Proteomics
www.proteomics-journal.com

# Viral Short ORFs and Their Possible Functions

*Yaara Finkel, Noam Stern-Ginossar, and Michal Schwartz**

**Definition of functional genomic elements is one of the greater challenges of the genomic era. Traditionally, putative short open reading frames (sORFs) coding for less than 100 amino acids were disregarded due to computational and experimental limitations; however, it has become clear over the past several years that translation of sORFs is pervasive and serves diverse functions. The development of ribosome profiling, allowing identification of translated sequences genome wide, revealed wide spread, previously unidentified translation events. New computational methodologies as well as improved mass spectrometry approaches also contributed to the task of annotating translated sORFs in different organisms. Viruses are of special interest due to the selective pressure on their genome size, their rapid and confining evolution, and the potential contribution of novel peptides to the host immune response. Indeed, many functional viral sORFs were characterized to date, and ribosome profiling analyses suggest that this may be the tip of the iceberg. Our computational analyses of sORFs identified by ribosome profiling in DNA viruses demonstrate that they may be enriched in specific features implying that at least some of them are functional. Combination of systematic genome editing strategies with synthetic tagging will take us into the next step—elucidation of the biological relevance and function of this intriguing class of molecules.**

## 1. Introduction

The big challenges of the post-genomic era, after completing the sequencing of genomes of a wide range of organisms and viruses, are the annotations of functional units in these genomes, including identification of protein coding sequences. Open reading frames (ORFs) are defined as DNA sequences with translation potential, consisting of a string of in-frame codons flanked by a start codon and a stop codon. Traditionally, the definition of ORFs was sequences that potentially code for peptides of 100 amino acids (aa) or more, this length limit originates from bioinformatic approaches that were used for annotations. Since it was assumed that the majority of coding genes would code for larger proteins and a stretch of 100 codons without a stop codon provides a statistically significant signal, using 100 aa as a cut-off provided a straightforward strategy to annotate genomes. However, in recent years there have been accumulating evidence for the prevalence of functional translation events from short ORFs (sORFs) encoding proteins with a length of 100 aa or less.[1,2]

sORFs were shown to carry out diverse functions and were involved in many biological processes. An important functional subclass of translated sORFs are upstream ORFs (uORFs) that are located at the 5′ leader sequence of mRNAs, upstream of the initiation codon of the main coding ORF. These have been shown to regulate translation efficiency via different mechanisms.[3] In most of the cases studied to date, uORFs are *cis* acting and regulate their downstream ORF. Some examples of uORFs were described, termed peptoswitches, that respond to environmental cues.[1] In these cases, small molecules bind to the nascent peptide leading to inhibition of translation of the main ORF by different mechanisms including non-sense mediated decay[4] and ribosome stalling.[5] In addition, molecular functions were also assigned to sORF-encoded peptides (SEPs), these include regulation of post-translational modifications,[6] regulation of metabolite transport,[7] and inhibition of kinase activity.[8] As expected by their different molecular functions, SEPs are involved in various biological processes including cell communication, signal transduction, transcriptional regulation, and metabolism.

A major advancement in the field was the introduction of ribosome profiling, a powerful experimental strategy to map translation events in vivo.[9] This technique is based on early studies showing that each ribosome physically protects a portion of its mRNA template from nuclease digestion.[10,11] The advent of high-throughput sequencing offered the opportunity to analyze all ribosome-protected fragments in living cells, thereby providing a snapshot of translation in vivo (**Figure 1**). Ribosome footprints can indicate the positions of ribosomes with subcodon precision. Such high-resolution information identifies the precise boundaries of translated regions as well as the specific reading frame in which translation occurs. While the ribosome decodes only a single codon at a time, it protects a much larger footprint, typically 28–32 nucleotides.[9] The signature of the triplet genetic code is present in these larger footprints, as the footprint positions show a three-nucleotide periodicity reflecting the translocation steps of the ribosome. This bias in footprint position provides a robust statistical signal that indicates which specific reading frame is being translated on an mRNA.[9,12–16]

Y. Finkel, Dr. N. Stern-Ginossar, Dr. M. Schwartz
Department of Molecular Genetics
Weizmann Institute of Science
Rehovot, Israel
E-mail: michalsc@weizmann.ac.il

**DOI: 10.1002/pmic.201700255**

 

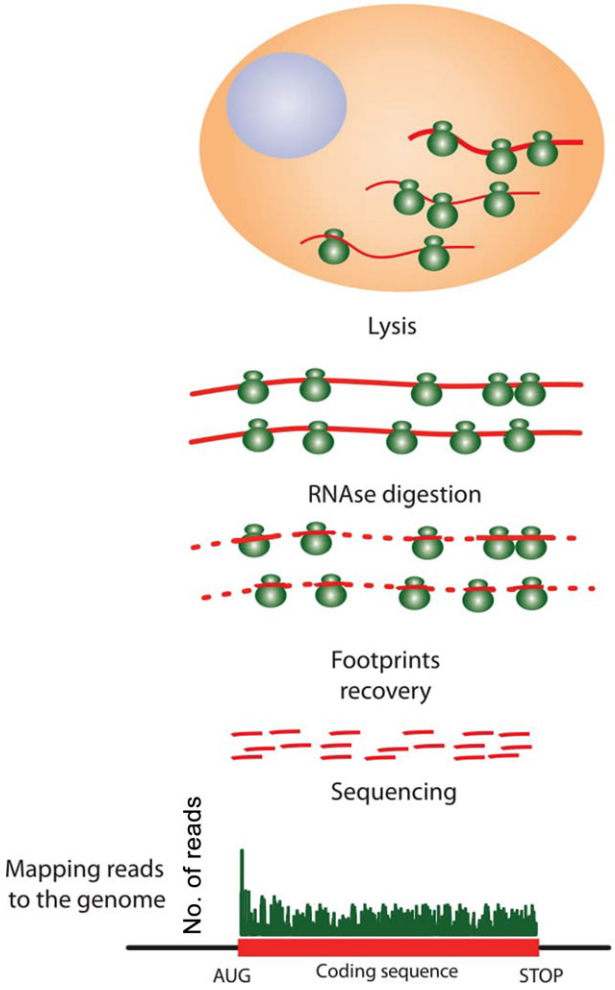


**Figure 1.** Ribosome profiling allows identification of translated regions. Cells are lysed and subjected to nuclease digestion. Prior to lysis, cells may be treated with cycloheximide to freeze ribosomes on their mRNA targets. Ribosome footprints are isolated and converted to deep sequencing libraries. Reads are then mapped to the genome, this facilitates mapping of translated ORFs in an unbiased quantitative manner.

Translational start sites can be mapped more directly by performing ribosome profiling under conditions that preferentially arrest initiating ribosomes. Harringtonine and lactimidomycin are two drugs that preferentially target eukaryotic initiating ribosomes.[12,13,17] Both of these drugs lead to strong accumulation of ribosomes precisely at translation initiation sites and depletion of ribosomes over the body of the message. Though their effects are similar, they act through very different mechanisms, and their combination helps to preclude drugs-related artifacts and results in robust start site detection.[13] Harringtonine binds to the peptidyltransferase center in disassembled large ribosomal subunits.[18,19] Lactimidomycin, in contrast, is related to cycloheximide and both bind to the same site on the large ribosomal subunit, but lactimidomycin displays a marked preference for an empty E site, which occurs only in the initiating ribosomes.[20,21] A similar effect can also be achieved by treatment with puromycin, which drives premature termination, thereby removing most elongating ribosomes and leaving footprints predominantly at sites of initiation.[22] Furthermore, puromycin can be combined with lactimidomycin in order to stabilize initiating ribosomes while depleting elongating ribosomes in cell lysates.[23]

Precise mapping of translation initiation and reading frames on RNA provides substantial information that allows accurate analysis of translation events in an unbiased manner. These approaches facilitated new studies that re-annotated the coding capacity of numerous organisms, revealing widespread translation outside of annotated protein coding sequences.[16,24–27] Ribosome profiling proved a sensitive methodology with high discovery rate[2] providing a robust platform for systematic analysis of translated ORFs. The outstanding challenge is to define the functionality of newly identified translation events and to discriminate between ORFs that provide regulatory or protein-based functions to products of random translational events.

On the computational side, approaches that rely on cross-species comparisons and conservation, that are adjusted to sORFs, are a valuable tool that could identify sORFs that are likely to produce functional peptides.[28,29]

An additional relevant technique, mass spectrometry, is a powerful tool to directly detect proteins and peptides. Detection of SEPs could provide insight into their stability and abundance and therefore can point to functionality. Improvements on the experimental and computational aspects of proteomic approaches[30] assisted in discovery of tens of novel SEPs.[31] For this, Slavoff et al. used peptidomics, a mass spectrometry–based approach that is augmented for preservation and enrichment of small peptides, mainly by reducing proteolysis during sample processing. Generation of a dataset based on the human transcriptome taking into account out-of-frame alternative translation products has led to the discovery of 1259 alternative proteins, many of which derived from sORFs in different tissues and cell lines.[32] Interestingly, many of these alternative proteins were found to be secreted. Thus, computational and proteomic approaches still have limitations when applied to search for SEPs, however they are constantly upgraded and improve our discovery capacity of the variety of these peptides. In this review, we will concentrate on sORFs in viruses focusing on known functions of virally encoded short proteins and how viral systems—due to their high transcriptional levels, manipulatable genomes, and measurable phenotypes—could provide opportunities to better decipher sORFs functions.

## 2. Virally Encoded Short Proteins

Viruses are essentially infectious units that replicate inside a living cell and their life cycle is tailored to support genome amplification and transmission to new hosts. An inevitable requirement of this life cycle is small genome size. Effective strategies to produce diverse functions and to maintain small genome is to produce small proteins and to evolve sophisticated gene expression regulation mechanisms. Studying sORFs in viruses also has the advantage of their rapid and restricting evolution which would imply that sequence conservation is highly indicative of conserved function. Indeed numerous viral functional sORFs were identified and characterized.[33] In addition, application of

ribosome profiling on infected cells revealed unanticipated complexity in viruses coding capacity, with many novel putative sORFs.

### 2.1. Short Transmembrane Proteins

A major group of viral SEPs encode short transmembrane proteins.[34] These may be advantageous for the virus as they form stable structures to support membrane-related functions at a minimal burden on genome size. Indeed, short transmembrane proteins were identified in a range of both RNA and DNA animal viruses and were shown to be involved in various viral processes including entry, genome replication, particle assembly and release as well as interfere with host processes. Some of these transmembrane proteins belong to a group of proteins named viroporins[35,36] which contain hydrophobic regions that upon oligomerization form aqueous pores in the host cellular membranes. Viroporins were mainly implicated in viral assembly and release.[35] For example, a role in viral release was demonstrated for the human immunodeficiency virus (HIV) vpu protein. Vpu is a ~80 aa viroporins that oligomerizes to form a selective ion channel[37–39] that may enhance viral release through its ion channel activity and/or by inhibiting tetherin, a cellular factor which inhibits viral release.[40,41] Another extensively characterized viroporin is the 97 aa M2 influenza A virus (IAV) protein which oligomerizes to form an ion channel.[42] During infection M2 localizes to the Golgi where it plays an important role in viral replication and assembly by perturbing protein trafficking due to its effect on the ion gradient in the secretory pathway.[43] Furthermore, M2 is essential for viral release, mediating membrane curvature and scission.[44,45] M2 may also have a role in viral entry by mediating virion acidification[46] which is required for the uncoating of the virus.[47] Other short viral transmembrane proteins, which are not part of the viroporin family as they do not oligomerize, are vital for viral entry. The vaccinia 35 aa O3L protein has a hydrophobic domain and is incorporated into the membrane of the mature virion. Importantly, O3L associates with the virus entry/fusion complex and is essential for viral entry,[48] a function which is conserved across many other poxviruses.[49] Interestingly, vaccinia encodes many more short transmembrane proteins that were shown to have a role in viral entry and biogenesis.[34]

Short transmembrane proteins were also shown to have effects on host processes. By forming aqueous pores in cellular membranes viroporins perturb membrane permeability leading to alterations in cellular ionic homeostasis and thus to cytopathic effects.[35] In addition, some of the short transmembrane viral proteins interfere with various cellular processes. A key role for the HIV vpu proteins is the downregulation of CD4 by targeting newly synthesized CD4 molecules in the ER and mediating its proteosomal degradation. This process is critical for the virus as CD4 expression on the cell surface interferes with HIV viral propagation.[50] Short viral transmembrane proteins can also induce cell death. M2, the IAV viroporin, inhibits autophagosome degradation which compromises the survival of infected cells.[51] Viruses of the paramyxovirus family encode short hydrophobic (SH) integral membrane proteins 44–60 aa long, which were shown to inhibit apoptosis.[52,53] Papillomaviruses encode e5 proteins, ranging in size from 40 to 85 aa long transmembrane oncoproteins.[34] The well-studied Bovine papillomavirus E5 protein induces stable transformation of cultured fibroblasts by strongly and specifically activating the platelet-derived growth factor $\beta$ receptor (PDGF$\beta$-R). Human papillomaviruses E5 proteins were also shown to have some transforming capabilities as well as immune evasion function via downregulation of MHC class I.[54]

### 2.2. Upstream Open Reading Frames

uORFs are sORFs that due to their location upstream of a primary ORF could serve as means of regulating its translation and several examples were found in viruses. In human cytomegalovirus (HCMV), the viral UL4 protein translation is regulated by translation of a uORF.[55] This regulation is robust despite the inefficient usage of its AUG, probably through ribosome stalling during translation termination.[56] Remarkably, the coding information of the uORF is essential for translation inhibition of UL4 implicating that this regulation is mediated by nascent peptide translated from the uORF.[55] In another virus from the herpesviridae, Kaposi's sarcoma-associated herpesvirus (KSHV), ORF35 and ORF36 protein products are translated from a polycistronic transcript, ORF35–37. The translation of these two proteins is regulated by translation of two uORFs, which inhibit translation from the adjacent ORF35. Interestingly, the second uORF which overlaps ORF35 start codon allows translation of the downstream ORF36 via a reinitiation mechanism and is essential for viral propagation.[57,58]

A similar case was demonstrated for hepatitis B virus (HBV), where the polymerase gene is preceded by and partially overlaps the core gene, which is preceded by an upstream AUG, on the pregenomic RNA. Translation from the first AUG was shown to inhibit translation from the core initiation site while allowing reinitiation of translation from the polymerase initiation site.[59,60] Also in mouse hepatitis virus (MHV) a uORF that is present in many coronaviruses was shown to repress translation of the downstream ORF1.[61] mRNAs of the Ebola virus (EBOV) have long 5' UTRs, some of which contain upstream AUGs. Translation from the AUG preceding the L protein coding region, suppresses translation of the primary ORF encoding the L protein and is interestingly responsive to cellular stress. Mutations in this uORF drastically attenuate viral growth, indicating the importance of its regulatory function.[62] More complex regulation mechanism was reported in the simian immunodeficiency virus (SIV), where translation from a number of different uORFs present in different splice variants of mRNAs encoding the *rev* and *env* genes regulate these genes to different extent.[63,64]

Another translation regulation mechanism that is active in different viruses is ribosome shunting. Ribosome shunting is a mechanism in which cap-dependent translation starts from a short uORF located upstream of a long 5′ UTR stretch that forms a large stem and loop structure. Upon termination, the ribosome is able to bypass the stem and loop structure to resume scanning just 3′ of it, allowing translation from a downstream AUG.[33] This mechanism was extensively studied in the cauliflower mosaic virus (CaMV) 35 S RNA,[65,66] but was also demonstrated in additional viruses.[33]

Functional regulatory uORFs were discovered in different viruses from different families, both DNA and RNA and of different sizes, implying that this regulatory mechanism is probably wide spread. Importantly, these uORFs are found in other strains and sometimes other viruses from the same family, and furthermore, some were shown to be essential for viral propagation, demonstrating their functional importance.

### 2.3. Additional Functional Viral sORFs

Besides the two groups of functional sORFs described above, that have several examples in several viruses, there are additional examples of functional viral SEPs with various molecular and cellular roles. The HIV Vpr is 96 aa long, conserved between all HIV and SIV and important for viral replication. This small protein has been implicated in different processes during the viral life cycle, including reverse transcription, nuclear import of viral DNA in non-dividing cells, induction of cell cycle arrest and apoptosis.[67] The PB1-F2 ~90 aa long protein encoded by IAV localizes to mitochondria and also induces apoptosis.[68] Intriguingly, this conserved protein is translated from a +1 reading frame of the PB1 transcript which encodes one of the polymerase subunits.[68] In KSHV, a 3 kb polyadenylated RNA is transcribed from the opposite strand of the replication and transcription activator (RTA) encoding ORF50. It has been annotated as a non-coding RNA following identification because no large ORF was found in the transcript.[69] However, later this transcript was found to encode a 48 aa long peptide, designated viral small peptide 1 (vSP-1) which interacts with RTA and prevents its degradation through the ubiquitin–proteasome pathway, facilitating the virus gene expression and lytic replication.[70] A different example is the recent exciting discovery of a communication system in phages that relies on a small peptide secreted to the medium. A 43 aa long peptide is translated from the phage *aimP* locus and this sORF is then processed to form the mature short communication peptide.[71] Significantly, homologs of the *aimP* gene as well as other components of this communication pathway were found in many other *Bacillus* phages demonstrating the prevalence of this system.

### 2.4. Novel translated sORFs Discovered by Ribosome Profiling

As discussed above ribosome profiling has great potential for depicting the full variety of translated ORFs and has been successfully applied to a number of viruses in the past few years. HCMV was the first virus to be analyzed by ribosome profiling and revealed numerous previously unidentified ORFs, a large number of which are sORFs.[13] A class of uORFs located upstream of canonical ORFs were found, two of them were shown to downregulate translation efficiency of the downstream ORF using a reporter. Interestingly, changes in the 5′ ends of these transcripts along infection led to inclusion of the uORFs in the transcripts, therefore reducing translation from these ORFs at late stages of infection. Translation of many other sORFs was demonstrated, initiating within known ORFs or encoded by distinct transcripts. Multiple sORFs were found to be translated from transcripts previously annotated as non-coding, more than ten were translated from the long non-coding RNA b2.7, four of which are highly conserved across different HCMV strains. Translation of two short proteins encoded by RNA1.2 and additional sORFs were confirmed by mass spectrometry. Intriguingly, initiation of translation from a near-cognate codon was observed for both long and short ORFs. Similarly, in another DNA double stranded herpes virus, KSHV, ribosome profiling identified many uORFs which are widely spread in the KSHV genome.[72] Significantly, 24 out of the 85 annotated genes contained 1–6 uORFs, encoding peptides consisting of <100 aa, translating either in or out of frame and in many cases from a non-canonical start codon. The KSHV data also supported the existence and regulatory role of two uORFs previously described.[57,58] Ribosome profiling analysis was also done on the well characterized bacteriophage lambda during lysogeny and at different time points along the lytic process, revealing translation of tens of previously uncharacterized sORFs.[73] In vaccinia virus (VACV), a prototype poxvirus, ribosome profiling revealed translation from 596 unannotated ORFs that add to the 162 annotated ORFs.[74] Many of these novel ORFs are sORFs and include uORFs, truncated ORFs that resulted from translation initiation from non-annotated initiation site, ORFs that result from frameshifting, and ORFs from regions annotated as non-coding. In MHV, an RNA virus of the Coronavirus family, triplet phasing of the ribosome profiling data allowed precise determination of translated reading frames, revealing several translated sORFs upstream of, or within, annotated virus protein coding regions, some of which are conserved in different MHV strains.[14] One of them, a previously reported uORF upstream of ORF1[61] was confirmed and furthermore, a potential role for it in temporal regulation of replication protein synthesis was suggested.

Overall, all ribosome profiling analyses performed in viruses revealed a wealth of unknown translated sORFs, indicating that this is a prevalent phenomenon and demonstrating the discovery power of this technique. The fact that an ORF is being translated does not necessarily prove that it is functional, however there is a growing list of functional sORFs, thus it is likely that several of these novel translation units indeed have a role. Many of the peptides produced may be non-functional and may be rapidly degraded, nevertheless, ribosome association or the act of translation may have a role as has been shown for many uORFs. Significantly, dissecting the variety of viral peptides produced during infection is important, as even the non-functional peptides could be an important part of the immunological repertoire of the virus as major histocompatibility complex (MHC) class I bound peptides. This is evident from the robust cellular immune response that was reported for T cells from human HCMV-positive donors to peptides translated from several novel sORFs identified by ribosome profiling, including some that are translated from the beta 2.7 transcript, a designated long non-coding RNA.[75]

### 2.5. Analysis of Putative sORFs That Were Discovered by Ribosome Profiling

Since many putative sORFs were identified by ribosome profiling in double stranded DNA viruses—HCMV,[13] KSHV,[72] and VACV[74]—we set out to examine whether their sequence can

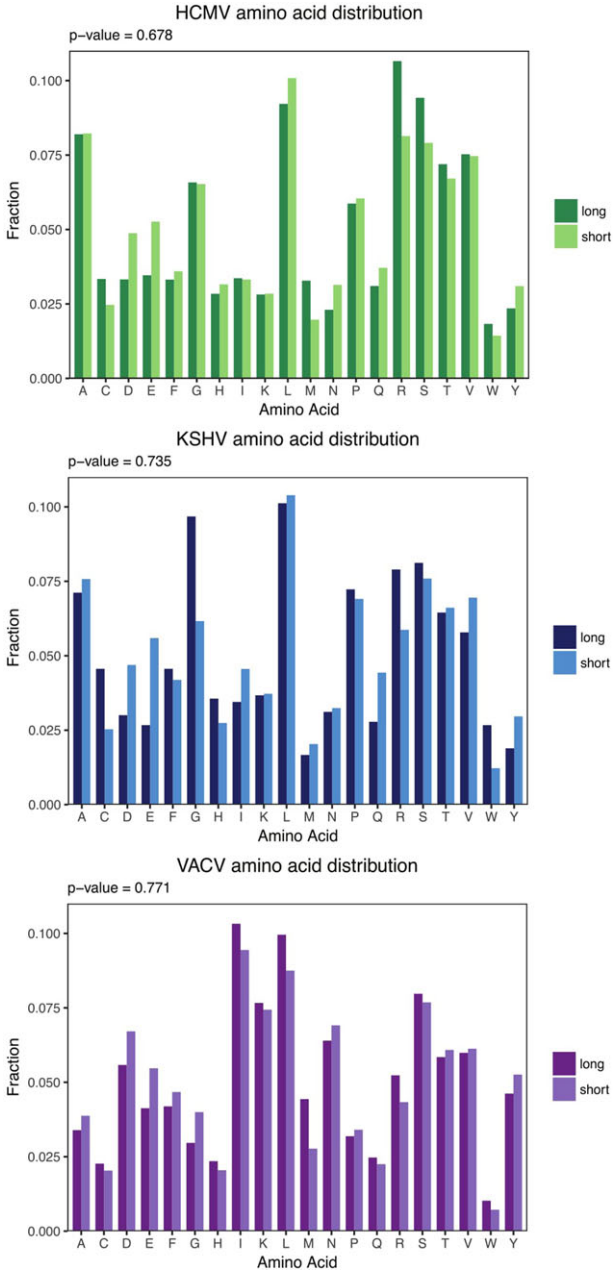


**Figure 2.** Distribution of amino acids in short and long ORFs of several DNA viruses. The frequency of amino acids in all short (20 aa> and <100 aa, light color) and long (>100 aa, dark color) viral ORFs is shown for HCMV, KSHV, and VACV. The statistical significance of the differences in amino acid distribution was calculated using Composition Profiler's relative entropy function, and the *p*-values are presented.

indicate anything about their potential functions. To this end, we applied several sequence-based analyses that gave us a broad view on the properties of these sORFs.

First, we compared the amino acid composition of sORFs to the composition of long ORFs in the same virus using Composition Profiler,[76] showing that there is no significant difference between the groups (**Figure 2**). This similarity is probably driven by the GC content and the codon usage of each virus, and further suggests similar amino acid selection rates for long and sORFs and no enrichment for specific amino acids in sORFs.

As discussed above, there are many cases of secreted short proteins. In order to test if these newly discovered sORFs are enriched for signal peptides we used signalP 4.1, a neural network–based method that predicts the probability of the presence of a signal peptide in a protein sequence.[77] The ratio of sORFs that contained a signal peptide was compared to the distribution of these ratios in a set of random sequences of the same length and aa composition. The long canonical ORFs showed significant enrichment in signal peptides prediction compared to shuffled sequences ($p$-value $< 1 \times 10^{-5}$). In the sORFs, KSHV had no predicted signal peptides, while in the shuffled sequences some signal peptides were generated. In sORFs from HCMV and VACV, we found that signal peptides are weakly enriched compared to what is observed in the set of random sequences ($p$-value: VACV 0.016, HCMV 0.2, **Figure 3**A), suggestive of selection for sORFs which are destined for the secretory pathway.

We next used the TMHMM 2.0 to test viral sORFs for enrichment in transmembrane domains.[78] This analysis revealed that similarly to the signal peptide prediction, the long canonical ORFs had a higher number of predicted transmembrane domains compared to random sequences ($p$-value: VACV 0.02, HCMV and KSHV $< 1 \times 10^{-5}$) while for sORFs significantly high numbers of transmembrane domains were found only for HCMV ($p$-value $= 1.201 \times 10^{-5}$, Figure 3B).

Whether sORFs are translated as a regulatory mechanism or into a functional protein, they can generate peptides that will be presented on MHC molecules and be recognized by the immune system. To explore if there is any selection for the immunogenicity of viral sORFs we used netMHC 4.0, which is an artificial neural network trained platform for predicting MHC class I binding affinities of peptides, based on their sequence.[79,80] The number of fragments that strongly bind MHC-I (high binders) was counted for each ORF. To take into account the effect of the ORF length, we calculated the ratio between the number of high binders in each ORF and its random set as a function of its length. We calculated the spearmen correlation and although some specific sORFs showed deviation from the median of random sequences, we found no significant correlation (**Figure 4**), suggesting that MHC-I presentation is not a dominating negative selection force for sORFs translation in these viruses.

It is noteworthy that in ribosome profiling as with other high-throughput methodologies, different approaches may introduce background noise originating from technical issues and thus might lead to biases in the detection of ORFs. False detection of spurious ORFs as well as falsely missing real ORFs are bound to occur. The analyses presented above were performed on data from three ribosome profiling studies. These studies used similar experimental approaches to generate the libraries but different computational approaches to predict the translated ORFs. Stern-Ginossar et al. and Arias et al. used a machine learning approach and the false negative rate was assessed as 13 and 36% for the HCMV and KSHV data, respectively and 1% false positive rate for the HCMV data.[13,72] Yang et al. used a rule-based approach hence it is harder to assess the false discovery rate.[74] We estimate that the use of different computational approaches in these studies has no major impact on the analyses we performed, however, further studies defining new viral sORFs will

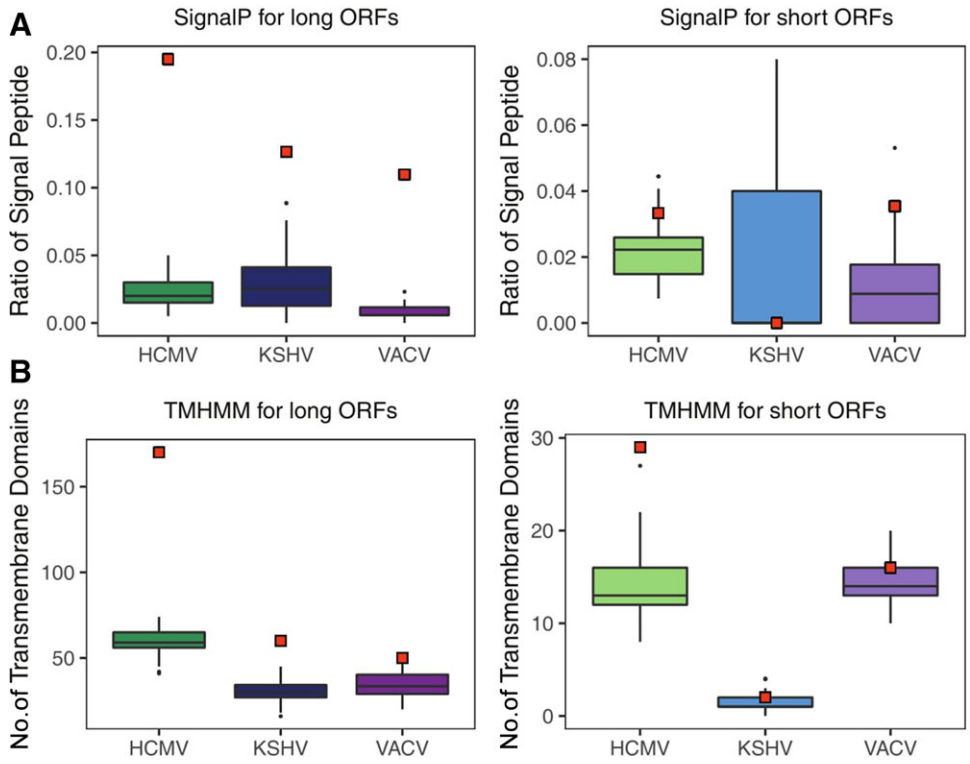


**Figure 3.** Functional predictions of non-canonical viral short ORFs that were discovered by ribosome profiling. Bioinformatic tools were used to predict protein characteristics for short and long viral ORFs, as annotated by ribosome profiling in HCMV, KSHV, and VACV. The results were compared to a distribution created by performing the same predictions of a set of 100 random sequences of the same size and aa composition. A) The presence or absence of a signal peptide in each ORF was predicted using SignalP. The fraction of signal peptide containing ORFs in each group is marked by an orange square, and the distribution of the sums in the random set is shown is shown in boxplots. B) The number of transmembrane domains in each ORF group was predicted using TMHMM. The sum of domains for each ORF group is shown as an orange square, and the distribution of the sums in the random set is shown in boxplots.

help to shed more light on the characteristics of these fascinating group of molecules.

## 3. Conclusions and Future Perspectives

Recent advances in computational and experimental techniques have revealed wide spread translation outside of canonical ORFs. It has been demonstrated that translated sORFs have essential roles during viral infection, however, the overwhelming majority of them remain to be characterized. To date, biological roles have been assigned to a small fraction of the translation products that have been mapped and a huge amount of work remains to be done to prove their existence and elucidate their functions. The outstanding challenge is to discriminate between ORFs that provide regulatory or protein-based functions to random translational events, and to identify their roles. Using sequence predictions, we show that viral sORFs are enriched for specific functional features, suggesting that some of these translation products may act at the protein level. Notably, differences between the functional enrichments we discovered in the different viruses can stem from genuine biological differences or may be due to variations in experimental approaches that were used to generate these datasets. Designing mutations that will distinguish between the function of the translated product and the act of translation itself and studying their phenotypic consequences could provide an important platform for future studies. Advancement in gene editing strategies and use of reporters provide powerful strategies to study the effects of mutating these translated regions and validating their expression.

Importantly, widespread translation outside of annotated protein coding genes was also found in many organisms including mammalians cells. As for many molecular biology principles, viral infection could provide a powerful model for studying functions of translated sORFs, due to robust expression levels and quantifiable phenotypes.

## 4. Methods

### 4.1. Data

Sequences of viral sORFs (encoding for peptides shorter than 100 aa) were obtained from three published papers presenting ribosome profiling data of DNA viruses[13,72,74] 515 HCMV, 50 KSHV, and 506 VACV sORFs were analyzed. Sequences of all previously annotated KSHV and VACV proteins longer than 100 aa were obtained from Uniprot manually annotated protein lists.[81] The full list of HCMV proteins longer than 100 aa of HCMV was taken from Stern-Ginossar et al.[13]

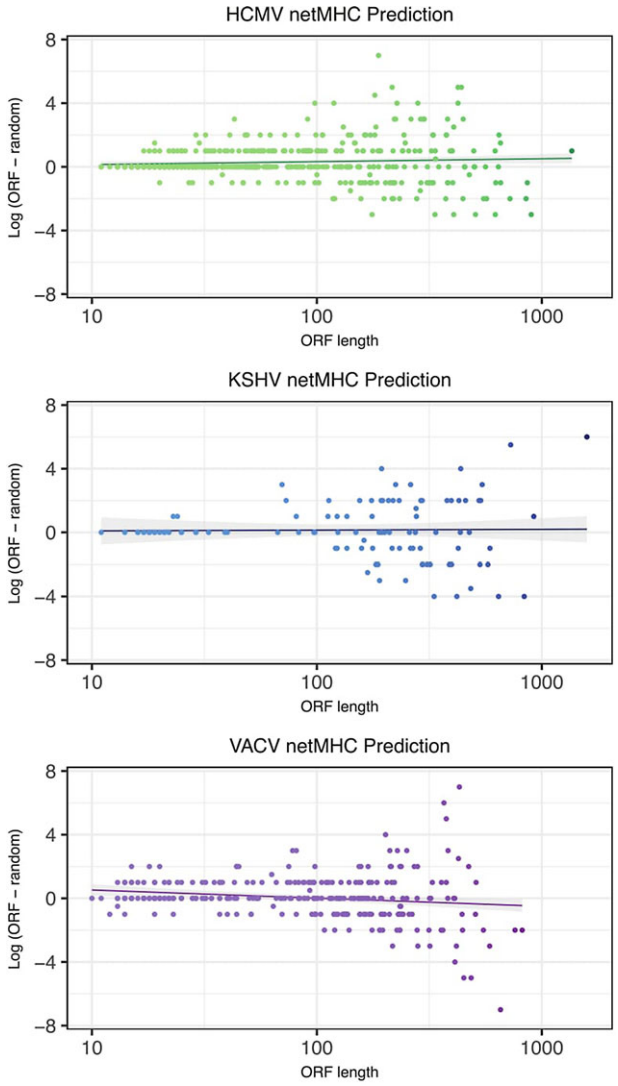


**Figure 4.** Prediction of MHC-I peptide presentation of viral ORFs. For each ORF, the number of nine aa long fragments that have the ability to strongly bind MHC-I molecules was predicted using netMHC. The difference between the number of predicted peptides in each ORF and the median of predicted peptides in a set of random sequences of the same size and aa composition is presented for HCMV, KSHV, and VACV.

### 4.2. Analysis

We calculated the frequency of amino acids in short and long ORFs for each virus. This was done by dividing the number of appearances of each amino acid in the sequences by the total combined length of the sequences. To test the significance of the difference between the frequencies in the short and long ORFs the relative entropy function in the Composition Profiler web-based tool was used, designed to compare aa distributions.[76]

Predictions of signal peptides and transmembrane domains were performed using SignalP[77] and TMHMM[78] web-based tools, respectively, using default parameters. Additionally, we used netMHC[80] to predict the number of all possible putative peptides that efficiently bind MHC-I molecules, using the recommended peptide length—9 aa. The predictions were done on the sets of short and long ORFs for each virus. For statistical comparison, the same tests were performed on random sets containing 100 versions for each sORF, created by shuffling the order of amino acids.

## Acknowledgements

N.S.G. acknowledges funding from the Israeli Science Foundation (1073/14), the European Research Council starting grant (StG-2014-638142), and Marie Curie career integration grant (2013-631003).

## Conflict of Interest

The authors have declared no conflict of interest.

## Keywords

micropeptides, ribosome profiling, translation, uORF, virus

Received: June 26, 2017
Revised: November 6, 2017

[1] S. J. Andrews, J. A. Rothnagel, *Nat. Rev. Genet.* **2014**, *15*, 193.
[2] J. I. Pueyo, E. G. Magny, J. P. Couso, *Trends Biochem. Sci.* **2016**, *41*, 665.
[3] C. Barbosa, I. Peixeiro, L. Romão, *PLoS Genet.* **2013**, *9*, e1003529.
[4] A. Gaba, A. Jacobson, M. S. Sachs, *Mol. Cell* **2005**, *20*, 449.
[5] P. Fang, C. C. Spevak, C. Wu, M. S. Sachs, *Proc. Natl. Acad. Sci. U. S. A.* **2004**, *101*, 4059.
[6] T. Kondo, S. Plaza, J. Zanet, E. Benrabah, P. Valenti, Y. Hashimoto, S. Kobayashi, F. Payre, Y. Kageyama, *Science* **2010**, *329*, 336.
[7] E. G. Magny, J. I. Pueyo, F. M. G. Pearl, M. A. Cespedes, J. E. Niven, S. A. Bishop, J. P. Couso, *Science* **2013**, *341*, 1116.
[8] J. I. Pueyo, E. G. Magny, C. J. Sampson, U. Amin, I. R. Evans, S. A. Bishop, J. P. Couso, *PLoS Biol.* **2016**, *14*, e1002395.
[9] N. T. Ingolia, S. Ghaemmaghami, J. R. S. Newman, J. S. Weissman, *Science* **2009**, *324*, 218.
[10] S. L. Wolin, P. Walter, *EMBO J.* **1988**, *7*, 3559.
[11] J. A. Steitz, *Nature* **1969**, *224*, 957.
[12] N. T. Ingolia, L. F. Lareau, J. S. Weissman, *Cell* **2011**, *147*, 789.
[13] N. Stern-Ginossar, B. Weisburd, A. Michalski, V. T. K. Le, M. Y. Hein, S.-X. Huang, M. Ma, B. Shen, S.-B. Qian, H. Hengel, M. Mann, N. T. Ingolia, J. S. Weissman, *Science*. **2012**, *338*, 1088.
[14] N. Irigoyen, A. E. Firth, J. D. Jones, B. Y. W. Chung, S. G. Siddell, I. Brierley, *PLoS Pathog.* **2016**, *12*, e1005473.
[15] G.-L. Chew, A. Pauli, J. L. Rinn, A. Regev, A. F. Schier, E. Valen, *Development* **2013**, *140*, 2828.
[16] A. A. Bazzini, T. G. Johnstone, R. Christiano, S. D. MacKowiak, B. Obermayer, E. S. Fleming, C. E. Vejnar, M. T. Lee, N. Rajewsky, T. C. Walther, A. J. Giraldez, *EMBO J.* **2014**, *33*, 981.
[17] S. Lee, B. Liu, S. Lee, S.-X. Huang, B. Shen, S.-B. Qian, *Proc. Natl. Acad. Sci. U. S. A.* **2012**, *109*, E2424.
[18] M. Fresno, A. Jiménez, D. Vázquez, *Eur. J. Biochem.* **1977**, *72*, 323.
[19] F. Robert, M. Carrier, S. Rawe, S. Chen, S. Lowe, J. Pelletier, *PLoS One* **2009**, *4*, e5428.
[20] T. Schneider-Poetsch, J. Ju, D. E. Eyler, Y. Dang, S. Bhat, W. C. Merrick, R. Green, B. Shen, J. O. Liu, *Nat. Chem. Biol.* **2010**, *6*, 209.

[21] N. Garreau de Loubresse, I. Prokhorova, W. Holtkamp, M. V. Rodnina, G. Yusupova, M. Yusupov, *Nature* **2014**, *513*, 517.
[22] C. Fritsch, A. Herrmann, M. Nothnagel, K. Szafranski, K. Huse, F. Schumann, S. Schreiber, M. Platzer, M. Krawczak, J. Hampe, M. Brosch, *Genome Res.* **2012**, *22*, 2208.
[23] X. Gao, J. Wan, B. Liu, M. Ma, B. Shen, S.-B. Qian, *Nat. Methods* **2014**, *12*, 147.
[24] A. P. Fields, E. H. Rodriguez, M. Jovanovic, N. Stern-Ginossar, B. J. Haas, P. Mertins, R. Raychowdhury, N. Hacohen, S. A. Carr, N. T. Ingolia, A. Regev, J. S. Weissman, *Mol. Cell* **2015**, *60*, 816.
[25] G.-L. Chew, A. Pauli, A. F. Schier, *Nat. Commun.* **2016**, *7*, 11663.
[26] A. Raj, S. H. Wang, H. Shim, A. Harpak, Y. I. Li, B. Engelmann, M. Stephens, Y. Gilad, J. K. Pritchard, *Elife* **2016**, *5*, e13328.
[27] L. Calviello, N. Mukherjee, E. Wyler, H. Zauber, A. Hirsekorn, M. Selbach, M. Landthaler, B. Obermayer, U. Ohler, *Nat. Methods* **2015**, *13*, 165.
[28] M. C. Frith, A. R. Forrest, E. Nourbakhsh, K. C. Pang, C. Kai, J. Kawai, P. Carninci, Y. Hayashizaki, T. L. Bailey, S. M. Grimmond, *PLoS Genet.* **2006**, *2*, 515.
[29] S. D. Mackowiak, H. Zauber, C. Bielow, D. Thiel, K. Kutz, L. Calviello, G. Mastrobuoni, N. Rajewsky, S. Kempa, M. Selbach, B. Obermayer, *Genome Biol.* **2015**, *16*, 179.
[30] K. Krug, S. Nahnsen, B. Macek, *Mol. BioSyst.* **2011**, *7*, 284.
[31] S. A. Slavoff, A. J. Mitchell, A. G. Schwaid, M. N. Cabili, J. Ma, J. Z. Levin, A. D. Karger, B. A. Budnik, J. L. Rinn, A. Saghatelian, *Nat. Chem. Biol.* **2012**, *9*, 59.
[32] B. Vanderperre, J. F. Lucier, C. Bissonnette, J. Motard, G. Tremblay, S. Vanderperre, M. Wisztorski, M. Salzet, F. M. Boisvert, X. Roucou, *PLoS One* **2013**, *8*, e70698.
[33] A. E. Firth, I. Brierley, *J. Gen. Virol.* **2012**, *93*, 1385.
[34] D. DiMaio, *Annu. Rev. Microbiol.* **2014**, *68*, 21.
[35] J. L. Nieva, V. Madan, L. Carrasco, *Nat Rev Microbiol.* **2012**, *10*, 563.
[36] C. W. Sze, Y. J. Tan, *Viruses* **2015**, *7*, 3261.
[37] K. Strebel, T. Klimbait, M. A. Martin, *Science* **1988**, *241*, 1221.
[38] S. Wang, B. Huang, Z. Wang, Y. Liu, W. Wei, X. Qin, X. Zhang, Y. Dai, *Dalton Trans.* **2011**, *40*, 12670.
[39] U. Schubert, A. V. Ferrer-Montiel, M. Oblatt-Montal, P. Henklein, K. Strebel, M. Montal, *FEBS Lett.* **1996**, *398*, 12.
[40] M. E. González, *Viruses* **2015**, *7*, 4352.
[41] N. Roy, G. Pacini, C. Berlioz-Torrent, K. Janvier, *Front. Microbiol.* **2014**, *5*, 177.
[42] R. T. Cardé, A. M. Cardé, A. S. Hill, W. L. Roelofs, *J. Chem. Ecol.* **1977**, *3*, 71.
[43] R. M. Pielak, J. J. Chou, *Biochim. Biophys. Acta* **2011**, *1808*, 522.
[44] J. S. Rossman, X. Jing, G. P. Leser, R. A. Lamb, *Cell* **2010**, *142*, 902.
[45] K. L. Roberts, G. P. Leser, C. Ma, R. A. Lamb, *J. Virol.* **2013**, *87*, 9973.
[46] J. Wang, J. X. Qiu, C. Soto, W. F. Degrado, *Curr. Opin. Struct. Biol.* **2011**, *21*, 68.
[47] A. Helenius, *Cell* **1992**, *69*, 577.
[48] P. S. Satheshkumar, B. Moss, *J. Virol.* **2009**, *83*, 12822.
[49] P. S. Satheshkumar, B. Moss, *J. Virol.* **2012**, *86*, 1696.
[50] M. Dubé, M. G. Bego, C. Paquay, É. A. Cohen, *Retrovirology* **2010**, *7*, 114.
[51] M. Gannagé, D. Dormann, R. Albrecht, J. Dengjel, T. Torossi, P. C. Rämer, M. Lee, T. Strowig, F. Arrey, G. Conenello, M. Pypaert, J. Andersen, A. García-Sastre, C. Münz, *Cell Host Microbe* **2009**, *6*, 367.
[52] Y. Lin, A. C. Bright, T. A. Rothermel, B. He, *J. Virol.* **2003**, *77*, 3371.
[53] R. L. Wilson, S. M. Fuentes, P. Wang, E. C. Taddeo, A. Klatt, A. J. Henderson, B. He, *J. Virol.* **2006**, *80*, 1700.
[54] A. Venuti, F. Paolini, L. Nasir, A. Corteggio, S. Roperto, M. S. Campo, G. Borzacchiello, *Mol. Cancer* **2011**, *10*, 140.
[55] C. R. Degnin, M. R. Schleiss, J. Cao, A. P. Geballe, *J. Virol.* **1993**, *67*, 5514.
[56] J. Cao, A. P. Geballe, *J. Virol.* **1995**, *69*, 1030.
[57] L. M. Kronstad, K. F. Brulois, J. U. Jung, B. A. Glaunsinger, *PLoS Pathog.* **2013**, *9*, e1003156.
[58] L. M. Kronstad, K. F. Brulois, J. U. Jung, B. A. Glaunsinger, *J. Virol.* **2014**, *88*, 6512.
[59] A. Chen, Y. F. Kao, C. M. Brown, *Nucleic Acids Res.* **2005**, *33*, 1169.
[60] L. Zong, Y. Qin, H. Jia, L. Ye, Y. Wang, J. Zhang, J. R. Wands, S. Tong, J. Li, *Virology* **2017**, *505*, 155.
[61] H.-Y. Wu, B.-J. Guan, Y.-P. Su, Y.-H. Fan, D. A. Brian, *J. Virol.* **2014**, *88*, 846.
[62] R. S. Shabman, T. Hoenen, A. Groseth, O. Jabado, J. M. Binning, G. K. Amarasinghe, H. Feldmann, C. F. Basler, *PLoS Pathog.* **2013**, *9*, e1003147.
[63] G. J. van der Velden, B. Klaver, A. T. Das, B. Berkhout, *J. Virol.* **2012**, *86*, 12362.
[64] G. J. van der Velden, M. A. Vink, B. Klaver, A. T. Das, B. Berkhout, *Virology* **2013**, *436*, 191.
[65] L. A. Ryabova, M. M. Pooggin, D. I. Dominguez, T. Hohn, *J. Biol. Chem.* **2000**, *275*, 37278.
[66] J. Fütterer, K. Gordon, H. Sanfaçon, J. M. Bonneville, T. Hohn, *EMBO J.* **1990**, *9*, 1697.
[67] C. A. Guenzel, C. Hérate, S. Benichou, *Front. Microbiol.* **2014**, *5*, 127.
[68] W. Chen, P. A. A. Calvo, D. Malide, J. Gibbs, U. Schubert, I. Bacik, S. Basta, R. O'Neill, J. Schickli, P. Palese, P. Henklein, J. R. R. Bennink, J. W. W. Yewdell, *Nat. Med.* **2001**, *7*, 1306.
[69] A. Saveliev, F. Zhu, Y. Yuan, *Virology* **2002**, *299*, 301.
[70] T. Jaber, Y. Yuan, *J. Virol.* **2013**, *87*, 3461.
[71] Z. Erez, I. Steinberger-Levy, M. Shamir, S. Doron, A. Stokar-Avihail, Y. Pekeg, S. Melamed, A. Leavitt, A. Savidor, S. Albeck, G. Amitai, R. Sorek, *Nature* **2017**, *541*, 488.
[72] C. Arias, B. Weisburd, N. Stern-Ginossar, A. Mercier, A. S. Madrid, P. Bellare, M. Holdorf, J. S. Weissman, D. Ganem, *PLoS Pathog.* **2014**, *10*, e1003847.
[73] X. Liu, H. Jiang, Z. Gu, J. W. Roberts, *Proc. Natl. Acad. Sci. U. S. A.* **2013**, *110*, 11928.
[74] Z. Yang, S. Cao, C. A. Martens, S. F. Porcella, Z. Xie, M. Ma, B. Shen, B. Moss, *J. Virol.* **2015**, *89*, 6874.
[75] N. T. Ingolia, G. A. Brar, N. Stern-Ginossar, M. S. Harris, G. J. S. Talhouarne, S. E. Jackson, M. R. Wills, J. S. Weissman, *Cell Rep.* **2014**, *8*, 1365.
[76] V. Vacic, V. N. Uversky, A. K. Dunker, S. Lonardi, *BMC Bioinformatics* **2007**, *8*, 211.
[77] T. N. Petersen, S. Brunak, G. von Heijne, H. Nielsen, *Nat. Methods* **2011**, *8*, 785.
[78] E. L. Sonnhammer, G. von Heijne, A. Krogh, *Proc. Int. Conf. Intell. Syst. Mol. Biol.* **1998**, *6*, 175.
[79] M. Andreatta, M. Nielsen, *Bioinformatics* **2015**, *32*, 511.
[80] M. Nielsen, C. Lundegaard, P. Worning, S. L. Lauemøller, K. Lamberth, S. Buus, S. Brunak, O. Lund, *Protein Sci.* **2003**, *12*, 1007.
[81] The UniProt Consortium, *Nucleic Acids Res.* **2017**, *45*, D158.

 