cDNA clones overlapping m01 or m170 were found in the cDNA cloning study and these two regions showed the lowest RPKM value in the RNASeq dataset of 132 and 141, respectively (Supplemental table 2). For comparison, well defined MCMV genes *m04* and *m138*, both represented with multiple clones in our cDNA library, have RPKM values of 14,137 and 16,935 respectively. However, cDNA library does not contain transcripts from well defined genes *m152* and *m157* (RPKM values of 5,328.9 and 3,075.04 respectively). This indicates that cDNA library does not have a complete coverage of all viral transcripts. Nevertheless, regions with low RPKM values and lack of cDNA clones should receive further attention to prove or disprove the existence of a gene predicted by *in silico* ORF analyses.

4.2.2 Transcripts overlapping more than 1 currently annotated gene

Several cDNA clones were isolated that overlapped more than one currently annotated gene. For example, four cDNA clones in our library overlap both the *m15* and *m16* genes. The longest of these clones is a 1673-bp long transcript, whereas current annotation predicts two genes of 908 bp (*m15*) and 632 bp (*m16*). Other examples include m19-m20, m25.1-m25.2, m42-m41, M50-M49, M55-M54, multiple spliced transcripts in M71-M75 and M76-M78 regions, M84-M85, M93-M94, M98-M99, multiple transcripts in M112-M114, m119 and m129-m131 regions, m155-m156, m160-m161, m167-m166 and m168-m169 (MAT). Although RNASeq protocol used in this work cannot be used for transcript reconstruction, in

many instances the shape of read alignments speaks in favor of single transcripts longer than currently annotated genes. M93-M94 is an illustrative example where in RNASeq alignment only one "transcript" histogram shape is visible. These data suggest that polycistronic transcripts are more widespread in MCMV than previously thought. This is in accordance with the recent finding of numerous ORFs translated from polycistronic transcripts in HCMV [127].

4.2.3 Transcripts shorter than currently annotated genes

In cDNA library analysis many smaller transcripts overlapping currently annotated genes have been isolated, for example *M35, M37, M69, M82,* etc. In the absence of confirmatory studies to define the precise ends of these clones, it is likely that many could represent premature truncations during the reverse transcription step of the cloning process or degradation of RNA prior to RT, but there is also the possibility that a subset of genes may require refinements to reflect smaller gene products.

4.2.4 Antisense transcripts

One of the main incentives for initiation of this study was a finding published by Zhang *et al.* [147] showing extensive antisense transcription in human cytomegalovirus transcriptome. In our cDNA library excluding 20 cDNA clones that mapped to intergenic regions, 275 (64%) of cDNA clones were in the sense (S) orientation, 39 (9%) were antisense (AS), and 114 (27%) overlapped more than one gene in both S and AS orientations relative to the original annotation provided by Rawlinson *et al.* [106]. These designations were re-evaluated according to the publication of new NCBI reference sequence (NC_004065) in which some putative genes were removed from the annotation. Some AS transcripts to these ORFs were revised as S transcripts due to the lack of evidence for the predicted sense transcript. For AS transcripts that overlapped two or more hypothetical proteins in both AS and S orientation, the AS designation was preserved. According to these criteria, 431 (97%) transcripts were in S orientation, and only 4 (0.09%) were in AS orientation, and 9 (2%) overlapped more than one gene in both S and AS orientations.

4.2.5 Novel spliced transcripts

In this study 22 novel spliced transcripts were cloned along with the spliced transcripts reported by others. A complete list of spliced transcripts detected in this research is provided in Table 5.

Table 5. Summary of spliced transcripts in MCMV transcriptome.

Overlapping gene(s) or known designation	Strand	Clone name	Total no. of clones found in this study	Genomic position (only positions of longest clone shown)	First report
m15, m16	+	E253	1	14635-15083 + 15622-15700	this study
m18	+	L123	1	17079-17188 + 17853-17957 + 18351-19285	this study
M33				41486-41519 + 41679-42780	[35]
M34	+	E196	1	44012-44242 + 44304-44516	this study
M36	-			49267-49036 + 48909-47621	[106]
m42, m41	-	IE106	1	55312-55123 + 54218-53678	this study
M44, M43	-	IE160		58976-58668 + 57157-56856 + 56667-56361	this study
m60				94984-95063 + 105879-106093	[114]
M71, IGR m74-M75	+	E180	1	102514-102829 +105879-106090	this study
M73	+	L33	1	103985-104549+ 105880-106093	[114] and this study
M73 longer	+			103700-104548+105879-106093	[114] and this study
M73.5	+	L443	4	103985-104160+ 105879-106093	[114] and this study
M73.5 longer	+			103700-104160+ 105879-106093	[114] and this study
M76(AS), M78(S)	+	E139	1	108476-108714 + 111789-----112593*	this study
M78	+	E41	1	111280-111409 + 111444-----112592*	this study

Overlapping gene(s) or known designation	Strand	Clone name	Total no. of clones found in this study	Genomic position (only positions of longest clone shown)	First report
M80	+	L116	1	114889-115148 + 115187-115396	this study
M92 , M93, M94	+	L14	1	134691-135369 + 135956-137333*	this study
M93, M94	+	L107	1	135978-136052 + 136181-136754	this study
M93, M94	+	L172	1	135978-136524 + 136651-137227	this study
M89	-			138283-137393 + 132771-131649	[106]
M102	+	IE224	1	145586-145908 + 147011-147682	this study
Stable 7.2kb intron				162090-161622 + 154365-153916	this study and [66]
Stable 8.0kb intron				162606-162415 + 154365-153916	[66]
Stable 7.2kb intron **IGR m106-m107, m106(S)**	-	E289 E206 L63	3	161905-161622 + 154368-153873, 161919-161622 + 154368-153886, 161904-161622 + 154368----153867*	this study and [66]
M112	+			163097-163889 + 163983-164159 + 164486-164505	[26, 106]
M112 Ex1, M113, M112 Ex2	+	E184	1	163778-163891 + 163983-164157	this study
M112 Ex1, M113, M112 Ex2, M112 Ex3 (last exon in IGR M112 Ex3-M114)	+	L2	1	163779-163891 + 163983-164160 + 164485-164582 + 164871-165510*	this study
M116	-	L29	8	168189-168091 + 168015-167555	this study

Overlapping gene(s) or known designation	Strand	Clone name	Total no. of clones found in this study	Genomic position (only positions of longest clone shown)	First report
m119, M118	-	E243		171957-171684 + 171585-171255	this study and Rawlinson et al.[1]
M102		IE224		145586-145908 + 147011-147682	this study
m123 Ex2,3,4				181766-181660 + 181562-181372 + 181249-179763	[106]
M122 Ex 5				179517-177983	[106]
M128 Ex3				186085-187296	[106]
m133 Ex1, m132 Ex2	-			189795-188881 + 188601-188382 +	[106] [68]
m133 Ex1, m132 Ex2, m131	-	IE138	1	189808-189499 + 188602-188269	this study
IGR m124.1 - m125, m123Ex2, m123 Ex3, m122 Ex5	-	E279	1	182798-182596 + 181770-181659 + 181562-181371 + 179520-179420	this study
m133 Ex1, m132 Ex2, m131	-	L78 IE208	2	189791-188880 + 188602-188407*	this study
m132 Ex2 - m131	-	E96 L102	2	188885-188695 + 188603-188292*	this study
m165, m164, m163	-	IE197	1	223828-223662 + 223593-----221832*	this study
m169(S) m168 (AS)	-	E125	139	229112-228325 +228247-227426	this study

Interestingly, many new spliced transcripts detected in this study are not only highly abundant but also have no known function. By far the most abundant transcript is spliced m169(S) m168 (AS) transcript (discussed in detail in chapter 4.2.10.5). Panels A in Figure 11 and Figure 12 shows that 31% of the viral cDNA clones and 41% of all viral reads from the RNASeq analysis mapped to this novel spliced transcript at the right end of the genome.

M116 is another highly abundant novel spliced transcript represented in the cDNA library with 8 clones and is discussed in detail in chapter 4.2.10.4.

4.2.6 Temporal analysis of cDNA clones

Generation of separate sub-libraries for 3 temporal classes of viral genes allowed us to monitor temporal expression of transcripts (Supplemental table 1).

Temporal assignment of cDNA clones in this study agrees with previous studies [67] with few discrepancies. Most discrepancies are likely a consequence of more time points used in the construction of libraries of this study. For instance, *m20(S), m19(AS), M49, M72(AS), M73, M78; M97; M113-M114; m133 Ex1, m132 Ex2* and *m131* were all detected 24 hours PI in Lacaze's study. In our study, the earliest detection was in the IE library. However, our IE library consists of 4, 8 and 12 hours PI time points, whereas Lacaze's study has only 2 time points for IE temporal class at 0.5 and 6.5 hours PI. Between 6.5 and 24 hours PI no time points were checked. Likewise, *m119.3* was detected at 48 hours PI at the earliest in Lacaze's study, whereas it can be found in early cDNA library in our study.

Interestingly, one of the most abundant clones in our cDNA library, M116 was not detected in Lacaze's study. In our study M116 is represented with 15 clones found in E and L libraries. As will be discussed later on, M116 is a novel spliced transcript. Lack of the detection by Lacaze's study may be a consequence of unfortunate probe design where a part of the probe aligns to the intronic sequence.

Also consistent with the Lacaze's study, the cDNA library analysis suggests very low levels of ie1 or ie3 expression. We detected only 2 clones of ie1 (m123Ex2) in the L library, no 128ex3 (ie2) clones, and just one 122ex5 (ie3) in the L library. By RNASeq analysis, an RPKM value of 5070 was found for gp114 (ie1/ie3 exon 2) and 1626 for gp120 (m128Ex3; ie2) (Supplemental table 2). For perspective, the average RPKM for all genes using annotation derived from the NCBI reference sequence (NC_004065) is 8015 (standard deviation of 38.536) [If most abundant transcript is excluded from this calculation, the average RPKM for all genes using the annotation derived from the NCBI reference sequence (NC_004065) is 4275 (standard deviation of 8.666)]. Therefore, relative to other viral genes, both the cDNA library analysis and the RNASeq analysis show average to below average

levels of transcription for ie1 and ie3 gene expression, consistent with the studies by Lacaze *et al.*

Lacaze *et al* have not detected *M44, M70, M75, m135, m143, m144, m153* and *m157*. We also have not detected either of these genes in the cDNA library (although we did detect one clone overlapping m45 and m44).

4.2.7 Analysis of viral gene expression

RNASeq data was used to assess expression levels of MCMV genes. Transcript levels were quantified as reads per kilobase of exon model per million mapped reads (RPKM; data shown in Supplemental table 2); a value that reflects molar concentration of a transcript in the starting sample by normalizing for gene length and total read number [96].

Figure 11 shows the alignment of RNASeq reads against MCMV genome visualized in Integrated Genomics Viewer (IGV). From this figure it is evident that (a) nearly whole genome is transcribed to some extent (best seen in panel C), and (b) viral transcripts vary greatly in their expression levels. This is most dramatically visible in Figure 11 panel A, where the expression of MAT transcript dwarfs the expression levels of all other genes.

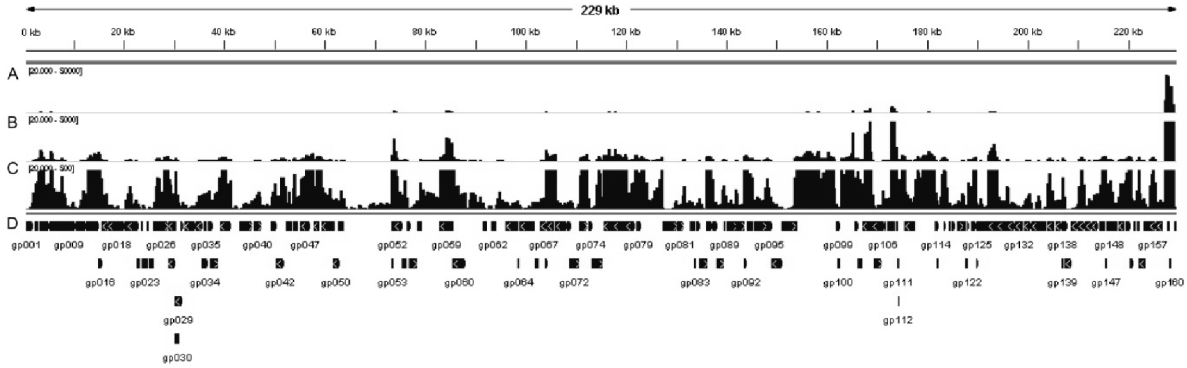


Figure 11. Transcriptional activity of MCMV. Whole genome visualization using IGV of RNASeq reads aligned to the MCMV genome (annotation and sequence from GenBank acc. no. NC_004065) showing different data ranges. (A) range of 20-50 000 reads, (B) range of 20–5000 reads, (C) range of 20–500 reads, (D) annotation. Lower data ranges (B and C) show that nearly whole MCMV genome is transcribed to some level.

Ideally, RNASeq data should be quantified using a definitive and verified annotation or annotation reconstructed from the sequencing data. Unfortunately, no software has yet been developed that can reconstruct transcripts from dense genomes that are the hallmark of

viruses. Therefore, both currently available annotations were used when calculating RPKM (shown in Supplemental table 2 and Figure 12) from RNASeq data bearing in mind their limitations.

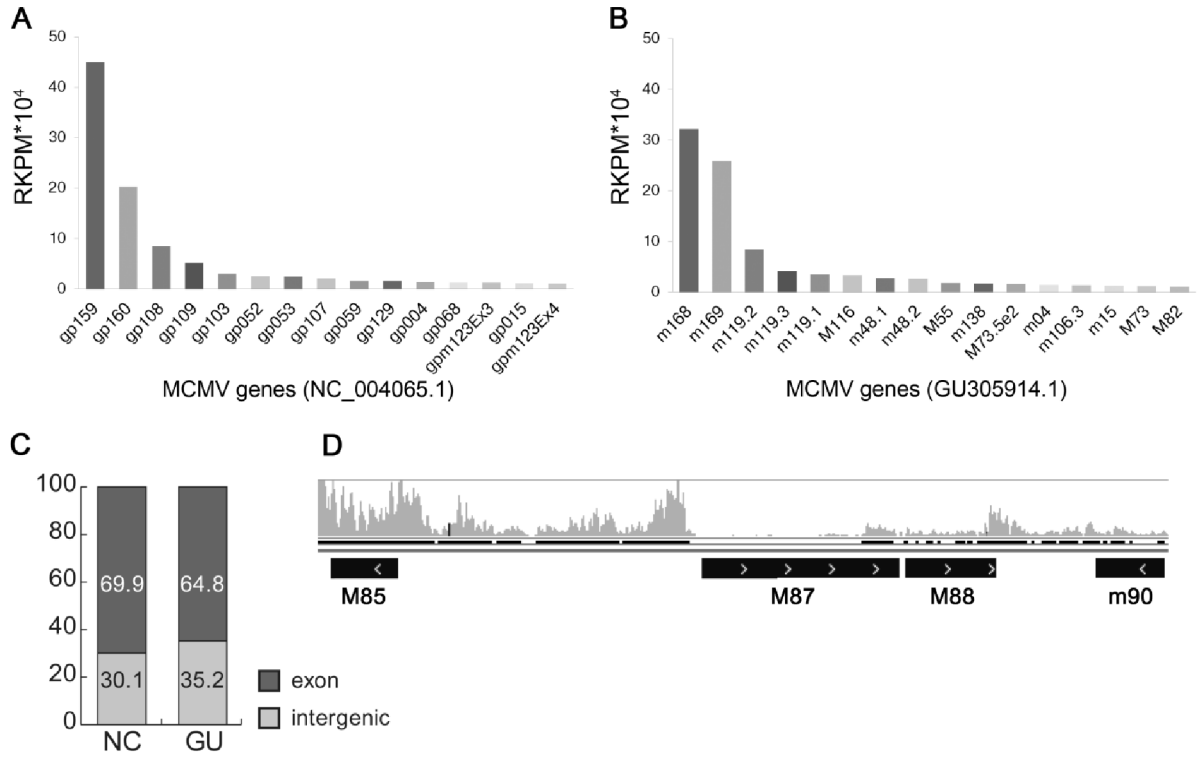


Figure 12. Quantization of transcript abundance varies with annotation. (A) The most expressed MCMV genes (RPKM >10 000) relative to NCBI NC_004065 and (B) GU305914.1. (C) Percentage of reads mapping to coding (exon) or intergenic regions using NC_004065.1 (NC) or GU305914.1 (GU). (D) Transcriptionally active but currently un-annotated region between M85 and M87.

As can be seen in Figure 12 30-35% of the reads mapped to intergenic regions, depending on the annotation. Since genome size of herpesviruses is limited, it is highly unlikely that precious genome space would be wasted on non-coding DNA and this finding is therefore likely a consequence of current annotations missing some genes and wrongly annotating others (Figure 12, panel D. and multiple transcripts detected in cDNA and listed in Supplemental table 1). In addition, currently used genomic maps are still showing only protein-coding genes. Illustrative examples are 8 and 7.2 kb long introns, also found in our cDNA library and detected in RNASeq (Figure 10) [66]. Confirming this, 35.2% of reads aligning to intergenic region using Rawlinson's modified annotation (GU305914.1) are reduced to 14% when the annotation is modified to correct MAT transcript structure.

Therefore, these findings highlight numerous incongruences with the current annotation for the MCMV genome.

A very interesting finding of this analysis is that most abundantly expressed genes have unknown functions. As can be seen in Figure 12, most highly expressed genes (Rawlinson's annotation) are *m168, m169, m119, M116* and *m48*, all genes with unknown functions and unconfirmed annotations. As will be discussed in detail in chapter 4.2.10.5, *m168* and *m169* are in fact part of a larger spliced transcript and are not separately transcribed of translated. Also highly expressed are the immune evasion genes *m04* and *m138,* glycoprotein *M55* (glycoprotein B), and additional genes of unknown functions (*M73* and *m15*).

4.2.8 Sensitivity of transcriptomic analysis

Estimating the coverage and sensitivity of transcriptomic analysis depends on the quality of annotation. Currently available annotations of MCMV genomes are not definitive and many ORFs have not been experimentally verified. Therefore, for cDNA library, the only coverage and sensitivity assessment available is in relation to the previously performed analyses.

A microarray study conducted by Tang *et al.* [133] identified novel ORFs and confirmed a few previously predicted ones [16]. Of these, in cDNA library we have detected m166.5(1 clone), m132.1 (5 clones) and m84.2 (2 clones).

Studies by both Lacaze *et al.* and Tang *et al.* [67, 133] failed to detect transcripts from numerous annotated genes. Lacaze *et al.* did not detect *M44, M70, M75, m135, m143, m144, m153* and *m157*. We also failed to isolate transcripts from these ORFs in our cDNA library with the exception of M44. Genes whose expression was not detected in the microarray analysis conducted by Tang and Maul [133] include *m01, m19, m26, m22, m69.1, m70, m117.1, m119.5, m126, m127, m129, m134, m144, m150, m165, m170.* Of these, we did detect one large clone overlapping *m129-131* and one clone overlapping *m150* (m150, m151(AS)) in the cDNA library.

Tang and Maul [133] reported the following ORFs as negative by both PCR and microarray analysis: *m21, m44.1, m58, m107, m124.1, m125, m130, m141.1, m148, m149, m151, m157* and m165.1. The cDNA library in this study did include *m107* (4 clones) and *m151*, however, the clone overlapping *m151* was in the antisense orientation to the predicted ORF.

In contrast to the cDNA library, most genes not detected by microarray analyses or not represented in this cDNA library were nevertheless detected by RNASeq analysis (Supplemental table 1 and Figure 10). The possible exceptions include *m01, m150* and *m170*, all of which have RPKM values below 200. Therefore, it can be concluded that RNASeq provides a more sensitive level of detection for analyzing viral gene expression. It is, however, important to note that TruSeq protocol for RNASeq used in this work is not strand specific (therefore cannot distinguish the original coding strand). As a consequence of this limitation it is not possible to exclude these low expression values as background noise due to DNA contamination.

4.2.9 Validation of RNASeq data

Although both cDNA library and RNASeq analysis gave concordant results, in order to further confirm these unexpected findings, comparison was made between these RNASeq data and recently published RNASeq analysis of the MCMV transcriptome using BAC-derived WT virus on NIH-3T3 fibroblasts [83]. This analysis used strand-sensitive RNASeq protocol, different sequencing platform (ABI SOLiD) and commercially available NIH3T3 cell line. Reads from this experiment were aligned against MCMV genome using Bowtie aligner and visualized in IGV alongside our RNASeq alignments. The comparison shown in Figure 13 clearly shows that the profiles obtained from these two different RNASeq experiments are remarkably similar despite differences in sequencing platforms and library generation approaches. Also, either seven or eight of the ten most abundant genes were identical in both datasets (Supplemental table 3). Minor differences in abundance of some transcripts can be attributed to the differences in the time points analyzed in these two studies as well as the fact that our analysis achieved an order of magnitude greater sequencing depth (compare the reads sequenced for each histogram set in Figure 13).

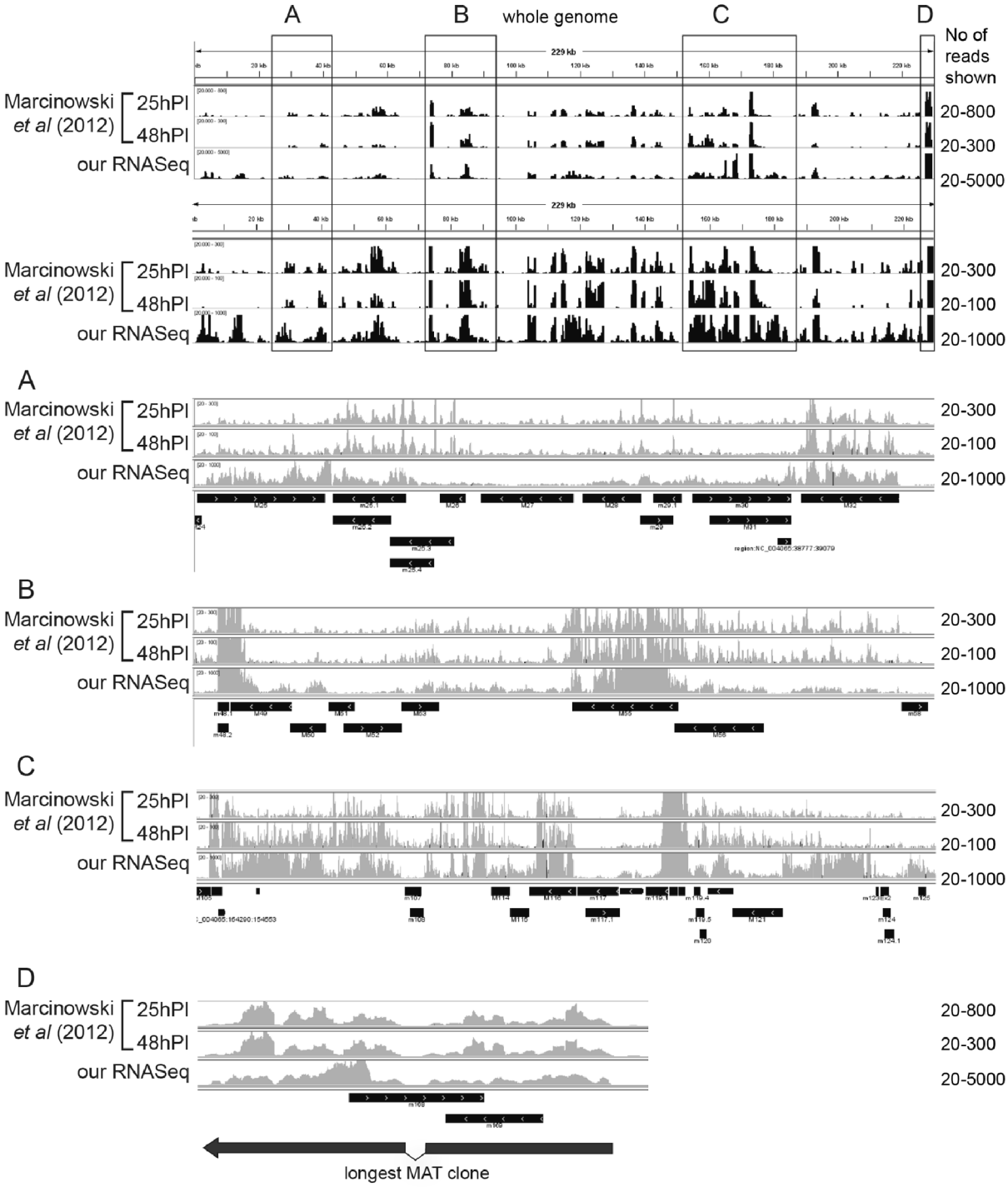


Figure 13. RNASeq profile comparison. RNASeq data from total RNA obtained from MCMV-infected NIH-3T3 fibroblasts from 25 and 48 h PI sequenced by Dölken group [83] (GSE35833) was aligned against MCMV genome (gB acc no NC_004065.1) using Bowtie aligner and visualized in IGV in comparison with our RNASeq data. The view of the complete genome is shown at the top with 4 areas magnified below (labeled A-D) and the number of displayed reads are noted on the side. Since viral genes display a wide range of expression levels, the whole genome view is shown in a wide data range (upper panel), more suitable for displaying highly transcribed regions, and a narrowed data range (lower panel) is more suitable for less transcribed regions. As can be seen, the profiles of the compared alignments are remarkably similar, the only differences being the abundance of certain

transcripts which are due to different time points analyzed in comparison with the pooled data of our RNASeq and significantly greater depth of at least one order of magnitude of our data in comparison with Marcinowski data.

Together these findings demonstrate that RNASeq analysis is a highly sensitive method for detection of viral gene expression during infection.

4.2.10 Validation of novel transcripts by Northern blot

Because cDNA cloning and RNASeq identified significant differences between the MCMV transcriptome and current annotations, in-depth analysis of several genomic regions by Northern analysis was performed to confirm these findings. cDNA clones were used to generate strand-specific riboprobes (described in chaper 3.2.10, oligonucleotides used to generate S riboprobes are listed in Table 4, chapter 3.1.9).

The following regions were analyzed: *m15-m16* and *m19-m20* as examples of regions where transcripts overlapping more than one currently annotated genes were found, *m71-m74* region as an example of transcriptionally complex region with new spliced transcript detected, and *M116* and *m168-m169* region where putative new spliced transcripts which differed significantly from current annotations were detected.

It is important to note that some discrepancies in sizes between bands in Northern blot in comparison with the expected and previously published data are due to “smiling” effect during separation of RNA by agarose gel electrophoresis.

4.2.10.1 Analysis of m15-m16 region

In the cDNA cloning study, 5 transcripts overlapping the predicted *m15* and *m16* ORFs were cloned, and one of these transcripts was spliced. The RNASeq profile also strongly indicated transcription that spans both predicted genes (Figure 14). In line with our cDNA library where only S transcripts were cloned, no antisense transcripts were detected in Northern analysis. The sense probe detected 5 bands: the strongest band at approximately 4.7 kb started to accumulate already at 10 hours PI, while other 4 bands became visible only at late time post infection. The clones isolated in the course of cDNA cloning study indicate that the transcripts in this region end at nucleotide position 15700 (nucleotide positions are relative to Smith reference sequence gB acc. no. NC_004065.1): the 3’ end of all cDNA clones ended at or

close to nucleotide position 15700, and RNASeq data alignment to MCMV genome shows a sudden drop in reads around this nucleotide position. If this is so, low-abundance band between 5 and 6 kb (1 in Figure 14, A) should initiate in *m11*, the most abundant band slightly below 4.7 kb (2 in Figure 14 A) should initiate in *m12*, band between 4.7 and 1.9 kb (3 in Figure 14A) should initiate in *m13* or *m14*, while the band slightly below 1.9 kb (4 in Figure 14A) corresponds to the longest unspliced cDNA clone detected in this region, E119 (1.67 kb). Multiple transcripts with alternative 5' ends were found when this region was analyzed by RACE in cells infected with wild isolates of MCMV (Alec Redwood, personal communication). The smallest band (5 in Figure 14A) corresponds in size to novel spliced transcript, E253 (566 bp). Since only one spliced transcript was cloned in the cDNA library, PCR analysis was performed using primers that flank the putative intron. As can be seen in Figure 14 B splicing could not be confirmed by PCR (Figure 14B). It is possible that the single spliced clone represents an aberrant transcript, a result of intra-molecular template switching during reverse transcription [29].

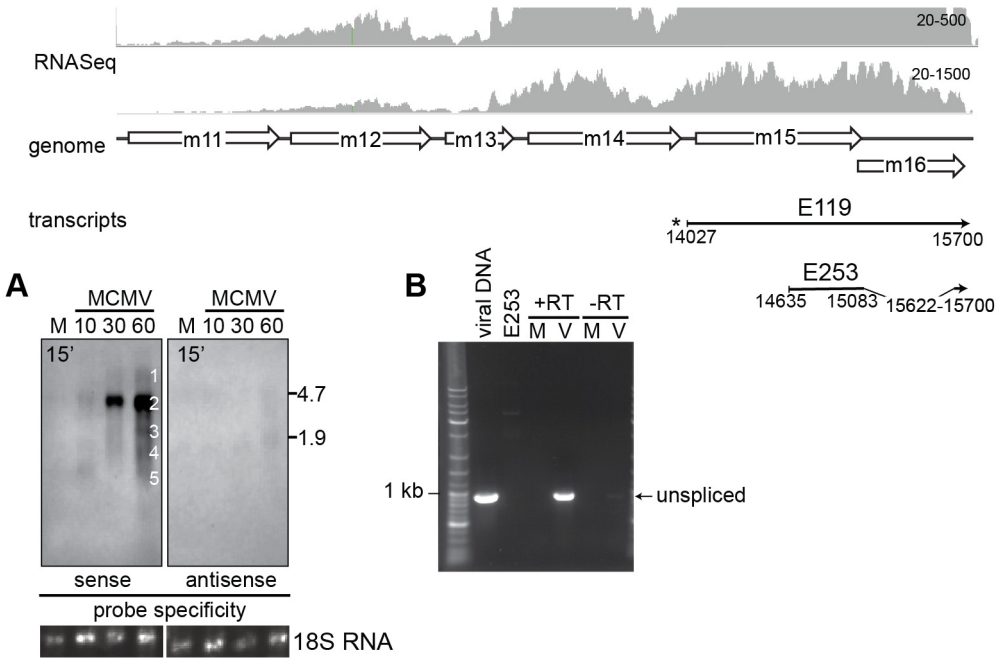


Figure 14. Analysis of transcription in *m15-m16* gene region by Northern blot (A) and PCR (B). Scheme of m15-m16 genomic region: predicted ORFs (Rawlinson's annotation) are depicted as empty arrows, thin black arrows show the longest transcripts cloned in our cDNA library. Clones used to generate probes are marked with asterisk. Arrowheads denote 3' ends of transcripts. The nucleotide coordinates relative to Smith sequence (NC_004065.1) of the isolated transcripts are given below the thin arrows, while the names of the clones are written above. Gray histograms show RNASeq reads aligned to MCMV genome, Smith sequence (NC_004065.1). For Northern blot analysis (A), Balb/c MEF cells were infected with BAC derived Smith virus and harvested at indicated times post

infection. Total RNA was separated using denaturing gel electrophoresis, transferred to nylon membrane and incubated with probes specific for S and AS transcripts. RNA integrity and loading were evaluated by inspecting 28S (not shown) and 18S rRNA bands under UV light after transfer to membrane. Maximal possible exposure times (noted on the blots) were used to ensure the detection of even low-abundance transcripts. For PCR analysis of putative splicing (B), the same RNA used in the Northern blot at 60 h PI was treated with DNaseI, reverse transcribed using oligo(dT) primers and then PCR amplified using primers that flank putative intron (listed in Table 3). No reverse transcriptase (-RT) controls were run in parallel. Viral DNA served as unspliced control.

4.2.10.2 Analysis of m19-m20 region

Four clones in total were isolated in the cDNA analysis overlapping this region; 3 overlapped both *m20* in sense orientation and *m19* in antisense, while one overlapped just *m19* in antisense. Similar to *m15-m16* region, using clone IE205 as a probe in Northern analysis 5 transcripts with differential temporal expression patterns were detected. Consistent with our cDNA library, no transcript was detected using an AS probe derived from clone IE205 or L57, which has a greater overlap with putative *m19*, indicating that *m19* is not transcribed (Figure 15A and B). Therefore, *m19* should be removed from the MCMV genome annotation. Also similar to *m15-m16* region, of 4 cDNA clones isolated in the cDNA library, ends of cDNA clones and transcription profile of RNASeq analysis speak in favor of 3' co-terminal transcription with the 3' end located at around nucleotide position 20430 (Supplemental table 1 and Figure 15).

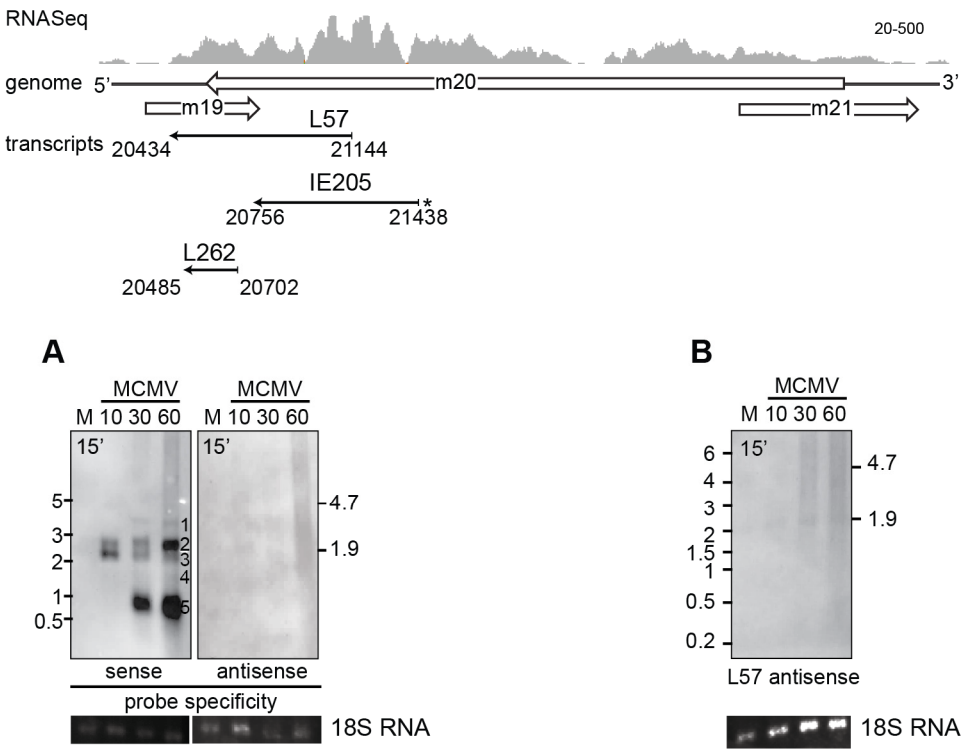


Figure 15. Analysis of transcription in *m19-m20* gene region by Northern blot. Scheme of *m19-m20* genomic region: predicted ORFs (Rawlinson's annotation) are depicted as empty arrows, thin black arrows show the longest transcripts cloned in our cDNA library. Clones used to generate probes are marked with asterisk. Arrowheads denote 3' ends of transcripts. The nucleotide coordinates relative to Smith sequence (NC_004065.1) of isolated transcripts are given below thin arrows, while the names of the cDNA clones are written above. Gray histograms show RNASeq reads aligned to MCMV genome, Smith sequence (NC_004065.1). For Northern blot analysis (A), Balb/c MEF cells were infected with BAC derived Smith virus and harvested at indicated times post infection. Total RNA was separated using denaturing gel electrophoresis, transferred to nylon membrane and incubated with probes specific for S and AS transcripts. RNA integrity and loading were evaluated by inspecting 28S (not shown) and 18S rRNA bands under UV light after transfer to the membrane. Maximal possible exposure times (noted on the blots) were used to ensure the detection of even low abundance transcripts. (A) Northern blot using IE205 as S and AS probe; (B) Northern blot using L57 antisense as a probe to validate that there is no transcription coming from + genomic strand in *m19* gene region.

The largest band at 4 kb (1 in Figure 15B) is detectable at 30 and 60 hours PI and based on its size in Northern analysis it should initiate in *M23*. Consistent with previous studies [133], no transcripts from *m20* to *m25* ORFs were cloned in the cDNA library, which can be explained by low abundance and size of this transcript, as well as by the propensity of cDNA libraries to enrich 3' ends. The band slightly smaller than 3 kb (2 in Figure 15A) shows a peak accumulation at 60 hours PI and is consistent with a transcript overlapping *m19-m21* (approx.

locations 20430-23220, 2.79 kb). The band of approximately 2 kb (3 in Figure 15A) shows peak accumulation 10 hours PI. Based on the RNASeq profile, this band could represent transcripts that initiate at nucleotide position 22060. Finally, late time points are dominated by smaller transcripts of approximately 1 kb (5 in Figure 15A; predicted start site at 21440) which correspond in size to the cDNA clones detected in our study.

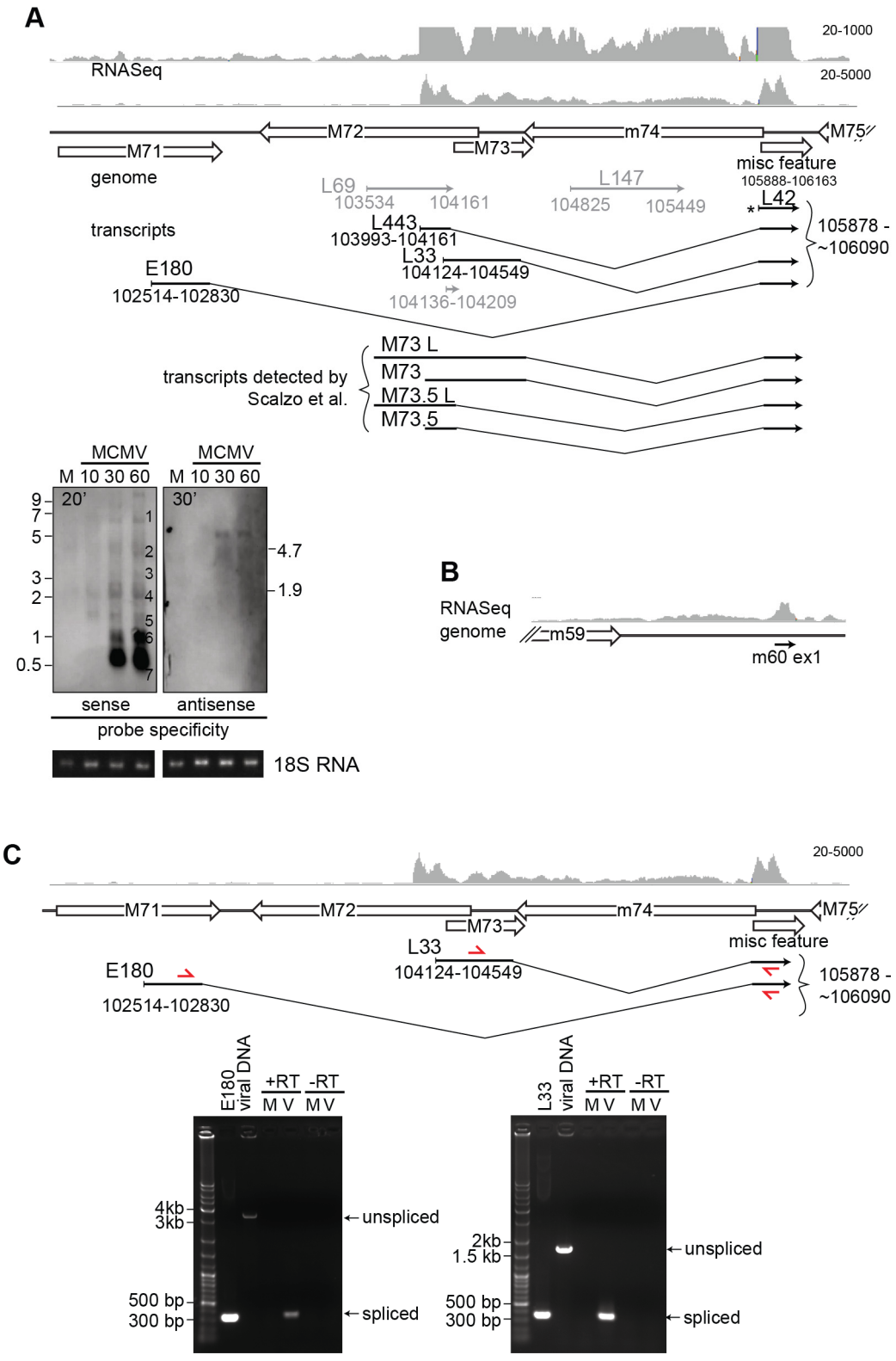
In short, Northern analyses of *m15-m16* and *m19-m20* gene regions detected multiple transcripts, very likely 3'-co-terminal, that exhibit different temporal expression patterns. Similarly to certain transcripts previously reported for MCMV and HCMV [10, 69, 114], smaller transcripts tend to accumulate at later time points.

4.2.10.3 Analysis of m71-m74 region

m71-m74 gene region was selected for further analysis by Northern blot and PCR for two reasons: (1) it had previously been shown to have a very complex transcriptional profile [105, 114] and (2) new, previously not reported splice transcript, clone name E180 (see Table 5 and Figure 16) was cloned in the cDNA study.

Figure 16A shows that cDNA library, RNASeq profile and the results of Northern blot with L42 as a probe all are in agreement with the findings of Scalzo et al. [114] of multiple spliced transcripts that share exon 2.

Bands 5-7 (Figure 16A) correspond in size to m60, m73 and m73.5 spliced transcripts previously reported by Scalzo *et al.* [114]. In cDNA cloning study, 4 isolated clones correspond to M73.5 transcripts (represented by the longest clone, L443) and 1 that corresponds to M73 (L33) (listed in Table 5). Transcripts corresponding to m60 were not isolated in the cloning study; however, the RNASeq profile in the region corresponding to m60 exon1 shows active transcription (Figure 16B). A band of 1.1 kb (5 in Figure 16A) corresponds to longer M73 and M73.5 transcripts, while unspliced versions of M73 and M73.5 are probably bands around 2 kb (4 in Figure 16A).


A
RNASeq
20-1000
20-5000
M71
M72
M73
m74
M75
misc feature
105888-106163
genome
L69
103534
104161
L147
104825
105449
L42
transcripts
L443
103993-104161
L33
104124-104549
105878 -
~106090
E180
102514-102830
104136-104209
transcripts detected by
Scalzo et al.
M73 L
M73
M73.5 L
M73.5
MCMV
M 10 30 60
20'
30'
9
7
5
3
2
1
0.5
1
2
3
4
5
6
7
4.7
1.9
sense
antisense
probe specificity
18S RNA
B
RNASeq
genome
m59
m60 ex1
C
20-5000
M71
M72
M73
m74
M75
misc feature
L33
104124-104549
E180
102514-102830
105878 -
~106090
E180 viral DNA
+RT
-RT
M V M V
4kb
3kb
500 bp
300 bp
unspliced
spliced
L33 viral DNA
+RT
-RT
M V M V
2kb
1.5 kb
500 bp
300 bp
unspliced
spliced

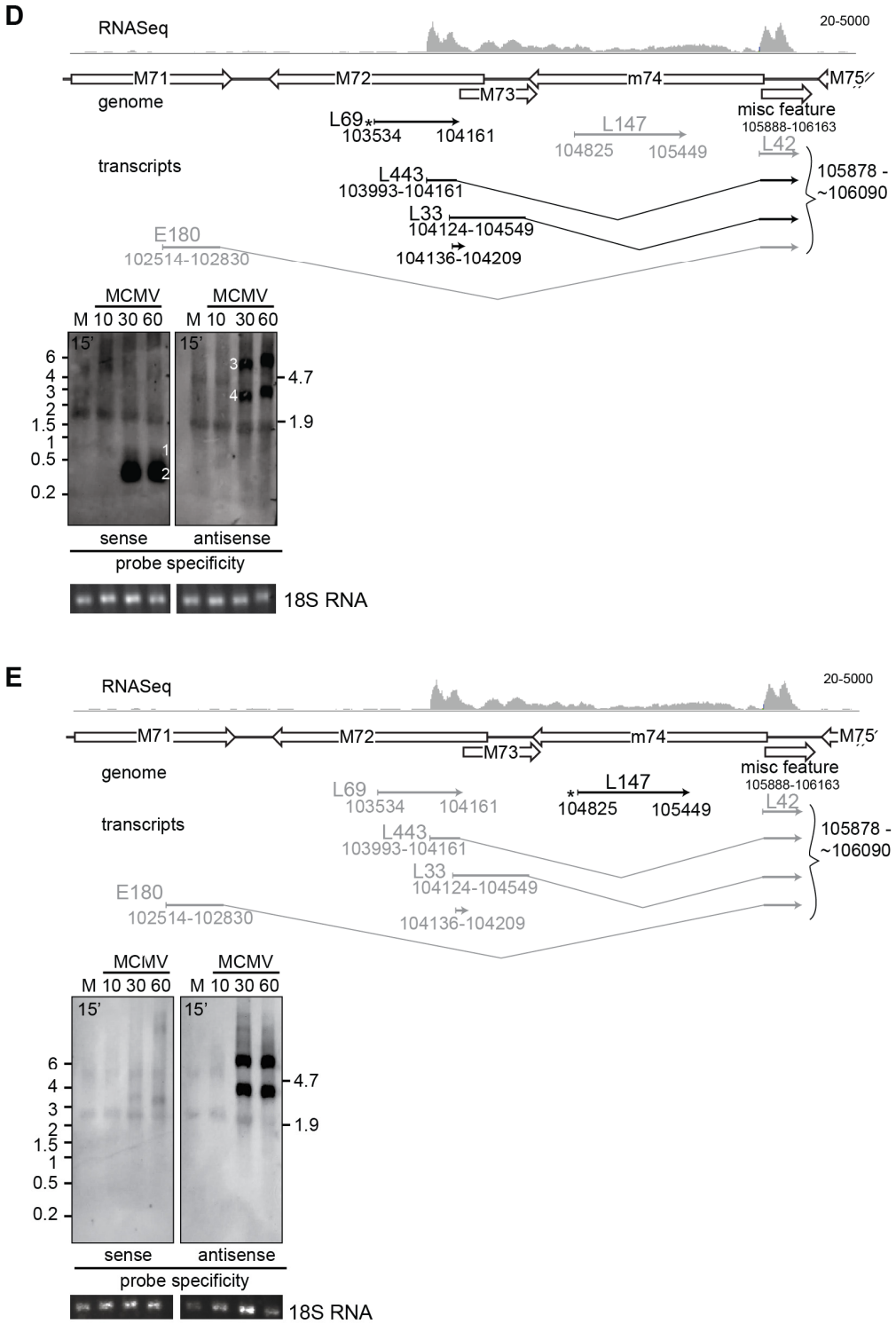


Figure 16. Analysis of transcription in *m71-m74* gene region by Northern blot and PCR. Scheme of *m71-m74* genomic region: predicted ORFs (Rawlinson's annotation) are depicted as empty arrows, thin black arrows show the longest transcripts cloned in our cDNA library. Arrowheads denote 3' ends of the transcripts. Clones used to generate probes are marked with asterisk. Transcripts that cannot be detected by probes are shown with thin gray arrows. The nucleotide coordinates relative to Smith sequence (NC_004065.1) of the isolated transcripts are given below thin arrows, while the names of the cDNA clones are written above. Gray histograms show RNASeq reads aligned to MCMV genome,

Smith sequence (NC_004065.1). For Northern blot analysis (A, C, D, E), Balb/c MEF cells were infected with BAC-derived Smith virus and harvested at indicated times post infection. Total RNA was separated using denaturing gel electrophoresis, transferred to nylon membrane and incubated with probes specific for S and AS transcripts. RNA integrity and loading were evaluated by inspecting 28S (not shown) and 18S rRNA bands under UV light after transfer to membrane. Maximal possible exposure times (noted on the blots) were used to ensure the detection of even low abundance transcripts. For PCR analysis (B) of putative splice variants, the same RNA used in Northern blot at 60 h PI was treated with DNaseI, reverse transcribed using oligo(dT) primers and then PCR amplified using primers that flank putative intron (listed in Table 3 and denoted by small half arrows). No reverse transcriptase (-RT) controls were run in parallel. Viral DNA served as unspliced control, while plasmid DNA harboring spliced transcripts served as positive, spliced control.

Adding to the already known transcriptional complexity of this region is the cloning of a novel spliced transcript, E180. Similar to other spliced transcripts in this region, this transcript shares exon 2 with other transcripts in this region. Its splice donor site is located at 102829 bp. Since only one clone of this transcript was isolated, 5' RACE analysis is needed to determine its exact 5' start site. In Figure 16A, spliced E180 is probably represented by the band around 0.5 kb (7 in Figure 16A), whereas unspliced version could be represented by band around 3 kb (3 in Figure 16A). To verify this spliced transcript, PCR analysis was done using primers that flank the intron (sequences of primers are listed in Table 3, results are shown in Figure 16C). Plasmid bearing spliced transcript E180 served as positive control of splicing version, while viral DNA served as positive control of unspliced isoform. Amplification by PCR using RNA isolated from MCMV-infected Balb/c MEFs at 60 h PI (the same RNA sample that was used for Northern analysis) confirmed that E180 splicing is a real, new spliced transcript transcribed from this region for which we propose designation M71S (M71 spliced). Additional PCR was performed to validate M73 spliced transcripts and detected both short (strong band) and long variant as reported by Scalzo *et al*. [114].

Antisense probe transcribed from L42 clone (Figure 16A) also detected 2 weak bands transcribed from – genomic strand around 5 and 3 kb in size which correspond to the transcripts previously reported by Rapp *et al.* [105]. Additional Northern blots using clones L69 (AS to *m72*) and L147 (AS to *m74*) shown in Figure 16 D and E, confirmed that 5-kb transcript starts in *m75* and ends in *m72*, and 3-kb transcript starts in *m74* and ends in *m72*. These results are in accordance with previous reports of transcription in this region [105, 114], where the 5- and 3-kb transcripts encode gH and dUTPase respectively. L147 probe (Figure 16E) also detected 3.5 kb band transcribed from + genomic strand which could represent unspliced novel transcript E180 and an additional, very large transcript coming from

+ genomic strand of unknown origin. The multitude of different transcripts detected in this region underscore the complex transcriptional pattern that seems to be a hallmark of herpesviruses.

4.2.10.4 Analysis of M116 region

Both cDNA library and RNASeq identified M116 as one of the most abundant transcripts. Twenty-three clones in total corresponding to M116 were isolated in cDNA library (5.1% of all cDNA clones isolated), 8 of which were spliced. Current annotations predict an ORF of 1.9 kb, whereas RNASeq profiles and cDNA clones isolated in this study detected a slightly shorter (1.6 kb) transcript with 81-bp intron. To confirm this finding, Northern and PCR analyses were performed (Figure 17).

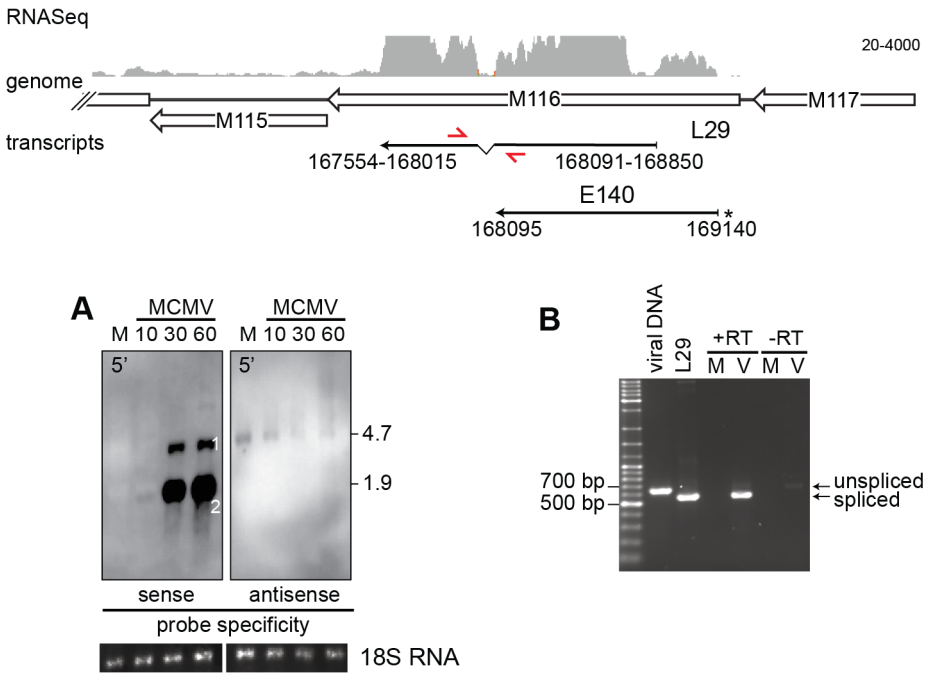


Figure 17. Analysis of transcription in *M116* gene region by Northern blot and PCR. Scheme of *M116* genomic region: predicted ORFs (Rawlinson's annotation) are depicted as empty arrows, thin black arrows show the longest transcripts cloned in our cDNA library. Arrowheads denote 3' ends of the transcripts. Clones used to generate probes are marked with asterisk. The nucleotide coordinates relative to Smith sequence (NC_004065.1) of the isolated transcripts are given below thin arrows, while the names of the cDNA clones are written above. Gray histograms show RNASeq reads aligned to MCMV genome, Smith sequence (NC_004065.1). For Northern blot analysis (A), Balb/c MEF cells were infected with BAC-derived Smith virus and harvested at indicated times post infection. Total RNA was separated on denaturing gel electrophoresis, transferred to nylon membrane and incubated with probes specific for S and AS transcripts. RNA integrity and loading were evaluated by inspecting

28S (not shown) and 18S rRNA bands under UV light after transfer to membrane. Maximal possible exposure times (noted on the blots) were used to ensure even low abundance transcripts. Due to strong smiling effect, detected band sizes were estimated by comparison to ribosomal bands and not RNA ladder. For PCR analysis (B) of putative splice variants, the same RNA used in Northern blot at 60 h PI was treated with DNaseI, reverse transcribed using oligo(dT) primers and then PCR amplified using primers that flank putative intron (listed in Table 3 and denoted as half-arrows). No reverse transcriptase (-RT) controls were run in parallel. Viral DNA served as unspliced control, while plasmid DNA harboring spliced transcript served as positive, spliced control.

Using L29 cDNA clone as probe, 2 bands transcribed – genomic strand (in S orientation to predicted *M116* ORF) were detected (Figure 17A): strong band that corresponds in size to M116 (size determined by comparison with ribosomal bands, not ladder due to intensive smiling effect) and starts to accumulate at IE times PI, and one larger transcript (approx. 3 kb) detectable at E and L times PI. Leatham *et al.* [70] detected a 3.2-kb band in homologous region in HCMV that encompasses *UL119-115* genes. No transcripts were cloned in cDNA analysis overlapping m117 region. As was already mentioned, cDNA libraries tend to over-represent 3' ends of the transcripts and very long transcripts are hard to clone. Therefore, it is possible that the bigger transcript detected in our Northern analysis starts in M117 and ends 3' co-terminally with shorter M116 spliced transcripts (size of the region is 3.3 kb). However, to confirm these speculations, additional Northern or 5'RACE analyses should be performed.

Due to the abundance of M116 and small intron size (83 bp), the signal on Northern blot was too strong to differentiate between bands corresponding to spliced and unspliced transcript variants. PCR analysis using primers that flank the putative intron (Figure 17B) was therefore performed. As can be seen in the agarose gel image, strong band corresponding to the spliced variant and a much weaker band corresponding to the unspliced variant were both detected at 60 h PI.

Current annotation for M116 predicts a protein of 645 amino acids (AA), whereas the splicing results in a novel truncated protein product of 400 AA.

4.2.10.5 Analysis of m168-m169 region

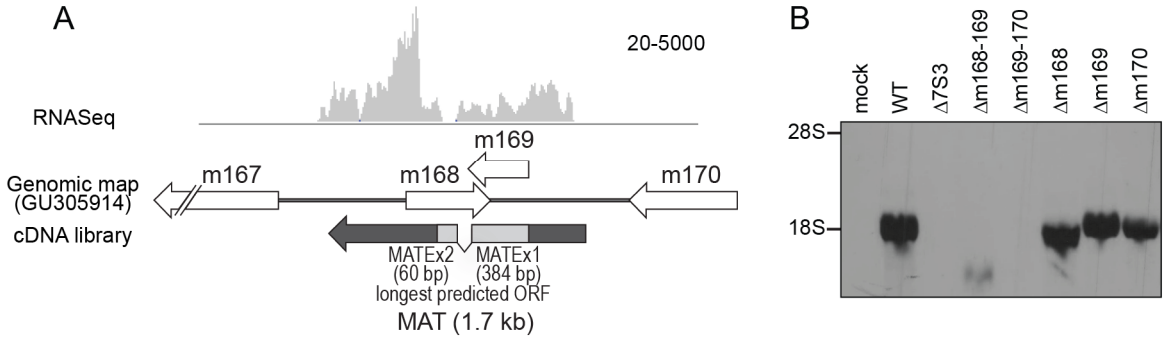


Figure 18. Analysis of transcription in *m168-m169* gene region by Northern blot. (A) Scheme of*m167-m170* genomic region (predicted ORFs are shown with white arrows) showing Rawlinson's modified annotation (gB acc no GU305914) compared to the longest *m168-m169* transcript detected in the cDNA library (E125) and to RNASeq data (gray histograms). Both cDNA and RNASeq data indicate a single, spliced transcript overlapping *m169* in sense and *m168* in AS. The longest predicted ORF is shown as light gray boxes. (B) Northern analysis of MAT RNA in MEF cells infected with various deletion mutants. Note that the single gene mutants are partial gene deletions resulting in the accumulation of truncated transcripts. Δ7S3 MCMV lacks genes from *m167-m170* gene region.

cDNA and RNASeq data both indicated that the expression of a single, 1.7-kb transcript from the right end of the genome dominates the transcriptome of MCMV (Figure 11A). As was already mentioned, 31% of all viral cDNA clones isolated in the cDNA cloning study and 41% of all viral RNASeq reads from the RNASeq analysis mapped to the m169-m168 region of the genome. Both analyses identified a single, spliced transcript whose structure is shown in Figure 18A. Due to its high abundance, it was given the name MAT for most abundant transcript. Using two longest MAT transcripts from the cDNA library, E125 and E134, a DNA Northern probe was generated and subsequent Northern analysis confirmed that a single transcript is transcribed from m169-m168 ORFs (Figure 18B). MAT transcript was detected in all temporal cDNA sublibraries: of 138 cDNA clones, 28 were detected in IE (20%) cDNA library, 57 in E (41%) and 53 in L (38%). This finding indicates that MAT transcript is continuously being transcribed throughout the infection.

The longest predicted ORF in the new MAT transcript matches the predicted *m169* ORF but extends also into *m168* in the second exon and encodes a protein of 147 AA (cca 17 kDa) and was confirmed by Western blot analysis (discussed in chapter 4.2.10.5).

4.3 THE HOST TRANSCRIPTOME

A major advantage of using next-generation sequencing to analyze virally infected cells is that in addition to following viral transcription, RNASeq also allows sequencing and quantification of host transcripts at the same time. Of 34 million RNASeq reads that have passed the filter, only 11% aligned to the MCMV and 67% to the mouse genome (unaligned sequencing reads represent reads that could not be unambiguously aligned to target genomes (some reads spanning exon-exon junctions, reads from highly repetitive regions) or reads coming from adapters).

In order to determine which mouse genes have been affected by MCMV infection, RNA from mock infected and MCMV-infected Balb/c MEF cells was sequenced, and differentially expressed (DE) murine genes were determined by calculating RPKM (reads per kilobase per million mapped reads; [96]) using SAMMate with EdgeR [145]. This analysis identified 10748 statistically significant ($p<0.05$) genes altered by infection. Mm9 (NCBI Build 37) assembly of mouse genome used for the alignment of RNASeq data to mouse genome and in SAMMate for calculation of gene expression levels contains 36678 "genes" (included are non-coding RNAs and ORFs encoding putative proteins). Therefore, nearly 30% of all murine genes are significantly changed as a consequence of the infection. The top induced, upregulated, repressed and downregulated genes are presented in Table 6-Table 9 and are discussed bellow.

4.3.1 Mouse genes induced by the infection

Genes induced by the infection are those which are not transcribed in non-infected MEF but get induced to transcription by MCMV infection. In the infected MEF, 283 (0.84%) of mouse genes were induced. Top 20 are shown in Table 6.

Table 6. Top 20 mouse genes induced by MCMV infection ($p<0.05$). Genes associated with genetic networks identified by IPA are shown in bold.

Gene	Full name	Fold change
Ankrd34b	ankyrin repeat domain 34B	34.4
Ifnb1	interferon beta 1	34.1
Foxa1	forkhead box A1	34.1
Spint1	serine protease inhibitor, Kunitz type 1	33.8
Lin28b	lin-28 homolog B	33.8
En2	homeobox protein engrailed-2	33.3

Gene	Full name	Fold change
Hrk	harakiri, BCL2 interacting protein (contains only BH3 domain)	33.0
Insm1	insulinoma-associated 1	33.0
Pyhin1	pyrin and HIN domain family, member 1; ifi-209; interferon-inducible protein 209	32.9
Tnfsf10	tumor necrosis factor (ligand) superfamily, member 10	32.9
Gabrq	gamma-aminobutyric acid (GABA) A receptor, subunit theta	32.7
1110032F04Rik	RIKEN cDNA 1110032F04 gene	32.6
Cnpy1	canopy 1 homolog	32.5
Slc35d3	solute carrier family 35, member D3; Frcl1	32.4
Esx1	extraembryonic, spermatogenesis, homeobox 1; Spx1	32.4
Esrp1	epithelial splicing regulatory protein 1; Rbm35a	32.3
Tbx21	T-box transcription factor TBX21, T-bet	32.3
Trim71	tripartite motif-containing 71; Lin41	32.2
Trp73	transformation related protein 73	32.0
Cpne5	copine V; A830083G22Rik	32.0
Cdh7	cadherin-7	32.0

Consistent with the expected host response to the infection, among top induced genes are genes associated with interferon response: *Interferon β* (*Ifnb1)* and interferon-inducible *pyhin 1* (alternative names: *ifi-209, ifix*), and genes associated with the induction of apoptosis: *Hrk* and *Tnfsf10* (TRAIL). Transcription factors associated with immune response or development are also a prominent group among top induced genes and include: *Foxa1, En2, Insm1*, *Tbx21*, [aka *T-bet*], and *Trp73.* Interestingly, *Trp73* is also involved in cellular response to stress and is recognized as one of tumor suppressor genes. *Trim71* and *Cpne5* play a role in the development of the nervous system.

4.3.2 Mouse genes upregulated by the infection

Genes upregulated by the infection are genes which are expressed in mock infected MEF but whose expression levels (aka transcription) increased in the infected cells as compared to mock-infected cells and these are the largest group of DE genes. In total 7591 genes were upregulated by the infection (70% of all DE genes). Of that, 1143 (10% of all DE genes) had log fold change of at least 2. Top 20 upregulated genes are shown in Table 7.

Table 7. Top 20 mouse genes upregulated by MCMV infection (p<0.05). Genes associated with genetic networks identified by IPA are shown in bold.

Gene	Full name	Fold change
Art3[1]	ADP-ribosyltransferase 3	8.7
Cxcl10	chemokine (C-X-C motif) ligand 10	8.5
Ccl5	chemokine (C-C motif) ligand 5, RANTES	8.5
Trank1	tetratricopeptide repeat and ankyrin repeat containing 1	8.4
Cxcl9	chemokine (C-X-C motif) ligand 9; Mig	8.1
Rsad2	radical S-adenosyl methionine domain containing 2; virus inhibitory protein	7.9
Dsg2	desmoglein 2	7.9
Mx1	myxovirus (influenza virus) resistance 1	7.1
Ugt8	UDP galactosyltransferase 8A	7.0
Cxcl11	chemokine (C-X-C motif) ligand 11; interferon-inducible T-cell alpha chemoattractant	6.9
Tex16	testis expressed gene 16	6.9
Gpr50	G protein-coupled receptor 50; melatonin-related receptor	6.8
Jag2	Jagged2	6.7
Oasl1	2'-5' oligoadenylate synthetase-like 2	6.5
Cited1	Cbp/p300-interacting transactivator with Glu/Asp-rich carboxy-terminal domain 1; Msg1	6.5
Kcnq2	potassium voltage-gated channel, subfamily Q, member 2	6.5
Map3k9	mitogen-activated protein kinase kinase kinase 9	6.4
Gbp5	guanylate binding protein 5	6.3
Pou4f1	POU domain, class 4, transcription factor 1; Brn3	6.2
Ina	internexin neuronal intermediate filament protein, alpha; NF66	6.2

[1]Overlaps CXCL10 and CXCL11 therefore its upregulation may be due to this overlap

Top upregulated group of genes is largely dominated by chemokine ligands (*Cxcl10, Ccl5, Cxcl9* and *Cxcl11*) and genes with roles in cellular antiviral defense (*Oas1, Mx1, Gpb5* and *Rsad2* (*viperin*)). Similarly to induced genes, a lot of upregulated genes are associated with development and differentiation, especially development of nervous system (*Cited 1, Pou4f1, Jag2, Ina*).

4.3.3 Mouse genes repressed by the infection

Genes repressed by the infection are genes whose active transcription in mock-infected MEF was completely silenced as a consequence of MCMV infection. Of all DE genes, only 15 genes have been found to be repressed and are listed in Table 8. It is important to note that sequencing of total RNA that was used in this RNASeq analysis is less sensitive for the

detection of downregulated and repressed transcripts than sequencing of newly made RNA probably due to long half-life of RNA in mammalian cells [83]. Namely, some cellular RNAs are very stable with long half-life. In the cases of such RNAs, using total RNA as opposed to using newly made RNAs downregulation will not be as noticeable due to preexisting stable RNAs.

Table 8. Genes repressed by the MCMV infection (p<0.05). Genes associated with genetic networks identified by IPA are shown in bold.

Gene	Full name	Fold change
Npy6r	neuropeptide Y receptor Y6	-30.6
Rxfp1	relaxin/insulin-like family peptide receptor 1	-30.3
Gm15411[2]	predicted gene 15411	-29.6
Mc2r	melanocortin 2 receptor, adrenocorticotropic hormone receptor	-29.3
Gm867	predicted gene 867	-29.3
4933400A11Rik	RIKEN cDNA 4933400A11 gene	-29.3
AC159008.1 (Musd2)	Mus Musculus type D-like endogenous retrovirus 2	-29.3
A530013C23Rik[4]	RIKEN cDNA A530013C23 gene	-29.1
Cd200r3	CD200 receptor 3	-29.1
Antxrl	anthrax toxin receptor-like	-29.1
8030423F21Rik	RIKEN cDNA 8030423F21 gene	-29.1
Mup3	major urinary protein 1	-29.1
Gm10689	predicted gene 10689	-29.1
4930455H04Rik	RIKEN cDNA 4930455H04 gene	-29.1
4930412B13Rik	RIKEN cDNA 4930412B13 gene	-29.1

[1]Total number of host genes repressed in the infection with p<0.05
[2]lincRNA

4.3.4 Mouse genes downregulated by the infection

Genes downregulated by the infection have expression profiles opposite of induced genes: their expression levels are negatively influenced by the infection, leading to lower transcription post infection. Of all DE genes, 2859 genes exhibited downregulated expression

(27% of all DE genes); however, only 228 had log fold change of -2 or smaller (8%). Top 20 genes downregulated by the infection are shown in Table 9.

Table 9. Top 20 downregulated mouse genes (p<0.05). Genes associated with genetic networks identified by IPA are shown in bold.

Gene	Full name	Fold change
Ggt2	gamma-glutamyltransferase 2	-5.6
Scara5	scavenger receptor class A member 5; testis expressed scavenger receptor	-5.1
Il1r2	interleukin 1 receptor, type II	-4.7
E230015J15Rik	RIKEN cDNA E230015J15 gene	-4.5
Gm12963[1]	predicted gene 12963	-4.4
Gpr165	G protein-coupled receptor 165	-4.3
Clec3b	C-type lectin domain family 3, member b	-4.3
Gm15883[1]	Predicted gene 15883	-4.2
Palmd	Palmd	-4.2
Agtr2	angiotensin II receptor, type 2	-4.2
Gm16890[2]	Dsec\GM16890	-4.1
Ahnak2	AHNAK nucleoprotein 2	-4.0
Cyp2f2	cytochrome P450, family 2, subfamily f, polypeptide 2	-3.9
Gm10544[2]	predicted gene 10544	-3.9
Gstm6	glutathione S-transferase, mu 6	-3.8
Gm12575[2]	predicted gene 12575	-3.8
mmu-mir-685.1[3]	microRNA 685	-3.8
Olfr1314	olfactory receptor 1314	-3.7
Snord15a	small nucleolar RNA, C/D box 15A	-3.7
Olfr78	olfactory receptor 78	-3.7

[1]antisense transcripts
[2]lincRNA
[3] microRNA record discontinued

The top downregulated and all repressed (see above) genes are of unknown relevance to infection, though many are receptor or cell surface molecules (*Npy6R, Rxfp, Mc2r, Cd200r3, Antxrl, Scara5, Il1r2, Agtr2, GPR165*, the olfactory receptor genes, *Olfr1314* and *Olfr78* and the lectin or lectin-like genes *Clec 3b* and *Reg3A*). Interestingly, among top repressed and downregulated genes many are noncoding transcripts including small nucleolar RNA (*Snord15A*), miRNA (mmu-mir-685.1), 4 long intergenic noncoding RNAs (lincRNAs: *Gm10544, Gm15411, Gm16890, Gm12575*), the miscellaneous RNA, *4930412B13Rik*, and 2

antisense transcripts (*Gm12963, Gm15883*). The relevance of these non-coding transcripts to MCMV infection is unknown but underscores the advantage of RNASeq over protein-coding oriented microarrays.

4.3.5 Validation of RNASeq analysis of host genes by Western blot

A primary caveat of any transcriptomic analysis is determining whether changes in gene transcript levels are also reflected at the protein level. While a recent paper by Schwanhäusser *et al*. [116] found a much better correlation between transcript and protein levels for mammalian cells than previous analyses, the correlation is still pretty poor (around 40%). Cells themselves regulate protein levels not only at transcript levels but also at post-transcriptional, translational and posttranslational levels. For instance, genes involved in cell adhesion, phosphorylation, proteolysis, integrin-mediated signaling and defense response have been found to have stable RNAs but unstable proteins. In addition, herpesviruses can exert their influence on host proteins on all these levels as well [28, 124, 131].

Many DE genes identified in RNASeq analysis have either been previously reported as impacted by CMV or are targeted by other herpeviruses (e.g. induction of interferon and interferon-inducible genes is a well known feature of CMV [86]; induction of viperin was shown for HCMV [118]) and thus needed no further confirmation. Protein levels of several genes which were found to be induced or upregulated in the infected MEF cells and with no known relevance to MCMV infection were analyzed: notch ligands Delta 1 and Jagged 2, homeobox containing transcriptional factor Engrailed 2 and E3 ubiquitin-protein ligase Trim71. Protein levels of all these proteins correlated with their transcript levels in the infected Balb/c fibroblasts and are shown in Figure 19. Additionally, protein levels of *Jag2* were also correlated with transcript levels when immortalized endothelial cell line SVEC was used (data not shown).

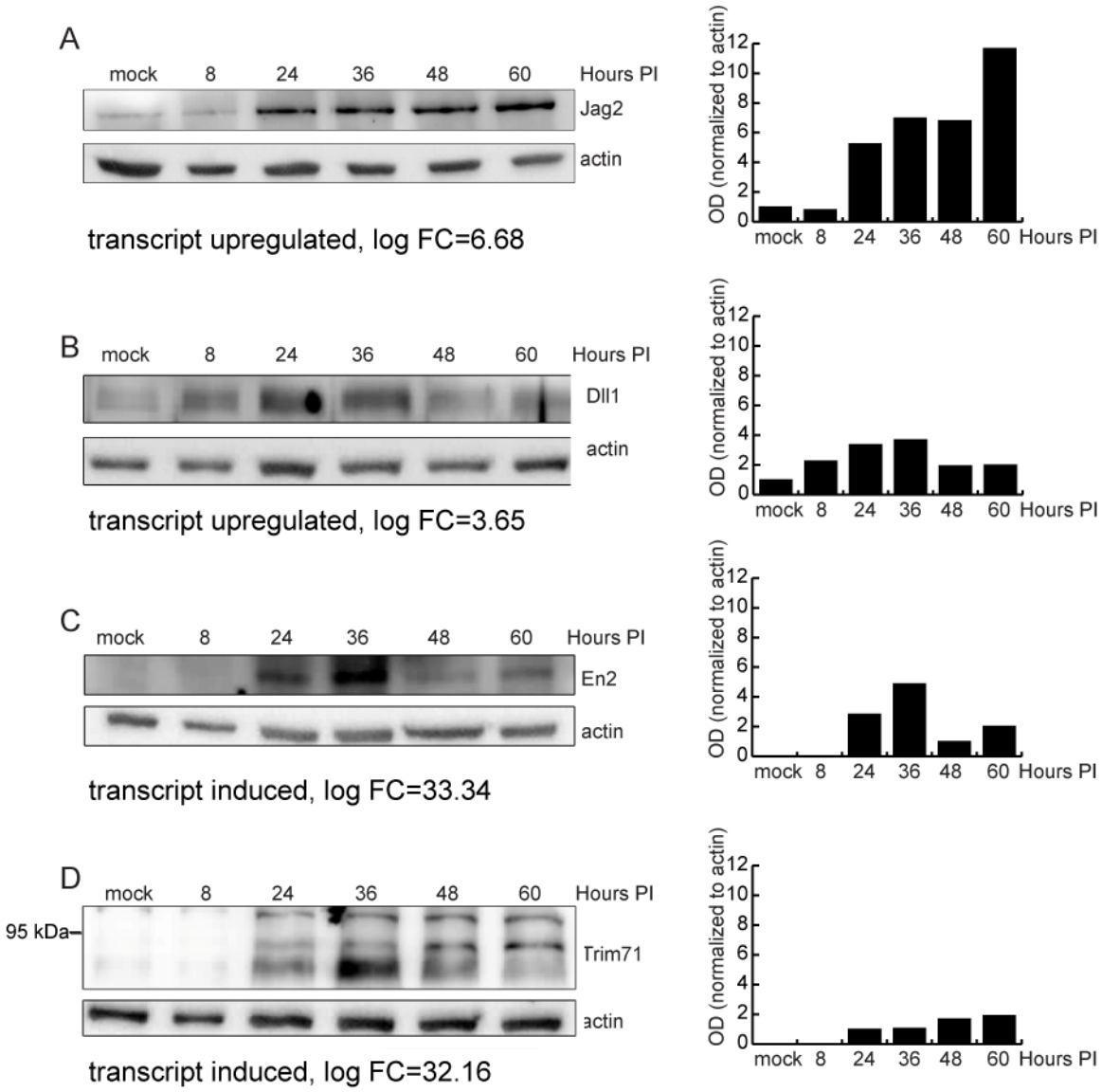


Figure 19. Validation of RNASeq analysis of host genes by Western blot. (A) Immunoblot analysis of MEF.K (A) or Balb/c MEF (B-D) cell lysates infected with wild-type MCMV. Cell lysates were separated by SDS-PAGE, transferred to PVDF membrane, and probed with antibody to Jag2 (A), Delta 1 (B), En2 (C) or Trim71 (D). Monoclonal antibody to actin was used as loading control. Bar charts represent relative quantification of proteins using ImageJ. In the case of Trim71 (D), where anti-Trim71 antibody detected multiple bands, the bars show quantification of the middle band.

4.3.6 Gene networks altered by MCMV

Differential expression analysis using SAMMate identified 10748 statistically significant ($p<0.05$) differentially regulated genes in the infected MEF. Such long list is nearly impossible to analyze on gene-by-gene basis and, while lists of most highly differentially regulated genes can be very informative, they do not give the full picture. Genes do not work in isolation but rather form pathways and networks. Several small misbalances in the

expression if genes involved in one particular pathway or network can have as much influence as strong differential regulation of just one gene, especially if regulatory genes impacting multiple pathways are targeted. Gene networks offer one way of understanding and resolving such complex interactions and answer the question of what regulatory relationships exist between significantly perturbed genes in a particular dataset. Gene network analysis as well as functional analyses of gene networks were performed using Ingenuity's IPA Core analysis on differentially expressed genes identified by SAMMate using fold change cut-off of 2.

Three analyses were performed: analyses with a whole dataset of differentially expressed genes including genes induced/repressed by the infection and genes up- or down-regulated (aka differentially regulated (DR)) in the course of the infection. Such strategy was selected for the following reason: in our dataset, most DE genes fall into the category of differentially regulated, while a smaller portion are induced or repressed. On the other hand, induced/repressed genes have a bigger fold change score. Therefore, in order to avoid introduction of any biases, IPA analyses were performed on all DE genes but also on differentially regulated and induced/repressed gene sets in isolation.

When all differentially expressed (DE) genes were analyzed (induced, repressed up- and down-regulated), 3 top scoring gene networks were all associated with immune and antimicrobial response. Top 10 scoring gene networks are shown in Figure 20.

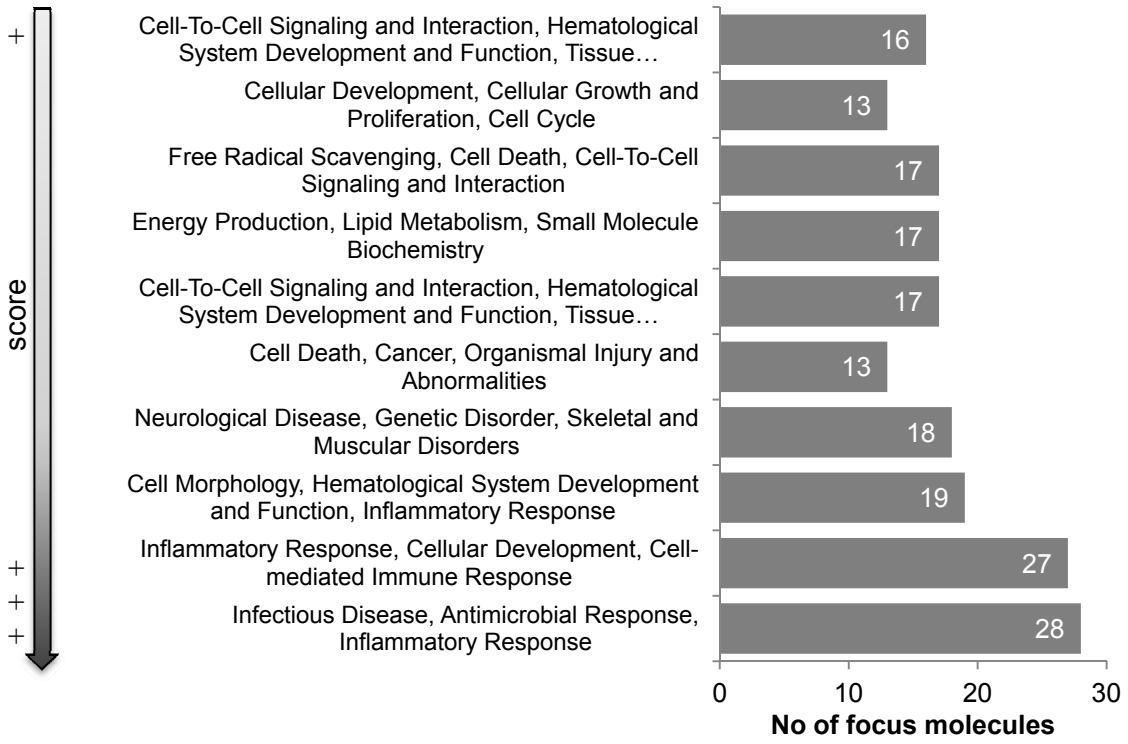


Figure 20. Top 10 scoring networks associated with DE genes. Top scoring networks are shown at the bottom.

Similar top networks were identified when only differentially regulated (up- and down-regulated) genes were analyzed. Similarities between the findings for all differentially regulated genes and up/down-regulated is not surprising since in our dataset the majority of DE genes fall into the category of up/down-regulated rather than induced-repressed. Top 10 scoring gene networks for differentially regulated genes are shown in Figure 21. Merged graphical representation of top 3 scoring networks and molecular relationships between genes in those networks in DR genes dataset is shown in Figure 22.

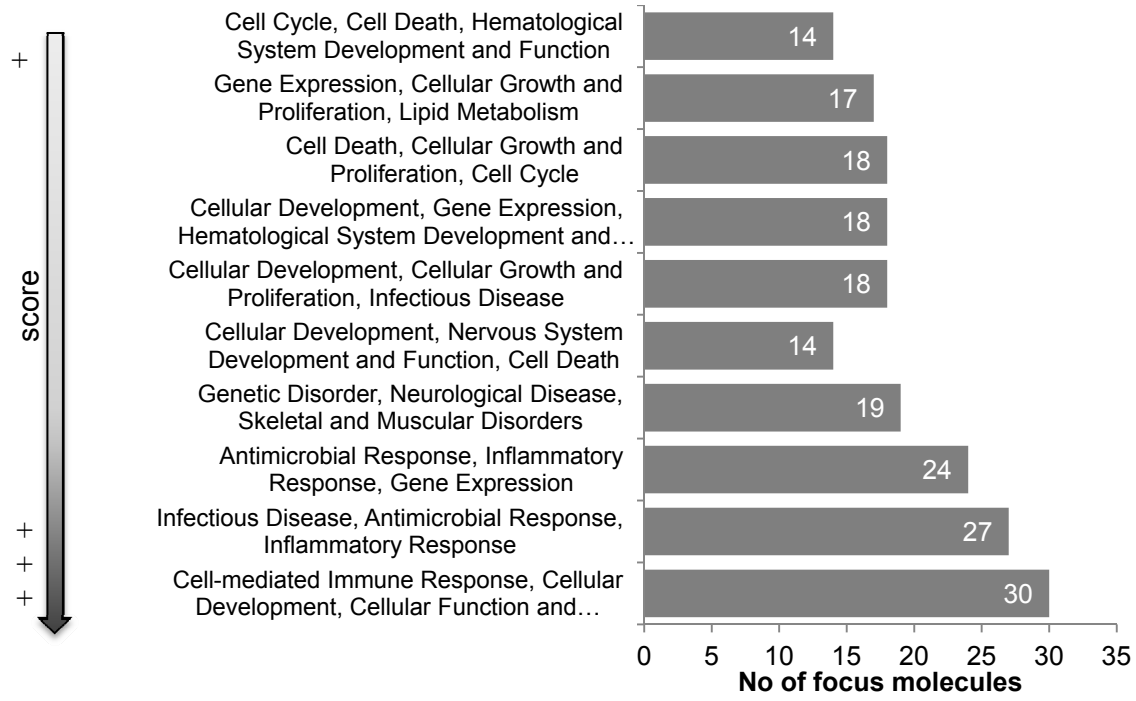


Figure 21. Top 10 scoring networks associated with differentially regulated genes. Top scoring networks are shown at the bottom.

Also identified for DE and DR gene networks were those associated with neurological disease, skeletal and muscular disorders, hematological development, cell cycle and development and lipid metabolism; all of which are known targets or consequences of cytomegalovirus infection.

When gene network analysis was conducted with only induced and repressed genes, the top networks identified were predominantly oriented towards development and included cellular development, cell-mediated immune response, cellular function and maintenance, gene expression and embryonic development.

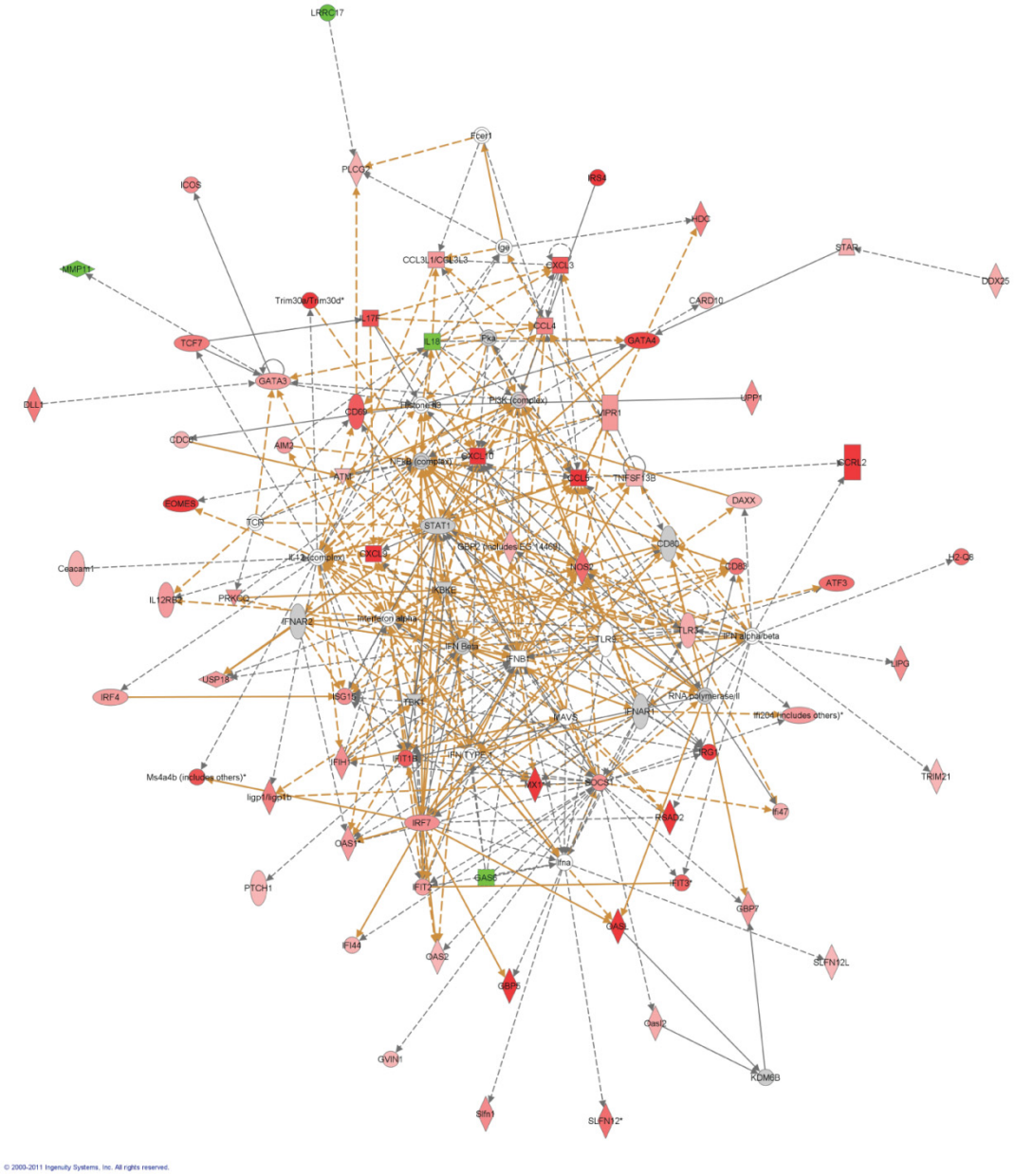


Figure 22. Graphical representation of top 3 genetic networks identified for DR genes. A fold change cut-off of 2.0 was set to identify genes whose expression was significantly differentially regulated. These genes, called focus genes, were overlaid onto a global molecular network developed from the information contained in the Ingenuity knowledge base. Networks of these focus genes were then algorithmically generated based on their connectivity. Genes or gene products are represented as nodes, and the biological relationship between two nodes is shown as an edge (line). All edges are supported by at least 1 reference from the literature, from a textbook, or from canonical information stored in the Ingenuity knowledge base. Human, mouse, and rat orthologs of a gene are stored as separate objects in the Ingenuity knowledge base, but are represented as a single node in the network. The intensity of the node color indicates the degree of up - (red) or down - (green) regulation.

4.3.7 Functional analysis of gene networks

Functional analysis of networks identifies biological functions and diseases that are associated with genes in the top networks. Canonical pathway analysis answers which well characterized signaling and metabolic pathways are most perturbed in the analyzed dataset.

Molecular, cellular and developmental functions associated with gene networks identified for DE genes dataset are depicted in Figure 23A. Strikingly, a strong bias for developmental functions can readily be observed. In addition to genes with functions important for the development and immune response, IPA identified cardiovascular disease, genetic disorders and skeletal and muscular disorders as top bio-functions connected with diseases and disorders altered by MCMV infection. While MCMV involvement in cardiovascular disease is a subject of intensive research, potential involvement in skeletal and muscular disorders is not so well documented. Nervous system development and function is at the top of the list of physiological and developmental biofunctions, followed by organismal and tissue development and, surprisingly, behavior with 92 associated differentially regulated genes. Among molecular and cellular functions, cell growth and proliferation were the top ranked perturbed functions, consistent with known effects of lytic MCMV infection of cells.

DE genes associated with well described, canonical pathways from Ingenuity's library were also evaluated (Figure 23 B). The pathways most affected by MCMV were G-protein coupled receptor signaling, pathogenesis of multiple sclerosis and GABA receptor signaling.

Gene network and functional analyses have pointed out both known and expected consequences of infection whose relevance to MCMV infection is well documented and also functions and diseases which were not so far associated with CMV infection.

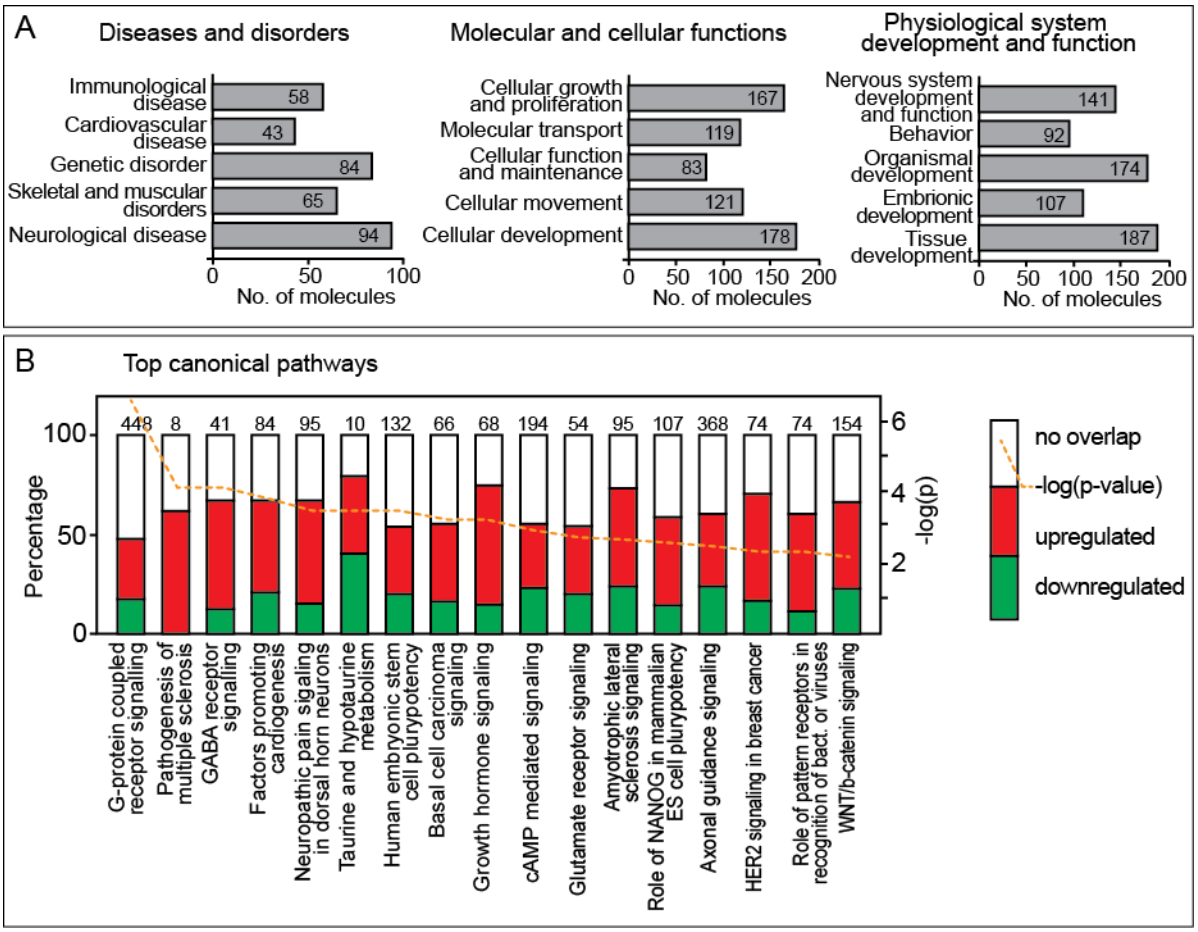


Figure 23. IPA functional analysis of gene networks in DE mouse gene dataset in MCMV infection. Differentially expressed genes were identified by SAMMate and analyzed with IPA Core Analysis with fold change ratio cut-off of 2. Shown are top diseases and disorders, molecular and cellular functions and physiological system development and functions (A) and top canonical pathways (B) of DE genes.

4.3.8 GO enrichment analysis of DE genes

IPA analysis requires an arbitrary cutoff threshold and analyzes the data by filtering out genes with unknown functions and/or relationships with other genes. In order to avoid missing some potentially interesting biological functions, gene ontology enrichment analysis was performed on all DE genes using GOrilla software [41, 42]. Gene ontology (GO) is a result of a major bioinformatics initiative that strives to standardize the representation of gene and gene product attributes across species and databases. Basically, a gene is given a list of attributes pertaining to its function (or supposed function), sub-cellular localization, involvement in pathways, etc. using controlled vocabulary of terms. Gene ontology enrichment analyses GO terms for genes in a dataset and reports whether particular term is overrepresented. GOrilla tool [42] offers the analysis of ranked lists where it identifies enriched terms at the top of the

given list and threshold is determined by the data rather than by the user, thus eliminating user biases. Two GOrilla ranked list analyses were performed – for induced/upregulated genes and downregulated/repressed.

When induced/upregulated genes were analyzed, enriched biological processes included developmental processes, cell and neuron differentiation, transcription, G-protein coupled signaling, reproductive process and regulation of ion transport, while most enriched gene functions were nucleic acid-binding transcription factor activity, ion channel activity, neurotransmitter receptor activity and cytokine activity. Genes downregulated/repressed during MCMV infection were associated with cell adhesion, motility, extracellular matrix organization, regulation of developmental processes, cell communication and proliferation of biological processes. One unexpected process associated with downregulated/repressed genes was sensory perception of smell. Functions associated with genes in downregulated/repressed group included molecular transducer activity, receptor binding, ion channel activity, activity of various enzymes and enzyme inhibitors, activity of several growth factors and neuropeptide receptor activity.

Altogether, GOrilla analyses support the results of the Ingenuity pathway analysis but also suggest novel processes regulated in the infected cells, notably suggesting that infection leads to a restructuring of the extracellular environment of the infected cells.

4.4 ANALYSIS OF MOST ABUNDANT TRANSCRIPT (MAT)

One unexpected finding of MCMV transcriptome was the domination of a single transcript of unknown function, MAT. This finding was even more interesting in the light of our previous observations that mutant viruses lacking *m168-m170* predicted ORFs are significantly attenuated *in vivo* (Marina Babić Čač, unpublished results, PhD thesis) in NK-cell-dependent manner. This phenomenon was in part explained by the finding that 3'UTR of this transcript binds cellular microRNA miR-27 [72, 84]. In addition, this region was found to be necessary, along with viral protein gp34/m04, for efficient recognition of infected cells by natural killer (NK) cells via activating Ly49 receptors (Marina Babić Čač, unpublished results, PhD thesis). To gain a deeper insight into this interesting genomic region, MAT transcript and its coding potential was further analyzed.

4.4.1 MAT is transcribed and gives rise to low-abundance protein

MAT transcript is 1.7 kb long transcript overlapping putative *m168* ORF in antisense orientation and *m169* in sense orientation (Figure 18). The longest predicted ORF overlaps in frame with the predicted *m169* ORF but extends into predicted *m168*. This ORF should give rise to a protein of 147 AA, of which the first 127 residues match the predicted *m169* protein sequence. To determine if this ORF is translated, monoclonal antibody (mAb) was prepared to the protein sequence predicted for ORF *m169*. Western blot analysis using Balb/c (Figure 24), C57Bl/6 MEF cells (data not shown) and macrophage cell line Raw 264 (data not shown) infected with a panel of deletion mutants confirmed that the longest predicted ORF is indeed translated and gives rise to 17 kDa protein. Immunoblot with mAb for MCMV protein *m04* was used as a control of successful infection, whereas staining with α-actin mAbs was used as loading control.

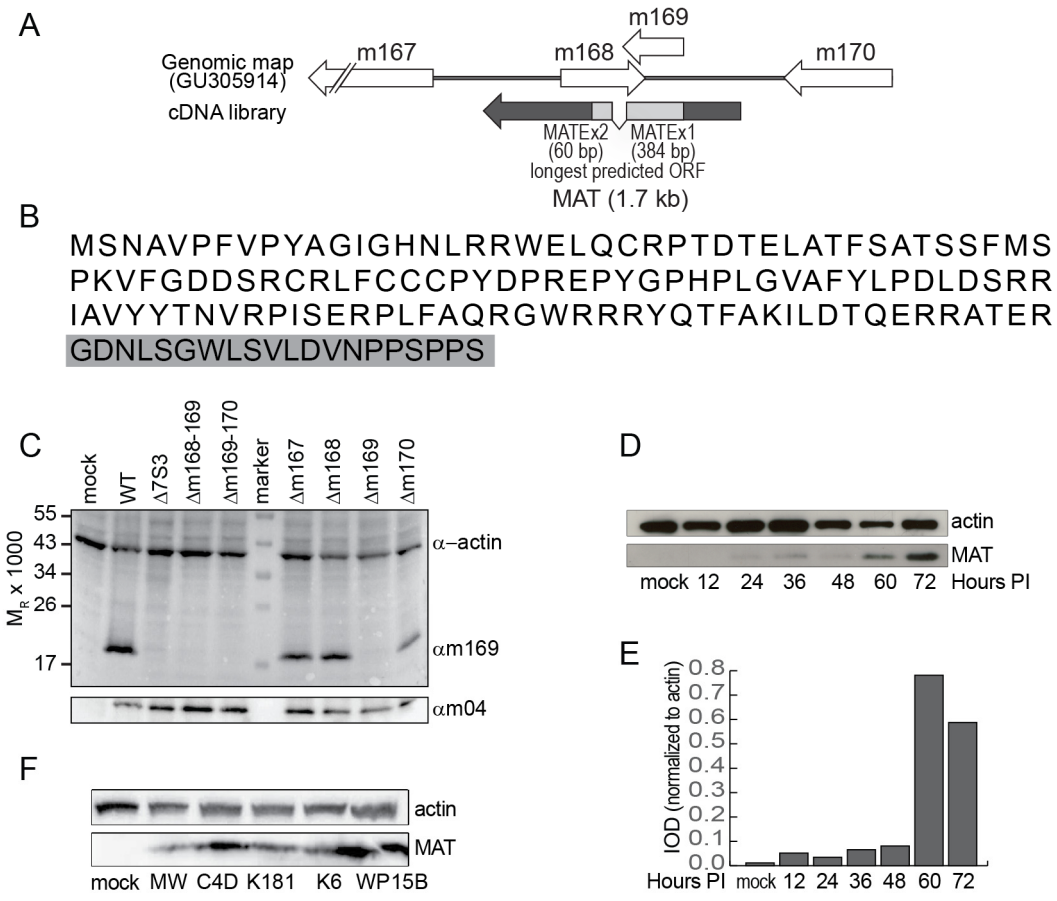


Figure 24. Detection and characterization of MAT protein. (A) Schematic representation of MAT (dark gray arrow) in relation to Rawlinson's annotation (white arrows). The longest predicted ORF (light gray boxes) overlaps completely predicted *m169* ORF and extends into *m168* in AS orientation. (B) Translation of putative MAT protein. The first 127 residues match a truncated *m169* translation

and the 20 C-terminal residues highlighted in gray are derived from exon 2, mapping to the *m168* gene. (C) Immunoblot analysis of Balb/c MEF cell lysates probed with monoclonal antibody generated to the predicted *m169* ORF, monoclonal antibody to actin (45 kDa band, loading control) or monoclonal antibody to viral gene *m04* (serves as control of infection). (D) Immunoblot analysis of MAT protein accumulation in the infected cells over 72 hours and (E) relative quantitation. (F) Immunoblot analysis of MAT protein from the cells exposed to wild virus isolates.

The *m169* mAb detected MAT protein in fibroblasts infected with Smith strain WT MCMV as well as 4 other field isolates indicating that this protein is conserved among wild strains of MCMV. BLAST analysis [94, 148] of nucleotide sequence encoding MAT protein showed 99% conservation among all sequenced wild isolates of MCMV (Table 10).

Table 10. BLASTn analysis of MAT ORF. MAT ORF sequence was analyzed in nucleotide BLAST against nucleotide collection. Query coverage is 96% due to splicing.

Description	Max score	Total score	Query cover	E value	Max ident	Accession
Murid herpesvirus 1, Smith strain, complete genome	710	710	96%	0.0	100%	GU305914.1
Murine cytomegalovirus (strain K181), complete genome	710	710	96%	0.0	100%	AM886412.1
Murid herpesvirus 1 strain NO7, complete genome	699	699	96%	0.0	99%	HE610455.1
Murid herpesvirus 1 strain C4D, complete genome	693	693	96%	0.0	99%	HE610456.1
Murid herpesvirus 1 strain N1, complete genome	693	693	96%	0.0	99%	HE610454.1
Murid herpesvirus 1 strain C4B, complete genome	693	693	96%	0.0	99%	HE610452.1
Muromegalovirus WP15B, complete genome	693	693	96%	0.0	99%	EU579860.1
Murid herpesvirus 1 strain C4C, complete genome	688	688	96%	0.0	99%	HE610453.1
Murid herpesvirus 1 strain AA18d, complete genome	688	688	96%	0.0	99%	HE610451.1
Muromegalovirus C4A, complete genome	682	682	96%	0.0	99%	EU579861.1
Muromegalovirus G4, complete genome	682	682	96%	0.0	99%	EU579859.1

Interestingly, while MAT transcript is highly abundant and detectable in all temporal cDNA libraries, MAT protein is first detectable at 24 h PI and reaches its maximal amounts at very late times post infection.

In addition to previously published findings that this transcript regulates cellular levels of miR-27 [72, 84], our findings demonstrate that the MAT gene region generates a single transcript with both noncoding and protein-coding functions.

4.4.2 MAT protein is cytoplasmic protein

Our transcriptomic analysis analyzed only polyadenylated transcripts; thus MAT is also polyadenylated transcript. Libri *et al.* [72] confirmed polyadenylation of MAT and by using *in situ* hybridization showed that MAT localized in cytoplasm. In order to determine localization of MAT protein, proteins from nuclear and cytoplasmic cell fractions were separated using PARIS cell fractionation kit (Ambion) and analyzed by immunoblot using mAbs against MAT, m04 and actin. As can be seen in Figure 25, MAT protein could only be detected in cytoplasmic fraction.

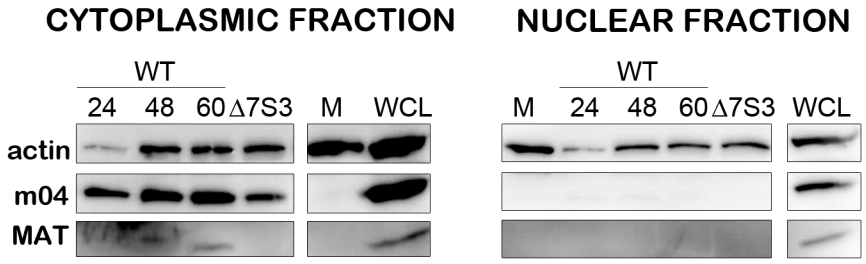


Figure 25. Localization of MAT protein. Immunoblot analysis of cytoplasmic and nuclear fractions of MEF cells infected with WT and Δ7S3 (lacks m167-m170 ORFs). Whole cell lysate (WCL) was used as a positive control of immunoblot assay. MAT protein could only be detected in cytoplasmic fraction.

4.4.3 Regulation of MAT protein expression

Despite very high abundance of MAT, MAT protein becomes detectable only at 24 hours PI and accumulates at low levels, as was shown in Figure 24D and E. One possible explanation for such low protein levels is regulation of MAT transcript abundance by cellular miR-27 [84]. Marcinowski *et al.* [84] have shown that when binding site for miR-27 is mutated (m169-mut virus), MAT transcript levels are increased twofold in comparison with cells

infected with WT MCMV at 24 hours PI due to loss of transcript regulation by mir-27 microRNA. The difference in MAT transcript abundance between cells infected with WT and m169-mut virus was lost by 48 hours PI. Interestingly, no differences in MAT protein amounts between WT and m169-mut viruses were observed at any time points tested (Figure 26) presumably since MAT protein gets translated late in the infection when regulation of transcript abundance by miR-27 no longer plays a role. Another possible explanation for the low levels of MAT protein is that MAT protein is rapidly degraded. To test for that, cells infected with WT and m169-mut virus were treated with the inhibitor of lysosomal degradation (leupeptin), irreversible proteasomal inhibitor lactacystin and reversible proteasomal inhibitor MG132 12 hours before cell collection and the resultsing immunoblot is shown in Figure 26.

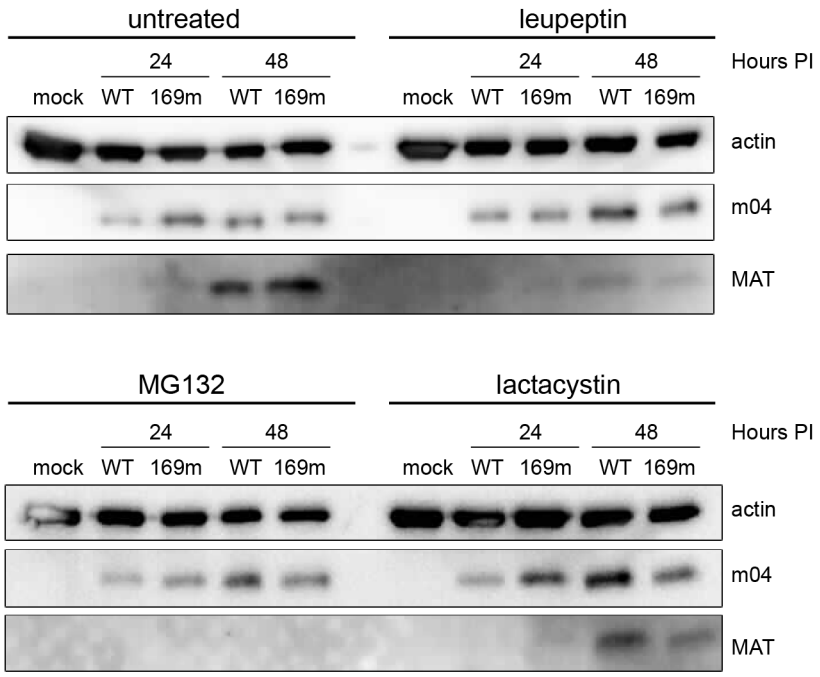


Figure 26. MAT protein abundance is not regulated by miR-27 or by rapid degradation. (A) Immunoblot analysis of MAT protein in WT or m168mut MCMV-infected Balb/c MEF cells at 24 and 48 h post infection. Inhibitors were added 12 hours before cell collection to avoid the influence of their cytotoxic effect on the cells.

As can be seen in Figure 26, the treatment of cells with either lysosomal or proteasomal inhibitors did not result in increased accumulation of MAT protein, indicating that MAT protein is not regulated by rapid degradation. Interestingly, in cells treated with MG132, MAT

protein translation was completely abrogated, whereas treatment with leupeptin and lactacystin resulted in the diminished levels of MAT protein. Removal of miR-27 binding site in MAT 3'UTR had no impact of MAT protein accumulation in the cells treated with inhibitors.

Finally, viruses with deletion of either MAT 5'UTR or intron were generated to test whether 5'UTR or intron regulate MAT protein accumulation. Immunoblot analysis of MAT protein levels at 16 hours (data not shown) and 48 hours PI (Figure 27) showed that MAT protein levels are regulated by MAT's 5'UTR.

As can be seen in Figure 27 when 5'UTR of the transcript was deleted, the amount of MAT protein increased by several orders of magnitude at both 16 h PI (data not shown) and 48 h PI. In WT MCMV, MAT is hardly detectable at 16 hours PI.

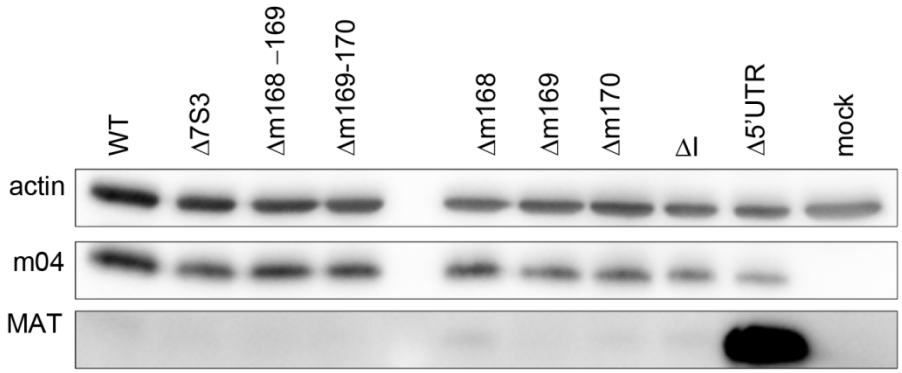


Figure 27. MAT protein accumulation is regulated by its 5'UTR. Immunoblot analysis of MAT protein levels in cells infected with different MCMV deletion mutants at 48 h PI.

4.4.4 MAT 5'UTR contains potential uORFs and is highly variable among field isolates

Translation starts by binding of translation initiation complex to 5' cap structure on mRNA. Then the initiation complex, comprised of 40S ribosomal subunit, initiator tRNA, GTP and several initiation factors, scans the mRNA for start codons. The scanning process may be hampered by long 5'UTRs, especially if they form secondary structures and/or contain AUGs (reviewed in [101]) or even upstream ORFs (uORFs) [95].

Having found that the 5'UTR regulates translation of MAT protein, sequence analysis of 5'UTR was performed. MAT transcript contains long 5'UTR (>400 bp), riddled with AUG (6

found) and GUG start codons (15 found) in sense orientation, and potentially encodes a small additional, upstream ORF (Figure 28). A potential uORF is 264 bp long if it starts with AUG or 321 if non-canonical start codon GUG is used.

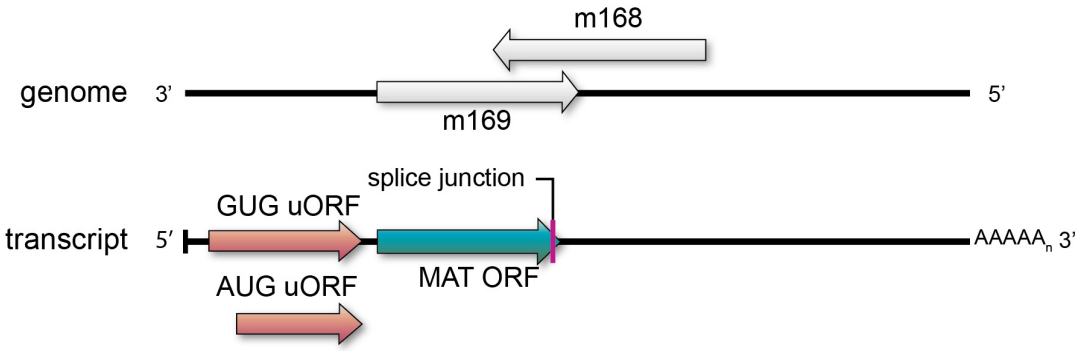


Figure 28. Schematic representation of MAT transcript structure and locations of putative uORFs.

Interestingly, unlike MAT protein, which is well conserved among different MCMV strains (99% sequence identity, see chapter 4.4.1), 5'UTR is highly variable with max sequence identity going as low as 80% for some field MCMV isolates (Table 11).

Table 11. BLASTn analysis of MAT 5'UTR. 5'UTR sequence of MAT consensus sequence was analyzed using nucleotide BLAST [94, 148] against nucleotide collection.

Description	Max score	Total score	Query cover	E value	Max ident	Accession
Murid herpesvirus 1 strain Smith, complete genome	747	747	100%	0	100%	GU305914.1
Murine cytomegalovirus (strain K181), complete genome	747	747	100%	0	100%	AM886412.1
Muromegalovirus G4, complete genome	176	275	44%	1.00E-40	95%	EU579859.1
Murid herpesvirus 1 strain N1, complete genome	628	628	97%	8.00E-177	95%	HE610454.1
Murid herpesvirus 1 strain C4D, complete genome	601	601	97%	2.00E-168	94%	HE610456.1

Description	Max score	Total score	Query cover	E value	Max ident	Accession
Murid herpesvirus 1 strain C4B, complete genome	601	601	97%	2.00E-168	94%	HE610452.1
Murid herpesvirus 1 strain AA18d, complete genome	156	245	44%	2.00E-34	92%	HE610451.1
Muromegalovirus C4A, complete genome	156	223	45%	2.00E-34	92%	EU579861.1
Murid herpesvirus 1 strain C4C, complete genome	158	244	46%	5.00E-35	91%	HE610453.1
Murid herpesvirus 1 strain NO7, complete genome	473	473	97%	4.00E-130	89%	HE610455.1
Muromegalovirus WP15B, complete genome	261	261	97%	3.00E-66	80%	EU579860.1

4.4.5 5'UTR is responsible for recognition of infected cells by activating Ly49 receptors

Natural killer (NK) cells play an important role in virus control at early times after infection. Their importance in CMV pathogenesis and infection is perhaps best underscored by the numerous evasion mechanisms developed by CMVs to evade NK cell control (reviewed in [74]). NK cells survey their surroundings via panel of activating and inhibitory receptors and the decision whether an NK cell will be activated or not depends on the balance of signals coming from these receptors. Ly49 is a family of NK cell receptors containing inhibitory and activating members. Inhibitory Ly49 receptors screen the cells for the presence of MHC I and thus play a role in "missing-self" recognition [7]. Activating Ly49 receptors, on the other hand, recognize viral proteins or viral proteins in addition to MHC I [6, 60, 122]. We have previously shown that, in addition to Ly49H which recognizes virally encoded m157 protein, activating Ly49P, L and D2 specifically recognize MCMV-infected cells [60]. MCMV-encoded m04/gp34 was shown to be necessary but not sufficient for successful recognition via activating Ly49P, L and D2 receptors (Marina Babić Čač, PhD thesis). MAT transcript was identified as additional, necessary requirement for efficient activation of Ly49 P, L and D2 receptors.

To show which region of the MAT transcript is needed for recognition by activating Ly49 receptors, C3H MEF was infected with a panel of MAT deletion mutants and then incubated with Ly49P and L reporter cells, as described in chapter 3.2.14. As can be seen in Figure 29, this analysis identified MAT 5'UTR as the region crucial for recognition of infected cells by activating Ly49 receptors.

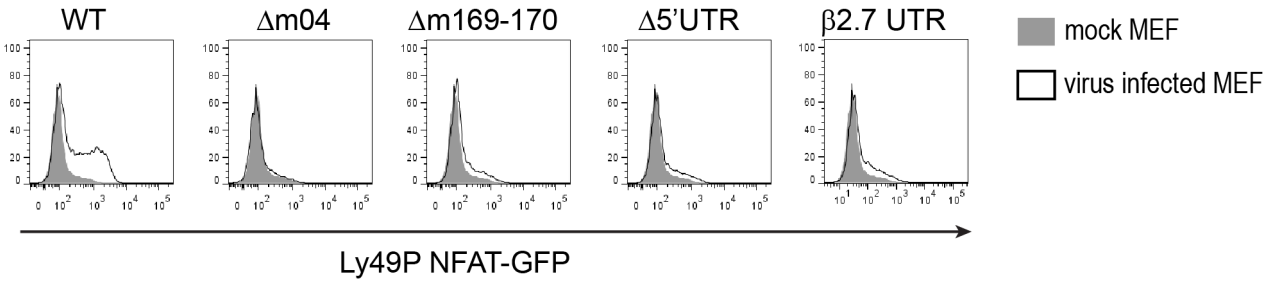


Figure 29. MAT 5'UTR is needed for recognition of infected cells by activating Ly49 receptors. MCMV-infected C3H MEF was coincubated with Ly49P and Ly49L (not shown) for 24 hours. Activation resulted in GFP expression, which was measured by flow cytometry. Gray filled histograms show reporter cells incubated with mock-infected MEF that did not result in the activation of reporter cells and consequent expression of GFP. Empty histograms overlaid over gray histograms represent reporter cells incubated with infected MEF. Activation can clearly be seen in the cells incubated with WT MCMV infected MEF. In contrast, incubation with Δm04, Δm169-170 (no MAT transcript), Δ5'UTR (virus that expresses MAT transcript without 5'UTR) or β2.7 UTR (virus where MAT 5'UTR is replaced with 5'UTR of HCMV transcript β2.7) infected MEF failed to activate reporter cells.

We have previously shown that cells infected with Δm169-170 could not activate Ly49P or L reporter cells (Marina Babić Čač, PhD thesis). As was shown in Northern blot analysis of this region (Figure 18), deletion of *m169* and *m170* ORFs (Δm169-170 virus) results in the destruction of MAT transcript, probably since this deletion encompasses the start signal of the transcript and part of the promoter. However, deletion of *m168* and *m169* preserves the transcript, although at significantly lower levels (Figure 18). Cells infected with Δm168-169 virus could activate reporter cells but at a much lower level, indicating a role for 5'UTR (Marina Babić Čač, PhD thesis). Finally, deletion of just *m169* ORF results in the activation of reporter cells comparable to that of WT virus (data now shown). Since the removal of 5'UTR results in significant upregulation of MAT protein levels, virus in which MAT's 5'UTR was replaced with the 5'UTR of HCMV β2.7 transcript was also tested in reporter cell assay (Figure 29). Although MAT protein levels in β2.7 UTR virus are comparable to the levels observed in WT virus (Lars Dölken, personal communication), this virus was not able

to activate reporter cells. This indicates that the activation of reporter cells is a function of 5'UTR but is not connected with its function as regulator of MAT protein expression.

4.4.6 Field MCMV isolates cannot activate reporter cells

Since 5'UTR, part of MAT transcript needed for recognition of MCMV infected cells via activating Ly49 receptors, was shown to be highly variable among different MCMV strains, several filed isolates were tested in reporter cell assay. As can be seen in Figure 30A, only G4 virus isolate could activate reporter cells. In addition to the viruses depicted in Figure 30, field isolates K6 and C4D were also tested and showed phenotype comparable to that of K181.

Since m04 and MAT 5'UTR are highly variable among different field isolates (Table 11 and [24]), it is impossible to assess whether the inability of MCMV field isolates to activate reporter cells is due to m04/gp34 or MAT 5'UTR. To address this problem, reporter cell assay was performed on MEFs co-infected with two viruses. To test the influence of variability of m04/gp34, MEF was infected with equal amounts of Δm04 MCMV Smith virus and field isolate. As can be seen in Figure 30.B, only cells infected with G4 virus could activate reporter cells when co-infected with Smith Δm04 virus. Of all field isolates so far published, m04/gp34 from G4 field isolate is the most similar and highly related to that of Smith strain m04/gp34. Based on the results of this experiment, it can be deduced that the inability of field isolates to activate reporter cells is mostly due to the variability in their *m04* ORF.

To test theinfluence of MAT 5'UTR, MEF cells were coinfected with equal amount of Smith strain Δ5'UTR and field isolates. Under these conditions, the majority of MAT 5'UTRs of the field isolates were able to activate reporter cells (C4D and K6 were also tested but are not shown). The two exceptions were WP15B and C4C. Of all the viruses tested, MAT 5'UTRs of these two viruses differ the most from Smith's MAT 5'UTR (Table 11). These results indicate that due to the recognition by host immune cells, MAT 5'UTR and m04/gp34 are under strong selective pressure, which resulted in the emergence of "escape" strains that avoid recognition via Ly49P, L and D2.

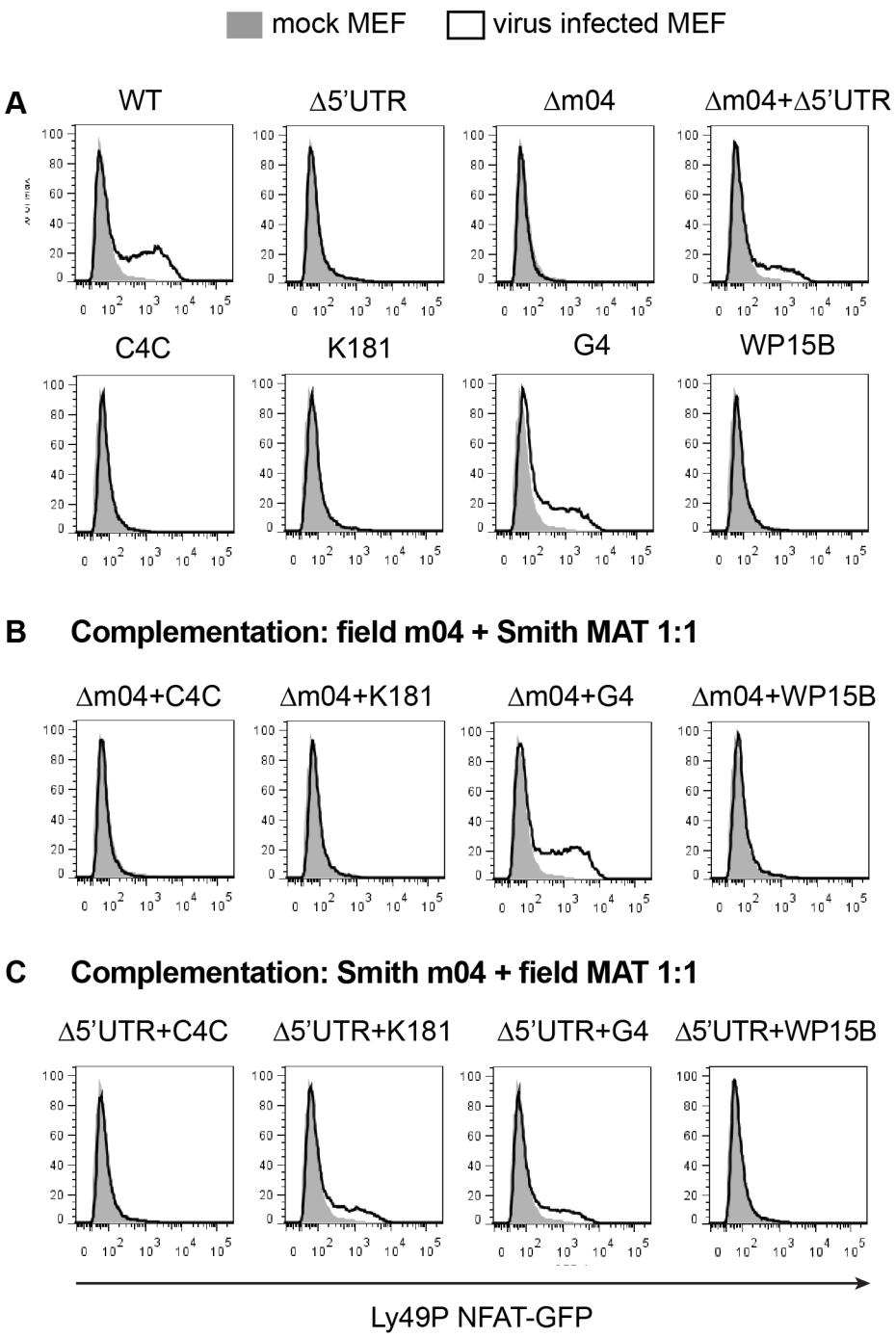


Figure 30. Analysis of the ability of field isolates to activate Ly49P reporter cells. Activation of reporter cells results in GFP expression, which was measured by flow cytometry. Gray filled histograms show reporter cells incubated with mock-infected MEF that did not result in activation of reporter cells and consequent expression of GFP. Empty histograms overlaid over gray histograms represent reporter cells incubated with infected MEF. MCMV-infected C3H MEF (1PFU/cell) was co-incubated with Ly49P and Ly49L (not shown) for 24 hours. In complementation experiments, 0.5 PFU/cell of Smith strain and 0.5 PFU/cell of field isolate were used giving a total of 1 PFU/cell. 1PFU/cell was used in all single infection experiments. (A) Analysis of the ability of field isolates to activate reporter cells. Only G4 could activate reporter cells in single virus infections. Δm04+Δ5'UTR coinfection was performed as control for complementation assays shown in B and C. (B)

Complementation assay to test the role of m04 variability in activating Ly49 receptor recognition of infected cells. MEF cells were co-infected with Smith Δm04 and field isolates in equal amounts (0.5 PFU/cell). m04 from most field isolates is not recognized by Ly49P and L (not shown), notable exception being G4. (C) Complementation assay to test the role of MAT 5'UTR variability in activating Ly49 receptor recognition of infected cells. MEF cells were co-infected with Smith Δ5'UTR and field isolates in equal amounts (0.5 PFU/cell). WP15B and C4C, field isolates with the most variable MAT 5'UTR, were not able to activate reporter cells. All other field isolates were.

4.4.7 WP15B and C4C have dominant negative phenotype

Under natural conditions, most wild mice are not infected by just one strain of virus but are co-infected by multiple MCMV strains [87]. Multiple co-infections have also been observed in humans and also included other, non-viral pathogens. Multiple strains co-infecting one host can interact in a positive (complementation) or negative (competition) way. McWhorter *et al.* [87] have shown fierce competition between strains within a host that differed in their ability to ligate activating Ly49H receptor. Most field isolates, with the exception of G4, are unable to ligate Ly49 P or L due to variability in both m04 and MAT 5'UTR. We therefore asked the question whether co-infection of field isolates with virus that can activate Ly49P or L will be beneficial or detrimental to co-infecting viruses with regard to Ly49 recognition. In addition to 1:1 ratio of co-infection, where cells were infected with equal amount of field and Smith MCMV, 4:1 (four times more Smith than field MCMV; 0.8 PFU/cell Smith + 0.2 PFU/cell field MCMV) and 1:4 (4 times more field than Smith MCMV; 0.2 PFU/cell Smith + 0.8 PFU/cell field MCMV) viral ratios were also used (Figure 31).

Co-infection of K181 with WT Smith MCMV resulted in successful activation of reporter cells. This was expected as MAT 5'UTR of K181 field isolate was previously shown to be able to activate reporter cells (Figure 30) due to high degree of similarity to Smith WT MCMV (Table 11). In contrast to that, cells co-infected with WP15B and WT Smith MCMV were unable to activate reporter cells even when the amount of Smith virus particles was four times higher than that of WP15B (4:1 ratio). Co-infection with Smith and C4C strains gave similar results, although at Smith to C4C ratio of 4:1 some activation of reporter cells could be seen.

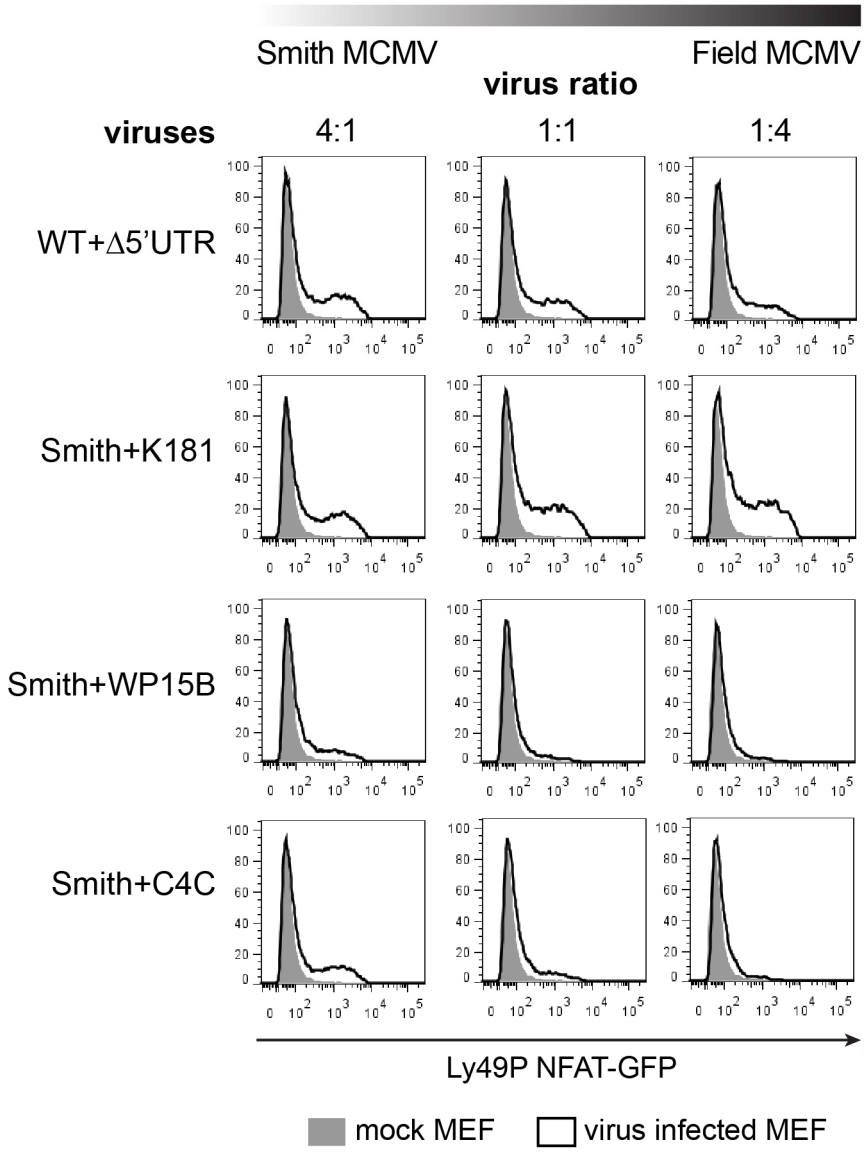


Figure 31. WP15B and C4C, but not K181 display dominant negative phenotype when co-infected with Smith MCMV. Activation of reporter cells results in GFP expression, which was measured by flow cytometry. Gray filled histograms show reporter cells incubated with mock-infected MEF that did not result in the activation of reporter cells and consequent expression of GFP. Empty histograms overlaid over gray histograms represent reporter cells incubated with infected MEF. Total amount of viral particles per cell in all complementation experiments was 1 PFU/cell. Co-infection of Smith MCMV (WT MCMV) with Δ5'UTR expectedly resulted in activation as did K181 whose MAT 5'UTR is highly similar to that of Smith strain. Interestingly, co-infection with Smith and WP15B resulted in the lack of activation even when cells were co-infected with four times more Smith WT MCMV, indicating that mutations in 5'UTR of WP15B and C4C exhibit dominant negative phenotype. Similar results were obtained with Ly49L reporter cells (not shown).

These findings underscore the importance of MAT 5'UTR in viral pathogenesis and immune evasion, and indicate that this region has been under strong selective pressure that resulted in

the acquisition of multiple mutations in different field isolates. Some mutations resulted in the generation of a dominant negative variant that could act beneficially to the virus possessing MAT 5'UTR which can be recognized by Ly49 P and L.

5. DISCUSSION

Human cytomegalovirus is an important human pathogen infecting a significant part of human population depending on socioeconomic status. Immunocompetent individuals effectively control the virus and show minor or no symptoms upon primary infection. The virus persists in the infected individual for life and may reactivate following immune suppression [91]. In immunocompromised patients (AIDS, transplant or cancer patients), primary infection or virus reactivation is associated with a variety of serious and often life-threatening conditions involving numerous organs and tissues. In transplant patients, HCMV is a primary cause of graft loss [12, 14]. Recently persistent HCMV infection has also been linked to atherosclerosis and some cancers [125, 126]. Congenital HCMV infection causes devastating disease with long-term neurological sequelae [19] and is in fact the main viral cause of congenital infections [15]. While therapies do exist, they are toxic and not suitable for long term application. All currently approved antivirals (ganciclovir, foscarnet, and cidofovir) target single gene – viral DNA polymerase, and unfortunately the effectiveness of these therapies is threatened by the appearance of resistant strains [79]. Several vaccines are under development and entering clinical trials; however, it seems so far that none have managed to raise long-lasting protective immunity in majority of patients [129].

Development of better vaccines and new therapies relies heavily on good understanding of the target pathogen. A major obstacle to HCMV research is its strict species specificity, which precludes the use of HCMV in animal models. Nevertheless, the use of murine CMV and other animal CMVs has significantly advanced our understanding of these viruses; many new viral genes and their functions, especially immune evasion genes, have been characterized thanks to the studies of MCMV and the development of tools that allowed us to generate mutant viruses missing specific genes. Generation and accuracy of mutant viruses relies heavily on genomic maps and, as was discussed in the introduction, current genomic maps mostly show coding features, missing regulatory non-coding transcripts. Although two annotations of MCMV currently exist (modified Rawlinson's and reference GenBank annotation), they are hardly definitive as major parts of the annotated ORFs have been predicted by *in silico* analyses with limited experimental confirmation and known non-coding RNAs are missing.

For these reasons, in the course of this PhD work, comprehensive analysis of MCMV transcriptome during lytic infection was performed using two approaches: classical cDNA

cloning and sequencing of viral transcripts, and next-generation sequencing of polyadenylated RNA (RNASeq).

The combination of these two approaches was used to construct a map of MCMV transcriptome and identified numerous differences between the detected transcripts and two currently used annotations. Although the results obtained by cDNA library and RNASeq diverged dramatically from the currently used annotations, these two approaches yielded remarkably complementary data despite different biases in each of these methods. Biases introduced by cDNA libraries include selection bias for isolating transcripts with long tracts of adenosines during cDNA library construction [147], while RNASeq results may be influenced by GC content, bias in the sites of fragmentation, primer affinity and transcript-end effects [117]. cDNA and RNASeq findings have been further corroborated by comparison with independent RNASeq study performed by Dölken group [83] as well as by Northern analysis and RT-PCR in certain complex regions.

In the course of this study, several novel transcripts have been identified that can be grouped into four categories: (1) transcripts overlapping more than 1 annotated gene, (2) novel spliced transcripts, (3) transcripts from areas previously designated as non-coding, and (4) antisense transcripts. Similar results and discrepancies between the currently used annotations and the detected transcripts were found in the study of HCMV transcriptome [147]. HCMV cDNA study, however, detected a significantly higher proportion of antisense transcription (>50% of all cDNA clones analyzed were in antisense orientation to the known or predicted genes) than was detected in MCMV cDNA analysis. Depending on the annotation used, in our cDNA analysis 0.09% (NC_004065) or 9% (GU305914.1) of all clones were in antisense orientation, while 27% (NC_004065) or 2% (GU305914.1) overlapped more than one gene in both sense and antisense orientation. cDNA analyses are only semi-quantitative and while recent strand specific analysis of HCMV transcriptome [45] did detect antisense transcription, antisense transcripts were transcribed at significantly lower levels than their sense counterparts. Our cDNA analysis, lack of antisense transcription in most Northern analyses and strand-specific RNASeq analysis performed by Dölken group [83] all indicate a similar low level antisense transcription in MCMV. Further analyses with longer sequencing reads and deeper coverage will likely resolve these inconsistencies in the future.

HCMV cDNA analysis [147] was among the first analyses that pointed out incredible complexity of herpesviral transcriptomes. Recent study utilizing ribosomal footprinting has

identified 751 translated proteins from HCMV genome [127]. This is several times more proteins than is predicted by current genomic maps. This discrepancy is, at least in part, a consequence of the polycistronic nature of HCMV transcripts, which appear to code for many more ORFs than previously predicted (internal in-frame or out-of-frame ORFs, uORFs) as well as ORFs coming from antisense transcripts or dedicated short transcripts. Our analysis demonstrated that the MCMV transcriptome is similarly complex: several regions where multiple 3' co-terminal transcripts were expressed in different temporal phases have been detected in this analysis. Transcripts with alternative 5' ends have a potential to code for truncated protein forms or even completely new proteins, as described for HCMV. In addition, our analysis has identified several regions with transcripts overlapping more than one annotated genes which also have the potential to encode multiple proteins. Polycistronic transcripts have previously been described for certain MCMV transcripts [105], while Stern-Ginossar study [127] showed that polycistronic transcripts are a widespread feature of HCMV transcriptome. All of these findings suggest that the size and complexity of the MCMV proteome, like the MCMV transcriptome, is currently underestimated.

Another feature of HCMV that seems to be shared by MCMV is abundant transcription of non-coding RNAs. Analyses of HCMV transcriptome show that over half of all transcribed polyadenylated transcripts are non-coding [45, 147]. Both our RNASeq and cDNA analyses show intense transcription in previously described stable MCMV introns and in intergenic regions, consistent with abundant ncRNAs reported for HCMV and MCMV [66].

Forty-two spliced transcripts were cloned in the course of this study, 22 of which were novel spliced transcripts. Of these, 3 have been further confirmed by RT-PCR and Northern analysis, and the existence of one was disproved following further analysis. While additional analyses are needed to confirm or disprove the remaining 18, this finding nevertheless underscores underestimated complexity of MCMV transcriptional products and is in line with recent findings of widespread splicing in HCMV transcriptome [45].

The complexity of virus transcriptome has a profound implication for future CMV studies, especially studies utilizing deletion mutants. The functions of many MCMV genes have been elucidated by using deletion mutants [49]. However, in a transcriptionally complex region of the genome any deletion will likely impact multiple transcripts and possibly multiple proteins resulting in complex phenotypes. In the future, transcriptomic maps will be needed in addition to genomic maps. Furthermore, the discrepancies between the currently used genomic maps

and the observed transcripts underscore the need for better annotation of MCMV. Genomic maps are not only used in mutant virus generation but also in various quantitative analyses of gene expression (microarray and RNASeq). Although RNASeq has now successfully been applied to *ab initio* genome reconstruction of eukaryotic transcriptomes [50], condensed microbial genomes are still too complex for currently available bioinformatics tools. Until better tools are available, RNASeq analyses must rely on the comparison to existing gene annotation and other experimental methods for gene structure prediction and quantization. While definitive transcriptomic map of MCMV is still pending, combination of cDNA analysis and RNASeq facilitated reconstruction of several well expressed transcripts and thus the results presented here represent an important first step in the re-annotation of the MCMV genome and underscore the utility of transcriptome studies in validating and refining genome annotations.

Quantitative analysis of RNASeq data revealed that transcription of individual viral transcripts varies by several orders of magnitude (Figure 11 and Supplemental table 2) and identified a striking abundance of single, novel spliced transcript MAT. Furthermore, most other top expressed viral genes following MAT are novel transcripts with unknown functions. These results highlight fundamental gaps in our understanding of basic MCMV biology.

Further analyses of MAT transcript revealed that this 1.7 kb long transcript encodes at least one protein of approximately 17 kDa. This finding, along with recent reports of MAT serving as a sponge for cellular micro-RNA miR-27b [72, 84], make MAT the first viral transcript that has both coding and non-coding functions. Unlike the transcript that is highly abundant and can be found in all temporal cDNA sub-libraries, MAT protein starts to accumulate only late in the infection in cytoplasm. Such a poor translation of MAT protein is a consequence of MAT's long 5'UTR; in mutants where 5'UTR has been deleted, MAT protein becomes detectable already at 16 hours PI and at significantly higher levels than in wild-type virus.

Long 5'UTRs that contain numerous start codons and possible uORFs are often found in transcripts encoding regulatory proteins like proto-oncogenes, growth factors, their receptors, and homeodomain proteins [101]. Analysis of conserved domains using ELM or CD search in PubMed did not identify any domains that could indicate its function (data not shown); however, MAT protein nucleotide sequence is well conserved in all published MCMV strains, and a protein could be detected by Western blot in all field isolates. This conservation, as well as transcript abundance, indicates that it must play a role in the infection. Interestingly, in

addition to its role as MAT protein translation regulator, 5'UTR was also found to be a necessary viral factor for the NK cell recognition of MCMV-infected cells via activating Ly49 receptors. Unlike the rest of the MAT transcript, which is well conserved in all sequenced MCMV strains, 5'UTR is highly variable. This variability results in the inability of activating Ly49 receptors to recognize 5'UTRs of most field isolates. Even more interesting is the fact that co-infection of WT virus with either of WP15B of C4C field isolates that contain MAT 5'UTR which do not engage in activating Ly49 receptors results in dominant negative phenotype. Coinfections with multiple strains of viruses are common among wild mice and in humans. In a recent work, McWhorter *et al.* have shown fierce competition within host between different MCMV strains that differed in their ability to bind activating Ly49H receptor [87]. Based on reporter cell assay results in co-infection, viruses can also cooperate, not just compete.

RNASeq analysis allowed us to analyze transcriptomic response of host cells to infection. There were 10748 genes differentially regulated in response to infection. Number of mouse genes is estimated to 33,207 in mouse genome build used in this work (mm9) [43] making 31% of mouse genes differentially regulated as a consequence of infection. Many of the top upregulated and induced genes and gene networks were associated with immune responses to infection, including interferon and interferon-inducible genes such as *phyin1*, a potential activator of p53 [23], the inflammasone regulator *Gpb5* [121] and *Rsad2* (aka viperin), also known to be induced by HCMV [118].

MCMV encodes virus-derived chemokine homolog encoded by *m131/m129* genes [80, 98] and one chemokine receptor homolog, M33 [20]. Inflammatory chemokine ligand genes as well as chemokine receptors are highly upregulated during infection, suggesting a remarkably complex interplay between MCMV-derived and host-derived chemokine signaling during infection. Induction of inflammatory gene networks by MCMV also lends credence to the hypothesis that inflammatory responses link CMV infection to chronic diseases, such as chronic allograft rejection, cardiovascular disease, and cancer [14, 125, 126]. One of the top diseases associated with DE genes in infected fibroblasts identified by IPA was multiple sclerosis. Balb/c mice are resistant to MOG (myelin oligodendrocyte glycoprotein)-induced experimental autoimmune encephalomyelitis (EAE). Unpublished data now indicate that after MCMV infection, this resistance is lost (Mijodrag Lukić, personal communication).

Numerous transcription factors are also induced or upregulated by infection including insulinoma-associated 1 (*Insm1*). Recently, *Insm1* has been found to be strongly upregulated by HSV-1 infection and shown to promote HSV gene expression, probably by binding the HSV-infected cell protein (ICP)0 promoter [56]. This raises the intriguing possibility that INSM1 plays a similar role in promoting virus gene expression during MCMV infection. Another transcription factor induced at the transcript and protein level is engrailed-2 (En2). This transcription factor is key to patterning cerebellar foliation during development [25]. We previously described a profound dysregulation of cerebellar development in brains of neonatal mice infected with MCMV [62], suggesting a possible physiological link to regulation of this gene. GABA receptor, *Gabrq*, was also among top induced genes. Glutamate receptor signaling was also identified as significantly impacted canonical pathway in our dataset. In the developing brain GABA and glutamate receptors influence neuronal proliferation, migration, differentiation or survival processes [78]. Whether and how these observations relate to our previous findings that MCMV infection of neonates results in decreased granular neuron proliferation and migration [62] are important areas for future study and may impact our understanding of neurological damage and sequelae associated with HCMV in congenitally infected infants.

Many top regulated genes, especially downregulated and repressed ones, are associated with functions whose roles in infection are obscure, including many genes of unknown function. Many downregulated or repressed genes are cell surface molecules, or host lincRNAs, antisense RNAs or small nucleolar RNAs. Regulation of lincRNAs has recently been observed during infection with severe acute respiratory syndrome coronavirus (SARS-CoV) and influenza virus, and has been suggested to impact host defenses and innate immunity [100]. Further studies to identify the functions of these downregulated and repressed genes and noncoding RNAs during MCMV infection may well provide novel insights into the virus-host molecular interface as well as possible therapeutic targets.

This analysis also revealed immunological disease, cardiovascular disease, genetic disorders and skeletal and muscular disorders as top bio-functions connected with genes altered by MCMV infection. While MCMV involvement in cardiovascular disease is a subject of intensive research, potential involvement in skeletal and muscular disorders is not well documented but may be relevant to the novel observation that MCMV infection of mice with a heterozygous *Trp53* mutation develops rhabdomyosarcomas at high frequency [102].

A primary caveat of RNASeq analysis is determining whether changes in gene transcript levels are also reflected at the protein level. This is particularly important as herpesviruses can control protein accumulation at the post-transcriptional, translational, and post-translational levels [28, 124, 131]. Many of differentially regulated genes detected in this study have previously been associated with MCMV infection. To test how well transcriptomic data correlate with protein levels, differentially regulated genes whose relevance to MCMV infection was not previously shown were selected. For all differentially regulated genes tested: notch ligands Delta 1 and Jagged 2, homeobox containing transcriptional factor Engrailed 2 and E3 ubiquitin-protein ligase Trim71 changes in protein levels correlated with changes at transcript levels.

Notch signaling is a highly conserved signaling pathway that plays important roles in development, including neurogenesis and differentiation of immune cell subsets [34]. Jagged 2 is also upregulated by alphaherpesviruses, HSV-1 and Pseudorabies viruses [107]. KSHV and EBV also exploit the notch signaling pathway to facilitate aspects of their life cycle [52] and notch signaling is proposed to influence HSV-2-induced interferon responses [130]. We show for the first time that a betaherpesvirus, MCMV, also influences notch signaling. Dysregulation of Jagged2 as a consequence of MCMV infection is highly interesting since Jagged2 plays a role in important processes affected by CMV including inner ear development [97, 150], generation of motor neurons [104] and differentiation of immune cell subsets [9, 63].

To summarize, this study has refined the understanding of MCMV gene expression and opened numerous new areas of research. Transcriptomic analysis of MCMV indicated that there are numerous gaps in our knowledge of MCMV genes and their viral products, and showed an urgent need for better genomic maps. Analysis of host transcriptome, while confirming many previous findings, also identified numerous virus and host genes of unknown function that are differentially regulated during infection as well as gene networks whose relevance to the infection is still unknown.

6. CONCLUSIONS

The MCMV transcriptome diverges substantially from that predicted by two currently used annotations indicating an urgent need for newer genomic map of CMV based on experimentally detected transcripts. This work presents an important first step towards this goal. Although almost all of the genome of MCMV is transcribed, levels of transcription of different viral genes vary by several orders of magnitude. The majority of the most abundantly transcribed viral genes are of unknown function, and many are new transcripts detected in this study. The most abundant transcript (MAT) identified in this study has at least 3 functions: (1) its 5'UTR is involved in NK cell recognition of infected cells via activating Ly49 receptors, (2) it encodes at least 1 protein, and (3) it contains binding site for cellular micro-RNA miR27 in its 3'UTR. MAT is the first viral transcript so far described that has both coding and non-coding functions.

Twenty-two novel spliced transcripts have been detected, indicating that splicing is more widespread than previously thought. In contrast, antisense transcription is present in MCMV transcriptome but at much lower levels than anticipated based on previous studies of HCMV transcriptome.

Infection of primary fibroblasts with CMV results in differential expression of nearly a third of host genes. While many detected deregulated genes were those whose relevance to the infection was already known and verified, a significant number were unexpected and clustered in biological pathways and gene networks yet unconnected to CMV infection. Such analysis has the potential to identify new conditions and diseases influenced by CMV as well as point out potential targets for treatment.

7. REFERENCES

[1] Abi-Rached, L. and P. Parham, *Natural selection drives recurrent formation of activating killer cell immunoglobulin-like receptor and Ly49 from inhibitory homologues.* J Exp Med, 2005. 201(8): p. 1319-32.

[2] Anders, D.G., J.A. Kerry, and G.S. Pari, *DNA synthesis and late viral gene expression.* 2007.

[3] Andrei, G., E. De Clercq, and R. Snoeck, *Drug targets in cytomegalovirus infection.* Infect Disord Drug Targets, 2009. 9(2): p. 201-22.

[4] Ansorge, W.J., *Next-generation DNA sequencing techniques.* N Biotechnol, 2009. 25(4): p. 195-203.

[5] Arapovic, J., et al., *Promiscuity of MCMV immunoevasin of NKG2D: m138/fcr-1 down-modulates RAE-1ε in addition to MULT-1 and H60.* Mol Immunol, 2009. In press.

[6] Arase, H., et al., *Direct recognition of cytomegalovirus by activating and inhibitory NK cell receptors.* Science, 2002. 296(5571): p. 1323-6.

[7] Babic, M., et al., *Cytomegalovirus immunoevasin reveals the physiological role of "missing self" recognition in natural killer cell dependent virus control in vivo.* J Exp Med, 2010. 207(12): p. 2663-73.

[8] Bankier, A.T., et al., *The DNA sequence of the human cytomegalovirus genome.* DNA Seq, 1991. 2(1): p. 1-12.

[9] Beck, R.C., et al., *The Notch ligands Jagged2, Delta1, and Delta4 induce differentiation and expansion of functional human NK cells from CD34+ cord blood hematopoietic progenitor cells.* Biol Blood Marrow Transplant, 2009. 15(9): p. 1026-37.

[10] Biegalke, B.J., et al., *Characterization of the human cytomegalovirus UL34 gene.* J Virol, 2004. 78(17): p. 9579-83.

[11] Black, D.L., *Protein diversity from alternative splicing: a challenge for bioinformatics and post-genome biology.* Cell, 2000. 103(3): p. 367-70.

[12] Boeckh, M. and A.P. Geballe, *Cytomegalovirus: pathogen, paradigm, and puzzle.* J Clin Invest, 2011. 121(5): p. 1673-80.

[13] Britt, B., *Maturation and egress.* 2007.

[14] Britt, W., *Manifestations of human cytomegalovirus infection: proposed mechanisms of acute and chronic disease.* Curr Top Microbiol Immunol, 2008. 325: p. 417-70.

[15] Britt, W., *Cytomegalovirus*, in *Infectious Diseases of the Fetus and Newborn Infant*, J.S. Remington, et al., Editors. 2010, Elsevier: Philadelphia. p. 706-756.

[16] Brocchieri, L., et al., *Predicting coding potential from genome sequence: application to betaherpesviruses infecting rats and mice.* J Virol, 2005. 79(12): p. 7570-96.

[17] Brune, w., H. Hengel, and U. Koszinowski, *A mouse model for cytomegalovirus infection.*, in *Current protocols in immunology*. 1999., John Wiley & Sons: New York. p. 19.17.11-19.17.13.

[18] Buck, A.H., et al., *Post-transcriptional regulation of miR-27 in murine cytomegalovirus infection.* RNA, 2010. 16(2): p. 307-15.

[19] Cannon, M.J., *Congenital cytomegalovirus (CMV) epidemiology and awareness.* J Clin Virol, 2009. 46 Suppl 4: p. S6-10.

[20] Case, R., et al., *Functional analysis of the murine cytomegalovirus chemokine receptor homologue M33: Ablation of constitutive signaling is associated with an attenuated phenotype in vivo.* Journal of Virology, 2008. 82(4): p. 1884-1898.

[21] Chandriani, S., Y. Xu, and D. Ganem, *The lytic transcriptome of Kaposi's sarcoma-associated herpesvirus reveals extensive transcription of noncoding regions, including regions antisense to important genes.* J Virol, 2010. 84(16): p. 7934-42.

[22] Chee, M.S., et al., *Analysis of the protein-coding content of the sequence of human cytomegalovirus strain AD169.* Curr Top Microbiol Immunol, 1990. 154: p. 125-69.

[23] Chen, Z., et al., *Stabilization of p53 in human cytomegalovirus-initiated cells is associated with sequestration of HDM2 and decreased p53 ubiquitination.* J Biol Chem, 2007. 282(40): p. 29284-95.

[24] Cheng, T.P., et al., *Stability of murine cytomegalovirus genome after in vitro and in vivo passage.* J Virol, 2010. 84(5): p. 2623-8.

[25] Cheng, Y., et al., *The Engrailed homeobox genes determine the different foliation patterns in the vermis and hemispheres of the mammalian cerebellum.* Development, 2010. 137(3): p. 519-29.

[26] Ciocco-Schmitt, G.M., et al., *Identification and characterization of novel murine cytomegalovirus M112-113 (e1) gene products.* Virology, 2002. 294(1): p. 199-208.

[27] Clark, M.B., et al., *The reality of pervasive transcription.* PLoS Biol, 2011. 9(7): p. e1000625; discussion e1001102.

[28] Clyde, K. and B.A. Glaunsinger, *Getting the message direct manipulation of host mRNA accumulation during gammaherpesvirus lytic infection.* Adv Virus Res, 2010. 78: p. 1-42.

[29] Cocquet, J., et al., *Reverse transcriptase template switching and false alternative transcripts.* Genomics, 2006. 88(1): p. 127-31.

[30] Compton, T. and A. Feire, *Early events in human cytomegalovirus infection.* 2007.

[31] Cook, C.H. and J. Trgovcich, *Cytomegalovirus reactivation in critically ill immunocompetent hosts: a decade of progress and remaining challenges.* Antiviral Res, 2011. 90(3): p. 151-9.

[32] Corbett, A.J., et al., *Functional consequences of natural sequence variation of murine cytomegalovirus m157 for Ly49 receptor specificity and NK cell activation.* J Immunol, 2011. 186(3): p. 1713-22.

[33] Costa, F.F., *Non-coding RNAs: lost in translation?* Gene, 2007. 386(1-2): p. 1-10.

[34] Dallman, M.J., et al., *Notch: control of lymphocyte differentiation in the periphery.* Curr Opin Immunol, 2005. 17(3): p. 259-66.

[35] Davis-Poynter, N.J., et al., *Identification and characterization of a G protein-coupled receptor homolog encoded by murine cytomegalovirus.* J Virol, 1997. 71(2): p. 1521-9.

[36] Davison, A.J., *Comparative analysis of the genomes.* 2007.

[37] Davison, A.J., *Overview of classification.* 2007.

[38] Davison, A.J. and D. Bhella, *Comparative genome and virion structure.* 2007.

[39] Del Val, M., et al., *Efficient processing of an antigenic sequence for presentation by MHC class I molecules depends on its neighboring residues in the protein.* Cell, 1991. 66(6): p. 1145-53.

[40] Desrosiers, M.P., et al., *Epistasis between mouse Klra and major histocompatibility complex class I loci is associated with a new mechanism of natural killer cell-mediated innate resistance to cytomegalovirus infection.* Nat Genet, 2005. 37(6): p. 593-9.

[41] Eden, E., et al., *Discovering motifs in ranked lists of DNA sequences.* PLoS Comput Biol, 2007. 3(3): p. e39.

[42] Eden, E., et al., *GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists.* BMC Bioinformatics, 2009. 10: p. 48.

[43] Eppig, J.T., et al., *The Mouse Genome Database (MGD): from genes to mice--a community resource for mouse biology.* Nucleic Acids Res, 2005. 33(Database issue): p. D471-5.

[44] French, A.R., et al., *Escape of mutant double-stranded DNA virus from innate immune control.* Immunity, 2004. 20(6): p. 747-56.

[45] Gatherer, D., et al., *High-resolution human cytomegalovirus transcriptome.* Proc Natl Acad Sci U S A, 2011. 108(49): p. 19755-60.

[46] Gibbons, J.G., et al., *Benchmarking next-generation transcriptome sequencing for functional and evolutionary genomics.* Mol Biol Evol, 2009. 26(12): p. 2731-44.

[47] Graveley, B.R., *Alternative splicing: increasing diversity in the proteomic world.* Trends Genet, 2001. 17(2): p. 100-7.

[48] Gresham, D., M.J. Dunham, and D. Botstein, *Comparing whole genomes using DNA microarrays.* Nat Rev Genet, 2008. 9(4): p. 291-302.

[49] Gutermann, A., et al., *Strategies for the identification and analysis of viral immune-evasive genes--cytomegalovirus as an example.* Curr Top Microbiol Immunol, 2002. 269: p. 1-22.

[50] Guttman, M., et al., *Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs.* Nat Biotechnol, 2010. 28(5): p. 503-10.

[51] Hasan, M., et al., *Selective down-regulation of the NKG2D ligand H60 by mouse cytomegalovirus m155 glycoprotein.* J Virol, 2005. 79(5): p. 2920-30.

[52] Hayward, S.D., J. Liu, and M. Fujimuro, *Notch and Wnt signaling: mimicry and manipulation by gamma herpesviruses.* Sci STKE, 2006. 2006(335): p. re4.

[53] Hirsch, P.R., *Detection of microbial DNA sequences by colony hybridization*, in *Molecular Microbial Ecology Manual*, G.A. Kowalchuk, et al., Editors. 2004, Kluwer Academic. p. 345-356.

[54] Hughes, T.R. and D.D. Shoemaker, *DNA microarrays for expression profiling.* Curr Opin Chem Biol, 2001. 5(1): p. 21-5.

[55] Jonjic, S., et al., *Dissection of the antiviral NK cell response by MCMV mutants.* Methods Mol Biol, 2008. 415: p. 127-49.

[56] Kamakura, M., et al., *Herpes simplex virus induces the marked up-regulation of the zinc finger transcriptional factor INSM1, which modulates the expression and localization of the immediate early protein ICP0.* Virol J, 2011. 8: p. 257.

[57] Karre, K., *Natural killer cell recognition of missing self.* Nat Immunol, 2008. 9(5): p. 477-80.

[58] Kattenhorn, L.M., et al., *Identification of proteins associated with murine cytomegalovirus virions.* J Virol, 2004. 78(20): p. 11187-97.

[59] Kavanagh, D.G., U.H. Koszinowski, and A.B. Hill, *The murine cytomegalovirus immune evasion protein m4/gp34 forms biochemically distinct complexes with class I MHC at the cell surface and in a pre-Golgi compartment.* J Immunol, 2001. 167(7): p. 3894-902.

[60] Kielczewska, A., et al., *Ly49P recognition of cytomegalovirus-infected cells expressing H2-Dk and CMV-encoded m04 correlates with the NK cell antiviral response.* J Exp Med, 2009. 206(3): p. 515-23.

[61] Kleijnen, M.F., et al., *A mouse cytomegalovirus glycoprotein, gp34, forms a complex with folded class I MHC molecules in the ER which is not retained but is transported to the cell surface.* EMBO J, 1997. 16(4): p. 685-94.

[62] Koontz, T., et al., *Altered development of the brain after focal herpesvirus infection of the central nervous system.* J Exp Med, 2008. 205(2): p. 423-35.

[63] Koyanagi, A., C. Sekine, and H. Yagita, *Expression of Notch receptors and ligands on immature and mature T cells.* Biochem Biophys Res Commun, 2012. 418(4): p. 799-805.

[64] Krmpotic, A., et al., *MCMV glycoprotein gp40 confers virus resistance to CD8+ T cells and NK cells in vivo.* Nat Immunol, 2002. 3(6): p. 529-35.

[65] Krmpotic, A., 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): p. 211-20.

[66] Kulesza, C.A. and T. Shenk, *Murine cytomegalovirus encodes a stable intron that facilitates persistent replication in the mouse.* Proc Natl Acad Sci U S A, 2006. 103(48): p. 18302-7.

[67] Lacaze, P., et al., *Temporal profiling of the coding and noncoding murine cytomegalovirus transcriptomes.* J Virol, 2011. 85(12): p. 6065-76.

[68] Lagenaur, L.A., et al., *Structure and function of the murine cytomegalovirus sgg1 gene: a determinant of viral growth in salivary gland acinar cells.* J Virol, 1994. 68(12): p. 7717-27.

[69] Leach, F.S. and E.S. Mocarski, *Regulation of cytomegalovirus late-gene expression: differential use of three start sites in the transcriptional activation of ICP36 gene expression.* J Virol, 1989. 63(4): p. 1783-91.

[70] Leatham, M.P., P.R. Witte, and M.F. Stinski, *Alternate promoter selection within a human cytomegalovirus immediate-early and early transcription unit (UL119-115) defines true late*

transcripts containing open reading frames for putative viral glycoproteins. J Virol, 1991. 65(11): p. 6144-53.

[71] Lenac, T., et al., *The herpesviral Fc receptor fcr-1 down-regulates the NKG2D ligands MULT-1 and H60.* J Exp Med, 2006. 203(8): p. 1843-50.

[72] Libri, V., et al., *Murine cytomegalovirus encodes a miR-27 inhibitor disguised as a target.* Proc Natl Acad Sci U S A, 2012. 109(1): p. 279-84.

[73] Lindberg, J. and J. Lundeberg, *The plasticity of the mammalian transcriptome.* Genomics, 2010. 95(1): p. 1-6.

[74] Lisnic, V.J., A. Krmpotic, and S. Jonjic, *Modulation of natural killer cell activity by viruses.* Curr Opin Microbiol, 2010. 13(4): p. 530-9.

[75] Liu, F. and Z. Hong Zhou, *Comparative virion structures of human herpesviruses.* 2007.

[76] Lodoen, M., et al., *NKG2D-mediated natural killer cell protection against cytomegalovirus is impaired by viral gp40 modulation of retinoic acid early inducible 1 gene molecules.* J Exp Med, 2003. 197(10): p. 1245-53.

[77] Lodoen, M.B., et al., *The cytomegalovirus m155 gene product subverts natural killer cell antiviral protection by disruption of H60-NKG2D interactions.* J Exp Med, 2004. 200(8): p. 1075-81.

[78] Lujan, R., R. Shigemoto, and G. Lopez-Bendito, *Glutamate and GABA receptor signalling in the developing brain.* Neuroscience, 2005. 130(3): p. 567-80.

[79] Lurain, N.S. and S.W. Chou, *Antiviral Drug Resistance of Human Cytomegalovirus.* Clinical Microbiology Reviews, 2010. 23(4): p. 689-712.

[80] MacDonald, M.R., et al., *Spliced mRNA encoding the murine cytomegalovirus chemokine homolog predicts a beta chemokine of novel structure.* J Virol, 1999. 73(5): p. 3682-91.

[81] Makrigiannis, A.P., et al., *Class I MHC-binding characteristics of the 129/J Ly49 repertoire.* J Immunol, 2001. 166(8): p. 5034-43.

[82] Maniatis, T. and B. Tasic, *Alternative pre-mRNA splicing and proteome expansion in metazoans.* Nature, 2002. 418(6894): p. 236-43.

[83] Marcinowski, L., et al., *Real-time transcriptional profiling of cellular and viral gene expression during lytic cytomegalovirus infection.* PLoS Pathog, 2012. 8(9): p. e1002908.

[84] Marcinowski, L., et al., *Degradation of cellular mir-27 by a novel, highly abundant viral transcript is important for efficient virus replication in vivo.* PLoS Pathog, 2012. 8(2): p. e1002510.

[85] Marioni, J.C., et al., *RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays.* Genome Res, 2008. 18(9): p. 1509-17.

[86] Marshall, E.E. and A.P. Geballe, *Multifaceted evasion of the interferon response by cytomegalovirus.* J Interferon Cytokine Res, 2009. 29(9): p. 609-19.

[87] McWhorter, A.R., et al., *Natural killer cell dependent within-host competition arises during multiple MCMV infection: consequences for viral transmission and evolution.* PLoS Pathog, 2013. 9(1): p. e1003111.

[88] Messerle, M., et al., *Cloning and mutagenesis of a herpesvirus genome as an infectious bacterial artificial chromosome.* Proc Natl Acad Sci U S A, 1997. 94(26): p. 14759-63.

[89] Mettenleiter, T.C., B.G. Klupp, and H. Granzow, *Herpesvirus assembly: a tale of two membranes.* Current Opinion in Microbiology, 2006. 9(4): p. 423-429.

[90] Mettenleiter, T.C., et al., *The way out: what we know and do not know about herpesvirus nuclear egress.* Cellular Microbiology, 2013. 15(2): p. 170-178.

[91] Mocarski, E.S. and C.T. Courcelle, *Cytomegaloviruses and their replication*, in *Fields Virology*, D.M. Knipe and P.M. Howley, Editors. 2001, Lippincot Williams & Wilkins: Philadelphia. p. 2629-74.

[92] Mocarski Jr, E., *Betaherpes viral genes and their functions.* 2007.

[93] Mocarski Jr, E.S., *Comparative analysis of herpesvirus-common proteins.* 2007.

[94] Morgulis, A., et al., *Database indexing for production MegaBLAST searches.* Bioinformatics, 2008. 24(16): p. 1757-64.

[95] Morris, D.R. and A.P. Geballe, *Upstream open reading frames as regulators of mRNA translation.* Mol Cell Biol, 2000. 20(23): p. 8635-42.

[96] Mortazavi, A., et al., *Mapping and quantifying mammalian transcriptomes by RNA-Seq.* Nat Methods, 2008. 5(7): p. 621-8.

[97] Murata, J., K. Ikeda, and H. Okano, *Notch signaling and the developing inner ear.* Adv Exp Med Biol, 2012. 727: p. 161-73.

[98] Noda, S., et al., *Cytomegalovirus MCK-2 controls mobilization and recruitment of myeloid progenitor cells to facilitate dissemination.* Blood, 2006. 107(1): p. 30-8.

[99] Paterson, Y. and H.M. Cooper, *Production of antibodies*, in *Current protocols in immunology*. 1995, John Wiley & Sons, Inc.

[100] Peng, X., et al., *Unique signatures of long noncoding RNA expression in response to virus infection and altered innate immune signaling.* MBio, 2010. 1(5).

[101] Pickering, B.M. and A.E. Willis, *The implications of structured 5' untranslated regions on translation and disease.* Semin Cell Dev Biol, 2005. 16(1): p. 39-47.

[102] Price, R.L., et al., *Cytomegalovirus infection leads to pleomorphic rhabdomyosarcomas in Trp53+/- mice.* Cancer Res, 2012. 72(22): p. 5669-74.

[103] Pyzik, M., et al., *Distinct MHC class I-dependent NK cell-activating receptors control cytomegalovirus infection in different mouse strains.* J Exp Med, 2011. 208(5): p. 1105-17.

[104] Rabadan, M.A., et al., *Jagged2 controls the generation of motor neuron and oligodendrocyte progenitors in the ventral spinal cord.* Cell Death and Differentiation, 2012. 19(2): p. 209-219.

[105] Rapp, M., et al., *Expression of the murine cytomegalovirus glycoprotein H by recombinant vaccinia virus.* J Gen Virol, 1994. 75 (Pt 1): p. 183-8.

[106] Rawlinson, W.D., H.E. Farrell, and B.G. Barrell, *Analysis of the complete DNA sequence of murine cytomegalovirus.* J Virol, 1996. 70(12): p. 8833-49.

[107] Ray, N. and L.W. Enquist, *Transcriptional response of a common permissive cell type to infection by two diverse alphaherpesviruses.* J Virol, 2004. 78(7): p. 3489-501.

[108] Redwood, A.J., et al., *Use of a murine cytomegalovirus K181-derived bacterial artificial chromosome as a vaccine vector for immunocontraception.* J Virol, 2005. 79(5): p. 2998-3008.

[109] Reusch, U., et al., *A cytomegalovirus glycoprotein re-routes MHC class I complexes to lysosomes for degradation.* EMBO J, 1999. 18(4): p. 1081-91.

[110] Robinson, J.T., et al., *Integrative genomics viewer.* Nat Biotechnol, 2011. 29(1): p. 24-6.

[111] Roizman, B. and P.E. Pellett, *The Family Herpesviridae: A Brief Introduction*, in *Fields virology*, D.M. Knipe, et al., Editors. 2007, Lippincott Williams & Wilkins. p. 2480-2499.

[112] Ross, D.S., et al., *The epidemiology and prevention of congenital cytomegalovirus infection and disease: activities of the Centers for Disease Control and Prevention Workgroup.* J Womens Health (Larchmt), 2006. 15(3): p. 224-9.

[113] Sambrook, J. and D.W. Russel, *Molecular Cloning: A Laboratory Manual*. 3 ed. 2001: Cold Spring Harbor Laboratory Press.

[114] Scalzo, A.A., et al., *The murine cytomegalovirus M73.5 gene, a member of a 3′ co-terminal alternatively spliced gene family, encodes the gp24 virion glycoprotein.* Virology, 2004. 329(2): p. 234-50.

[115] Schena, M., et al., *Quantitative monitoring of gene expression patterns with a complementary DNA microarray.* Science, 1995. 270(5235): p. 467-70.

[116] Schwanhausser, B., et al., *Global quantification of mammalian gene expression control.* Nature, 2011. 473(7347): p. 337-42.

[117] Sendler, E., G.D. Johnson, and S.A. Krawetz, *Local and global factors affecting RNA sequencing analysis.* Anal Biochem, 2011. 419(2): p. 317-22.

[118] Seo, J.Y., et al., *Human cytomegalovirus directly induces the antiviral protein viperin to enhance infectivity.* Science, 2011. 332(6033): p. 1093-7.

[119] Shabalina, S.A. and N.A. Spiridonov, *The mammalian transcriptome and the function of non-coding DNA sequences.* Genome Biol, 2004. 5(4): p. 105.

[120] Shendure, J., *The beginning of the end for microarrays?* Nat Methods, 2008. 5(7): p. 585-7.

[121] Shenoy, A.R., et al., *GBP5 promotes NLRP3 inflammasome assembly and immunity in mammals.* Science, 2012. 336(6080): p. 481-5.

[122] Smith, H.R., et al., *Recognition of a virus-encoded ligand by a natural killer cell activation receptor.* Proc Natl Acad Sci U S A, 2002. 99(13): p. 8826-31.

[123] Smith, L.M., et al., *Laboratory strains of murine cytomegalovirus are genetically similar to but phenotypically distinct from wild strains of virus.* J Virol, 2008. 82(13): p. 6689-96.

[124] Smith, R.W., S.V. Graham, and N.K. Gray, *Regulation of translation initiation by herpesviruses.* Biochem Soc Trans, 2008. 36(Pt 4): p. 701-7.

[125] Soderberg-Naucler, C., *Does cytomegalovirus play a causative role in the development of various inflammatory diseases and cancer?* J Intern Med, 2006. 259(3): p. 219-46.

[126] Stassen, F.R., T. Vainas, and C.A. Bruggeman, *Infection and atherosclerosis. An alternative view on an outdated hypothesis.* Pharmacol Rep, 2008. 60(1): p. 85-92.

[127] Stern-Ginossar, N., et al., *Decoding human cytomegalovirus.* Science, 2012. 338(6110): p. 1088-93.

[128] Stinski, M.F. and J.L. Meier, *Immediate-early viral gene regulation and function.* 2007.

[129] Sung, H. and M.R. Schleiss, *Update on the current status of cytomegalovirus vaccines.* Expert Rev Vaccines, 2010. 9(11): p. 1303-14.

[130] Svensson, A., et al., *Inhibition of gamma-secretase cleavage in the notch signaling pathway blocks HSV-2-induced type I and type II interferon production.* Viral Immunol, 2010. 23(6): p. 647-51.

[131] Taddeo, B., W. Zhang, and B. Roizman, *The virion-packaged endoribonuclease of herpes simplex virus 1 cleaves mRNA in polyribosomes.* Proc Natl Acad Sci U S A, 2009. 106(29): p. 12139-44.

[132] Tandon, R. and E.S. Mocarski, *Viral and host control of cytomegalovirus maturation.* Trends Microbiol, 2012. 20(8): p. 392-401.

[133] Tang, Q., E.A. Murphy, and G.G. Maul, *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): p. 6873-82.

[134] Voigt, V., et al., *Murine cytomegalovirus m157 mutation and variation leads to immune evasion of natural killer cells.* Proc Natl Acad Sci U S A, 2003. 100(23): p. 13483-8.

[135] Wagner, M., et al., *Major histocompatibility complex class I allele-specific cooperative and competitive interactions between immune evasion proteins of cytomegalovirus.* J Exp Med, 2002. 196(6): p. 805-16.

[136] Wagner, M., et al., *Systematic excision of vector sequences from the BAC-cloned herpesvirus genome during virus reconstitution.* J Virol, 1999. 73(8): p. 7056-60.

[137] Wagner, M. and U.H. Koszinowski, *Mutagenesis of viral BACs with linear PCR fragments (ET recombination).* Methods Mol Biol, 2004. 256: p. 257-68.

[138] Walker, M.S. and T.A. Hughes, *Messenger RNA expression profiling using DNA microarray technology: diagnostic tool, scientific analysis or un-interpretable data?* Int J Mol Med, 2008. 21(1): p. 13-7.

[139] Wang, E.T., et al., *Alternative isoform regulation in human tissue transcriptomes.* Nature, 2008. 456(7221): p. 470-6.

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

[141] White, E.A. and D.H. Spector, *Early viral gene expression and function.* 2007.

[142] Whitley, R.J., *Herpesviruses.* 1996.

[143] Wilhelm, B.T. and J.R. Landry, *RNA-Seq-quantitative measurement of expression through massively parallel RNA-sequencing.* Methods, 2009. 48(3): p. 249-57.

[144] Wilusz, J.E., H. Sunwoo, and D.L. Spector, *Long noncoding RNAs: functional surprises from the RNA world.* Genes Dev, 2009. 23(13): p. 1494-504.

[145] Xu, G., et al., *SAMMate: a GUI tool for processing short read alignments in SAM/BAM format.* Source Code Biol Med, 2011. 6(1): p. 2.

[146] Yokoyama, W.M., *Production of monoclonal antibodies.* Curr Protoc Cytom, 2006. Appendix 3: p. Appendix 3J.

[147] Zhang, G., et al., *Antisense transcription in the human cytomegalovirus transcriptome.* J Virol, 2007. 81(20): p. 11267-81.

[148] Zhang, Z., et al., *A greedy algorithm for aligning DNA sequences.* J Comput Biol, 2000. 7(1-2): p. 203-14.

[149] Ziegler, H., et al., *A mouse cytomegalovirus glycoprotein retains MHC class I complexes in the ERGIC/cis-Golgi compartments.* Immunity, 1997. 6(1): p. 57-66.

[150] Zine, A., T.R. Van De Water, and F. de Ribaupierre, *Notch signaling regulates the pattern of auditory hair cell differentiation in mammals.* Development, 2000. 127(15): p. 3373-83.

8. LIST OF FIGURES AND TABLES

8.2 FIGURES

8.3 TABLES

9. SUPPLEMENTAL MATERIALS

Supplemental table 1. Summary of MCMV transcripts identified in this study compared to the currently used annotations and previous temporal analysis. Sequenced cDNA clones were aligned to MCMV genome [GenBank accession number NC_004065.1] and this sequence entry was used to determine genomic locations.

Overlapping Genes NA	Overlapping Genes RA	Start[1]	End[1]	Range start[2]	Range end[2]	ORF Length[3]	Strand	No. of Clones	Libraries	time detected[4]
gp004	m04	3250	4102	3250	4102	852	+	7	4IE,E,L	6.5
gp006	m06	5319	6260	5319	6365	941	+	6	4IE,E,L	6.5
gp008	m08	7679	8440	7679	8440	761	+	1	L	6.5
gp015, gp016 spliced	m15, m16 spliced	14635, 15622	15083, 15700	14635, 15622	15083, 15700	448+78	+	1	E	6.5
gp015, gp016	m15, m16	14027	15700	14772	15699	1673	+	4	2L, E, IE	6.5
gp017	m17	16032	15704	16032	15704	328	-	2	IE	6.5
gp018 (AS)	m18 AS spliced	17079, 17853, 18351	17188, 17957, 18777	17079, 17853, 18351	17188, 17957, 18777	109+104+426	+	1	L	6.5
gp018 (AS)	m18 AS	18927	19285	18927	19285	358	+	1	L	6.5
gp019 (AS)	m19 AS	20702	20485	20702	20485	217	-	1	L	
gp020 (S), gp019 (AS)	m20(S), m19(AS)	21144	20434	21438	20434	710	-	3	IE,E,L	24
gp026	M25	27240	28285	26206	28959	1045	+	6	4L,2E	24
gp027, gp028	m25.1	29893	29169	30293	29169	724	-	3	2E,1L	6.5
gp027, gp028	m25.1, m25.2	30321	29128	30321	29128	1193	-	1	E	6.5

Overlapping Genes NA	Overlapping Genes RA	Start[1]	End[1]	Range start[2]	Range end[2]	ORF Length[3]	Strand	No. of Clones	Libraries	time detected[4]
gp34 (AS)	m29 (S), m29.1 (AS)	36135	35849	36135	35849	286	-	1	E	24
gp036	m30, M31	37055	37626	37055	37626	571	+	1	E	6.5
gp038	M32	40210	39324	40886	39195	886	-	6	3E, 3L	24
gp039	M34 spliced	44012, 44304	44242, 44516	44012, 44304	44242, 44516	230+212	+	1	E	6.5
gp040	M35	47052	47522	47052	47522	470	+	2	E	24
gp040 (AS), IGR[5] gp040-gp041	M36, M36 Ex2 (S/AS)	47794	47533	47794	47533	261	-	1	L	6.5
gp041	M37	50148	49411	50148	49391	737	-	2	1IE, 1L	6.5
gp045	m41	54217	53677	54217	53677	540	-	2	L	6.5
gp045, gp046	m42, m41	54863	53699	54863	53678	1164	-	4	3L, IE	6.5
gp045, gp046	m42, m41 spliced	55312, 54218	55123, 53678	55312, 54218	55123, 53678	189+540	-	1	IE	6.5
IGRgp046-047, gp047	m42	54842	54508	54842	54508	334	-	1	1L	6.5
gp047	M43	56402	55336	57157	55336	1066	-	9	6E, 2L, 1IE	6.5
gp047, gp048	M44, M43 spliced	58976, 57157, 56667	58668, 56856, 56361	58976, 57157, 56667	58668, 56856, 56361	308+301+306	-	1	IE	6.5
IGR gp048-gp049	M45, M44	59414	59271	59414	59271	143	-	1	L	6.5
IGR gp048-	M45	60126	59270	60126	59270	856	-	1	E	6.5

Overlapping Genes NA	Overlapping Genes RA	Start[1]	End[1]	Range start[2]	Range end[2]	ORF Length[3]	Strand	No. of Clones	Libraries	time detected[4]
gp049, gp049										
gp051	M47(AS), M46(S)	63968	63506	63968	63506	462	-	1	L	
gp053	m48.2 (S), m48.1 (AS)	73939	73535	73939	73526	404	-	3	2L,1E	6.5
gp053, gp054	M49 (S), m48.2 (S), m48.1 (AS)	74312	73526	74312	73526	786	-	18	10L, 8E	
gp054	M49	74304	73885	74818	73885	419	-	2	1E, 1L	24
gp054	M50, M49	75506	74851	75506	74851	-655	-	1	L	24
gp058	M53	78853	79534	78853	79534	681	+	4	3L, 1E	24
gp058 AS	M53 AS	80332	79485	80332	79485	847	-	1	IE	nd
IGR gp058-gp059	M54	80561	79589	80561	79589	972	-	2	1IE, 1L	6.5
go059, IGR gp058-gp059	M55, M54	84005	82893	84005	82893	1112	-	2	L	24
gp060	M69	96916	96080	96916	96080	836	-	1	E	24
gp066, IGR AS gp069-gp070	M71, IGR m74-M75	102514, 105879	102830, 106090	102514, 105879	102830, 106090	316+211	+	1	E	24
gp067	M72(AS), M73(S)	103709	104265	103534	104265	556	+	2	1E, 1L	24

Overlapping Genes NA	Overlapping Genes RA	Start[1]	End[1]	Range start[2]	Range end[2]	ORF Length[3]	Strand	No. of Clones	Libraries	time detected[4]
gp067 (AS), IGR gp069-gp070 (AS)	M72 (S), M73(AS), IGR m74-M75	103993, 105878	104161, 106090	103993, 105878	104161, 106095	168+213	+	4	L	
gp068 (S), IGR gp069-gp070 (AS)	M72 (S), M73(AS), IGR m74-M75 (alternate splice)	104124, 105878	104549, 106089	104124, 105878	104549, 106089	425+211	+	1	L	
gp068	M72 (AS)	104136	104209	104136	104209	73	+	1	L	24
gp069 (AS)	M74 (AS)	104825	105449	104825	105449	624	+	1	L	48
IGR gp069-gp070 (AS)	IGR m74-M75	105878	106095	105878	106095	217	+	1	L	
gp071,gp073	M76(AS), M78(S)	108476, 111789	108714, 112145	108476, 111789	108714, 112145	238+356	+	1	E	
gp073	M78	111752	112593	110933	111866	841	+	8	7IE, 1L	24
gp073	M78 spliced	111280, 111444	111409, 111710	111280, 111444	111409, 111710	129+266	+	1	E	24
IGR gp073-gp074	M79 (AS)	112418	112595	112418	112595	177	+	1	L	
gp075	M80	114322	115140	113589	115524	818	+	8	4E, 2L, 2IE	6.5
gp075	M80 spliced	114889, 115187	115148, 115396	114889, 115187	115148, 115396	259+209	+	1	L	6.5
gp076	M82	117413	116486	117413	115526	927	-	4	3E, 1L	6.5

Overlapping Genes NA	Overlapping Genes RA	Start[1]	End[1]	Range start[2]	Range end[2]	ORF Length[3]	Strand	No. of Clones	Libraries	time detected[4]
IGR gp078-gp079, gp079	IGR M84-85, M85	122735	121931	122735	121931	804	-	2	L	24
gp081	M88	131047	131370	131047	131370	323	+	1	L	24
gp084, gp085	M92 , M93, M94	134691, 135956	135369, 136399	134691, 135956	135369, 136399	678+443	+	1	L	
gp085	M93	135978	136146	135978	136146	168	+	1	L	24
gp085, gp086 spliced 1	M93, M94 spliced1	135978, 136181	136052, 136754	135978, 136181	136052, 136754	74+573	+	1	L	24
gp085, gp086 spliced 2	M93, M94 spliced2	135978, 136651	136524, 137227	135978, 136651	136524, 137227	546+576	+	1	L	24
gp085, gp086	M93, M94	136264	137333	135978	137333	1069	+	4	2L, IE, E	6.5
gp086	M94	136587	137345	136487	137345	758	+	3	2L, E	6.5
gp088, gp089	M95, M96	139307	139980	139307	139980	673	+	1	L	24
gp089	M96	139628	139967	139628	139967	339	+	1	E	24
IGR gp089-gp099, gp099	M97	139995	140880	139995	140880	885	+	1	IE	24
gp092	M98, M99	143462	144147	143462	144150	685	+	8	5L, 3E	6.5
gp093	M100	145355	144169	145355	144160	1184	-	6	3E, 3L	24
gp094	M102 spliced	145586, 147011	145908, 147682	145586, 147011	145908, 147682	322+671	+	1	IE	6.5
gp094	M102	147128	148034	147128	148169	906	+	3	2E, IE	6.5
gp095	M103	148772	148169	148772	148169	603	-	2	L	24
gp097	M105	153268	153874	153268	153874	606	+	2	L	6.5
gp098	m106	154101	154073	154101	154073	28	-	1	E	24

Overlapping Genes NA	Overlapping Genes RA	Start[1]	End[1]	Range start[2]	Range end[2]	ORF Length[3]	Strand	No. of Clones	Libraries	time detected[4]
gp098, IGR gp098-gp099	IGR m106-m107, m106(S)	161719	153867	161719	153867	7852	-	1	L	24
gp098, IGR gp098-gp099	IGR m106-m107, m106(S) spliced	161919, 154368	161622, 153886	161904, 154368	161622, 153886	297+482	-	3	3E, 1L	
IGR gp098-gp099	IGR m106-107	161357	160933	161357	160933	424	-	1	L	
IGR gp098-gp099, gp099	m108 - m106	162228	160670	162228	160670	1558	-	1	E	24
gpM112, gpM113	M112, M113 spliced	163778, 163983	163891, 164157	163778, 163983	163891, 164157	113+174	+	1	E	24
gpM112, gpM113	M112 Ex1, M113, M112 Ex2, M112 Ex3 (last exon in IGR M112 Ex3-M114)	163779, 163983, 164485, 164871	163891, 164160, 164582, 165510	163779, 163983, 164485, 164871	163891, 164160, 164582, 165510	112+177+97+639	+	1	L	
gpM113	IGR-m112Ex3-M114	164516	164581	164516	164581	-65	+	1	E	
gpM113	M113	164877	165502	164877	165502	-625	+	1	L	48

Overlapping Genes NA	Overlapping Genes RA	Start[1]	End[1]	Range start[2]	Range end[2]	ORF Length[3]	Strand	No. of Clones	Libraries	time detected[4]
gpM113, IGR gpM113, gp101	M113, M114	165020	165504	165020	165504	-484	+	1	IE	24,48
gp101	M114	166087	165497	166087	165497	590	-	1	L	24
gp101, gp102	M114, M115	166679	165935	166679	165935	744	-	1	L	24
gp103	M116 spliced	168850, 168015	168091, 167554	168685, 168015	168091, 167555	759+461	-	8	5E, 3L	ND
gp103	M116	169140	168095	169140	167261	1045	-	15	9L, 6E	ND
gp106	m119, M118	171957, 171585	171684, 171255	171957, 171585	171684, 171255	273+330	-	1	E	6.5
gp107, gp108	m119.1,	173217	172789	173217	172789	428	-	1	L	24
gp107, gp108, gp109, IGR gp109-gp110	m119.3, m119.2, m119.1	173897	172792	173897	172792	1105	-	7	4E, 2L, 1IE	6.5
gp108, gp109 (AS)	m119.2, m119.3 AS	173154	173578	173154	173578	-424	+	1	IE	6.5
gp107, gp108, IGR gp108-gp109	m119.3, m119.2	173899	172973	173905	172790	926	-	10	6L, 4E	6.5
gp109	m119.3	173892	173576	173892	173576	316	-	2	E	48
IGR gp108-gp109	IGR m119.3 - m119.4	173902	173737	173902	173737	165	-	1	E	24

Overlapping Genes NA	Overlapping Genes RA	Start[1]	End[1]	Range start[2]	Range end[2]	ORF Length[3]	Strand	No. of Clones	Libraries	time detected[4]
gp108, gp109, gp110, gp111 (AS), gp112	m120(S), m119.5(AS), m119.4 (S),m119.3(S), m119.2 (S)	174546	173131	174546	173131	1415	-	2	IE, E	24
gpM122Ex5	M122 Ex5	178405	177900	178405	177900	505	-	1	L	24
gpm123Ex4	m123 Ex4	180323	179554	180323	179554	769	-	2	L	6.5
gpM122Ex5, gpm123Ex4, gpm123Ex3, gp114, gp114ex2, gp115	IGR (m124.1 and m125), m123Ex2, m123 Ex3,m122 Ex5	182798, 181562, 181770, 179520	182596, 181371, 181659, 179420	182798, 181562, 181770, 179520	182596, 181371, 181659, 179420	202+191+111+100	-	1	E	24
gp121, gp122	m131 - m129	188054	187318	188054	187318	736		1	E	
gpm132Ex2, gp124	m133 Ex1, m132 Ex2, m131	189791, 188602	188880, 188407	189791, 188602	188880, 188407	991+195	-	3	2IE, 1L	24
gp123, gpm132Ex2	m132 Ex2 - m131	188885, 188603	188695, 188292	188885, 188603	188695, 188292	190+311	-	2	E, L	24
gp124	m133 Ex1	189793	188949	189793	188949	844	-	1	IE	
gp128 AS	m137 AS	191105	191373	191105	191373	268	+	1	IE	6.5
IGR gp128-gp129, gp129	m138, m137	193025	192162	193025	192162	863	-	1	L	6.5

Overlapping Genes NA	Overlapping Genes RA	Start[1]	End[1]	Range start[2]	Range end[2]	ORF Length[3]	Strand	No. of Clones	Libraries	time detected[4]
gp129	m138	193289	192188	193986	192160	1101	-	12	6E, 3IE, 3L	
gp130	m139	194810	194091	194810	194091	719	-	1	IE	24
gp133	m142	200648	199635	200648	199635	1013	-	2	IE, L	
IGR gp135, gp136	m145	204650	203973	204650	203973	677	-	6	3E, 3L	24
gp140 AS	m149(AS), m150 (S)	208012	207467	208012	207467	545	-	1	L	nd
gp141(AS), gp142(AS)	m150, m151 AS	208477	210069	208477	210069	-1592	+	1	IE	nd
gp142	m151	209564	208963	209564	208963	601	-	1	E	nd
gp145	m154	213864	212909	213864	212909	955	-	1	IE	6.5
gp146	m155	215486	214468	215486	214373	1018	-	3	2IE, L	24
gp147	m156, m155	215873	215098	215873	215098	775	-	2	E	24
IGR gp149-gp150, gp150	m159 A	218327	218054	218327	218054	273	-	1	L	24
gp150	m159 B	219397	219132	219397	219132	265	-	1	L	
gp151	m160	219890	219460	219890	219460	430	-	1	L	6.5
gp151	m160, m161	220641	219677	220641	219677	964	-	1	L	24
gp154	m163	222281	221832	222281	221832	449	-	1	L	6.5
gp154	m164 - m162	222465	221878	222465	221878	587	-	1	IE	6.5
gp154, gp155	m164, m163	222616	221986	222616	221832	630	-	14	10L, 3E, 1IE	6.4
gp157	m166	225639	224735	225650	224331	904	-	5	3IE, E, L	6.5

Overlapping Genes NA	Overlapping Genes RA	Start[1]	End[1]	Range start[2]	Range end[2]	ORF Length[3]	Strand	No. of Clones	Libraries	time detected[4]
gp157, gp158	m167, m166	226145	225250	226145	225250	895	-	1	L	6.5
gp158, gp159 (AS), gp160	IGR m167-m168, m168 (AS), m169, IGR m169-m170	229086, 228247	228325, 227426	229112, 228247	228325, 227426	761+821		138	28IE, 57E, 53L	6.5

[1]Start and End values were derived from the longest clone in the group.
[2]Start range and end range were derived from all clones belonging to a group.
[3]Based on longest clone; plus signs indicate spliced genes and exon lengths are given.
[4] Earliest time post-infection transcript as detected by [67]
[5]IGR, Intergenic region

Supplemental table 2. Comparison of cDNA library and RNASeq quantifications to currently used annotations

Modified Rawlinson's annotation (GU305914.1)				**Current GenBank (NC_004065.1)**				**cDNA library***			
ORF	**Start**	**End**	**RPKM**	**ORF**	**Start**	**End**	**RPKM**	**clone name**	**Start**	**End**	**No. of clones**
m01	468	870	132.0	gp001	480	836	106.8	ND			
m02	1033	2013	717.1	gp002	999	1979	615.3	ND			
m03	2270	3109	3668.2	gp003	2236	3102	3358.7	ND			
m04	3267	4063	14134.8	gp004	3270	4070	12977.7	m04	3250	4102	7
m05	4179	5200	3321.1	gp005	4185	5210	3066.0	ND			
m06	5291	6336	9753.0	gp006	5300	6337	9084.1	m06	5319	6260	6
m07	6463	7407	2140.7	gp007	6463	7407	1985.2	ND			
m08	7459	8529	3080.1	gp008	7459	8529	2856.4	m08	7679	8440	1
m09	8632	9513	533.3	gp009	8632	9513	494.6	ND			
m10	9624	10499	879.6	gp010	9624	10499	815.7	ND			
m11	10715	11614	530.0	gp011	10715	11614	491.5	ND			
m12	11686	12504	2414.8	gp012	11686	12504	2239.4	ND			
m13	12599	13000	2341.2	gp013	12599	13000	2171.2	ND			
m14	13085	13990	7662.4	gp014	13085	13990	7105.9	ND			
m15	14085	15065	12104.5	gp015	14085	15065	11225.3	m15, m16 spliced	14635, 15622	15083, 15700	1
								m15, m16	14027	15700	4
m16	15044	15676	9934.3	gp016	15044	15676	9212.7	ND			
m17	15749	16951	1918.8	gp017	15749	16951	1779.4	m17	15704	16032	2
m18	17071	20193	453.2	gp018	17071	20193	420.3	m18 AS spliced	17079, 17853, 18351	17188, 17957, 18777	1

Modified Rawlinson's annotation (GU305914.1)				Current GenBank (NC_004065.1)				cDNA library*			
ORF	**Start**	**End**	**RPKM**	**ORF**	**Start**	**End**	**RPKM**	**clone name**	**Start**	**End**	**No. of clones**
								m18 AS	18927	19285	1
m19	20338	20781	1454.8	gp019	20338	20781	1349.2	m19 AS	20485	20702	1
m20	20579	23044	1744.2	gp020	20802	23045	1596.4	m20(S), m19(AS)	20434	21144	3
m21	22644	23333	702.0	gp021	22645	23334	650.4	ND			
m22	23585	23899	492.1	gp022	23586	23900	457.6	ND			
M23	23777	24952	390.2	gp023	23778	24953	361.8	ND			
m23.1	24825	25160	319.4	gp024	24826	25161	296.2	ND			
M24	25147	26118	391.3	gp025	25148	26119	361.7	ND			
M25	26014	28812	3910.7	gp026	26015	28813	3628.5	M25	27240	28285	6
m25.1	28997	30601	3449.5	gp027	28998	30602	3198.9	m25.1	29893	29169	3
m25.2	28997	30280	3040.5	gp028	28998	30281	2819.7	m25.1, m25.2	30321	29128	1
m25.3	30244	31656	2262.4	gp029	30245	31657	2084.1	ND			
m25.4	30244	31215	2908.6	gp030	30245	31216	2681.3	ND			
M26	31346	31924	953.7	gp031	31347	31925	883.1	ND			
M27	32247	34295	489.6	gp032	32247	34295	454.1	ND			
M28	34486	35778	1621.1	gp033	34486	35778	1503.4	ND			
m29	35747	36475	2233.5	gp034	35747	36730	2025.9	m29 (S), m29.1 (AS)	36135	35849	1
m29.1	36030	36661	1559.6	gp035	36109	36660	1474.8	ND			
m30	36885	39071	2273.0	gp036	36884	37729	1737.5	m30, M31	37055	37626	1
M31	37281	39071	2301.0	gp037	37279	38829	1786.8	ND			
M31b*	38777	39079	4046.6					ND			

Modified Rawlinson's annotation (GU305914.1)				Current GenBank (NC_004065.1)				cDNA library*			
ORF	**Start**	**End**	**RPKM**	**ORF**	**Start**	**End**	**RPKM**	**clone name**	**Start**	**End**	**No. of clones**
M32	39283	41439	5069.7	gp038	39280	41436	4701.7	M32	40210	39324	6
M34	43086	45650	832.6	gp039	43083	45647	771.8	M34 spliced	44012, 44304	44242, 44516	1
M35	45912	47471	1865.8	gp040	45909	47468	1730.5	M35	47052	47522	2
								M36, M36 Ex2 (S/AS)	47533	47794	1
M37	49444	50481	2704.7	gp041	49441	50478	2513.4	M37	49411	50148	2
M38	50465	51958	3172.9	gp042	50462	51955	2924.6	ND			
m38.5c	51783	52523	2932.3					ND			
m39	52487	53203	1441.7	gp043	52484	53200	1334.3	ND			
m40	53268	53633	867.3	gp044	53265	53630	818.9	ND			
m41	53786	54202	4871.9	gp045	53783	54199	4491.5	m41	53677	54217	2
m42	54355	54846	1419.9	gp046	54352	54843	1317.5	m42, m41	53699	54863	4
								m42, m41 spliced	55123, 53678	55312, 54218	1
								m42	54508	54842	1
M43	55354	57147	6149.4	gp047	55351	57144	5688.1	M43	55336	56402	9
M44	57888	59123	6801.6	gp048	57885	59120	6306.9	M44, M43 spliced	58668, 56856, 56361	58976, 57157, 56667	1
m44.1*	58759	60108	5943.6					ND			
m44.3*	59144	59428	5685.0					ND			
								M45, M44	59271	59414	1
M45e1*	59518	62160	2159.3	gp049	59515	62876	1985.3	M45	59270	60126	1

Modified Rawlinson's annotation (GU305914.1)				Current GenBank (NC_004065.1)				cDNA library*			
ORF	**Start**	**End**	**RPKM**	**ORF**	**Start**	**End**	**RPKM**	**clone name**	**Start**	**End**	**No. of clones**
M45e2*	62773	62880	1256.3					ND			
m45.1	59520	63042	1904.2	gp050	61764	63038	1114.0	ND			
m45.2*	62810	62890	1321.6					ND			
M46	63044	63928	1783.1	gp051	63040	63924	1643.3	M47(AS), M46(S)	63506	63968	1
m48.1	73566	73877	27485.1	gp052	73562	73873	25442.3	m48.2 (S), m48.1 (AS)	73535	73939	3
m48.2	73575	73871	26546.0	gp053	73571	73867	24422.7	M49 (S), m48.2 (S), m48.1 (AS)	73526	74312	18
M49	73923	75533	6832.4	gp054	73919	75529	6441.5	M49	73885	74304	2
M50	75505	76455	2487.3	gp055	75501	76451	2301.5	M50, M49	74851	75506	1
M51	76519	77220	385.4	gp056	76515	77216	355.2	ND			
M52	76919	78471	807.3	gp057	76915	78468	746.2	ND			
M53	78465	79462	2223.8	gp058	78461	79462	2060.3	M53	78853	79534	4
								M54	79589	80561	2
M55	83004	85811	17682.9	gp059	83003	85816	16371.3	M55, M54	82893	84005	2
M56	85711	88107	2684.9	gp060	85716	88112	2457.9	ND			
m58	91756	92459	331.9	gp061	91761	92465	307.2	ND			
m59	93236	94393	354.7	gp062	93241	94263	234.4	ND			
M69	96284	98812	983.1	gp063	96193	98721	928.2	M69	96916	96080	1
m69.1	98621	98979	833.8	gp064	98530	98889	895.0	ND			
M70	99101	101995	538.0	gp065	99010	101904	505.1	ND			

Modified Rawlinson's annotation (GU305914.1)				**Current GenBank (NC_004065.1)**				**cDNA library***			
ORF	**Start**	**End**	**RPKM**	**ORF**	**Start**	**End**	**RPKM**	**clone name**	**Start**	**End**	**No. of clones**
M71	101994	102893	1195.1	gp066	101903	102802	1045.6	M71, IGR m74-M75	102514, 105879	102830, 106090	1
M72	103122	104327	6160.0	gp067	103031	104236	5134.0	M72(AS), M73(S)	103709	104265	2
								M72 (AS), M73(S), IGR m74-M75	103993, 105878	104161, 106090	4
								M72 (AS), M73(S), IGR m74-M75 (alternate splice)	104124, 105878	104549, 106089	1
								M72 (AS)	104136	104209	1
M73	104191	104609	11454.0	gp068	104100	104519	12725.6	ND			
M73.5e2*	105888	106160	15955.1					ND			
m74	104587	105903	8313.3	gp069	104496	105812	7405.2	ND	104825	105449	1
M75	106205	108382	965.8	gp070	106110	108287	890.9	IGR m74-M75	105878	106095	1
M76	108479	109242	1165.2	gp071	108384	109148	1122.5	M76, M78	108476, 111789	108714, 112145	1
M77	109026	110912	655.2	gp072	108931	110817	609.4	ND			
M78	111084	112498	7801.8	gp073	110989	112404	7082.7	M78	111752	112593	8

Modified Rawlinson's annotation (GU305914.1)				Current GenBank (NC_004065.1)				cDNA library*			
ORF	**Start**	**End**	**RPKM**	**ORF**	**Start**	**End**	**RPKM**	**clone name**	**Start**	**End**	**No. of clones**
								M78 spliced	111280, 111444	111409, 111710	1
M79	112737	113513	379.9	gp074	112639	113415	376.4	M79 (AS)	112418	112595	1
M80	113512	115607	3387.8	gp075	113414	115507	2876.9	M80	114322	115140	8
								M80 spliced	114889, 115187	115148, 115396	1
M82	115812	117611	10135.5	gp076	115711	117507	9864.5	M82	116486	117413	4
M83	117718	120147	9517.6	gp077	117614	120043	9022.0				
M84	120186	121949	4283.2	gp078	120082	121845	3692.9				
M85	122293	123228	5326.3	gp079	122189	123124	5060.8	IGR M84-85, M85	122735	121931	2
M87	127487	130267	424.8	gp080	127383	130163	372.9				
M88	130347	131626	1293.5	gp081	130243	131523	913.5	M88	131047	131370	1
				42.8 UL89 (CHS)	131649	132774	1332.9	ND			
m90	133020	133976	924.2	gp082	132920	133876	924.8	ND			
M91	133768	134172	652.8	gp083	133668	134072	631.7	ND			
M92	134175	134867	1313.6	gp084	134075	134767	1083.4	M92 , M93, M94	134691, 135956	135369, 136399	1
M93	134833	136379	2831.7	gp085	134733	136280	2321.3	M93	135978	136146	1
								M93, M94 spliced1	135978, 136181	136052, 136754	1
								M93, M94 spliced2	135978, 136651	136524, 137227	1

Modified Rawlinson's annotation (GU305914.1)				Current GenBank (NC_004065.1)				cDNA library*			
ORF	**Start**	**End**	**RPKM**	**ORF**	**Start**	**End**	**RPKM**	**clone name**	**Start**	**End**	**No. of clones**
								M93, M94	136264	137333	4
M94	136334	137370	7094.4	gp086	136234	137271	7099.3	M94	136587	137345	3
M83Ex1*	137487	138380	962.6	gp087	137390	138283	911.0	ND			
M95	138379	139632	1709.6	gp088	138282	139535	1330.6	M95, M96	139307	139980	1
M96	139632	140021	2662.1	gp089	139535	139924	3205.9	M96	139628	139967	1
M97	140238	142168	1393.3	gp090	140141	142072	1250.3	M97	139995	140880	1
M98	142198	143883	2967.4	gp091	142101	143786	2332.6	M98, M99	143462	144147	8
M99	143820	144158	8110.6	gp092	143723	144061	8801.5	ND			
M100	144393	145508	4238.1	gp093	144296	145411	4549.7	M100	144276	145355	6
M102	145693	148131	1825.6	gp094	145596	148034	1604.8	M102 spliced	145586, 147011	145908, 147682	1
								M102	147128	148034	3
M103	148279	149232	2387.7	gp095	148182	149135	2621.2	M103	148169	148772	2
M104	149210	151324	578.6	gp096	149113	151227	559.1				
M105	151125	153971	921.3	gp097	151028	153874	327.2	M105	153268	153874	2
m106	154010	154453	5578.6	gp098	153913	154356	6366.2	m106	154073	154101	1
m106.1*	154293	154553	4605.4					ND			
m106.3*	155878	156015	12851.2					ND			
								IGR m106-m107, m106(S)	153867	161719	1

Modified Rawlinson's annotation (GU305914.1)				Current GenBank (NC_004065.1)				cDNA library*			
ORF	**Start**	**End**	**RPKM**	**ORF**	**Start**	**End**	**RPKM**	**clone name**	**Start**	**End**	**No. of clones**
								IGR m106-m107, m106(S) spliced	161622, 153886	161919, 154368	3
								IGR m106-107	160933	161357	1
m107	162083	162777	477.9	gp099	161983	162678	427.3	m108 - m106	160670	162228	1
m108	162310	162870	449.3	gp100	162210	162770	433.0	ND			
								M112, M113 spliced	163778, 163983	163891, 164157	1
								M112 Ex1, M113, M112 Ex2, M112 Ex3 (last exon in IGR M112 Ex3-M114)	163779, 163983, 164485, 164871	163891, 164160, 164582, 165510	1
								IGR-m112Ex3-M114	164516	164581	1
								M113	164877	165502	1
				M112	163097	164511	3880.7	ND			1

Modified Rawlinson's annotation (GU305914.1)				**Current GenBank (NC_004065.1)**				**cDNA library***			
ORF	**Start**	**End**	**RPKM**	**ORF**	**Start**	**End**	**RPKM**	**clone name**	**Start**	**End**	**No. of clones**
				38.3 UL113 P	163983	165079	6694.9	ND			
								M113, M114	165020	165504	
M114	165696	166484	4704.2	gp101	165596	166384	3765.7	M114	165497	166087	1
M115	166484	167308	3667.1	gp102	166384	167208	3703.0	M114, M115	165935	166679	1
M116	167305	169242	32531.8	gp103	167205	169142	30352.0	M116 spliced	168091, 167554	168850, 168015	8
								M116	168095	169140	15
m117	169313	171010	527.9	gp104	169213	170910	461.5	ND			
m117.1	169641	171055	553.2	gp105	169541	170956	443.8	ND			
M118*	171080	172045	1983.9	gp106	170980	171945	2035.3	m119, M118	171684, 171255	171957, 171585	1
m119.1	172156	173091	35140.0	gp107	172056	172991	20919.6	m119.1	172789	173217	1
m119.2	173122	173490	84676.3	gp108	173022	173390	84488.3	m119.3, m119.2, m119.1	172792	173897	7
								m119.2, m119.3 AS	173154	173578	1
								m119.3, m119.2	172973	173899	10
m119.3	173510	173821	42123.7	gp109	173410	173721	51339.2	m119.3	173576	173892	2

Modified Rawlinson's annotation (GU305914.1)				Current GenBank (NC_004065.1)				cDNA library*			
ORF	**Start**	**End**	**RPKM**	**ORF**	**Start**	**End**	**RPKM**	**clone name**	**Start**	**End**	**No. of clones**
								IGR m119.3 - m119.4	173737	173902	1
m119.4	174154	174435	4566.6	gp110	174054	174335	3031.5	m120(S), m119.5(AS), m119.4 (S),m119.3(S), m119.2 (S)	173131	174546	2
m119.5	174254	174589	4656.2	gp111	174154	174489	4333.9	ND			
m120	174399	174674	4275.6	gp112	174299	174574	4434.7	ND			
m120.1*	174740	175825	6140.0					ND			
M121	175779	177875	1921.9	gp113	175679	177775	1653.2	ND			
				gpM122Ex5	177980	179517	6791.6	M122 Ex5	177900	178405	1
				gpm123Ex4	179760	181249	12281.9	m123 Ex4	179554	180323	2
m123.1	181963	182319	1292.9					ND			
				gpm123Ex3	181368	181562	10742.0	ND			

Modified Rawlinson's annotation (GU305914.1)				Current GenBank (NC_004065.1)				cDNA library*			
ORF	**Start**	**End**	**RPKM**	**ORF**	**Start**	**End**	**RPKM**	**clone name**	**Start**	**End**	**No. of clones**
m123Ex2	181756	181866	1644.7	gp114	181656	181766	5070.3	IGR (m124.1 and m125), m123Ex2, m123 Ex3,m122 Ex5	182596, 181371, 181659, 179420	182798, 181562, 181770, 179520	1
m124	182033	182380	1170.9	gp115	181933	182280	1222.8	ND			
m124.1	182111	182518	914.0	gp116	182011	182418	1127.1	ND			
m125	183536	183865	5488.5	gp117	183436	183765	5687.2	ND			
m126	184635	184910	951.9	gp118	184535	184810	969.8	ND			
m127	185290	185691	933.8	gp119	185190	185591	687.7	ND			
m128Ex3	186185	187399	1744.8	gp120	186085	187299	1626.3	ND			
m129	187447	187947	741.1	gp121	187347	187847	736.0	ND			
m130	187907	188380	1518.9	gp122	187807	188280	431.2	m131 - m129	188054	187318	1
m131	188126	188476	3692.3	gp123	188026	188376	1776.2	m133 Ex1, m132 Ex2, m131	188880, 188407	189791, 188602	3
				gpm132Ex 2	188379	188601	5905.9	m132 Ex2 - m131	188695, 188292	188885, 188603	2
m133Ex1*	188978	189895	5201.4	gp124	188878	189795	4812.8	m133 Ex1	188949	189793	1
m134	189968	190381	690.2	gp125	189868	190281	793.0	ND			
m135	189995	190321	691.6	gp126	189895	190221	698.5	ND			

Modified Rawlinson's annotation (GU305914.1)				Current GenBank (NC_004065.1)				cDNA library*			
ORF	**Start**	**End**	**RPKM**	**ORF**	**Start**	**End**	**RPKM**	**clone name**	**Start**	**End**	**No. of clones**
m136	190410	191171	1161.2	gp127	190310	191071	1089.9	ND			
m137	191188	192192	695.9	gp128	191088	192092	460.6	m138, m137	192162	193025	1
m138	192333	194042	16935.1	gp129	192233	193942	15867.6	m138	192188	193289	12
m139	194182	196116	1702.5	gp130	194082	196016	1507.9	m139	194091	194810	1
m140	196162	197616	1346.8	gp131	196062	197516	1436.9	ND			
m141	197805	199331	540.7	gp132	197705	199231	477.9	ND			
m142	199541	200848	1938.6	gp133	199441	200748	1712.4	m142	199635	200648	2
m143	201065	202694	984.8	gp134	200920	202593	883.6	ND			
m144	202843	203994	471.5	gp135	202742	203893	322.1	ND			
m145	204130	205593	3012.7	gp136	204029	205492	2912.9	m145	204650	203973	6
m146	205743	206876	1267.2	gp137	205642	206775	1160.1	ND			
m147	206963	207400	3879.8	gp138	206862	207299	3806.9	ND			
m148	207029	207388	4056.9	gp139	206928	207287	4314.0	ND			
m149	207427	208116	817.6	gp140	207326	208015	853.8	ND			
m150	207724	208890	199.4	gp141	207623	208789	171.5	m150, m151 AS	210069	208477	1
m151	208915	210084	233.7	gp142	208814	209983	218.4	m151	208963	209564	1
m152	210342	211478	5328.9	gp143	210241	211377	4968.7	ND			
m153	211688	212905	883.8	gp144	211587	212804	834.0	ND			
m154	213043	214149	2029.1	gp145	212942	214048	1846.9	m154	212909	213864	1
m155	214535	215668	6927.1	gp146	214434	215567	5951.9	m155	214468	215486	3
m156	215635	216078	5202.6	gp147	215534	215977	6135.2	m156, m155	215098	215873	2
m157	215996	216985	3075.0	gp148	215895	216884	2832.8	ND			
m158	217033	218103	936.0	gp149	216932	218002	899.7	ND			

Modified Rawlinson's annotation (GU305914.1)				Current GenBank (NC_004065.1)				cDNA library*			
ORF	**Start**	**End**	**RPKM**	**ORF**	**Start**	**End**	**RPKM**	**clone name**	**Start**	**End**	**No. of clones**
m159	218271	219467	3367.0	gp150	218170	219366	3072.7	m159 A	218054	218327	1
								m159 B	219132	219397	1
m160	219699	220625	6812.5	gp151	219598	220524	6581.7	m160	219460	219890	1
m161	220573	221250	1867.2	gp152	220472	221149	2601.1	m160, m161	219677	220641	1
m162	221287	221766	456.6	gp153	221186	221665	608.5	ND			
m163	221976	222515	4994.9	gp154	221875	222414	4771.9	m163	221832	222281	1
								m164 - m162	221878	222465	1
m164	222467	223750	1451.9	gp155	222366	223649	1510.4	m164, m163	221986	222616	14
m165	223381	224379	738.4	gp156	223280	224278	561.8	ND			
m166	224514	225662	4994.9	gp157	224413	225561	4637.0	m166	224735	225639	5
m167	225880	227190	1088.6	gp158	225779	227089	1007.5	m167, m166	225250	226145	1
m168	228021	228566	321627.4	gp159	227920	228465	450005.1				
m169	228411	228809	258885.9	gp160	228310	228708	202852.8	IGR m167-m168, m168 (AS), m169, IGR m169-m170	228325, 227426	229086, 228247	126
m170	229440	230147	141.1	gp161	229339	230046	127.1				

Supplemental table 3. Comparison of RPKM values in Marcinowski et al. (2012) and this RNASeq experiment. Top 10 genes in all lists are separated by double line. 8 out of 10 top genes are identical between our RNASeq data Dolken 25 hpi, while Dolken 48 hpi vs our RNASeq share 7 out of 10 top genes. Genes diverging between Dolken and our RNASeq data lists are marked in bold.

Dolken 25 hpi total RNA			**RNASeq data from this study**			**Dolken 48 hpi total RNA**		
ORF ID	**Reads Counts**	**RPKM**	**ORF ID**	**Reads Counts**	**RPKM**	**ORF ID**	**Reads Counts**	**RPKM**
m119.3	5658	73091.9	m168	427096	335276.6	m48.2	2648	99511.4
m119.2	6484	70823.5	m169	251224	269872.5	m48.1	2771	99127.3
m169	5785	58437.5	m119.2	75992	88269.7	m119.3	1903	68076.2
m48.2	4115	55843.7	m119.3	31964	43911.4	m119.2	2217	67058.0
m48.1	4288	55393.8	m119.1	79994	36631.3	m169	2327	65093.1
m168	7013	51769.3	**M116**	153335	33912.4	m168	2889	59056.3
m138	12250	28873.6	m48.1	20856	28651.5	**M94**	2828	30408.4
M94	7356	28563.1	m48.2	19175	27672.6	**m106**	1049	26369.6
m119.1	5155	22198.0	**M55**	120762	18433.3	M55	6278	24953.7
M44	6300	20543.9	m138	70431	17653.8	**M49**	2664	18456.5
M55	14159	20323.4	m04	27433	14734.7	m119.1	1465	17469.2
M43	8896	19986.3	m15	28880	12618.2	M99	521	17153.4
M99	1480	17596.4	M73	11700	11940.1	M96	599	17142.5
m106	1745	15840.7	M82	44371	10565.7	M85	1385	16515.3
M96	1440	14881.9	m16	15294	10355.9	m120	392	15852.2
M80	7444	14307.7	m06	24835	10166.9	M80	2849	15163.7
M78	4894	13930.3	M83	56249	9921.5	m163	700	14468.2
M49	5144	12869.6	m74	26628	8666.1	m119.5	419	13918.3
m166	3225	11312.8	M99	6687	8454.8	M116	2200	12670.1
m163	1506	11240.7	M78	26868	8132.8	m138	1844	12035.8

Dolken 25 hpi total RNA			RNASeq data from this study			Dolken 48 hpi total RNA		
ORF ID	**Reads Counts**	**RPKM**	**ORF ID**	**Reads Counts**	**RPKM**	**ORF ID**	**Reads Counts**	**RPKM**
M85	2507	10795.4	m14	16884	7987.6	M32	2195	11357.8
m120	703	10266.1	M94	17910	7395.5	M100	1129	11291.2
M116	4727	9830.9	m155	19105	7221.1	M43	1774	11036.8
m119.5	809	9704.4	M49	26770	7122.3	M78	1275	10049.8
M56	5592	9402.9	m160	15359	7101.6	M56	2153	10025.1
m131	818	9393.1	M44	20446	7090.2	M72	1034	9569.4
m25.3	3092	8819.8	M72	18068	6421.5	M44	1043	9418.4
M93	3333	8678.1	M43	26831	6410.4	m166	889	8635.6
m04	1707	8621.7	m106	6024	5815.3	m119.4	213	8430.3
m25.1	3286	8251.9	m125	4405	5721.4	M93	1152	8306.0
M98	3405	8139.9	m152	14736	5555.1	m25.3	999	7891.0
M100	2236	8075.5	M85	12125	5552.3	m25.1	1094	7607.7
m41	820	7925.7	m156	5618	5423.4	M35	1025	7333.5
m148	692	7747.6	M32	26596	5284.9	m25.2	834	7249.6
m25.2	2448	7684.4	m163	6560	5206.9	m25.4	631	7245.6
m25.4	1804	7480.5	m166	13958	5206.8	M121	1289	6860.7
M26	1048	7295.3	m41	4941	5078.7	M73	254	6749.9
m29	1310	7242.8	M114	9027	4903.8	m41	246	6584.3
m147	784	7214.4	m119.5	3805	4853.8	M95	728	6479.6
M114	1411	7207.9	m119.4	3132	4760.4	m74	726	6152.7
M38	2578	6954.9	M84	18376	4465.0	M98	927	6136.7
m160	1587	6900.1	m120	2870	4457.0	M114	423	5983.8
M32	3639	6799.7	M100	11503	4417.9	M53	517	5776.1
M72	2011	6720.9	m148	3552	4229.0	M83	1239	5690.8

Dolken 25 hpi total RNA			RNASeq data from this study			Dolken 48 hpi total RNA		
ORF ID	**Reads Counts**	**RPKM**	**ORF ID**	**Reads Counts**	**RPKM**	**ORF ID**	**Reads Counts**	**RPKM**
m154	1782	6488.2	M25	26622	4076.7	M38	757	5655.3
m29.1	1018	6481.9	m147	4133	4044.5	m131	172	5469.3
m155	1816	6454.5	m131	3152	3849.0	m04	366	5119.1
m119.4	445	6360.2	m03	7494	3823.9	M26	262	5050.5
M92	1062	6176.6	M115	7358	3822.8	m19	197	4952.2
m45.1	5342	6108.1	m25.1	13465	3595.9	m148	158	4898.5
m130	716	6088.3	M80	17278	3531.6	m147	192	4892.6
m145	2166	5963.2	m159	9802	3509.9	m45.1	1527	4834.9
M73	580	5566.0	m05	8263	3462.0	m155	462	4547.2
M53	1297	5232.8	M38	11529	3307.6	M84	691	4372.1
m161	851	5059.0	m08	8023	3210.8	m14	349	4299.4
M95	1563	5023.7	m157	7404	3205.5	m128Ex3	442	4060.3
M35	1800	4650.6	m25.2	9495	3169.6	m29	260	3980.7
m74	1512	4627.3	m145	10727	3140.6	m29.1	215	3790.9
m06	1157	4454.0	M98	12168	3093.4	m130	160	3767.5
M28	1409	4392.1	m25.4	6876	3032.1	M88	422	3676.8
M97	2042	4260.0	M93	10661	2951.9	M103	311	3638.5
M83	2511	4164.9	M37	6828	2819.5	M92	224	3607.7
m128Ex3	1222	4053.7	M56	15652	2798.8	M46	274	3455.6
m19	445	4039.6	M96	2525	2775.0	M28	389	3357.9
M121	2078	3994.0	M50	5753	2592.9	m59	346	3334.9
m13	386	3870.1	m12	4810	2517.3	m13	120	3331.7
M88	1219	3835.4	M103	5540	2489.0	m154	325	3276.8
m156	405	3676.5	m13	2289	2440.6	m156	129	3242.8

Dolken 25 hpi total RNA			RNASeq data from this study			Dolken 48 hpi total RNA		
ORF ID	**Reads Counts**	**RPKM**	**ORF ID**	**Reads Counts**	**RPKM**	**ORF ID**	**Reads Counts**	**RPKM**
M103	867	3663.0	M31	10023	2398.7	m167	380	3235.1
m140	1287	3565.1	m30	12090	2369.5	M37	294	3161.3
M91	358	3562.8	m25.3	7775	2358.5	m161	191	3144.2
M37	908	3525.7	m29	3960	2328.3	M25	774	3086.4
M102	2096	3463.7	M53	5403	2318.1	M69	693	3058.4
m162	405	3400.8	m07	4920	2231.5	M34	685	2980.7
m90	769	3238.7	m154	5463	2115.2	M105	756	2963.8
m17	966	3236.5	m142	6167	2020.9	m159	316	2946.5
m03	659	3162.0	M121	9802	2003.5	m20	637	2883.1
M46	686	3124.2	m17	5614	2000.2	M76	197	2874.2
m142	985	3035.2	m45.1	16325	1985.0	M97	496	2865.4
m20	1813	2963.2	m161	3079	1946.5	M71	230	2852.3
m08	777	2924.1	M35	7079	1945.0	m22	79	2799.2
m139	1364	2841.2	M102	10829	1903.0	m145	366	2790.3
M84	1237	2826.4	M46	3838	1858.8	m90	229	2670.8
m59	805	2801.9	m128Ex3	5156	1818.9	m69.1	86	2666.3
m164	871	2734.1	m20	10461	1818.2	M77	437	2584.8
m14	587	2611.4	M95	5214	1782.1	M91	92	2535.4
m42	304	2490.4	m139	8012	1774.7	m107	157	2517.7
m165	606	2444.9	m123Ex2	444	1714.5	m39	161	2506.2
m137	608	2438.4	M28	5098	1689.9	m21	152	2458.7
m159	708	2384.0	m29.1	2401	1625.8	m160	203	2444.1
m124.1	237	2341.3	m130	1751	1583.4	m126	59	2385.9
M77	1069	2283.3	m19	1571	1516.6	M23	248	2353.7

Dolken 25 hpi total RNA			RNASeq data from this study			Dolken 48 hpi total RNA		
ORF ID	**Reads Counts**	**RPKM**	**ORF ID**	**Reads Counts**	**RPKM**	**ORF ID**	**Reads Counts**	**RPKM**
M79	433	2246.1	m164	4534	1513.5	m165	207	2312.7
M76	421	2218.1	m39	2514	1502.9	m108	116	2307.8
m02	537	2206.3	m42	1699	1480.1	m142	268	2286.9
M31	956	2151.4	M97	6547	1452.5	m15	196	2230.0
m07	500	2132.5	m140	4766	1404.0	m140	290	2224.6
M87	1467	2126.1	M92	2214	1369.4	M75	425	2177.9
m30	1142	2104.6	M88	4030	1348.4	m02	191	2173.1
m152	592	2098.6	m123.1	1132	1347.8	m162	92	2139.2
M71	468	2095.9	m146	3495	1321.0	m40	68	2073.7
m167	673	2069.1	M71	2616	1245.8	M104	392	2068.7
M115	423	2066.6	m124	991	1220.6	m03	155	2059.5
M34	1315	2066.3	M76	2168	1214.7	m124.1	75	2051.7
m22	161	2060.0	m136	2152	1210.5	m164	236	2051.4
m124	167	1934.2	m167	3471	1134.8	m137	178	1976.8
M75	1014	1876.5	m143	3909	1026.6	M115	145	1961.7
M50	426	1805.5	M69	6047	1024.9	M52	265	1903.3
m123.1	160	1791.3	M75	5116	1006.8	m08	182	1896.7
m21	306	1787.4	M26	1343	994.2	M102	400	1830.5
M25	1207	1738.1	m126	639	992.3	M50	155	1819.1
M23	507	1737.6	m158	2438	975.7	M82	289	1792.0
M52	667	1730.0	m127	913	973.5	m139	305	1759.3
m12	345	1697.8	m90	2151	963.4	m06	163	1737.6
m15	403	1655.8	M105	6379	960.4	m123.1	55	1705.2
M69	1030	1641.5	m124.1	907	952.8	m124	52	1667.8

Dolken 25 hpi total RNA			RNASeq data from this study			Dolken 48 hpi total RNA		
ORF ID	**Reads Counts**	**RPKM**	**ORF ID**	**Reads Counts**	**RPKM**	**ORF ID**	**Reads Counts**	**RPKM**
M105	1148	1625.2	m153	2618	921.3	M70	422	1627.0
m127	160	1604.2	m10	1874	916.9	M31	261	1626.5
M104	834	1589.3	m40	772	904.1	m127	58	1610.3
M27	806	1585.5	m69.1	730	869.1	m07	134	1582.6
m05	401	1579.9	M34	5194	867.9	M87	390	1565.2
m143	620	1531.2	m149	1372	852.3	M79	106	1522.6
m39	269	1512.1	M52	3051	841.5	m30	298	1520.8
m40	125	1376.5	m129	903	772.5	m18	423	1511.7
m107	210	1216.1	m165	1794	769.7	m157	134	1510.7
m108	164	1178.3	m02	1711	747.6	m01	51	1405.5
m157	288	1172.5	m21	1178	731.8	m17	148	1373.1
m18	902	1164.1	m137	1701	725.5	M51	80	1271.9
M82	509	1139.7	m135	550	720.9	M24	110	1263.1
m144	325	1137.1	m134	695	719.5	M27	231	1258.3
M24	274	1136.2	M77	3007	683.0	m42	53	1202.3
m69.1	99	1108.4	M91	643	680.5	m12	88	1199.3
m141	416	1098.0	M104	2976	603.1	m16	67	1181.4
M51	174	999.0	m117.1	1905	576.6	m152	108	1060.2
m16	153	974.2	m141	2008	563.6	m143	149	1019.0
m126	61	890.8	M70	3788	560.8	m05	92	1003.7
m01	85	845.9	m09	1144	555.9	m144	96	930.1
m23.1	64	767.7	m11	1160	552.4	m123Ex2	9	905.0
m129	94	756.2	m117	2180	550.3	m146	91	895.7
m146	211	749.9	m22	377	513.0	m129	38	846.6

Dolken 25 hpi total RNA			RNASeq data from this study			Dolken 48 hpi total RNA		
ORF ID	**Reads Counts**	**RPKM**	**ORF ID**	**Reads Counts**	**RPKM**	**ORF ID**	**Reads Counts**	**RPKM**
M70	526	732.3	M27	2440	510.4	m58	52	823.2
m158	179	673.6	m107	809	498.2	m158	78	812.9
m136	124	655.9	m144	1321	491.5	m125	24	811.7
m153	188	622.1	m162	533	475.9	m23.1	24	797.2
m11	118	528.4	m18	3442	472.4	m141	107	782.1
m10	111	510.7	m108	613	468.3	m170	49	772.5
m123Ex2	14	508.4	M87	2873	442.8	m136	43	629.8
m134	51	496.5	M24	925	407.9	m153	67	614.0
m170	83	472.5	M23	1116	406.7	m10	37	471.4
m135	36	443.7	M51	658	401.8	m11	36	446.4
m58	74	423.1	M79	718	396.1	m134	15	404.4
m125	33	403.1	m59	999	369.8	m135	10	341.3
m09	70	319.9	m58	569	345.9	m149	14	226.5
m117.1	106	301.7	m23.1	261	332.9	m09	17	215.1
m149	40	233.7	m151	665	243.6	m117	30	197.2
m117	97	230.2	m150	566	207.9	m117.1	25	197.1
m151	42	144.7	m170	243	147.1	m151	15	143.1
m150	23	79.4	m01	130	137.6	m150	11	105.2