Dissertation
submitted to the
Combined Faculty of Natural Sciences and Mathematics
of the Ruperto Carola University Heidelberg, Germany
for the degree of
Doctor of Natural Sciences

Presented by
M.Sc. Shuai Ni
Born in: Jiuquan, China
Oral examination: 13th July 2020

The fate of RNA and RNA binding proteins in Sindbis virus infection

Referees:
Prof. Dr. Benedikt Bros
Prof. Dr. Jeroen Krijgsveld

Table of Contents

SUMMARY

RNA binding proteins (RBPs) accompany RNA throughout its whole life cycle. Therefore, the interaction of RBPs and target RNAs is particularly essential for post-transcriptional regulation. Not only can RBPs affect the RNA's expression, they can also control the localization, degradation, translation, and other activities of RNA. Capitalizing on recent advances in high-throughput sequencing, this thesis describes the use of transcriptomic and proteomic technologies to systematically study the interplay of RNA and RBPs under the context of viral infection. In brief, we infect the human cell line HEK293 with the Sindbis RNA virus, with the aim of demonstrating how the viral infection remodels the host transcriptome and proteome.

While it is commonly accepted that RBPs play a role in the regulation of gene expression, their contributions are still poorly understood. By using RNA interactome capture to track dynamic changes in RNA-binding proteome along the course of viral infection of Sindbis virus in human cells, we aim to assess the global impact of Sindbis virus infection on host transcriptome and proteome, and to identify host RBPs that interact with the Sindbis virus during its reproduction. This thesis reviewed the interplay dynamics between RNA and RBPs in human HEK293 cell line at three different viral infection stages. We observed a remodelling of binding activities of RBPs and the subsequent activation of the immune responses in the host cell. To our surprise, most RBPs demonstrating altered RNA binding did not show protein-level changes. Besides using statistical methods to evaluate the relative effects of different RNA processes, we also ddemonstrated that RNA degradation pathways had the biggest contribution to changes in RNA abundance change in SINV infected cells. Similar machinery may also apply to other alphaviruses, such as Chikungunya and Mayaro viruses, and thus we hope this study may contribute for the development of drugs to help solving public health problems caused by similar viruses in around the world.

Zusammenfassung

RNA-Bindungsproteine (RBPs) begleiten die RNA während ihres gesamten Lebenszyklus. Deshalb ist die Interaktion von RBPs und Ziel-RNAs für die posttranskriptionelle Regulation besonders wichtig. RBPs können nicht nur die Expression der RNA beeinflussen, sondern auch die Lokalisierung, den Abbau, die Translation und andere Aktivitäten der RNA kontrolieren. Diese Arbeit nutzt die jüngsten Fortschritte in der Hochdurchsatz-Sequenzierung und beschreibt den Einsatz von transkriptomischen und proteomischen Technologien zur systematischen Untersuchung des Zusammenspiels von RNA und RBPs im Kontext der Virusinfektion. Kurz gesagt, wir infizieren die menschliche Zelllinie HEK293 mit dem Sindbis-RNA-Virus, um zu zeigen, wie die virale Infektion das Transkriptom und Proteom des Wirts umgestaltet.

Es ist zwar allgemein anerkannt, dass RBPs eine Rolle bei der Regulierung der Genexpression spielen, aber ihre Beiträge sind noch immer wenig verstanden. Durch die Verwendung von RNA-Interaktom-Capture, um dynamische Veränderungen im RNA-bindenden Proteom zu verfolgen, entlang des Verlaufs der Virusinfektion des Sindbis-Virus in menschlichen Zellen, wollen wir die globalen Auswirkungen der Sindbis-Virusinfektion auf das Wirts-Transkriptom und Proteom abschätzen und Wirts-RBPs identifizieren, die während der Reproduktion des Sindbis-Virus mit diesem interagieren.In dieser Arbeit wurde die Interaktionsdynamik zwischen RNA und RBPs in der menschlichen HEK293-Zelllinie in drei verschiedenen viralen Infektionsstadien untersucht. Wir beobachteten eine Neumodellierung der Bindungsaktivitäten der RBPs und die anschließende Aktivierung der Immunantwort in der Wirtszelle. Zu unserer Überraschung zeigten die meisten RBPs, die eine veränderte RNA-Bindung zeigten, keine Veränderungen der Proteinebene. Neben der Verwendung statistischer Methoden zur Bewertung der relativen Auswirkungen verschiedener RNA-Prozesse zeigten wir auch, dass die RNA-Abbaustrukturen den größten Beitrag zu den Veränderungen der RNA-Häufigkeit in SINV-infizierten Zellen leisteten. Eine ähnliche Maschinerie kann auch auf andere Alphaviren wie Chikungunya- und Mayaro-Viren angewendet werden. Daher hoffen wir, dass diese Studie zur Entwicklung von Medikamenten beitragen kann, die zur Lösung von Problemen der öffentlichen Gesundheit beitragen, die durch ähnliche Viren in der ganzen Welt verursacht werden.

Acknowledgement

Firstly, I would like to express my sincere gratitude to my supervisor Dr. Bernd Fischer. It was an honor to be one of his Ph.D. students. The door to Bernd‘s office was always open and I miss our interesting and enlightening discussions. His modesty and enthusiasm for science encouraged me in all the time of my research. I could not imagine a better advisor and mentor for my Ph.D studies.

Besides my supervisor, I would like to thank the rest of my thesis committee: Prof. Benedikt Brors and Prof. Jeroen Krijgsveld, for their insightful comments and encouragement during my study. I am especially grateful for Benedikt‘s group members who warmly welcomed me during my last year in Heidelberg.

My deep appreciation also goes to Dr. Alfredo Castello, for his invaluable advice on my research and valued support during the various challenges I encountered. I recall how he once spent the whole morning on phone helping me go through materials for conference presentation, and it would not have been possible for me to have completed my doctoral work without him.

Next, I would like to thank my colleagues for all the fun we have had together on the Neckar river bank, in the old town bars, and during our excursions out of the city. Your stimulating discussions and valuable comments have also greatly improved this manuscript.

A very special gratitude goes out to Helmholtz International Graduate School at DKFZ, for helping me and providing the funding for our research.

Last but not the least, I would like to thank my mom and dad, for their unfailing support in every possible way throughout my years of study. This accomplishment would not have been possible without them. I would also like to thank my beloved wife Huang Xinxian. for all the sleepless nights and early mornings, for keeping me optimistic during my hard times. But most of all, thank you for being my best friend. I owe you everything. Thank you.

1. Introduction

The central dogma hypothesis, which was first proposed by Francis Crick in 1957, is now widely regarded as one of the most influential theories in biology. During the last decades, the content of central dogma has been profoundly expanded. RNA is one of central dogma's three essential macromolecules by virtue of its function storing and circulating genetic information between DNA and proteins. However, recent research suggests that RNA also takes part in various cell activities that involve its closest functional partners, the RNA binding proteins (RBPs). Capitalizing on recent advances in high-throughput sequencing, this thesis describes the use of transcriptomic and proteomic technologies to systematically study the interplay of RNA and RBPs under the context of viral infection. In brief, we infect the human cell line HEK293 with the Sindbis RNA virus, with the aim of demonstrating how the viral infection remodels the host transcriptome and proteome. In doing so we hope to uncover the fundamental processes driving remodelling and to provide an expanded view of mRNA-protein interactions in response to viral infection.

1.1) RNA and RNA-binding proteins

As an intermediate substance between DNA and protein, the primary cellular function of RNA is to transmit genetic information in the cell (Alberts B., 2002). A common metaphor considers the DNA genome as equivalent to a building projects blueprint, with the proteins as the functional 'bricks' that can do almost anything in the cell (Crick, F.H., 1958). In this process, RNA works as a transcript for each corresponding protein in the DNA genome. For an RNA to be transcribed and transported to the right place, the RNA must bond to several RBPs (Glisovic T., 2008). Thus, in most cases, the activity of RNAs in the cell are coupled with RBPs.

RBPs accompany RNA throughout its whole life cycle. Therefore, the interaction of RBPs and target RNAs is particularly essential for post-transcriptional regulation. Not only can RBPs affect the RNA's expression, they can also control the localization, degradation, translation, and other activities of RNA (Ross, A.F., 1997; Deshler J.O., 1998; Brewer G., 1991; Zhang, W., 1993;

Gualerzi, C.O., 1990). For example, the zip code binding protein (ZBP) recognizes a 54-nucleotide localization signal on β-actin mRNA which helps the β-actin messenger RNA (mRNA) to shuttle between a cell's nucleus and cytoplasm (Ross, A.F., 1997) (Figure 1). As most of RNA binding proteins are functional related, they represent a group of proteins with higher evolutionary conservativeness (Gerstberger, S., 2014). While the strategies that RBPs employ to bind to RNAs are not fully understood, there is evidence that alterations in RBPs expression or binding sites availability in target transcripts can contribute to human muscular atrophies, neurological disorders and various cancers (Castello, A., 2013). Therefore, the study of RNA-RBP interactions is critical for gaining a better understanding of gene expression regulation and RNA function in the cell. It is hoped a better understanding of the interplay of virus RNA and the host RBPs in our study will aid research and development efforts for a host of similar RNA viruses.

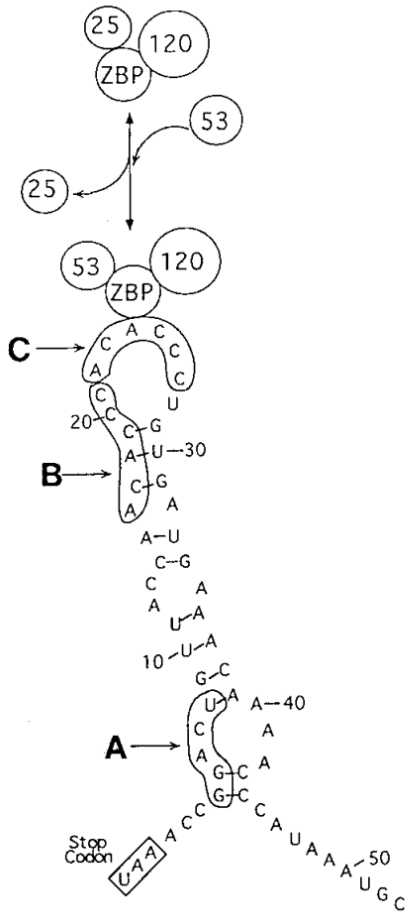

Figure 1, Hypothetical model of the ZBP protein binding to the 54 b-actin mRNA at the ACACCC motif, the binding site is proposed by mutating motifs in A, B and C regions. (Reprinted with permission from American Society for Microbiology, Ross, A.F., 1997)

1.1.1) RNA and its roles in the cell

The three main types of cellular RNAs are messenger RNA (mRNA), ribosomal RNA (rRNA), and transfer RNA (tRNA). Messenger RNA is the direct transcriptional product of RNA, and functions are an intermediate macromolecule between the DNA and the protein. Ribosomal RNA is the RNA component of ribosomes. There are four types of rRNA in humans that contributed to form either a large or small subunit. The 5S, 5.8S and 28S rRNA together forms the large subunit of ribosome, while the 18S takes part to assembl the small subunit. (Jeanteur, P., 1969; Aloni, Y., 1971; Aubert, M., 2018). The ribosome itself is formed from these two RNA subunits

along with their accompanying ribosomal proteins (Aubert, M., 2018). The process of translation is started by ribosomal assembly and attachment to the mRNA (Gualerzi, C.O., 1990). The tRNA molecules, each of which contains a 3-letter codon, then translate the mRNA sequence into a peptide sequence by carrying the codon-specific amino acid to the ribosome (Rich, A., 1976). This function makes tRNA essential to the process of mRNA translation.

Recent research has revealed that RNA has much broader functional and regulatory potential in the cell than anticipated. In 1967, Carl Woese et al., discovered that RNA could form a complex secondary structure which acted as a catalyst for specific biochemical reactions in the cell. Besides this catalytic potential, some types of RNA can also modulate translation (Cullen, B.R., 2004). These RNAs play a critical role in regulating gene expression, usually by suppressing protein synthesis by binding to the 3'UTR of a target mRNA (Zheng, B., 2017). Long non-coding RNAs (lncRNAs), so called because they typically surpass 200 nucleotides in length, are another type of recently discovered non-coding RNAs. While the function of more than 99% of lncRNAs are still unknown (Kung, J.T., 2013), described functions include recruitment and allosteric activation of proteins, recruitment of transcriptional regulators and inhibition of protein actions. (Long, Y.C., 2017). Moreover, in addition to being regulated by RBPs, some RNA may regulate the function of RBPs itself (Hentze M.W., 2018). These and other findings have revolutionized our knowledge of the function of RNA and indicate more diverse roles for RNA in the cell than previously thought.

Lastly, the genetic material in a large number of viruses is encoded by RNA rather than DNA, for instance in the case of influenza, severe acute respiratory syndrome and hepatitis C viruses (McGeoch, D., 1976; Marra, M.A., 2003; Takamizawa, A.C., 1991). The Sindbis virus used in this study is also an RNA virus. Using an RNA virus in our experiment helps us to quickly identify the set of RBPs that also potentially bind to viral RNAs.

1.1.2) RNA life cycle: from synthesis to decay

Multiple cellular processes carefully mediate the fate of an mRNA molecule throughout its lifespan (Bjork, P., L., 2015). During synthesis, an mRNA molecule will be adenylated, spliced,

and capped to become a mature mRNA (Peter, G., 1969). To be successfully translated into a protein, the mature mRNA will be then be transported from the nucleus to the cytoplasm. The mRNA can also be degraded to control the quality and the amount of mRNA in the cell (Bjork, P.L., 2015).

In eukaryotes, RNA synthesis is the process by which genetic information in DNA is transcribed into mature RNA (Geiduschek, E.P., 1961). It is orchestrated as a collaboration of multiple processes that is dependent on each other (Maniatis, T., 2002). It typically includes the following steps: the start of transcription, 5' capping, alternative splicing, and 3' polyadenylation (McCracken, S., 1997; Edery, I., 1985; Niwa, M., 1990). Multiple RNA binding proteins and RNA-protein complexes are involved in this process. The primary regulator of RNA synthesis is a multiprotein complex called RNA polymerase II (Hahn, S., 2004). RNA synthesis starts when RNA polymerase II attaches to the promoter region, which is located just upstream of the transcription start site (TSS) of the regulated genes (Hahn, S., 2004). The promoter contains a specific DNA motif that binds to RNA polymerase II to initiate transcription. RNA polymerase II then unwinds the double-strand DNA and starts to produce a complementary strand to the template DNA strand. The newly transcribed end of the transcript is called the 5' end. When the 5' end reaches a length of about 20–30 nucleotides, the capping enzyme complex, which sits on the RNA polymerase II, adds a 7-methyl-guanosine cap to the end (Rasmussen, E.B., 1993, Cho, E.J., 1997). The cap-binding complex is then attached to the capped region to protect the capped 5' end from degradation.

During transcription, the spliceosome, which consists of small nuclear RNA and splicing factors, recognizes intron-exon junction sites on each intron and cleaves the intron from the primary transcript (Will, C.L., 2011). After transcription is completed, a set of RBPs recognize and bind to a specific motif on the 3' end of the pre-mRNA, cleaving off a segment of the 3' end and adding multiple adenosine tail units (poly (A) tail) to the cleavage site. This process is called polyadenylation (Niwa, M., 1990). After polyadenylation, poly (A) tail-binding protein Pab1p binds to the poly (A) tail to form the protected 3' end (Zhao, J., 1999). After 5' end capping, splicing, and 3' end polyadenylation, the primary mRNA consists of a complex of mRNA and RNA binding proteins and is also called mature messenger ribonucleoprotein (mRNP).

The addition of the cap-binding complex and poly(A) tail binding protein Pab1p to both ends of the mRNA prevents the mRNA from been degraded. The mRNP is now ready to be transported from the transcription sites to the cytoplasm through nuclear pores formed by proteins assemblies termed Nuclear Pore Complexes (NPC). The mRNPs travel through NPCs mainly by diffusion (Cole, C.N., 2006) (Figure 2).

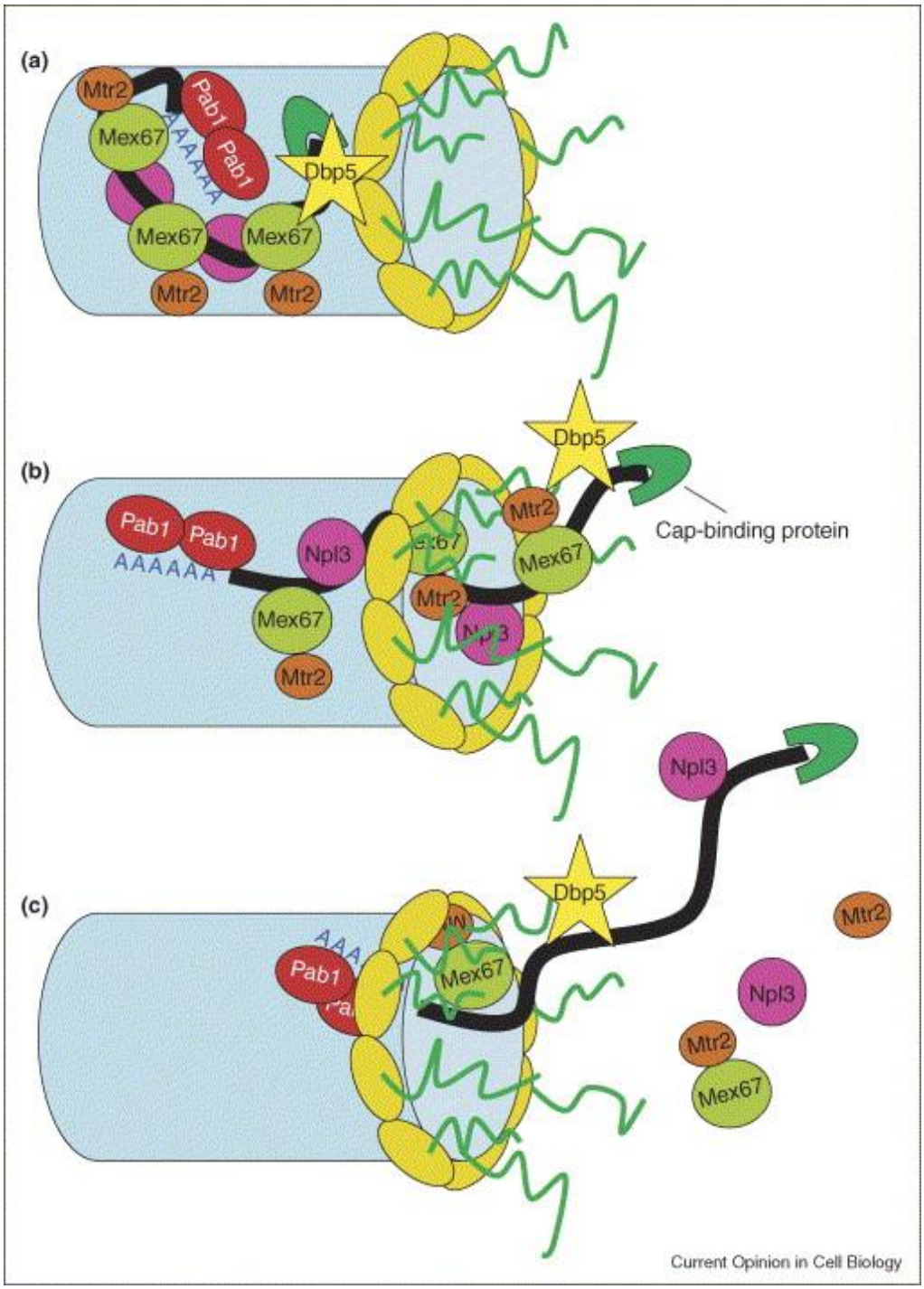

Figure 2, RNA binding protein Dbp5 facilitates the transportation of messenger RNA from the nucleus into the cytoplasm by going through the nuclear pore complex. (Reprinted with permission from Elsevier, Cole, C.N., 2006)

After mRNP is transported to the cytoplasm, it attaches to a ribosome and starts translation (Di Liegro, C.M, 2014). Translation starts when the ribosome assembles around the target mRNA and recruits the corresponding tRNAs. The tRNAs carry anticodons that are reverse complementary to each codon on mRNA. For each mRNA codon, the corresponding tRNA brings the codon-specific amino acid to the ribosome and leaves. The ribosome moves forward to the next mRNA codon in a 3' to 5' direction, linking up all amino acids carried in by tRNAs in the form of an amino acid chain, which can later be folded into a protein. The amino acid chain is released from the ribosome when one of 3 stop codons (i.e., UAG, UAA or UGA) on the mRNA is met.

The lifespan of mRNA in eukaryotic cells is controlled by a crucial post-transcriptional regulation mechanism called mRNA decay (Schoenberg, D.R., 2012). The mRNA decay can occur in different parts of a mature mRNA, it can happen from within the transcript sequence through endoribonuclease, or from both ends through exoribonuclease. Decay activities include decapping, deadenylation, exonucleolytic decay from both ends of the transcript, and endonucleolytic cleavage (Labno, A., 2016). It is also possible that the RNA molecule initiates the degradation even before the RNA is wholly synthesized (Morikawa, N., 1969). Various mRNA decaying mechanisms exist in the cell, mainly serve for two purposes, one that degrades transcripts explicitly with transcriptional error, preventing the production of potentially harmful proteins; and one which regulates the expression of normal mRNA as a post-transcriptional regulatory mechanism (He, F., 2015).

1.1.3) RNA binding proteins

RNA binding proteins bind to single or double-stranded RNA to form RNP complexes which exert various biological functions. The formation of these RNP complexes is prerequisite for RNA regulatory activities. Conventionally, RBP binds to RNA at the structurally well-defined RBDs, such as RNA recognition motif (RRM)3, hnRNP K homology domain(KH)4 or DEAD-box helicase domain (Figure 3). Later identification of novel RBPs, which is also called non-canonical RBPs, however, has shown that about half of the RNA binding sites are not canonical and map to intrinsically disordered regions of the protein (Castello, A., Fischer, B., 2016). These

Non-canonical RBPs lay various roles in RNA biology via alternative splicing and enzymatic activities, and are subject to posttranslational modifications such as lysine acetylation and tyrosine phosphorylation, adding another layer of regulation to RNA and RBPs interactions. (Castello, A., Fischer, B., 2016)

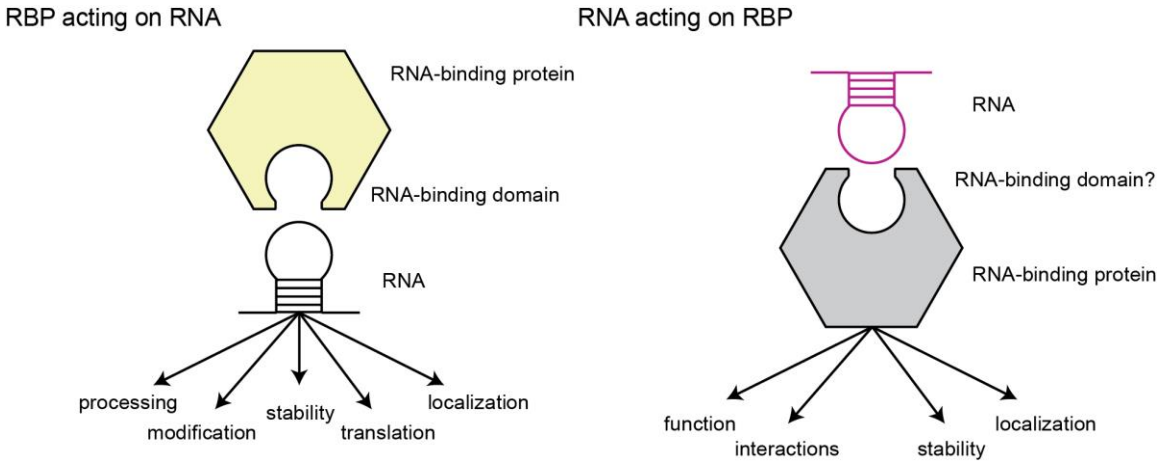

Figure 3, A typical view of RNA-RBP interaction and various RNA activities related to RNA binding. Canonical RBPs binds to RNA through an RNA-binding domain (Reprinted with permission from Elsevier, Hentze, M.W., 2018)

Multiple in vitro and in vivo approaches have been developed to screen the whole proteome for RBPs. (Baltz, A.G., 2012; Castello, A., 2017; Gerstberger, S., 2014). In vitro screening methods use an immobilized RNA or protein array as bait to incubate with cellular extracts for binding partners. In RNA arrays, the bound RBP is subsequently identified by quantitative mass spectrometry, while in protein arrays, cellular RNA is fluorescently labeled and a fluorescent intensity detector is used to identify different RBPs (Butter, F., 2009, Scherrer, T., 2010, Castello, A., 2016, Castello, A., 2017). In vivo methods use ultraviolet irradiation to cross-link neighboring RNAs and proteins in living cells, which are then identified using quantitative mass spectrometry. In terms of specificity, neither method is perfect; In vitro methods may also capture extraneous binding that does not appear in nature, while in vivo methods may identify indirect RNA-protein interactions. Nevertheless, both methods have contributed greatly to our knowledge of RBPs, with a recent compilation of the human RBPome reporting 1914 RBPs in total (Hentze, M.W., 2018)

Some RNA Binding Proteins are reported to have regulatory functions in viral infections, particularly in the case of RNA viruses (Li, Z., 2011). While the host cell dedicates more than 1,500 proteins to RNA metabolism, viral genomes typically encode only a dozen. RNA viruses have thus developed sophisticated mechanisms that hijack the host resources to support their replication (Leung, J.Y.S., 2011). To identify the important host RBPs needed for viral infection, we isolated the complete complement of RNA binding proteins (i.e. the Interactome) using "RNA interactome capture" across different viral infection stages, and compared these with the whole proteome at these time points.

1.2) The 'Omics' era in protein-RNA interactions

The rapid development of a new generation of high-throughput DNA sequencing technologies has produced a reduced sequencing cost and far greater accuracy. These developments have revolutionized biological research over the past years. From the sequence of the first individual human complete genome in 2007, to the first draft map of human proteome in 2014 (Levy, S., 2007, Kim, M.S., 2014), sequencing technologies have brought new perspectives to existing research fields, but also produced new research areas. As with the 'genome' and 'proteome', a collection of proteins that bind to RNA in a cell is termed an 'RBPome' or 'RNA interactome'. The RBPomes interplay with RNAs mediates post-transcriptional regulations in the cell.

The investigation of RNA-RBPs interactions has greatly benefited from advances in next-generation DNA sequencing and Mass-spectrometry based peptide sequencing. For instance, in interactome capture, identification of RBPome is determined by using mass-spectrometry to sequence the complete set of proteins in cross-linked protein complexes. On the other hand, if one is more interested in identifying the set of RNAs that bind to a protein, the RNA in the cross-linked complexes can be isolated and sequenced using 'CLIP' technologies (explained in 1.2.5). CLIP and RNA interactome capture (explained in 1.2.4) would not be possible without the recent advances in 'omics' technologies.

1.2.1) RNA sequencing

The popularity of RNA sequencing arises from the fact that knowledge of the transcriptome brings profound insights on the coding parts of a genome (Wang, Z., 2009a). RNA sequencing technology was developed in conjunction with the "sequencing by synthesis" approach in 2005, and is one of several "next-generation sequencing" techniques (Margulies, M., 2005, Matthew N.B., 2006). Further technical details of "sequencing by synthesis" are shown in Figure 4.

Prior to RNA sequencing, microarray hybridization was used to quantify RNA expression in parallel (Schena, M., 1995, Simon, M.D., 2013). However, this technology had certain disadvantages: First, microarray hybridization can only be used examine the expression of RNA with known sequences; Second, the accurate quantification of some sequences suffers from cross-hybridization with highly similar sequences; Third, the quantification is not likely to be exact in the case of extremely sparse or highly abundant transcripts (Casneuf, T., 2007). RNA sequencing, on the other hand, uses next-generation sequencing technology to directly determine the sequences from a complementary DNA (cDNA) library generated from a whole-cell transcriptome, providing a much higher resolution snapshot of the cell genetic makeup. As a consequence, RNA sequencing can address problems that befuddle traditional sequencing techniques, such as identifying alternative splicing events or new transcripts, discovering protein-coding mutations, and comparing allele-specific expression.

In our study, RNA sequencing is used to sequence the RNA level at Mock, 4 hpi, and 18 hpi. All types of RNA are sequenced in our experiment, including rRNA, mRNA, and non-coding RNAs. Since the Sindibis virus has an RNA genome, our study also examines the RNA production of the Sindbis virus.

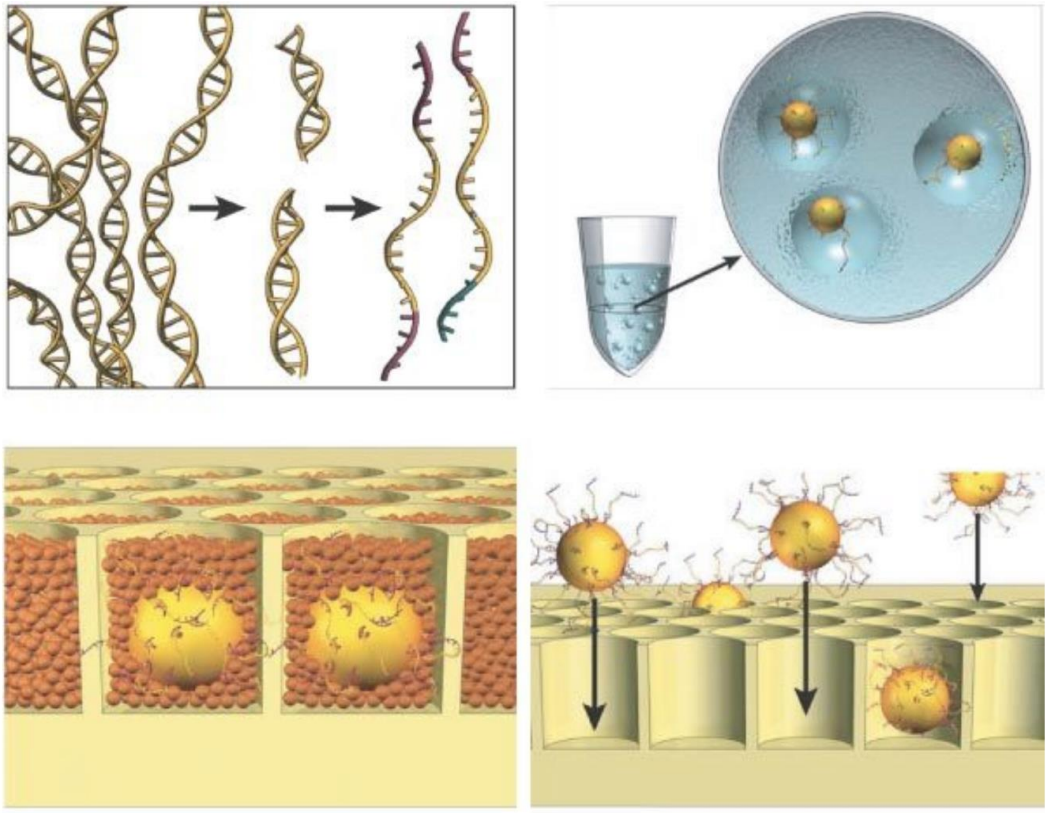

Figure 4. The widely used ‘sequencing by synthesis’ technique for high-throughput DNA sequencing. The genomic DNA is isolated, fragmented, and attached to beads that only allow one DNA fragment per bead. The beads are captured in oil emulsion droplets for PCR cloning of the attached fragments. Beads with cloned DNA fragments are then deposited into wells on a fiber optic slide for the PCR reaction. During a PCR reaction, the newly incorporated base will emit photons and pyrophosphates that can be detected for sequencing (Reprinted with permission from Springer Nature, Margulies, M., 2005).

1.2.2) Peptide sequencing

It is a little known fact that the development of the first robust peptide sequencing method, named Edman degradation (Edman, P., 1949), was invented 28 years earlier than the first universal nucleotide sequencing technique, Sanger sequencing (Sanger, F., 1977). The process of determining the peptide sequence was originally called ‘sequenator’ and gradually changed to the term ‘sequencing’ as nucleotide sequencing gained popularity in molecular biology research around the late 1980s. Edman’s sequenator works by repeatedly cleaving and identifying N-terminus amino acids. In each round of Edman degradation, the endmost N-terminus amino acid

of a peptide is cleaved and identified by electrophoresis or chromatography. However, the cellular proteome is complex entity with many types of protein modifications and intensities, hindering the efficiency of N-terminus cleavage, making Edman's approach ill equipped study the whole set of proteins in the proteome.

In 1981, another peptide identification method was developed that combined of ionization and mass spectrometry (termed the MS method) (Barber, M., 1981, Morris, H.R., 1981). Compared to Edman's sequenator, the MS method was more sensitive and had a much higher throughput. However, this method could only read short, fragmented pieces of proteins, and was not optimal for identifying a longer protein sequence as in Edman's approach. However, The MS method method did ease the way for future sequencing efforts: When whole-genome sequencing first emerged, the short peptide sequences uncovered using MS approaches were easily mapped to the existing genomic database, making the identification process much easier. (Kim, M.S., 2014)

The advances in MS-related technology gradually made MS an indispensable technique in proteome sequencing. In a classic proteome liquid chromatography-MS experiment, the purified protein sample has to be first digested into short peptide sequences, because the mass of the intact protein molecule exceeds the detection limit of the MS method. While different types of digesting enzymes with distinct cleavage preferences are available for protein digestion, in practice, trypsin is usually employed to do the task. Trypsin works by making a precise cut on the C-terminal side of Lysine and Arginine residue, producing peptides with appropriate lengths for MS detection and enhanced ionization properties (Huynh, M. L., 2009). After digestion, the mixture of peptides are separated according to their molecular size using high-pressure liquid chromatography (HPLC).The solution of separated peptide groups then sequentially goes through an ionization chamber, where they are ionized and gradually vaporize to generate isolated charged particles.

In the 1970s and 1980s much effort was devoted to the development of techniques that effectively ionized analytes in both liquid and solid forms (Macfarlane, R.D., 1976, Blakley, C.R., 1980, Heller, D.N., 1987, Dempster, A.J., 1921, Morris, H.R., 1981). Among these techniques, electrospray ionization (ESI) (Yamashita, M., 1984) and matrix-assisted laser desorption/ionization (MALDI) (Karas, M., 1987) drew most attention on account of their

advantages in sample mass limits and sensitivity. ESI was used to ionize samples in liquid form (Figure 5), while MALDI is used to ionize solid samples. These methods gradually become the most commonly used ionization techniques to date. In 2002, Fenn and Karas shared the 2002 Nobel Prize in Chemistry for development of these techniques.

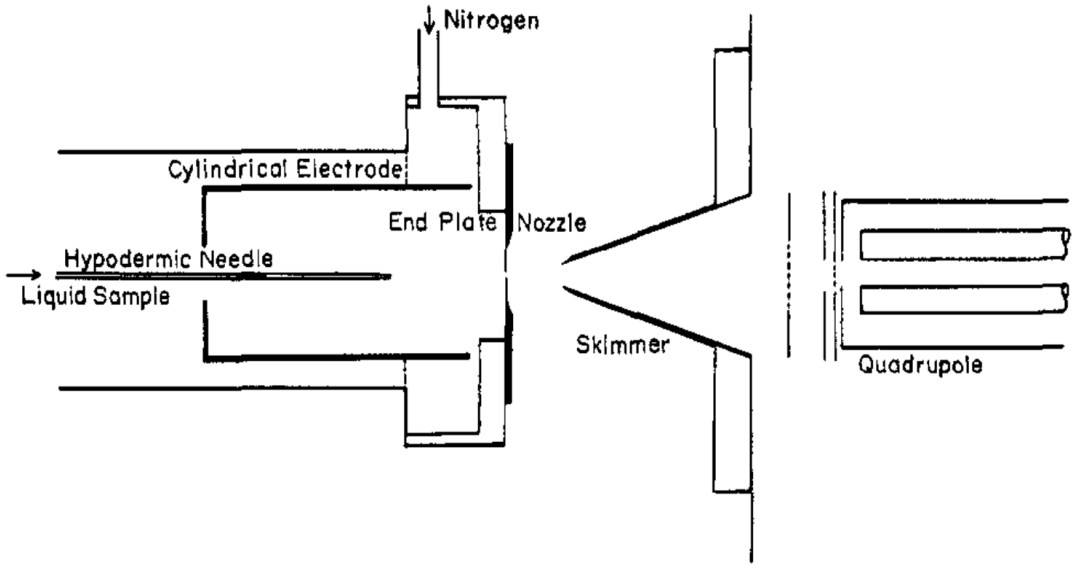

Figure 5, Schematic diagram of an electrospray. The liquid sample is sprayed from the needle, and then the droplets are evaporated under high voltage and temperature, leaving charged macromolecule ,bathed in gas for analysis. (Reprinted with permission from American Chemical Society, Yamashita, M., 1984)

After ionization, the ionized particles go through the analyzer, and enter either an electronic or magnetic field which alters their movement, allowing separation of particles based on their mass to charge ratio. For instance, time-of-flight mass spectrometry (TOF-MS) uses an electric field to accelerate the ionized particles, where the 'flight' time of a particle to the detector is related to the quality of an ion. When combined with HPLC, TOF-MS is ideal for the analysis of polar metabolites such as amino acids and other organic biological samples (Stewart, D., 2015). Here, the detector ascertains the amount of particles with different mass-charge ratio. The MS outcome for each peptide is a mass spectrum, which contains a measure of its relative abundance and the mass to charge ratio of all of its ionized particles.

1.2.3) Experimental and 'in silico' identification of RNA-protein interactions and RNA-binding domains

A range of methods have been developed to identify of RNA-protein interactions in the cell. These methods typically involve targeting a RNA of interest and using it as a bait to capture bound proteins bound. For example, proteins covalently bound to polyadenylated RNAs from cell lysis can be isolated by oligo(dT) selection and then identified using mass spectrometry based proteomic techniques. To date, over 1900 RBPs have been identified in humans (Hentze M.W., 2018) using protein microarrays (Tsvetanova, N.G., 2010, Scherrer, T., 2010, MacBeath, G., 2000) or target RNA pull-down followed by quantitative mass spectrometry (Butter, F., 2009, Tsvetanova, N.G., 2010, Treiber, T., 2017). In brief, a protein microarray is a set of fabricated proteins that are arranged in a grid pattern on a modified glass chip. These proteins can be probed using different sorts of fluorescently labeled RNA, and the signal can be spotted onto the microarray chip for RNA-protein immobilization. For example, Scherrer et al., used a yeast proteome microarray containing ~70% of the proteome (i.e., ~4,088 yeast proteins), to identify 180 mostly unannotated proteins that interacted with RNA (Scherrer, T., 2010).

Another approach for identifying RBPs is the use of stable isotope labeling of amino acids in cell culture (SILAC). In the SILAC method, two cell populations are fed growth medium containing amino acids labeled with non-radioactive (i.e.., stable) carbon-12 or carbon-13 isotopes (Ong, S.E., 2002). Non-binder proteins are expected to have an equal amount of control and bait eluate, thus having a heavy/light carbon isotope ratio around 1:1, while specific binders to the bait RNA will have different amounts in two populations, therefore, their heavy/light carbon isotope ratios are expected to be significantly differ from 1:1. By detecting differences in isotope abundance between control and bait eluate, Butter Falk and his co-worker were able to show that the HuR protein was a specific binder partner of Histone Deacetylase 2 (HDAC2) mRNA (Butter, F., 2009).

The more recent RNA interactome capture (RIC) methods has a higher sensitivity for RBP identification with lower background noise level compared with previous methods. RIC uses ultraviolet crosslinking of RBPs to RNA in vivo, followed by RNA pull-down and protein

identification using quantitative mass spectrometry. The crosslinking ensures robust binding of RBPs and RNAs, allowing for more stringent downstream purifying conditions that minimize contamination. The RNAs in pulled down mRNA-protein complexes are released by RNase treatment. The proteins are then cleaved into peptides and identified using MS (Castello, A., 2013). In two different studies, RIC yielded 860 and 791 RBPs from human HeLa and HEK293 cells, respectively (Baltz, A.G., 2012, Castello, A., 2012). A detailed description of RIC is shown in Figure 6.

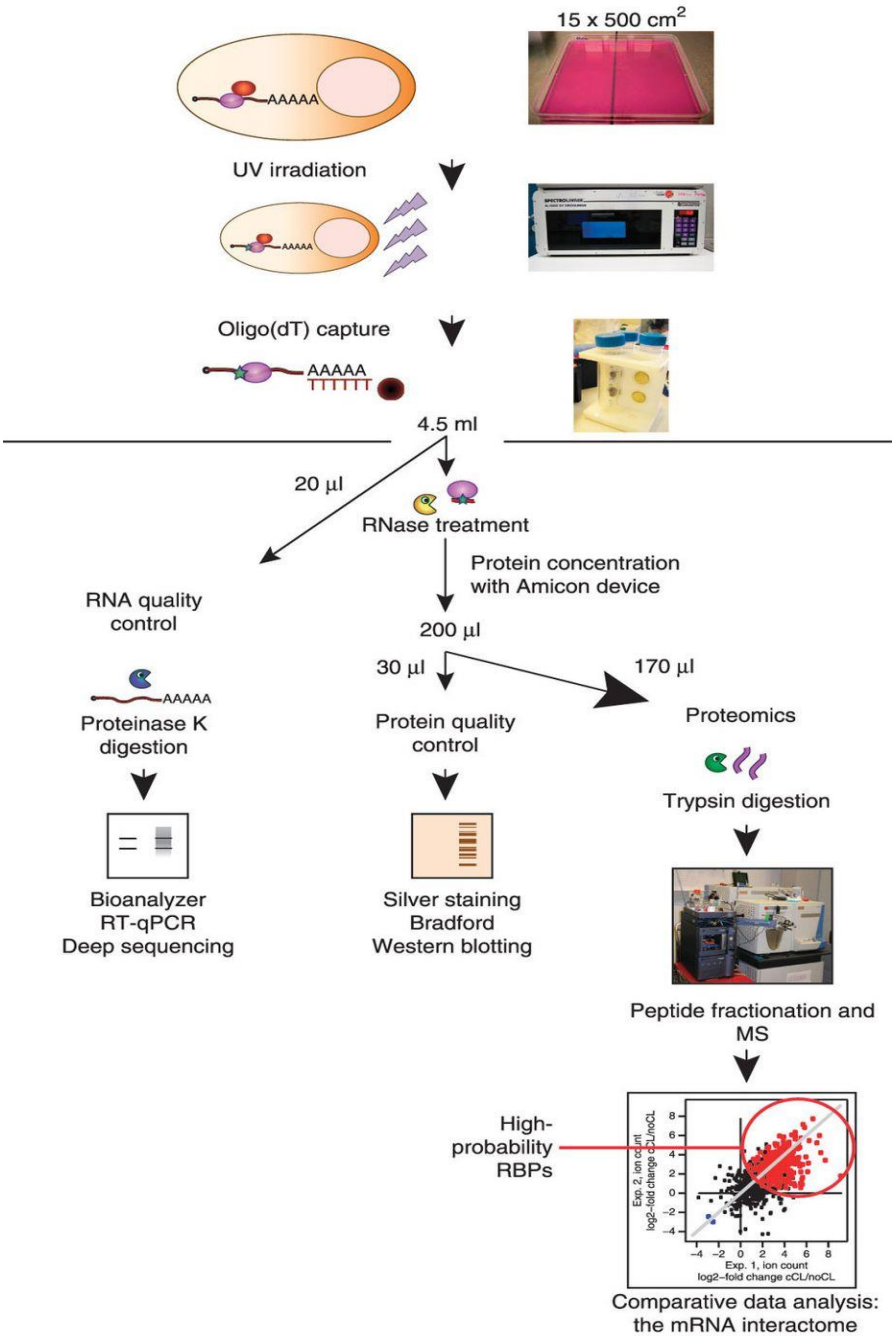

Figure 6, mRNA-protein interactions are preserved by performing UV cross-linking in vivo. Poly(A) RNA-protein complexes are captured by pull-down with oligo(dT) magnetic beads. Eluates are processed with proteinase K for RNA quality control, RNases for protein quality control, and RNases and trypsin for quantitative analysis. Comparative proteomic data analysis defines 'high-confidence' mRNA interactomes (Reprinted with permission from Elsevier, Castello, A., 2013).

To understand RNA-protein interaction and binding activities, both experimental and computational methods can be used to identify precise RBPs domains. A noteworthy experimental method which uses comparative proteomics was developed by Kramer K. et al. This method hydrolyses UV-cross-linked RNPs and their non-cross-linked controls using proteinase and nuclease enzymes that remove proteins and RNAs not in a binding position (Kramer, K., 2014). The resulting peptide–RNA complexes are then enriched and analyzed by ESI MS. Then, a quantitative comparison of peptide intensities in both conditions is conducted using a modified search algorithm that removes MS spectra that does not correspond to any spectra in the cross-linked samples. In Kramer's experiment, this allowed the identification of 257 cross-linked sites on 124 distinct RNA-binding proteins (Kramer, K., 2014). In 2016, Castello and co-workers published an extension of the RIC method that added a protease digestion step followed by a second round of oligo(dT) capture and mass spectrometry. They used this method to successfully identify 1,174 binding sites within 529 HeLa cell RBPs, as well as numerous RNA-binding domains (Castello, A., 2016).

In silico methods, on the other hand, make use of the protein sequence features from public databases (Wang, L., 2006, Terribilini, M., 2007) or the 3D structure of protein and binding RNAs to make predictions (Perez-Cano, L., 2010, Zhao, H., 2011). However, these prediction are complicated by the fact that about half of all RNA binding activities are driven by the intrinsically disordered regions (IDRs) of their constituent RBPs (Castello, A., 2016). That said, machine learning algorithms have shown a great potential for predicting binding sites given reliable training samples. My former colleague Mohamed Kammoun tested different machine learning algorithms including linear discriminant analysis, random forests, support vector classifiers, and convolutional neural networks to predict protein binding sites using an in vivo binding dataset of Hela cell (Kammoun, M., 2016). This dataset comprises of a collection of protein fragment sequences labelled as 'Bound' or 'Released' that represent binding or non-binding to RBPs. Then, using certain sequence features as input, the outcome of different algorithms can be evaluated to give the best performing model. We built an interactive web interface for predicted protein RNA binding-sites based on the best performing model. This application allows users to upload the protein sequences in a FASTA file format or simple copy and paste input sequences. Given this input the app generates a dataset with the probability for

each amino acid being evolved in binding and an accompanying visualization. The online version of the app "RBDetect" can be found in this link: (https://nishuai.shinyapps.io/RBDetect/). An example run of the app is shown in Figure 7.

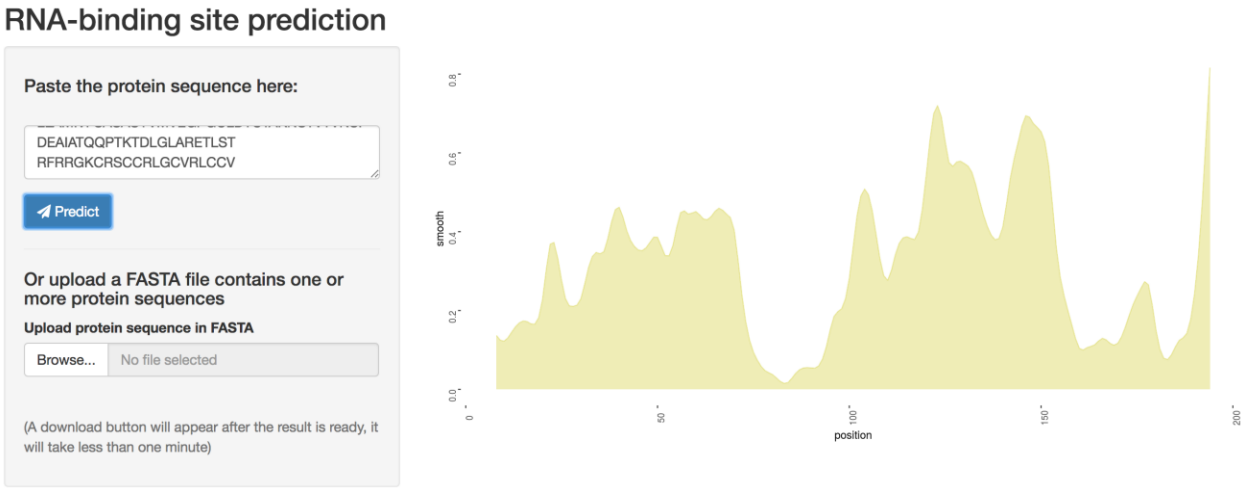

Figure 7, The web interface of "RBDetect" for RNA binding-sites prediction. The input panel on the left allows user to upload the input sequences. The plot on the right shows the prediction result from RBDetect. X axis represents the input sequence, Y axis represents the predicted probability for protein binding on each position.

1.2.4) Examining the RBP footprints on the RNA: the CLIP methods

Besides investigating proteins in this RNA-protein binding complex, the role of their RNA counterparts of RBPs are also a major interest for researchers. While there are several different approaches for identifying RNAs in the context of RNA-RBP interactions, the critical steps of remain the same. These are: 1) extraction of the protein-RNA complex of interest 2) isolation of the bound RNA and 3) further study of the now isolated RNA. CLIP methods are a collection of techniques that use ultraviolet cross-linking and immunoprecipitation for the extraction of the protein-RNA complexes, with the method itself is being named after the aforementioned analytical procedures. A variety of methods has arisen from different laboratories who have modified the classical CLIP approach in a bid to make it more precise and efficient.

In the next part of this thesis, we overview the pioneering attempts to study RNA-protein binding in using high-throughput sequencing. This technique employs immunoprecipitation of native endogenous mRNP complexes using a specific antibody for the protein of interest, which is termed RNA immunoprecipitation (RIP). The RIP is followed by purification and analysis of bound RNA using a cDNA microarray (RIP-chip) or with high-throughput sequencing (RIP-seq) (Tenenbaum, S.A., 2000, Brooks, S.A., 2000, Tenenbaum, S. A., 2002, Penalva, L., O. 2004). Compared with RIP, the CLIP methods have the advantage of initiating of cross-linking using UV 254-nm light prior to immunoprecipitation. This step ensures a more robust binding of RBPs and RNAs, enabling the discovery of kinetically unstable interactions and allowing for more rigorous downstream purification of the bound RNAs. For instance, cross-linking allows partial digestion of the non-binding part of RNA while retaining the core elements involved in protein binding, thus providing a more optimal way of locating binding sites on RNAs (Ule, J., 2003). Moreover, the separate digestion of single- and double-stranded RNA with specific nucleases in the absence of proteins allows inference of the secondary RNA structure adjacent to the binding site (Foley, S.W., 2016). The "clipped" RNA sequences are then investigated using a cDNA microarray or next-generation sequencing technologies. Compared with microarray techniques, next-generation sequencing allows for a more thorough inspection of bound RNAs, enabling the analysis of bound RNA on a genome-wide scale (Licatalosi, D.D., 2008). When coupled with high throughput sequencing, CLIP methods are referred to as CLIP-seq or high-throughput sequencing-CLIP (HITS-CLIP).

CLIP provides a feasible way for genome-wide profiling of binding RNAs and localization of the RBP recognition element (RRE) within target RNAs (Licatalosi, D.D., 2008; Van Nostrand, E.L., 2016). Hafner’s photoactivatable ribonucleoside enhanced cross-linking immunoprecipitation method (PAR-CLIP), further improves the UV cross-linking efficiency and the identification of RREs on bound RNAs (Hafner, M., 2010). PAR-CLIP utilizes a photo-active ribonucleic acid analog (4-thiouridine or 6-thioguanine) to insert nascent RNA transcripts into living cells, followed by cross-linkage using long-wavelength UV 365-nm light. The photoreactive nucleosides, which have not shown detectable toxic effects on cell growth, generate a characteristic sequence change upon cross-linking (T to C in 4-thiouridine and G to A in 6-thioguanine). This results in easy identification of cross-linking sites after the sequencing of

the isolated RNA fragments, and dramatically improves the resolution. Using PAR-CLIP, Ricardo et al., revealed that the DNA-binding protein CTCE (CCCTC-binding factor protein) could bind to a variety of RNAs in vivo and might regulate the p53 interaction with Wrap53 RNA (Saldaña-Meyer, R., 2014).

PAR-CLIP has certain advantages over CLIP method, but it also has some limitations. The insertion of nucleotide analogs is challenging to implement in animal models and clinical specimens, meaning the method can only be used at the cellular level. Furthermore, the insertion of analogs is reported to inhibit rRNA synthesis and cause a nucleolar stress response in cells, though it remains to be seen whether the effect is toxic enough to inhibit in vivo RNA binding activities (Burger, K., 2013).

The combination of high throughput sequencing with CLIP methods has the power to identify binding regions on a genome-wide scale. However, this method can only resolve about 30 nucleotides a time due to the current technical limits in high throughput sequencing (König, 2010). In their modified iCLIP method, König and colleagues reached a single-nucleotide resolution for the identifying binding regions, by taking advantage of the fact that during the identification of bound RNA after immuno-purification, the majority of reverse transcribing cDNAs will be truncated immediately before the cross-linked nucleotide (Urlaub, H., 2002). As shown in Figure 8, after immunoprecipitation and proteinase treatment, amino acid residues that have covalently attached to the RNA at the cross-link site remain in position. This blocks the transcription enzyme's activity and terminats the reverse transcription of the attached RNA, which is positioned precisely one nucleotide before the cross-linking site. iCLIP then captures the truncated cDNA, introducing two cleavable adapter regions with random barcode sequences, and evoking the circularization of the cDNA by annealing the two cleavage sites. The circularized cDNA library can then be sequenced using high-throughput sequencing (König, J.K., 2010). Using iCLIP, Wang et al., discovered that T-cell intracellular antigen 1 (TIA1) and TIA1-like 1 bind to the same positions on human RNAs, shedding light on the various functions of TIA1 and TIA1-like proteins in a cell (Wang, Z., 2010).

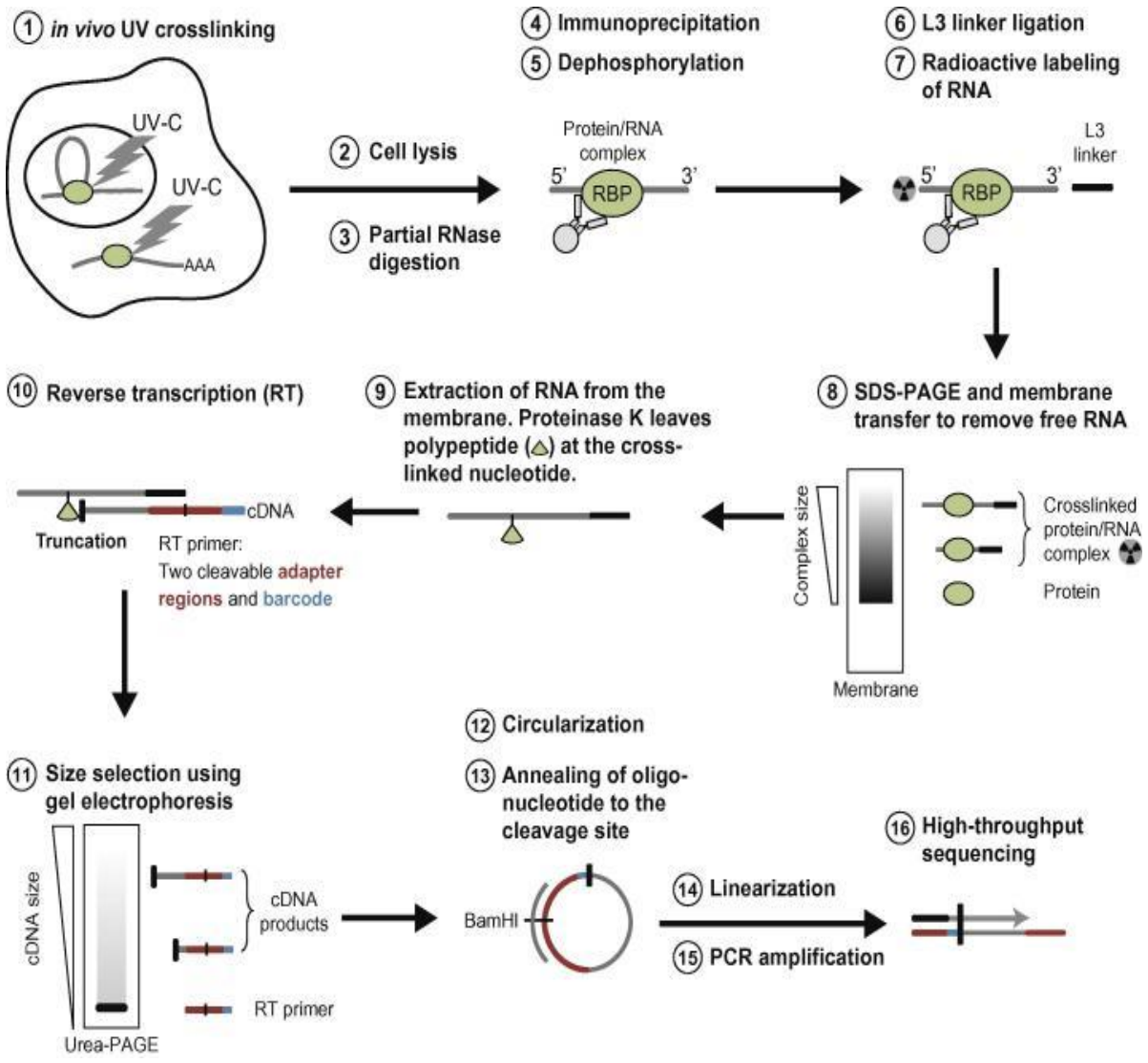

Figure 8, A schematic representation of the iCLIP protocol. After immuno-purification, the covalently linked RNA is cross-linked with the RBP and an RNA adaptor is added to its 3′ end. Proteinase K digests non-crosslinked parts of the protein, only leaving behind polypeptide fragments that are covalently attached to the RNA. Oftentimes, these residual fragments on the RNA will cause a truncation of reverse transcription (RT) at the cross-link site. The resulting cDNA molecules are circularized and prepared for high-throughput sequencing. The red bar shows the next nucleotide where cDNAs are truncated during reverse transcription (Reprinted with permission from Journal of Visualized Experiments, König, J., 2010).

Not only does iCLIP improves the resolution, but by introducing random tags on the adapters it reduces the effect of the PCR amplification preference. However, due to the stringent experimental conditions involved in iCLIP, laboratories often encounter technical problems generating the libraries, especially for those RBPs that lack a canonical RNA binding domain. Furthermore, the sequenced CLIP-seq libraries are often of extremely low complexity, which means a large proportion of CLIP-seq reads are actually PCR duplicates (Van Nostrand, E.L., 2016). The eCLIP method is a modified version of iCLIP, developed to improve the library complexity and simplify the iCLIP procedure. Ericet al., observed that circular ligation used in iCLIP is not as efficient as it could be. Therefore, rather than introducing the two cleavable adapters that later form the circularized cDNA molecules, the cDNA is ligated to single-stranded DNA adapters on the 3' end and converted to a linear double-stranded DNA library. As with iCLIP, the single-stranded DNA adaptor also contains a random-mer to distinguish it in case two identically sequenced reads arise from different RNA fragments or are generated as a PCR duplicate. Finally, the paired-end cDNA fragments are amplified and sequenced using high throughput sequencing (Van Nostrand, E.L., 2016). It was found that CLIP use decreases the required amplification by ~1,000-fold, decreasing discarded PCR duplicate reads by ~60% while maintaining single-nucleotide binding resolution. Consequently, eCLIP is a more suitable method when performing CLIP on a large scale.

1.2.5) The eCLIP database

A great deal of effort has been put into the improving methodological techniques for discovering RNA partners and pinpointing their precise binding regions to RBPs. These advances have allowed researchers to gain insights into the broader RNA regulatory network and determine functional roles associated with RNA-RBP interactions. To preserve the growing information generated in different labs with various CLIP methods, a database was required for the standardized sharing and archiving of information. The Encyclopedia of DNA elements (ENCODE) is a comprehensive database that logs various types of experimental data related to functional elements in the human genome (Consortium, E.P., 2012). ENCODE allows the user to upload identified RBP peaks determined with eCLIP so long as the experiment fulfills standard

requirements and follows a uniform data processing pipeline. In addition to eCLIP ENCODE also contains datasets generated by RIP-chip, RIP-seq, and iCLIP (Van Nostrand, E.L., 2016). The first eClip dataset in Encode was uploaded by Van Nostrand E.L. et al., who used eCLIP to determine 102 RBPs in K562 cells or HepG2 cells. In 2017, the eCLIP database contained more than 350 experiments, revealing upwards of 700000 RNS binding regions on RNA at single-nucleotide resolution, covering approximately 16.5% of the annotated mRNA transcriptome (Van Nostrand, E.L., 2018).

1.2.6) 'Omics' data analysis

Omics technologies have generated a considerable amount of biological data, making its storage and analysis more significant than ever before. Large international projects have pushed such as the "International Cancer Genome Consortium" (ICGC), "The Cancer Genome Atlas" (TCGA) and the "Human Cell Atlas" (HCA) (Zhang, J., 2011, Cancer Genome Atlas Research, 2013, Regev, A., 2017) have pushed need for data analysis pipelines even further. The ICGC and TCGA projects are expected to generate petabytes of data for thousands of whole genomes, while the HCA will map the transcriptome of billions of single cells in the human body. In fact the TCGA dataset, already harbors more than two petabytes publically available genomic data (Cancer Genome Atlas Research, 2013). In order to carry out practical in-depth research of this massive amount of omics data, customized data processing algorithms and statistical methods are essential for genomic, transcriptomic, or proteomic data analysis. As omics datasets can be prohibitively large, the analysis of such data requires significant computing resources leading researchers seek distributed computing methods to improve efficiency (Srimani, J.K., 2010). For example, The Genome Analysis Toolkit (GATK) makes use of a MapReduce framework for parallelized computing (McKenna, A., 2010). Tools such as these, which optimize the efficient mining of "omics" data will be essential for generating new breakthroughs in the field of bioinformatics (Dean, J., 2004, Quinlan, A.R., 2010).

Another important factor that can affect data processing efficiency is data quality, genomic sequencing data also suffers from quality issues. These issues usually come form multiple steps of biochemical reactions, which often lead to a certain level of uncontrollable bias. Biological

artifacts such as unstable enzyme activity, mismatch in reverse transcription, adaptor contaminations, and PCR amplification biases can have detrimental effects on the downstream analysis (Patel, R.K., 2012, Yang, X., 2013). The artifacts can result in poor quality reads, an excessive number of duplicated reads, sequence-specific duplication and so on. Handling these artifacts requires specific algorithms are used to control the quality of the sequencing data. For example, hash tables and heapsort algorithms are often used to efficiently detect unbalanced enrichment of specific sequences in a dataset (Marcais, G., 2011, Misra, S., 2011). Algorithm development is especially pertinent given the development of the single-molecule sequencing technologies by Pacific Biosciences and Oxford Nanopore, which produce a much longer readout and a lower base accuracy (Quail, M.A., 2012, Branton, D., 2008).

Genome and proteome data analysis often involves mapping the detected biological sequences to the reference genome or proteome. This step is needed to define where the sequences are originally located in the genome. Popular mapping applications generally use the hashing and/or the Burrows-Wheeler transform fast indexing algorithms. Hash-based methods build a hash table for the reference genome and compare it with hashed reads, or to first hash reads and then fit the hashes to the reference genome (Buhler, J., 2001, Lee, W.P., 2014, Li, H., 2008, Li, R., 2009). In another mapping algorithm SOAP, the reference genome and the sequencing reads are converted into a 2-bits-per-base encoding, and the number of mismatches between a read and the hashed reference is computed using bit operation (Li, H., 2008). Hash algorithms are often efficient for lookup of a match in constant time, but this come at the cost of high memory consumption. The Burrows-Wheeler transform, on the other hand, started out as a text compression algorithm. The higher the text repeatability, the greater the compression ratio, which overcomes the problem of large genome repeatability. For an exact sequence lookup using the Burrows-Wheeler transform, the match can be found within several calculations compatible with the length of a given sequence (Burrows, M., 1994). While, in theory, the lookup time is not as fast as for hash-based methods, this process consumes far less memory overall. As a consequence, software based on the Burrows-Wheeler transform is a bit more widely used in mapping.

A particular issue to be aware of when mapping RNAs to the reference genome is that of alternative splicing. In the process of transcription, genes are translated into pre-mRNA with all introns and exons from the original DNA sequence. To form a mature mRNA, all introns and sometimes one or more exons are cut out from the pre-mRNA molecule. This means that the mature mRNA preserves the order of exons but does not necessarily contain all exons from the gene. The length of introns in a gene ranges from 50 to 100,000 bases, so ordinary DNA mapping methods cannot distinguish mRNA reads that cross the exon-exon junction sites from regions in the reference genome, without addressing any skipped regions in the mRNA.

Therefore, mapping splicing reads to the reference genome might require a more sophisticated algorithm, but is normally worthwhile. The identified splicing reads provide direct evidence of alternative splicing events and their corresponding splicing sites. In general, two kinds of algorithms are designed to address this issue. The first constructs a reference library of all possible exon junctions for each transcript. This will, in theory, cover all possible junction patterns in mature mRNA. The mRNA reads can then be safely mapped to the library as unspliced DNA sequences. Because the 'junction library' size is usually much smaller than the size of the genome this "junction library" method can be used to find all known splicing events without consuming too much extra computing time and memory. However, this 'junction library' method is unable to find unknown exons. In contrast, the second algorithm maps the reads to the reference genome as unspliced DNA. Then, each unmapped read is split into smaller segments for another round of mapping (Wang, Z., 2009, C Trapnell, C., 2009). In the Spliced Transcripts Alignment to a Reference aligner (STAR) used in our analysis, all reads are mapped from one side to find the 'maximum mappable prefix,' which is used to realign the unmapped part of the read to the reference genome (Dobin, A., 2013).

One of the significant applications of RNA sequencing is the comparison of transcriptomic abundance levels in different conditions. This takes place in the following steps: First, after mapping RNA sequencing reads to the reference genome, the expression levels for each gene or isoform are estimated. Second, the mapped data is normalized and with the aid of statistical and machine learning methods and differentially expressed genes (DEGs) are identified. Finally, the relevance of the produced data is evaluated from a biological context (Bjork, P., 2015). Given

the increasing popularity of RNA-Seq technology a large choice of different software and analytical pipelines have now been developed to perform these types of analyses (Trapnell, C., 2009, Oshlack, A., 2010, Costa-Silva, J., 2017)

1.3) The intimate relationship between viruses and RNA

Humans have a long history of fighting with viruses. From the outbreak of swine flu in 2009, to the recent Ebola epidemic in 2013 (Centers for Disease Control and Prevention, 2012, Kaner, J., 2016), each outbreak of the virus pulls the world into a fresh wave of panic. While pandemics are rare, the resources spent on prevention in many countries often surpasses those spent on treatments. Depending on the primary genetic material, a virus can be classified into DNA or an RNA virus, with the majority of parasites that infecting humans, animals, and plants being RNA based (Domingo, E., 1997).

In contrast to DNA viruses, RNA viral replication can be considerably error-prone. This is mainly due to the absence of proofreading/repair and post replicative error correction mechanisms that tend exist in higher organisms (Domingo, E., 1996). This contributes to RNA viruses' possessing the highest mutation rate among all living beings (Drake, J.W., 1999. Moya, A., 2000). In addition, the probability of error increases with RNA virus's high replication rate. Indeed, it has been reported that a single infectious particle can produce, on average, 100,000 copies in 10 hours (Domingo, E., 1997). These features of the RNA virus contribute to its high adaptability towards environmental pressures at the population level. (Domingo, E., 1997)

1.3.1) Sindbis virus as a discovery model

The Sidnbis virus (SINV) is transmitted from mosquito to vertebrates, causing high fever, arthralgia (joint pain) and rash in humans. (Kurkela, S., 2008, Laine, M., 2004). Cases of SINV infection are mostly reported in northern Europe and South Africa. Although the fatality rates are rather low, it was recently found that 39% of patients had a chronic form of arthralgia that affected their daily lives even 6-8 months post infection (Gylfe, A., 2018). The Sidbis virus possesses a single-stranded RNA genome of approximately 11.7kb. It replicates in the cytoplasm

of the infected cell (Strauss, J.H., 1994), is highly tractable, and has a relatively well-understood life cycle that involves cellular factors common to other pathogenic alpha viruses such as the chikungunya virus and Venezuelan equine encephalitis virus (Carrasco, L., 2018). The SINV RNA genome was first sequenced in 1984. The genome is capped at its 5′ end and polyadenylated at the 3′ end, and the encoding part of the genome consists of about 11,700 nucleotides (Strauss, E.G., 1984). The genome contains two open reading frames separated by a UGA termination codon (Strauss, E.G., 1983). The first two-thirds of the genome from 5' end encode a non-structural polyprotein, which is later auto-catalytically cleaved into four non-structural proteins (Ding, M., 1989). The other third of the genome consist of a positive-sense RNA molecule generated from an internal promoter, that harbors a cap and a poly(A) tail, and is translated into structural proteins with the help of the non-structural proteins (Strauss, E.G., 1983) (Figure 9). SINV has been widely used as a model system in the laboratory to study viral translation and host viral interactions (Carrasco, L., 2018). In our study, we use SINV as a model for investigating if RNA binding proteins in the host cell are targeted by SINV to benefit its proliferation.

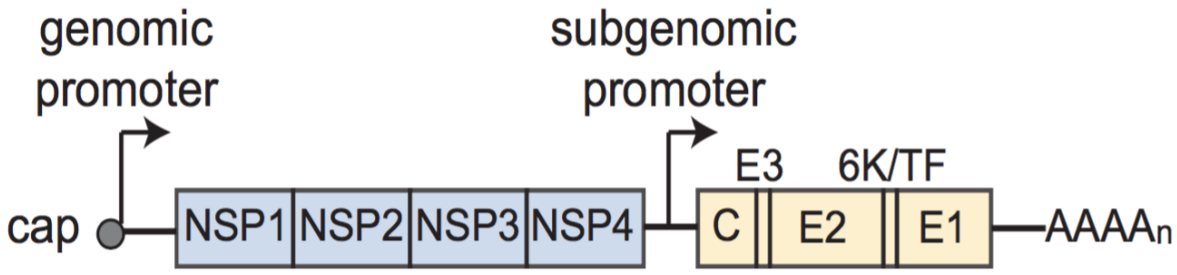

Figure 9, The structure of the Sindbis virus genome, The first two-thirds of the genome encode a non-structural polyprotein, the other third encodes structural proteins. The genome is capped at the 5′ end and polyadenylated at the 3′ end. (Garcia-Moreno, M., Mol Cell 2019)

1.3.2) Differential codon usage in virus infection

The translation of RNA to a protein requires a combination of three nucleotide bases which are collectively referred to as a codon. Codons determine the synthesis of an amino acid, and consist of a total of 64 combinations of three-base codons encoding 21 different amino acids in protein synthesis. Therefore multiple codons may encode a single amino acid. This phenomenon is referred to as the 'degeneration' of the genetic code (Pandit, A., 2011). Different codons that code for the same amino acid are called synonymous codons, while codons that code for different amino acids are called non-synonymous codons. Research suggest that in different organisms, synonymous codons are not used with equal frequencies; termed the 'codon bias' (Hershberg, R., 2008). The adaptability of a viruses to its host is significantly influenced by differences in codon usage bias (Kumar, N., 2016; Deka, H., 2014). Moreover, optimizing codon bias in vivo has been shown to enhance the efficiency and accuracy of protein expression (Zhou, J., 1999; Hershberg, R., 2008). Therefore, altering native codon usage might be a strategy for a virus to adapt to a new host. For example, the Asian lineage of the Zika virus (ZIKV) has reportedly been evolving during the second half of the 20th century to adapt to human codon usage bias to as a means of improving its fitness in the cell (Freire, C., 2018)

The codon usage preferences in the Sindbis virus and humans are markedly different. If there is a low concentration of preferred tRNAs that deposit appropriate amino acids for SINV, ribosome pausing or stalling of Sindbis virus replication is induced. Given reports that HIV can induce a reconfiguration of a cells tRNA pool to improve its translation efficiency, we wonder if SINV will behave similarly under the contest of intensive virus-host competitive pressure (Zhou, J., 1999).

1.4 Aim of study

While it is commonly accepted that RBPs play a role in the regulation of gene expression, their contributions are still poorly understood. By using RNA interactome capture (RNA-IC) to track dynamic changes in RNA-binding proteome along the course of viral infection of Sindbis virus in human cells, we aim to assess the global impact of Sindbis virus infection on host transcriptome and proteome and to identify host RBPs that interact with the Sindbis virus during its reproduction.

Previous studies have demonstrated that Sindbis virus infection can significantly inhibit the host RNA and protein synthesis on a global scale, however the specific functional processes exploited by the virus are not well characterized. Using external databases containing RBP-RNA interaction and innate RNA processing rates, we aim to discover the underlying regulatory mechanisms exploited by the virus to conquer the host cells defenses.

2. Proteome-wide analysis of RBP responses in Sindbis virus infected cells

While RBPs are critical for regulating gene expression during all viral infections, they are particularly important for RNA viruses. Given that the SINV genome only encodes seven proteins, it is impossible for the viruses own cellular machinery to fulfill all the tasks of RNA transcription, processing, and translation. Instead, SINV relies heavily rely on host RBPs for its replication in a host cell. It has been reported that SINV may induce a profound suppression of cellular protein synthesis in certain cell lines, as a means of inducing the protein-synthesizing machinery to preferentially translate viral RNA (Carrasco, L., 2018). Here, we aim to discover how virus infection changes the RNA-binding activities and functionality in HEK293 cells. More specifically, we address the following questions: 1) does SINV inhibit the host translation processes in HEK293 cells? 2) Are there any cellular RBPs exploited by SINV replication? 3) What is the potential functional relevance of the RBPs exploited by SINV replication?

2.1) Introduction and experimental design

To address these questions, we used 'RNA interactome capture' to isolate the "RNA binding proteome" at 4 and 18 hours post-infection (hpi) to capture different SINV infection stages, using uninfected cells as a control condition. The 4 and 18 hpi time points corresponded with two distinct infection stages of the SINV biological cycle; at 4 hpi, viral gene expression co-exists with host protein synthesis, while at 18 hpi the synthesized proteins are almost exclusively viral. In each condition, we took 3 biological replicates for more accurate and reliable measurement. Replicates of cell culture in each condition were labeled with three different amino acid isotopes using SILAC. This enabled the analyses of proteins from different replicates in one mass spectrometry run, minimizing potential batch effects. To correct for possible isotope-dependent effects in proteome MS quantitative analysis, SILAC labels in the three replicates were permutated between the three conditions (i.e., uninfected, 4 hpi and 18 hpi).

Next, labeled cells from the different infection stages were irradiated with UV light to induce covalent bonds between RNA and RBPs. The cells were then lysed, and RBP captured for

RBPome analyses (Figure 4). The lysed cells were then stored in aliquots for parallel transcriptomic and whole proteome analyses (Figure 10). Finally, equal amounts of lysate from each of the three conditions before the oligo(dT) capture (Figure 10), were combined and analyzed using quantitative proteomics.

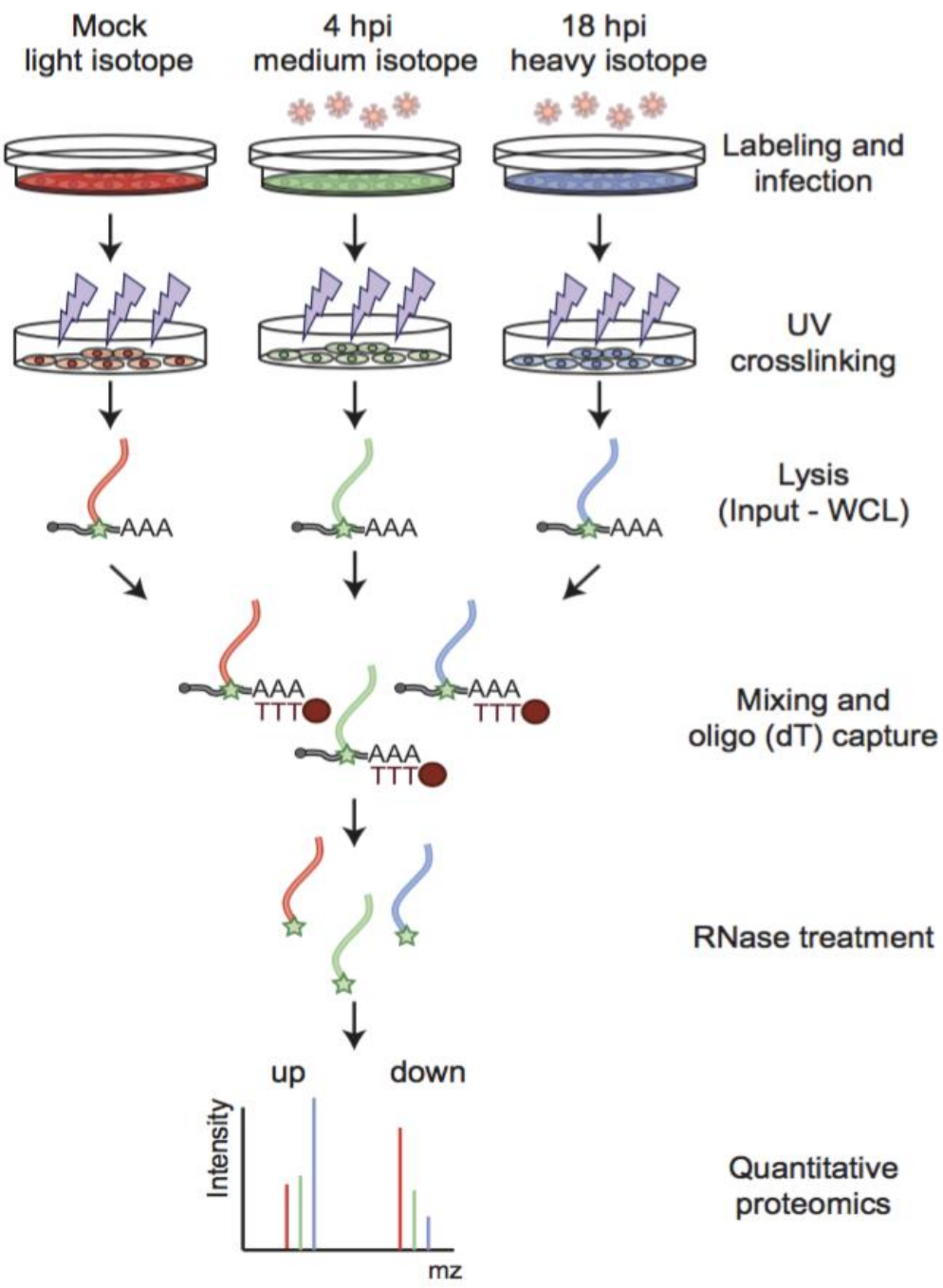

Figure 10. Schematic representation of RNA-IC combined with SILAC and virus infection. The isotope-labeled replicates are subjected to UV cross-linking to preserve the mRNA-protein interactions in vivo. Proteins bound to Poly(A) RNA are pulled down with oligo(dT) magnetic beads. Eluates are processed with proteinase K for RNA quality control, RNases for protein quality control, and RNases and trypsin for MS. (Garcia-Moreno, M., Mol Cell 2019)

2.2) Materials and methods

2.2.1) Generating human reference proteome and mapping peptide sequences

Human proteome and annotations were obtained from the R package mRNAinteractomeHeLa (http://www.hentze.embl.de/public/RBDmap/) (Castello. A., 2012). Specifically, the human reference proteome is downloaded from the Uniport database. The reference proteome was made by combining the Sindbis virus proteome with the Human proteome.

Mapping peptide sequences to the reference proteome was achieved using the 'mapPeptides' function in R package RBDmap (http://www.hentze.embl.de/public/RBDmap/). This function maps each peptide sequence in two steps. The first step indexes the reference proteome. This is done by sliding a window of 4-amino acid along all protein sequences in the reference proteome to generate a pool of all extant 4-letter subsequences in the proteome. Each 4-letter subsequence in the pool is indexed with a unique number. The number refer to the position of the protein in the reference proteome that contains that subsequence (Figure 11). The pool of 4-letter substrings is then converted into a hash table to ensure a search complexity of O(1), which means a 4-letter query sequence can always find the match in the pool with only a constant number of computational operations. Therefore, the output of indexing is a fast search hash table with all 4-letter amino acids each linked to a list of numbers indicating the position of their mother-protein in the reference proteome.

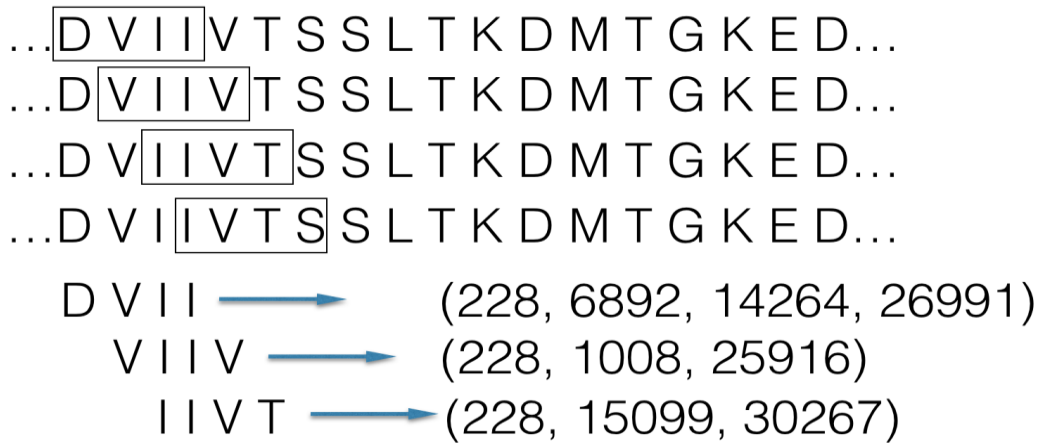

Figure 11. A window of 4-amino acids slides along the 228th protein sequence in the reference proteome. In addition to the 228th protein sequence, these 4-amino acids also appears in other protein sequences, as indicated by the numbers in the linked list.

The second step is to find matches for each query peptide sequence. The 'mapPeptides' function does this by extracting the first and last four amino acids from each query peptide as seeds, and using both seeds to locate proteins in the above-mentioned hash table. Only proteins concurrently matching both seeds are recorded for further investigation. After seed matching, the number of reference proteins that can potentially make a perfect match to the query peptide is greatly reduced. This means standard string matching can then be carried out to search for perfect matches to the query peptide in entire sequence.

This algorithm does not allow mismatches between the peptide sequence and the reference proteome, which means only perfect matches can be discovered. This is a good solution for the majority of cases because the query peptide sequence is often of much shorter length compared to the DNA sequencing output. One the other hand, allowing for mismatches may undermine the confidence of finding the exact sequence origin.

It is also possible to make modifications to the algorithm to allow for one mismatch during mapping. The cost is an increase in a tolerable amount of computing time during seed matching. Instead of having only two seeds, one can extract all possible 4-letter seeds from the query peptide sequence. For each seed, there will be a corresponding list of positions of the mother-

protein in the reference proteome. A perfect match ensures the position of the protein appears for all seeds. One mismatch will cause a 1 to 4 absence in the corresponding protein position. To counteract this, the lists of positions and identify positions that appear n to n-4 times are combined, n being the number of all possible 4-letter seeds from that peptide. It is safe to omit the string matching procedure as described in the original algorithm because the seed matching step guarantees to find all 1-mismatch and perfectly matched proteins.

The computing time of this modification largely depends on the average length of the query peptides, which determines n, i.e., the number of all possible 4-letter seeds generated. The increase in n will result in linearly increased computing time. Therefore, to avoid instances of a very large n from extremely long query peptide sequences, one can set a rule where, for query peptides extending more than 20 amino acids in length, 20 seeds will be randomly selected from the query sequence. This is followed by the same string matching procedure as in the original algorithm.

In our case, we only considered peptides that uniquely mapped to a single gene for downstream analysis. The mean intensity values of peptides mapped to the same gene are calculated to represent the intensity of that protein.

2.2.2) Proteome differential analysis

The aggregated mean log2-intensity ratio of each protein was tested for enrichment between the three biological replicates using a moderated t-test, which is implemented in the R/Bioconductor package Limma (Smyth, G.K., 2004). The p-values were corrected for multiple testing by controlling the false discovery rate using the Benjamini-Hochberg method. Results were visualized using the R package ggplot2 (Wickham, H., 2009).

For proteins whose intensity was ‘zero’ in one of the two conditions we applied a semi-quantitative approach which assumed that proteins without quantitative information were below the detection limit (Sysoev, V.O., 2016). The approach counted the number of replicates in each condition in which a given protein has an intensity value. When comparing two conditions and

three biological replicates, this leads to a matrix with 16 different groups (detected 0, 1, 2 or 3 times in condition one versus detected 0, 1, 2 or 3 times in condition 2). A protein is classified as a ‘dynamic RBP’ by the semi-quantitative method if an intensity value is assigned to it in 2-3 of the replicates in 1 of the 2 conditions, while only 1 or 0 intensity values are detected in the other condition. An accuracy assessment of semi-quantitative analysis for differential analysis can be found in 5.2.3.

The fraction of RNA-bound RBPs was determined by computing the ratio between the protein intensity of each RBP in the RNA-IC eluates at 18 hpi and then in the whole cell lysate. Hence, this calculation reflects the amount of protein crosslinked to RNA (RNA-IC), divided by the total amount of protein (whole cell lysate). It confirms that the changed binding for RBPs is not due to an overall change of protein abundance in the cell.

2.2.3) Gene set enrichment analysis

Gene set enrichment analysis is carried out by using package mRNAinteractomeHeLa (http://www.hentze.embl.de/public/RBDmap/) (Castello. A., 2012). Specifically, GO annotations are obtained from ‘GO.db’ package which contains a set of annotation maps describing the entire Gene Ontology. Gene set enrichment analysis is performed by applying Fisher's exact test to categories from Gene Ontology (GO) annotations with at least three annotated proteins. The functional enrichment of differently expressed proteins are carred out using Functional Enrichment of Significantly Changed Proteins (STRING).

2.3) Results

2.3.1) Dynamics of the RBPome in Sindbis infected cells

The RNA-IC experiments revealed that the RBPome of SINV-infected cells went through a pervasive remodeling. In particular, we identified a total of 794 RNA binding proteins; 91% of which were already annotated to the gene ontology term ‘RNA-binding’ or/and were previously

reported to be RBPs in eukaryotic cells by RNA-IC (Hentze M.W., 2018). Hence, the protein composition of our dataset closely resembles that of previously established RBPomes.

Most cellular RBPs remained unaltered at 4 hpi except for 17 RBPs (~2% of the identified RBPome) (Figure 12 and 13). Fifteen of these were detected exclusively by the semi-quantitative method due to the lack of protein intensity value in one condition, reflecting possible 'on-off' and 'off-on' states (Table S1). It is worth noting that the SINV capsid protein was already at high levels during this early stage of viral infection (Figure 12).

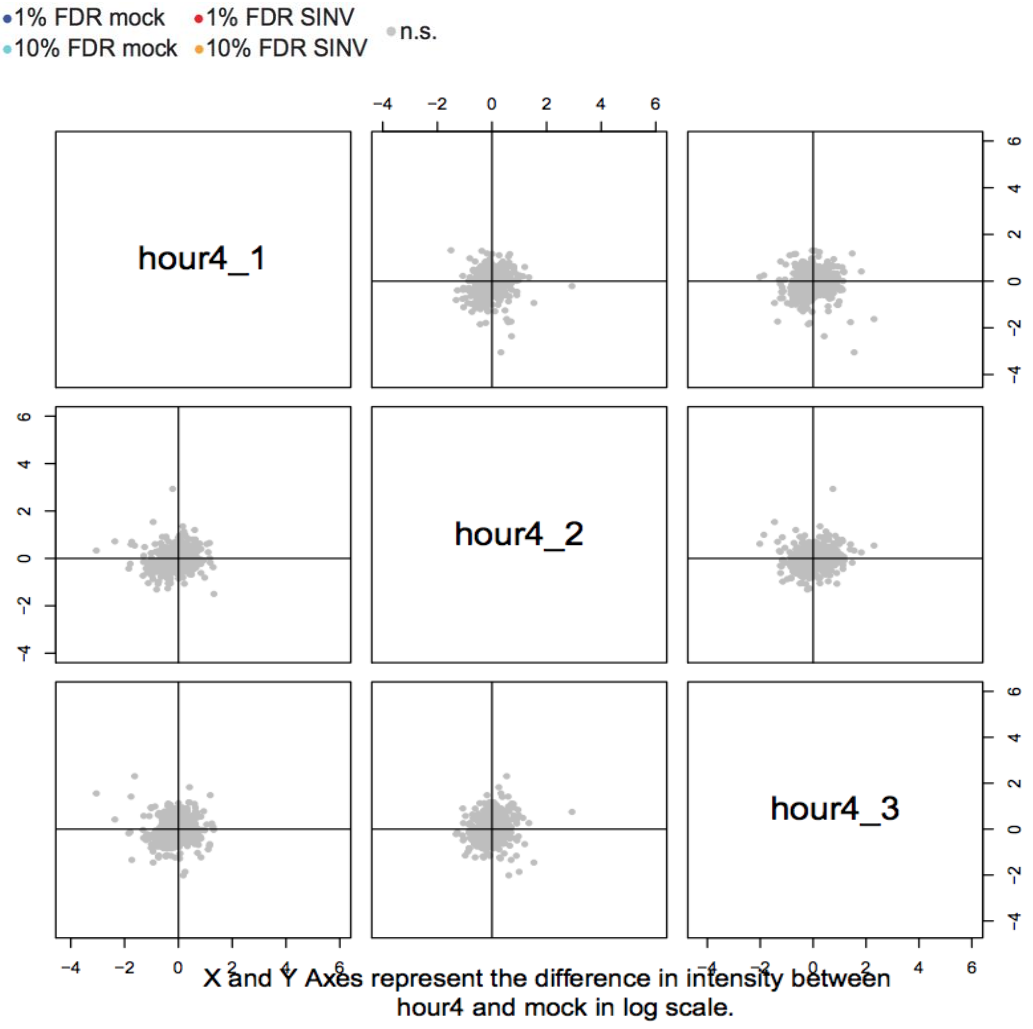

Figure 12. Scatter plot showing the intensity ratio between the 4 hours hpi and the uninfected condition for each protein (dots) in the eluates of two biological replicates of RNA-IC. (Garcia-Moreno M. et al. Mol Cell 2019)

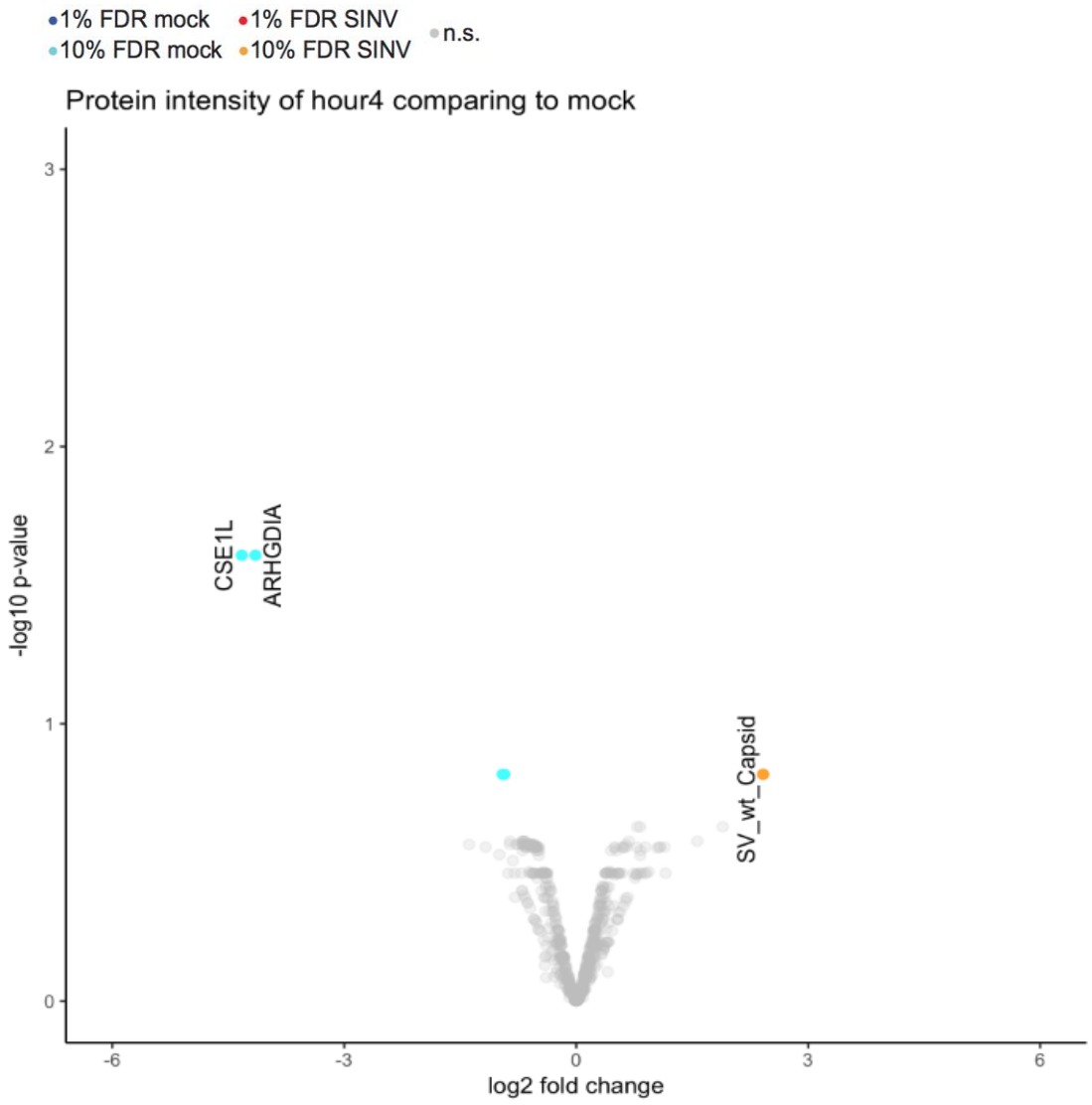

Figure 13. Volcano plot comparing the log2 fold change of each protein between 4 hpi and uninfected conditions and the p-value of this change across the three biological replicates. Sky blue/yellow points indicate proteins significantly enriched in uninfected condition (Mock)/4 hpi for at least two-fold change in intensity with 20% FDR. Names of proteins with at least two-fold change in intensity but higher FDR is also shown. (Garcia-Moreno M., Mol Cell 2019)

By contrast, SINV caused a pervasive remodeling of the RBPome at its later infection stage (Figure 14 and 15). Here, 236 RBPs (~30%) displayed altered RNA-binding activities (48 RBPs with 1% false discovery rate (FDR), 167 with 10% FDR and 21 by the semi-quantitative analysis)

(Table S2). RBPs with differential RNA-binding activity in SINV-infected cells are referred to here as ‘dynamic RBPs.’ It can also be seen that the amount of viral capsid protein dramatically increased throughout the infection (Figure 13 and 15), confirming the active replication of SINV in HEK293 cells.

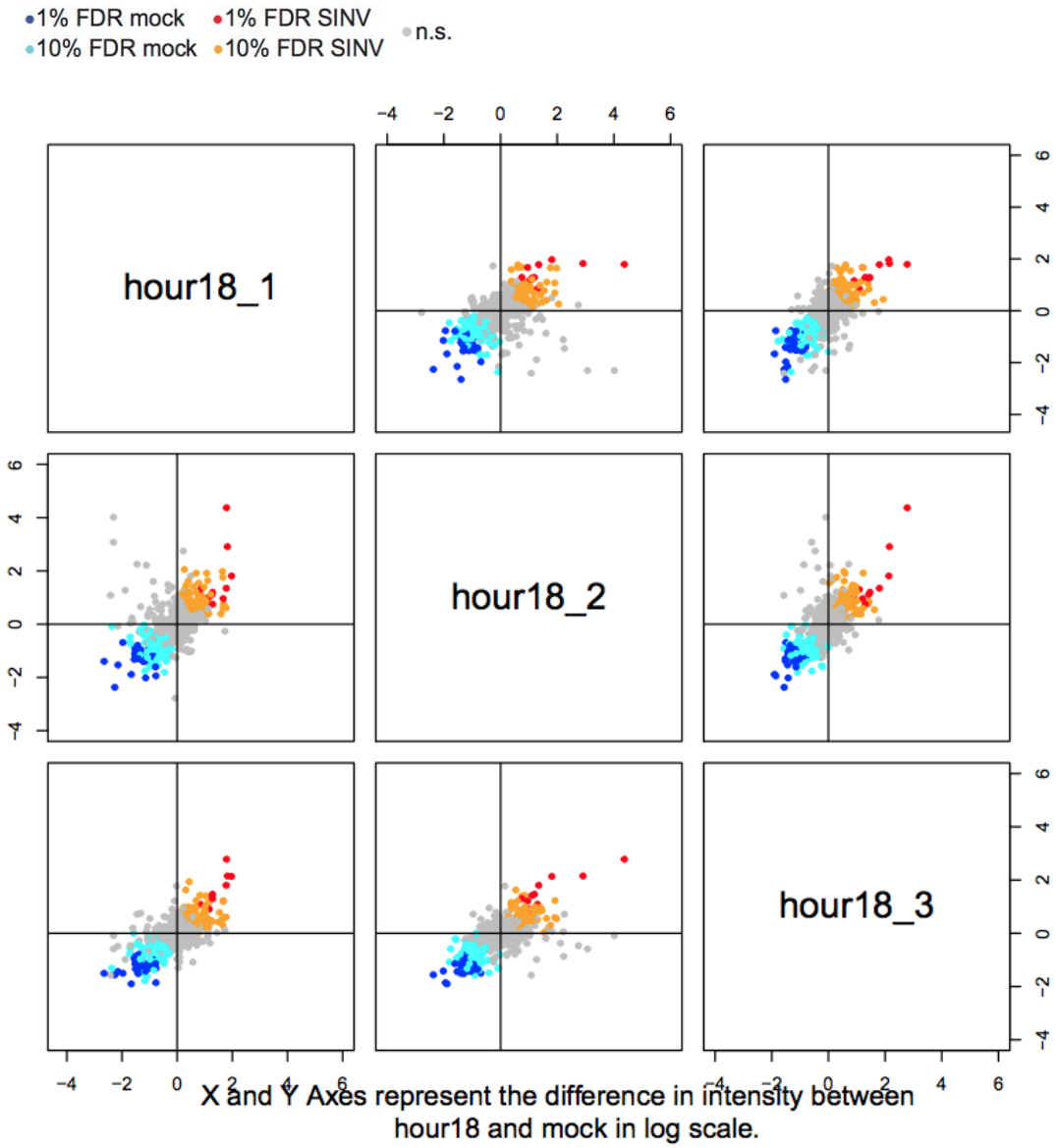

Figure 14. Scatter plot showing the intensity ratio between 18 hpi and the uninfected condition for each protein (dots) in the eluates of the two biological replicates of RNA-IC. Sky blue/yellow points indicate proteins significantly enriched in Mock/18 hpi with 10% FDR, Dark blue/red points indicate proteins significantly enriched in Mock/18 hpi with 1% FDR. (Garcia-Moreno M., Mol Cell 2019)

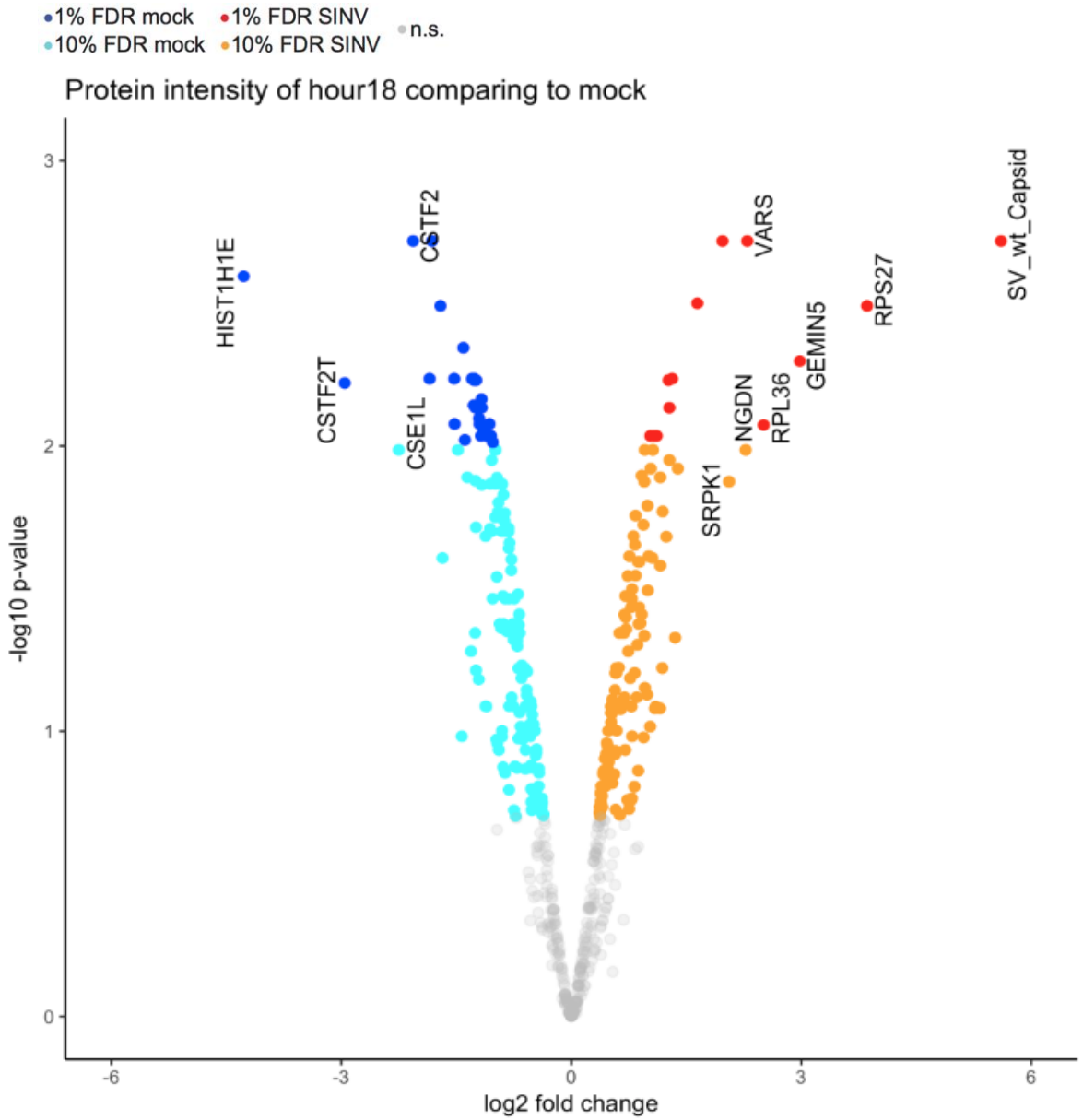

Figure 15. Volcano plot comparing the log2 fold change of each protein between the 18 hpi and uninfected conditions and the p-value of this change across the three biological replicates. Sky blue/yellow points indicate proteins significantly enriched in Mock/18 hpi with 20% FDR. Dark blue/red points indicate proteins significantly enriched in Mock/18 hpi with 1% FDR. Names of proteins with at least 2 fold change in intensity but higher FDR are also shown. (Garcia-Moreno M., Mol Cell 2019)

2.3.2) Virus infection turns off the nuclear RBPs and activates the cytoplasmic processes

Our data reveal an inhibition of the host translation processes in HEK293 cells and a global remodeling of the host RNA-binding proteome in response to infection. The results show that most RBPs inhibited by SINV at 18 hpi are linked to nuclear processes such as RNA processing and export. This finding is in good agreement with the previous reports of the inhibition of nuclear RNA metabolism by Sindbis virus (Gorchakov, 2005) (Figure 16). Conversely, most stimulated RBPs were cytoplasmic and are linked to protein synthesis, 5' to 3' RNA degradation, RNA transport, protein metabolism, and antiviral response (Figure 17). These findings suggest that viral infection by the Sindbis turns off the nuclear RBPs and activates the cytoplasmic processes.

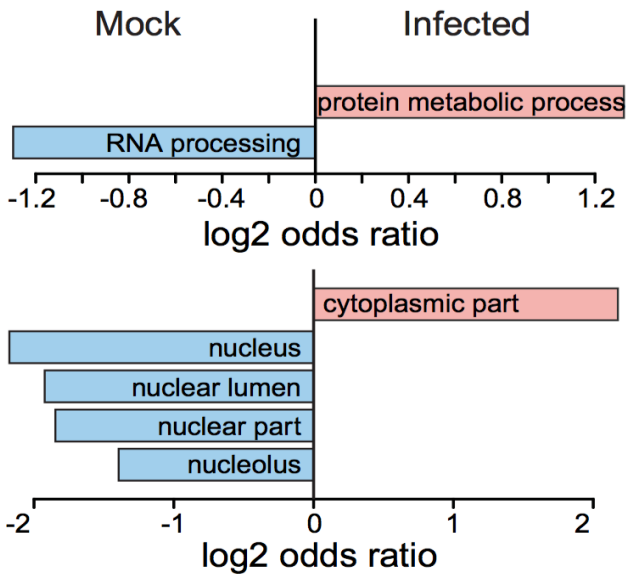

Figure 16. Molecular function (upper panel) and cellular component (bottom panel) gene ontology (GO) term enrichment analysis of the stimulated RBPs against those inhibited by SINV at 18 hpi. Blue bars represent enriched functions and cellular location for inhibited RBPs, salmon bars represent enriched functions and cellular locations for stimulated RBPs. (Garcia-Moreno M., Mol Cell 2019)

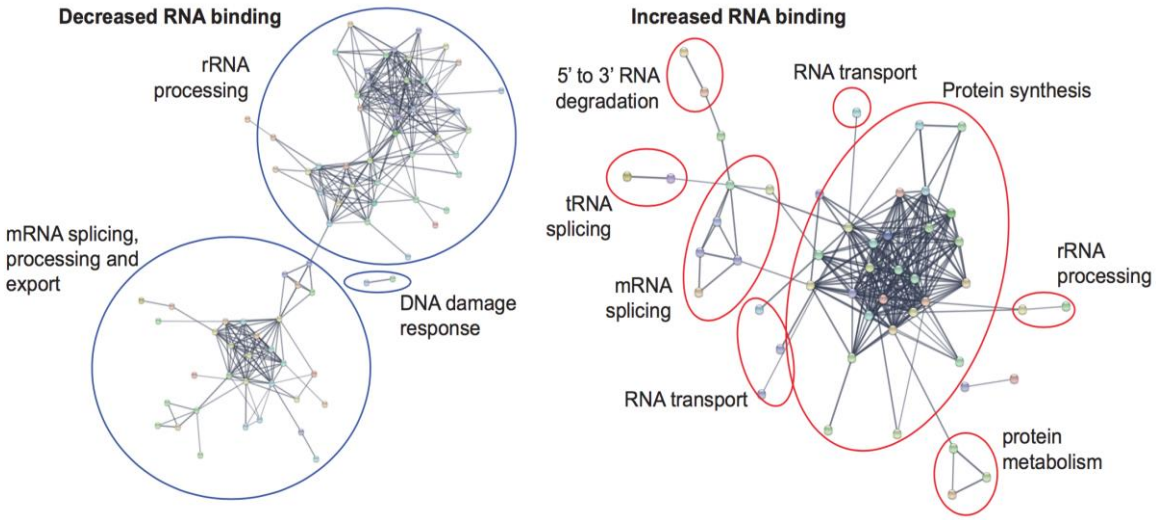

Figure 17. Functional Enrichment of Significantly Changed Proteins. The left enrichment plot shown in blue represents pathway analysis of proteins that demonstrate decreased binding, the right enrichment plot shown in red represents pathway analysis of proteins with increased binding (right) in the cell during viral infection. The plot is generated by STRING. (Garcia-Moreno M., Mol Cell 2019)

2.3.3) Changes in RBPome are not due to alterations in protein abundance

Changes detected by RNA-IC can be explained by various reasons. Besides changes in binding behavior, alterations in cellular protein abundance may also cause differential binding. To assess this possibility on a global level, we analyzed the total proteome (inputs of the RNA-IC experiments) using quantitative proteomics (Figure 18). We were surprised to see that SINV infection did not cause any noticeable alterations to host RBP levels, even at 18 hpi (Figure 18A and Table S3). The abundance of dynamic RBPs detected by RNA-IC also remained unaltered (Figure 18B). Therefore, whole proteome profiling infers that the changes in the RBPome are not due to alterations of relative protein abundance.

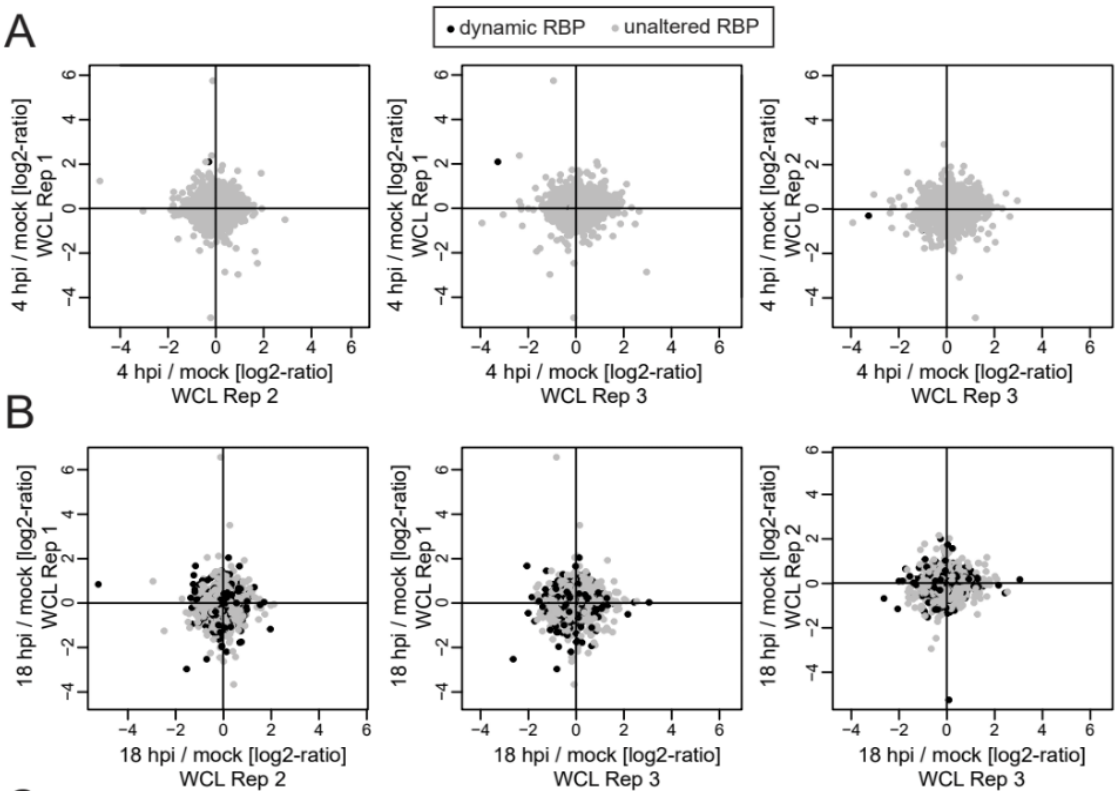

Figure 18. Scatter plot comparing the intensity of RBPs in the total proteome of two RNA-IC replicates from cells infected for 18 hours (grey dots) (A). Differentially expressed RBPs are highlighted as black dots (B). (Garcia-Moreno M. et al. Mol Cell 2019)

Surprised by the identification of the virus glycoprotein E2 in RNA-IC eluates, we estimated the proportion of protein-bound and unbound to RNA by normalizing the protein intensity reported in the RNA-IC experiment to that in the whole cell lysate (Figure 19). SINV NSP4 and NSP2 proteins were distributed within the top 50% of the identified RBPs, while Sindbis virus proteins E2 and NSP3 were present within the bottom 50%, suggesting lower affinity of the latter, or more transitory interactions with RNA (Figure 19).

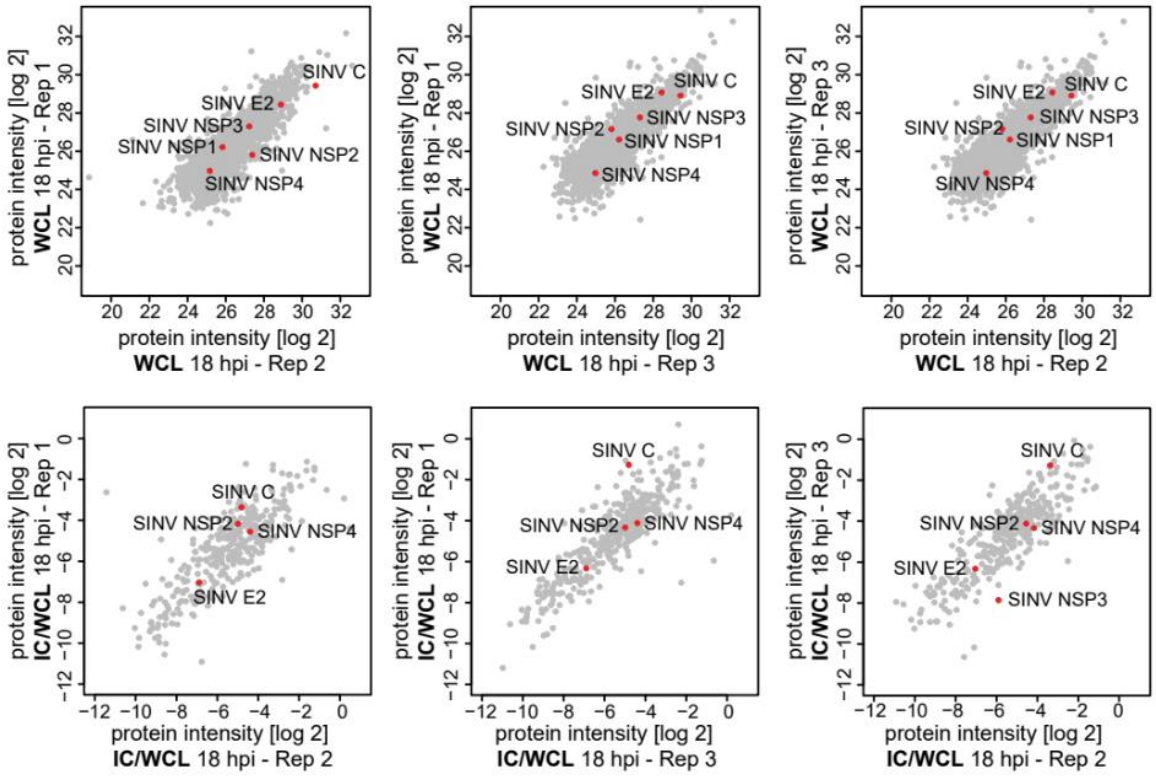

Figure 19. Scatter plot comparing the intensity of each protein in whole cell lysate from 18 hsi cells in 3 replicates (grey dots), red dots indicate the intensities of viral proteins (WCL). Normalized protein intensity reported in the RNA-IC experiment to that in the whole cell lysate (IC/WCL). (Garcia-Moreno M. et al. Mol Cell 2019)

2.4) Discussion

We used RNA-IC to demonstrate that SINV infection causes a massive remodeling of the host proteome and changes the protein binding activities in host RBPs. It shuts down nuclear activities and activates cytoplasmic activities. We believe this may be related to the cytoplasmic nature of the Sindbis virus. Since the replication of the Sindbis virus strongly relies on cellular RBPs, changes in RBPs binding may be due to elevated RBP-RNA interactions during Sindbis virus (SINV) infection. Furthermore, there is evidence to suggest that not only does SINV infection shuts off of host protein synthesis, it also initializes the global production of its proteins.

There are naturally a few limitations in these experiments we would like to address, first is the differential expression analysis in proteomics data. Many factors can contribute to a high missing rate in MS proteomics data, making it especially challenging to quantitatively assess differentially expressed proteins (Lazar, 2016). However, the semi-quantitative approach developed by Bernd Fischer and Vasiliy O. Sysoev has helped researcher to mostly overcome this issue (Sysoev, V.O., 2016). Among the total 253 differentially expressed RBPs at 4 hpi and 18 hpi, about 10% were discovered purely by semi-quantitative analysis.

Another important aspect of this study is the use of whole cell lysate as a negative control for in the RNA-IC experiment. Integrative analysis of the RNA-IC and whole cell lysate proteomics data shows that alteration in the proportion of proteins bound to RNA in host RBPs does not relate to changes in protein abundance. Therefore these findings are more likely to be explained by increased binding of the detected proteins to RNAs. However, this method does not discriminate between viral RNA binding or cellular RNA binding. To address this issue, we sequenced whole cell RNA for each of the 3 infection stages, to explore the interaction of viral RNA and cellular RBPs. Findings from these experiment are outlined in the next chapter.

3. Transcriptome analysis in Sindbis virus infected cells

3.1) Introduction

The RNA-IC experiments in SINV-infected cells revealed that the entire complement of cellular RBPs are remodeled upon infection. It is possible that the remodelling of the RBPome is due to alterations in the availability of RNA substrates. To address this question we profiled the transcriptome of uninfected versus SINV-infected cells by performing RNA sequencing analysis of the total RNA isolated from the RNA-IC input samples. This analysis showed how the activity of host RBPs may be altered by changes in the availability of their target RNAs. The analysis also demonstrates that alterations in RNA levels can be a consequence of increased cellular RNA degradation and shifts in codon usage.

3.2) Methods

3.2.1) Mapping RNA sequence to the reference genome

Version hg38 of the human genome was combined with the SINV sequence and used as our reference genome. RNA sequencing was then conducted in the Castello lab. The RNA sequencing reads were mapped to this reference genome using the ultrafast universal RNA-seq aligner STAR (Dobin, A., 2013). Reads mapping to each transcript were counted with 'featureCounts' function from the Subread software package (Liao, Y., 2014). Reads mapped to rRNA sequences are removed for further analysis. To avoid ambiguity, only uniquely mapped reads were considered for counting. Reads mapping to positive and negative strands of viral RNAs were separated using SAMtools 'view' utility (Li, H., 2009). In Illumina reverse paired-end sequencing, paired reads came from opposite strands. Therefore, reads with the second pair mapping to the positive strand, or with the first pair mapping to the negative strand, were both counted as mapping to the positive strand and vice versa. The total read counts mapping to each strand was compiled and counted using SAMtools merge and SAMtools depth, respectively (Li, H., 2009).

3.2.2) RNA count differential analysis

SINV infection is known to shut off transcription globally as a means of reducing the development of antiviral responses in a host cell. (Gorchakov, R., 2005). The global change in RNA abundance may bias (underestimate) differential expression results if normalization is carried out under the assumption overall RNA abundance remains unchanged. Therefore, we decided to normalize read counts in each condition to the corresponding rRNA expression by dividing by a factor proportional to the total rRNA read counts in the 3 conditions (0.899, 1 and 0.473 for Mock, 4 hpi and 18 hpi respectively). The R package DESeq2 (Anders, S., 2010) was used for differential gene expression analysis based on rRNA normalized read counts. As DESeq2 requires the read counts to be un-normalized integer values, rRNA normalized read counts were rounded to the closest integer to carry out "DESeqDataSet" differential analysis.

In DESeq2, a proper estimation of size factors can help to cancel out the error effects between technical replicates derived from the amount of pipetting or slightly different PCR circles. These errors can lead to proportional changes in the whole DNA yield, influencing between-sample read counts. This is further complicated by the fact that general amount of read counts between different infection times in our experiment are not expected to be the same, due to our observation that viral infection shuts down the expression of most genes, and this effect would not be canceled out. Therefore, we estimated the size factor of each sample separately in DESeq2, instead of pooling all the samples prior to estimating this parameter.

Differential RNA expression between infected (4 and 18 hpi) and uninfected cells (Mock) was visualized in MA-plots to show the log2 fold changes over the mean of normalized counts using DESeq2 (Anders, S., 2010). To visualize the overall effect of experimental covariates and potential batch effects, a principal component plot of the samples was generated using the plotPCA function in DESeq2 (Anders, S., 2010). Principal components (PC) of the variance stabilized expression of the top 500 genes with the highest expression variance among samples were extracted. As shown in Figure 20, the variance explained by the first and second PC (on the x and y-axes) accounted for 96% of the total variance, and, as expected, samples within the same

condition clustered better than those between conditions. We note that the first PC accounts for 94% of the total variance, and distinctly separates 18 hpi from the other samples (i.e., uninfected and 4 hpi), indicating that the cellular transcriptome is dramatically altered at 18 hpi.

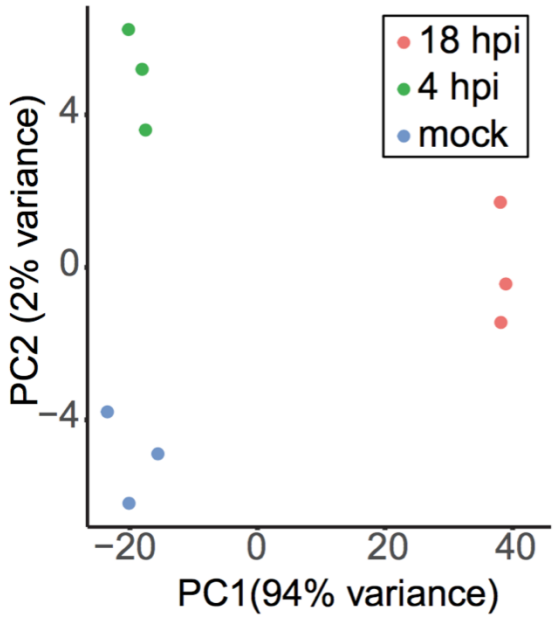

Figure 20, Principal component analysis of gene expression as profiled by RNAseq in uninfected and SINV-infected cells at 4 or 18 hpi. The 3 replicates of each condition are considered separately. Data shows that, first, the 23 replicates from the same condition cluster together (i.e., they are more similar to each other than to other conditions) and, second, that the transcriptome at 18 hpi strongly differs from the uninfected control condition and 4 hpi. (Garcia-Moreno M., Mol Cell 2019)

3.2.3) Linear regression model for explained variance

The levels of mRNA within a cell are determined by the combined modulation of synthesis, processing, and degradation rates of mRNA (Bjork, P., 2015; He, F., 2015). In order to analyze the relative contribution of RNA synthesis, processing and degradation to the alterations of RNA levels in SINV infected cells, we compared transcriptome changes observed at 4 and 18 hpi in our RNAseq dataset with freely available measures of the rates of synthesis, processing and degradation for all RNAs in HEK293 cells (Mukherjee, N., 2017). Specifically, the INSPEcT R package was used to infer the rates of different RNA processes. This R package is used for the comprehensive quantification of synthesis, processing and degradation rates of genes. The

estimation of differential RNA processes is based on a set of differential equations that describe the process of production, maturation, and degradation of pre-mRNA and mature mRNAs. For model simplicity, INSPEcT assumes that pre-mRNAs are not degraded, and mature mRNAs are immediately translocated from the nucleus to the cytoplasm (Pretis, S.D., 2015). These assumptions are not necessarily accurate but are acceptable in this particular context.

We used one-way ANOVA to evaluate the relative effects of the RNA processes (RNA degradation, processing, and synthesis) on transcriptomic changes in the different conditions (Welch, B. L., 1951). The effects are shown as explained variances in ANOVA for each RNA process in the two comparisons.

3.2.4) Predicting binding specificity with DeepBind

DeepBind is a software package that predicts sequence specificities of DNA- and RNA-binding proteins using deep learning methods; and it is reported to outperform other state-of-the-art models such as RankMotif++ and KmerHMM (Alipanahi, B., 2015). We used DeepBind to predict binding sites on the viral genome against 244 known human RNA-binding proteins in the DeepBind database. As suggested by DeepBind, the viral genome was chopped into 50-bp segments to achieve the best performance (Alipanahi, B., 2015). To correctly count binding sites sitting at chopping breakpoints, the same genome was rechopped into another set of 50-bp segments, starting from the 25th nucleotide of the original sequence instead of from the first base. The results were merged afterward. To fit our experimental model, only human-originated RNA-binding proteins in DeepBind database are used to make the prediction.

3.2.5) Codon usage bias in uninfected and infected cells

To get the preferred codon usage of Sindbis at 18 hpi, we counted the appearance of each 3-letter codon in the SINV 'genomic' and 'capsid' sequences and weighted the counts by their actual expression at 18 hpi. The tRNA annotation was downloaded from Genecode, and anticodons associated with each tRNA are converted to codons based on tRNA annotation. The real-time codon usage in the cell is calculated by counting the reads mapped to the tRNA annotation at

Mock, 4 hpi and 18 hpi. The mapped counts are normalized by the total amount of reads in each sequencing lane. To compare the codon usage bias in uninfected vs. infected cells, the F test was applied to the codon count between Mock, 4 hpi, and 18 hpi. Codons with an F test p-value smaller than 0.1 were chosen for comparison of the real-time codon usage shift between Mock and 18 hpi. We then tested the similarity of real-time codon usage in Mock, 4 hpi and 18 hpi using the preferred codon usage generated from viral sequences and their actual expression.

3.3) Results

3.3.1) Replication of Sindbis virus during infection

SINV produces two overlapping mRNAs: gRNA and sgRNA (Figure 9). Consequently, the read coverage was substantially higher in the last third of the gRNA, where these transcripts overlap (Figure 21). In agreement with published data we note that the sgRNA and gRNA from the positive strand were more abundant than those on the negative strand (Figure 21) (Gorchakov, R., 2005; Strauss, J.H., 1994). As the copy number of the negative strand is low and it lacks a poly(A) tail, this phenomenon should not contribute to our RNA-IC results.

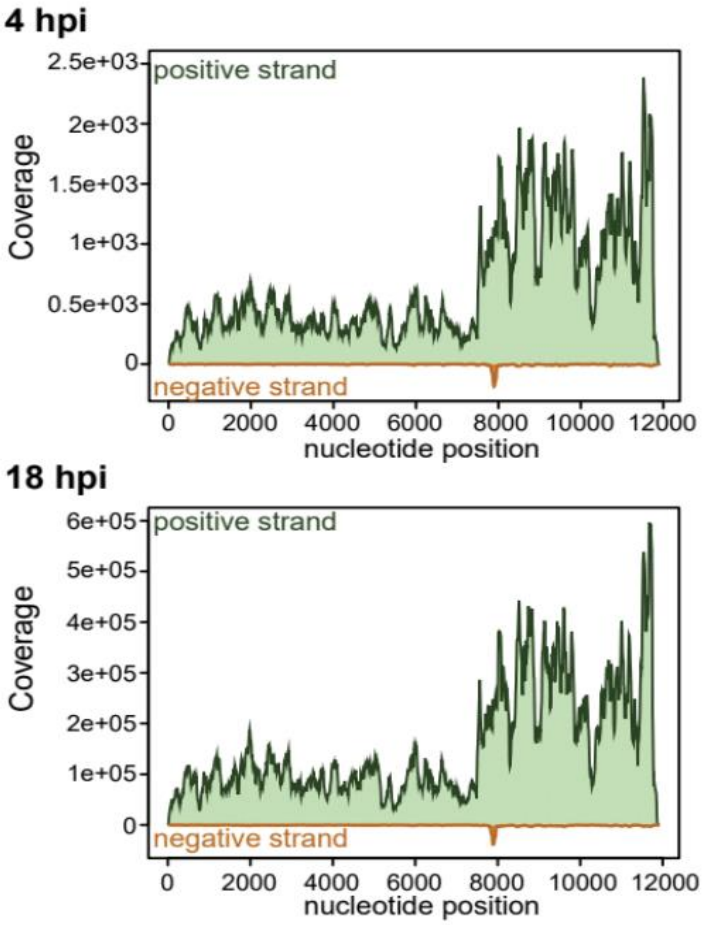

Figure 21. Read coverage of the positive and negative RNA strand of SINV in the RNAseq analysis at 4 and 18 hpi. Note that the y axis in both plots uses different scales to facilitate the visualization of the data. (Garcia-Moreno M., Mol Cell 2019)

During infection the amount of viral RNA greatly increased (Figure 21). Overall, viral RNA in the early stage of infection (4hsi) consisted of less than 0.7% of the total RNA. However, at the later stage of infection (18 hsi), more than 77.7% of the total RNA became viral RNA (Figure 22). This proportional increase in viral RNA is a consequence of not only of viral replication but also loss of host RNA, as outlined in section in 3.3.2.

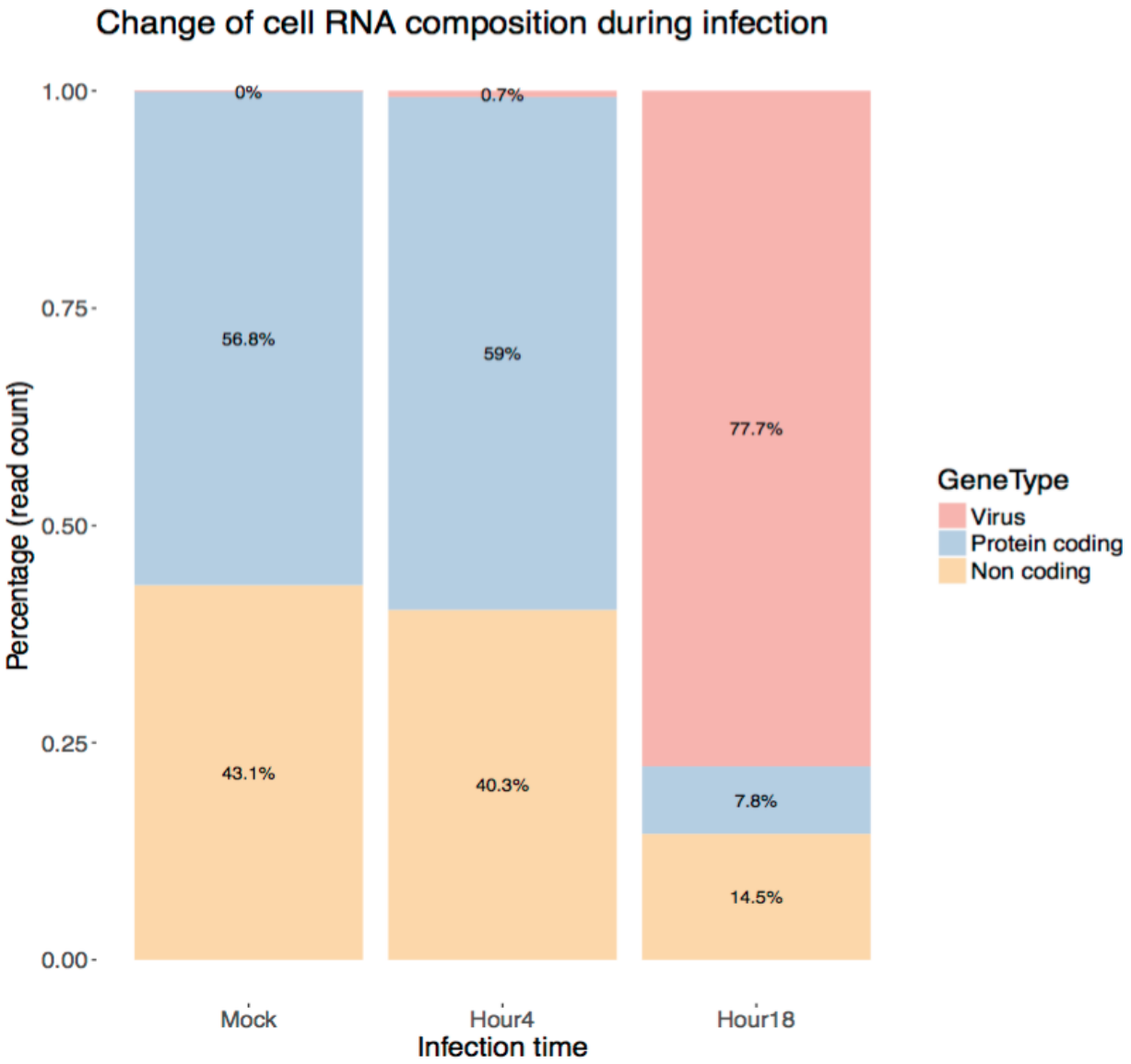

Figure 22. The composition change of different types of RNAs from different origin during viral infection. The x axis represents different infection times, while the y axis gives the relative abundance of 3 types of RNAs in the cell, i.e. protein-coding, non-coding and viral RNA. (Garcia-Moreno M., Mol Cell 2019)

3.3.2) Alterations of the transcriptome in Sindbis-infected cells

The host transcriptome responds dramatically to SINV infection. At 4 hours post SINV infection, the host transcriptome shows a relatively minor change in composition, with only 67 up and 177 down-regulated RNAs, respectively (Figure 23F). By contrast, profound changes in the cellular

transcriptome were observed at 18 hpi, with 12,372 differentially expressed RNAs detected ($p<0.1$; Figure 23G). Of these, 10,924 RNAs had significantly lower expression, and 1,448 RNAs were upregulated (Table S4). Hence, SINV infection causes a massive loss of host RNAs at its later infection stage.

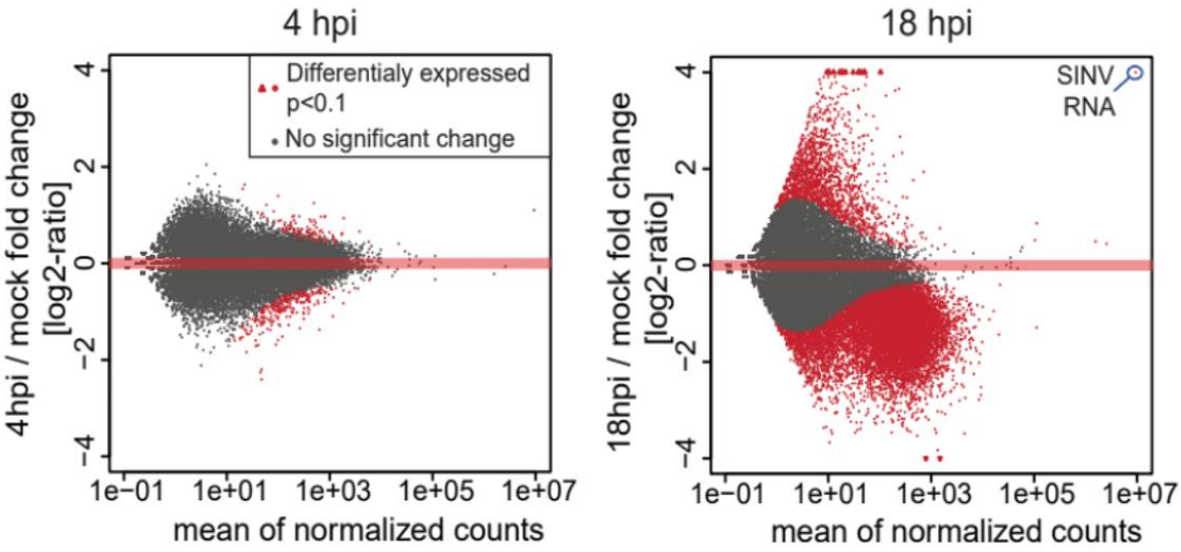

Figure 23. The amount of different types of RNA in different infection stages. The x axis represents the mean of normalized counts in the 3 replicates, while the y axis represents the log2 fold change of all transcripts between 4hpi (F) /18hpi (G) and the Mock condition. (Garcia-Moreno M., Mol Cell 2019)

Given the massive loss of cellular RNAs, we wondered if the proportion of different RNA species also changed during infection regardless of the overall inhibition of transcription. Figure 24 shows the composition of each type of RNA in the cell. The read counts in each RNA type are weighted by the total amount of reads for the three-time points per category. It can be seen that siRNA and rRNA molecules are among the most stable RNA species in the cell during infection. In contrast the abundance of protein-coding RNAs, lncRNAs, and miRNAs suffer the most from viral infection.

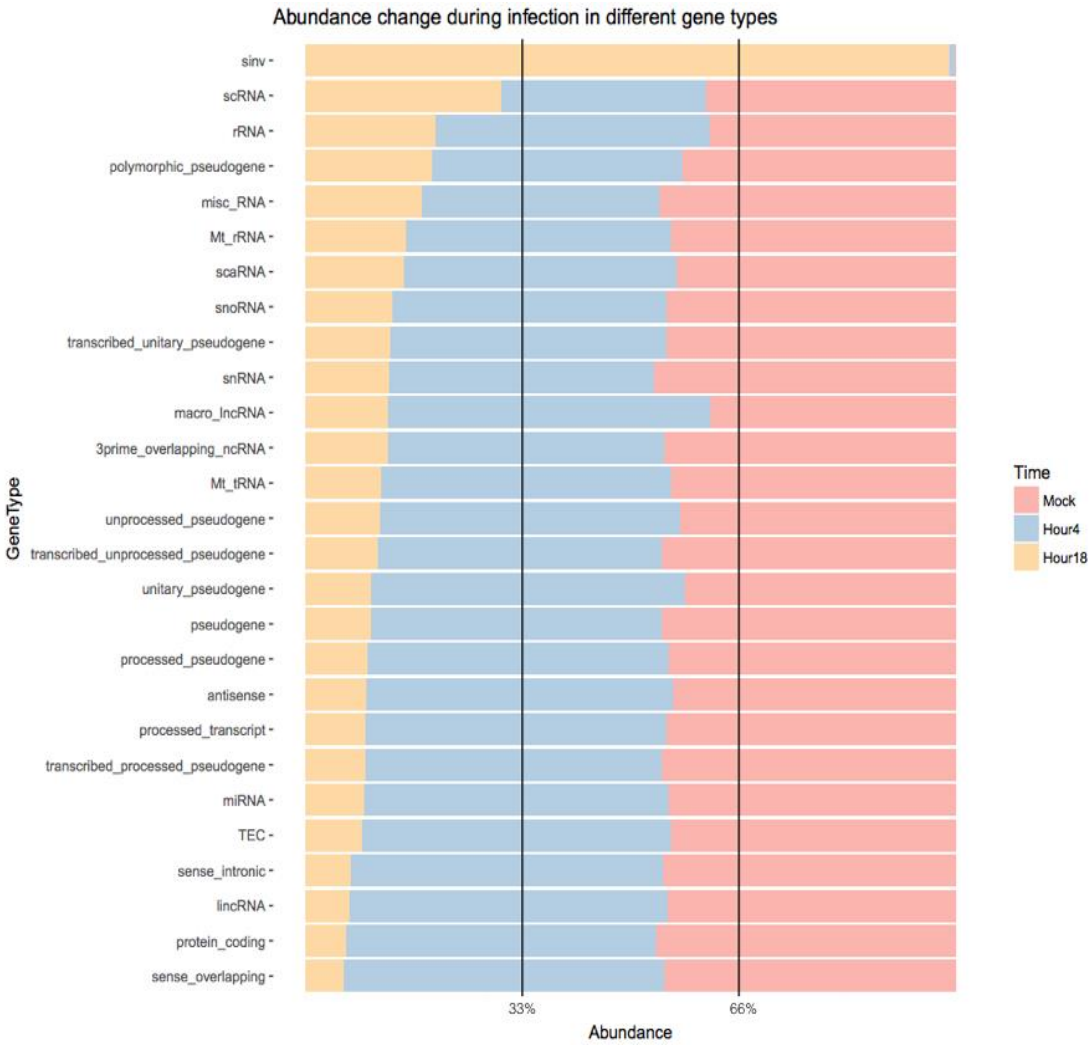

Figure 24. A detailed plot shows the abundance change of different types of RNA during infection. For each type of RNA, the relative abundance at Mock, 4hpi and 18hpi are shown in yellow, blue and red respectively. Read counts in each RNA type are weighted by the total amount of reads in the corresponding infection stages.

In the RNA-IC, we show that dynamic alterations of host RBPs at 18 hpi are not related to changes in protein abundance. Under the assumption that the binding affinity of RBPs did not change during infection, the altered binding of RBPs can, therefore, is more likely be explained by changes in the number of RNAs they bind to. In other words, altered binding is a consequence of altered RNA levels in the cell. Since a large proportion of cellular RNA at 18nhpi is accounted for by viral RNA, it is highly possible that some up-regulated host RBPs relocalize to bind viral

RNAs instead at this infection stage, while some down-regulated RBPs lose their binding partners.

In summary, the availability of cellular RNA is globally reduced upon infection, correlating with the emergence of viral RNA (Figure 22 and 23). We suggest that decreased availability of cellular RNA is expected to contribute to the inhibition of RNA-binding activity observed for 133 down-regulated RBPs at 18hpi, while some up-regulated host RBPs may relocalize to bind viral RNAs.

3.3.3) Immune response in HEK293 cell to viral infection

As we expected, upregulated RNAs were more enriched in genes related to antiviral activities, reflecting the activation of the host defense systems. Although not all antiviral related genes are up-regulated, compared to the 12% of up-regulated genes in differentially expressed genes on a global scale, 21.6% of genes are up-regulated in the same context for antiviral related genes. We selected two antiviral GO terms for further investigation (innate immune response: GO::0045087, which featured 687 genes, and defense response to virus: GO::0051607, which included187 genes), totaling 767 genes after removal of duplicates. In these GO terms, 255 genes were differentially expressed and 55 genes were up-regulated (Figure 25).

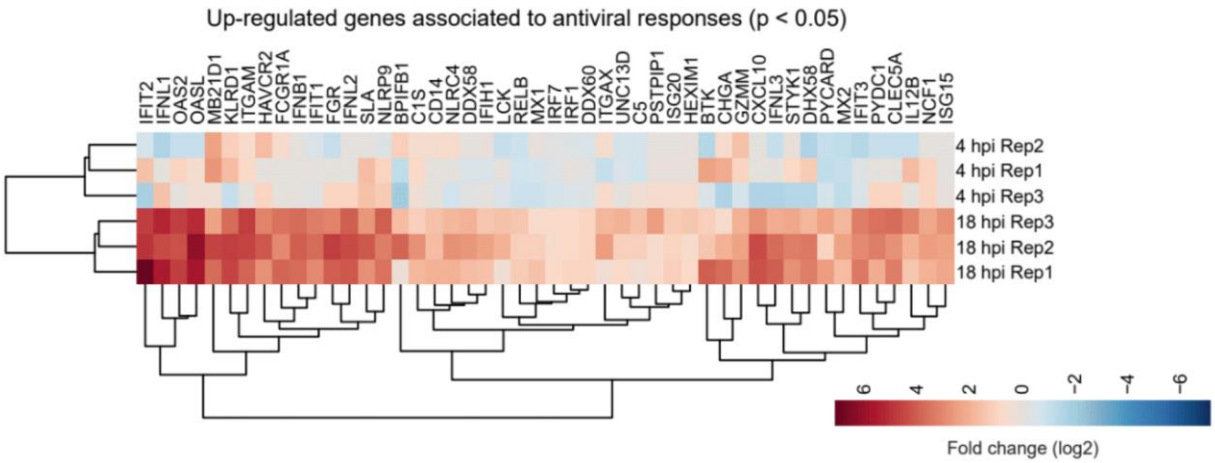

Figure 25. Heatmap showing the log2 fold expression change of 55 up-regulated genes determined by RNAseq analysis of RNAs enriched with p<0.05 in SINV-infected (4 or 18 hpi) versus uninfected HEK293 cells. Genes were annotated to 'viral defense' and 'innate immune response' gene ontology (GO) terms. These GO terms were statistically enriched in infected versus uninfected cells. Colors in the heatmap indicates the log2 fold change between 4hpi/18hpi to Mock, as red being up-regulated and blue been down-regulated. Note the presence of interferons (IFNs), interferon-stimulated genes (ISG), interferon-induced proteins (IFI), and interferon regulatory factors (IRF). (Garcia-Moreno M., Mol Cell 2019)

We defined the group of most differentiated genes where the log2 fold change was larger than 4 and the FDR was below 1%. In contrast, given the loss of host RNAs at 18 hpi (Figure 16), in terms of the most differentiated genes, up-regulated genes greatly outnumbered down-regulated genes (Figure 26). This may suggest that although the transcription of most genes has been inhibited, the virus may utilize some specific genes for its own replication. We also notice that some of the genes such as IFNL1, IFIT2, OASL, and CCL5 are related to cellular immune activities. This suggests that SINV infection triggers the host antiviral activities.

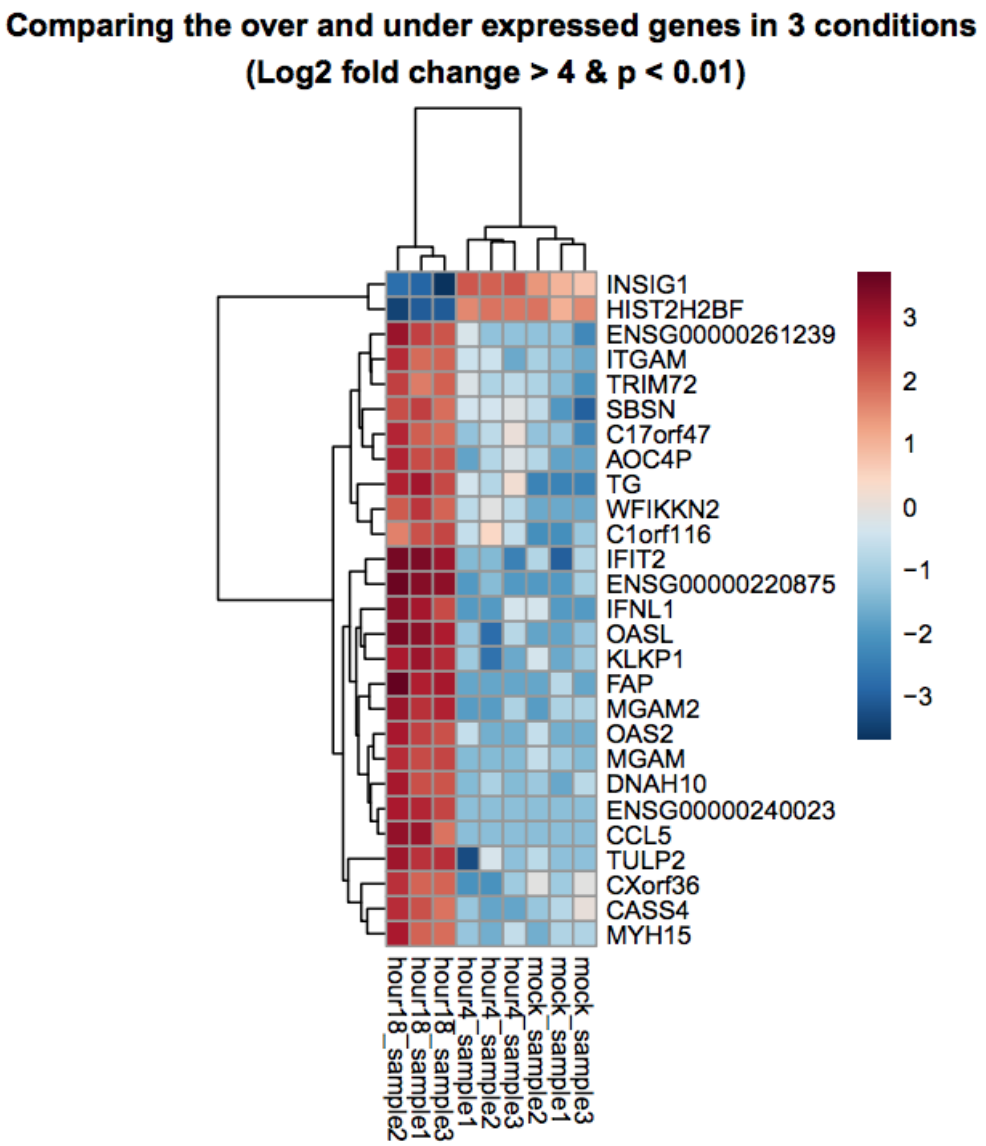

Figure 26. Heatmap of the most differentially expressed genes detected by RNAseq in SINV-infected (4 or 18 hpi) and uninfected HEK293 cells. This is defined as genes that over or under-expressed across all three conditions with log2 fold change > 4 and FDR < 0.01. (Garcia-Moreno M. et al. Mol Cell 2019)

3.3.4) Modeling alterations of transcriptome in Sindbis-infected cells

The loss of cellular mRNAs can be a vital driving force in shaping the RBPome in SINV-infected cells. However, it is unclear where transcriptome remodeling originated from and whether it

benefits viral infection. Alterations in RNA levels can be a consequence of reduced transcription and increased RNA degradation (Mukherjee, N., 2017). To explore which of these pathways contribute the most to RNA loss in SINV infected cells, we compared the fold change of each mRNA in our dataset to available data on the speed of synthesis, processing, and degradation of each transcript (Mukherjee, N., 2017) (Figure 27). The process of transcription can explain most of the differences between the uninfected and 4 hpi condition (Figure 28 and Table 1). However, RNA degradation accounted for more than 50% of the explained variance at 18 hpi. We reasoned that this phenomenon may be due to the combined effect of activation of the 5' to 3' RNA degradation machinery, as the exonuclease XRN1 and its interactor PAT1 homolog 1 (PATL1) are stimulated at 18 hpi (Table S1), and reduced transcriptional activity at this point (Gorchakov, R., 2005, Houseley, J., 2009).

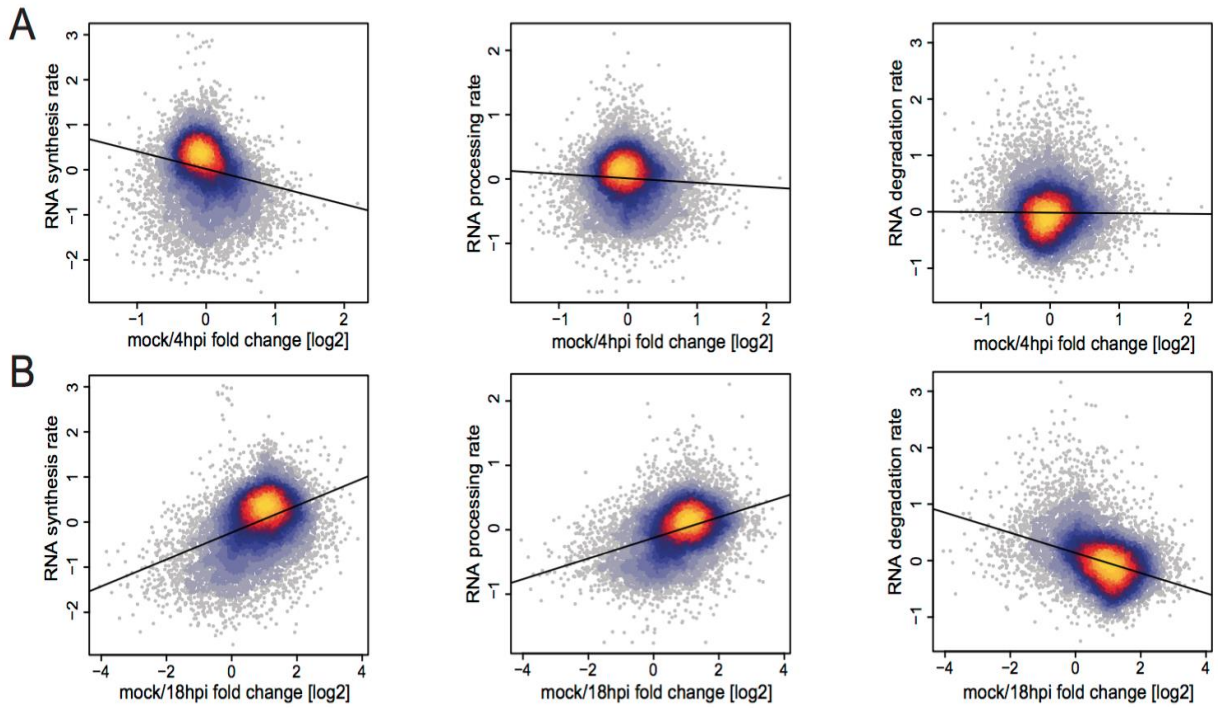

Figure 27. Analysis of the contribution of transcription, processing, and degradation to the transcriptome of SINV-infected cells. A) Plots representing the log2 fold change of cellular RNAs detected by RNAseq between uninfected and SINV-infected (4 hpi) cells, compared to rates of RNA synthesis (left), processing (middle)´, and degradation (right). These rates were determined separately in another study (Mukherjee et al., 2017). B) As in (A) but comparing 18 hpi and uninfected cells. (Garcia-Moreno M. et al. Mol Cell 2019)

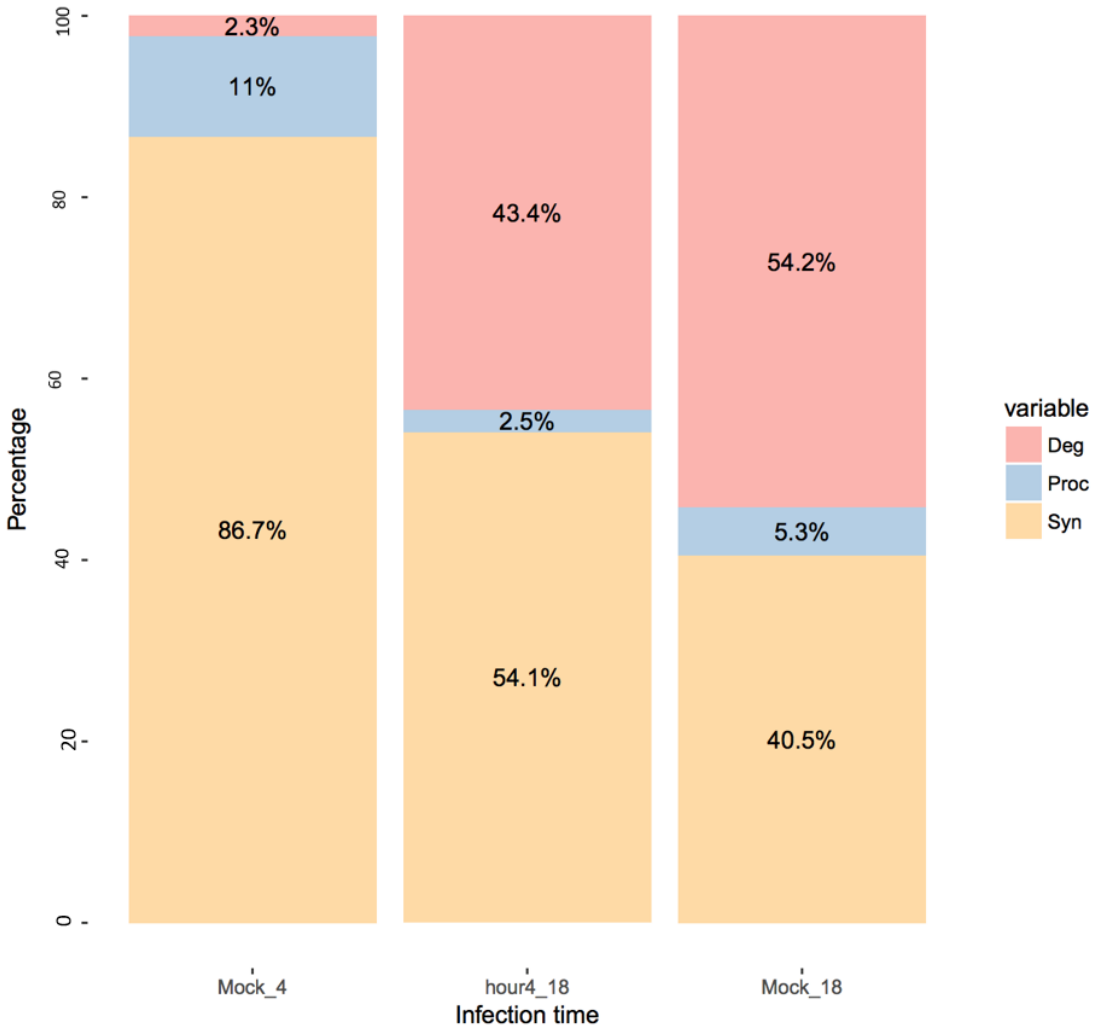

Figure 28. Percentage of variance in RNA expression explained by the three major RNA processes. At 18 hsi, the degradation rate explains a higher percentage of expression variance. The synthesis rate and processing rate also contribute to a proportion of the change in transcription. (Garcia-Moreno M., Mol Cell 2019)

Table 1. Percentage of variance in transcript abundance change between Mock and 18hpi as explained by altered RNA degradation, synthesis and processing rates in ANOVA one way analysis. Explained variance is represented in column 'Sum Sq'.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Deg	1	1334.3857	1334.3857428	2346.029	0.000000e+00
Syn	1	996.0430	996.0430065	1751.177	0.000000e+00
Proc	1	129.2109	129.2108693	227.170	6.179367e-51
Residuals	14036	7983.4639	0.5687848	NA	NA

3.3.5) RNA splicing and intron retention during SINV infection

Intron retention is a primary mechanism of gene expression regulation in eukaryotic organisms (Jacob, A.G., 2017). Although it is believed that intron retaining transcripts are non-functional because they are subject to nonsense-mediated decay, there is evidence that intron-retaining mRNAs can also play an essential role in diverse diseases (reviewed in Wong, J., 2015). One hypothesis is that if the splicing machinery is impaired by a viral infection, there would be a reduced number of splicing events happening at 18 hpi compared to the Mock condition. In other words, if there are the same number of reads per condition, then there will be more reads spanning exon-intron junctions at 18 hpi. The number of retained introns is considered a good proxy for an unsuccessful splicing event, and moreover provides information on the number of non-splicing reads that map to and span over exon-intron junctions. In regards to our hypothesis that viral infection may induce intron retention, we compared the number of successful and failed splicing events in the 3 infection stages. As shown in Figure 29, the number of failed splicing was significantly lower in the late stage of infection.

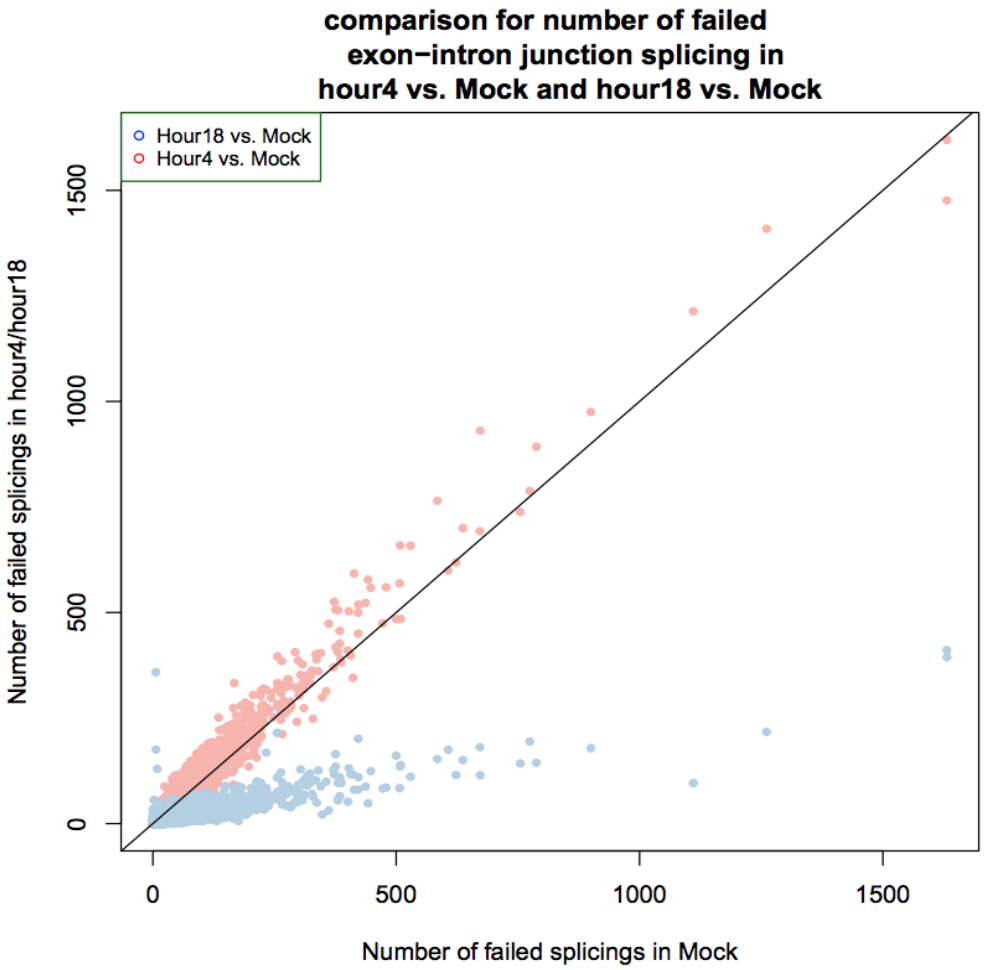

Figure 29: Scatter plot comparing the number of intron retention in the 4 hsi vs. uninfected (red dots) condition and 18 hsi vs. uninfected condition (blue dots). In 4 hpi vs. uninfected, the number of introns retained in the two conditions are relatively equal, while in 18 hpi, the number of intron retained was dramatically reduced compared to the Mock condition.

It is difficult to draw a firm conclusions from the above analysis given the massive inhibition of host RNA at 18 hpi. However, several splicing events on the virus genome can be observed at 18 hpi (Table 2). The most extensive splicing spanned around 4000 bases in length in the viral genome (Table 2). This suggests the possibility of various alternative virus protein isoforms being generated at 18 hpi.

Table 2. Most abundant splicing events in viral genome at 18hsiwith a read count coverage greater than 100

	chr	start	end	strand	motif	count_uni	overhang
1	negative	11486	11575	-	CT/AC	1172	39
2	negative	11497	11575	-	CT/AC	566	39
3	negative	6150	6200	-	CT/AC	230	39
4	negative	7594	11516	-	CT/AC	159	38
5	sinv_genomic	9286	9322	+	GT/AC	117	39
6	negative	10081	10116	-	CT/AC	105	36
7	sinv_genomic	8052	9090	+	GT/AC	103	38

3.3.6) Shift of codon usage in Sindbis-infected cells

There are some fundamental difference in codon usage between viruses and humans. Therefore, investigating the codon usage shift during infection will help us to understand how the Sindbis virus manages to thwart host defense mechanisms whilst exploiting host molecular resources under the pressure of tRNA pool discrepancy. We demonstrate that there is a change in codon usage between uninfected and infected stages. We also show that the preferred codon usage of viral genomic RNA could explain some of the observed shift in codon usage.

Figure 30 shows the similarity between real-time codon usage at Mock, 4 hpi and 18 hpi and the preferred codon usage for virus at 18 hpi. It can be seen that codon usage profile between Mock and 4 hpi has the highest correlation, while the codon usage profile at 18 hpi bears more resemblance to the preferred codon usage for virus replication. The changes in real-time codon usage at 18 hpi compare to Mock are also shown. These changes highly correlate with the differences in preferred viral codon usage at 18 hpi and real-time codon usage in Mock (cor = 0.513).

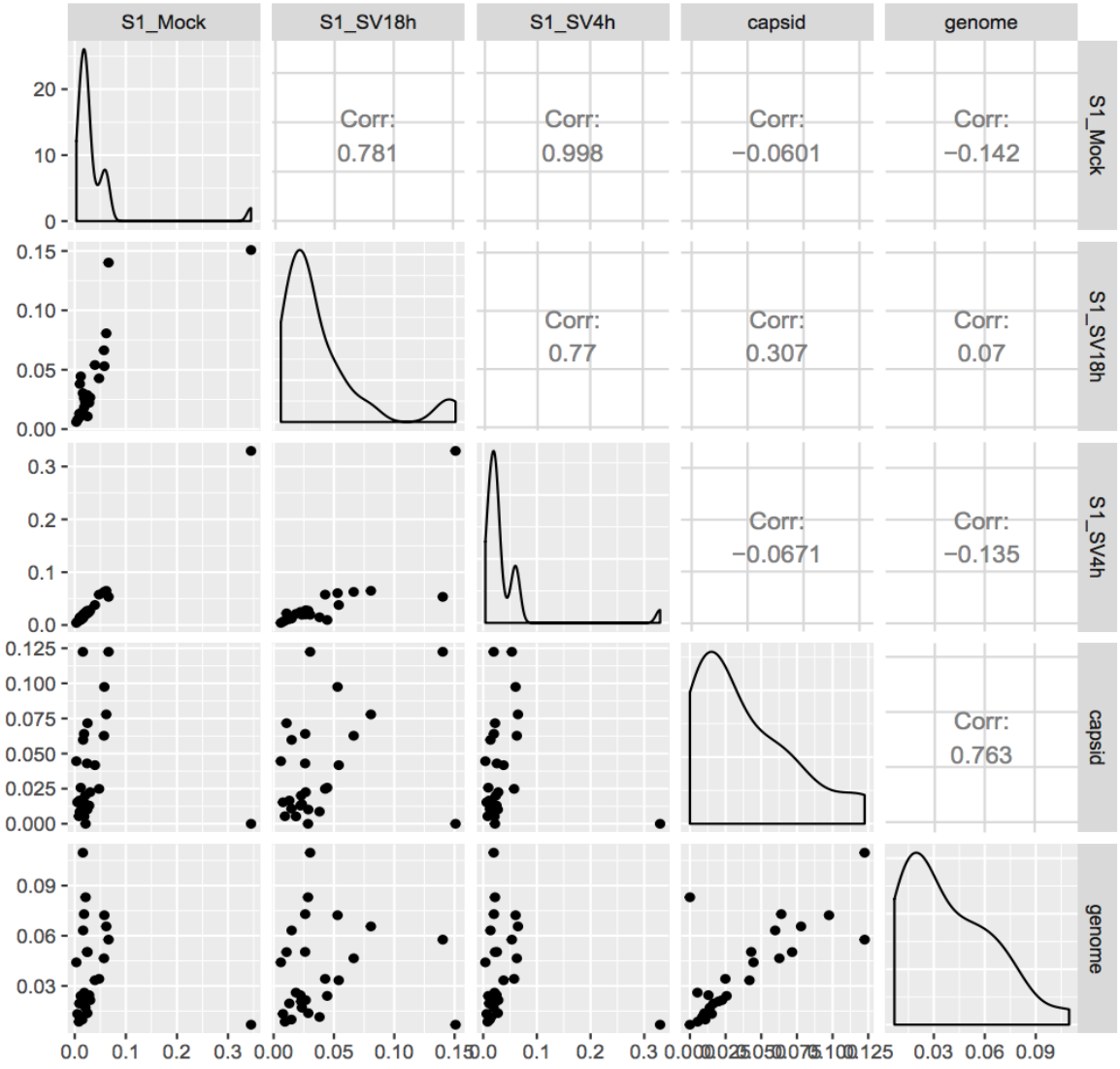

Figure 30. Similarity scatterplot of real-time codon usage and preferred codon usage for virus at 18 hpi. Each dot in the lower-left panel represents a codon. X-axis and y-axis represents the abundance of the corresponding tRNA in the cell.

3.3.7) Alterations in expression of protein-binding RNAs in eCLIP database

The first question about RBP binding will probably be “where do they bind?”. The large amount of eCLIP experiments on the eCLIP database provides a useful overview of RNA binding site locations. Experimental research shows that RBPs usually bind to RNA with a preference for 5′- or 3′-untranslated regions (Gebauer, F., 2012). The same pattern can be observed in the eCLIP

dataset. Most RNAs bind proteins at their 5′- or 3′-UTR, with the highest preference for the 5’ UTR. Figure 31 shows the distance distribution for RBP binding sites to their nearest 3’ UTR start site for 80 RBPs uploaded to eCLIP in 2016. It can be observed that a large proportion of binding sites on RNA tend to have a distance preference to their nearest 3’ UTRs. In our study, we use eCLIP data as a reference to uncover the expression of potential target RNAs for RBPs of interest. We also found that, in general, RNAs that bind to multiple RBPs tended have a higher expression in the cell.

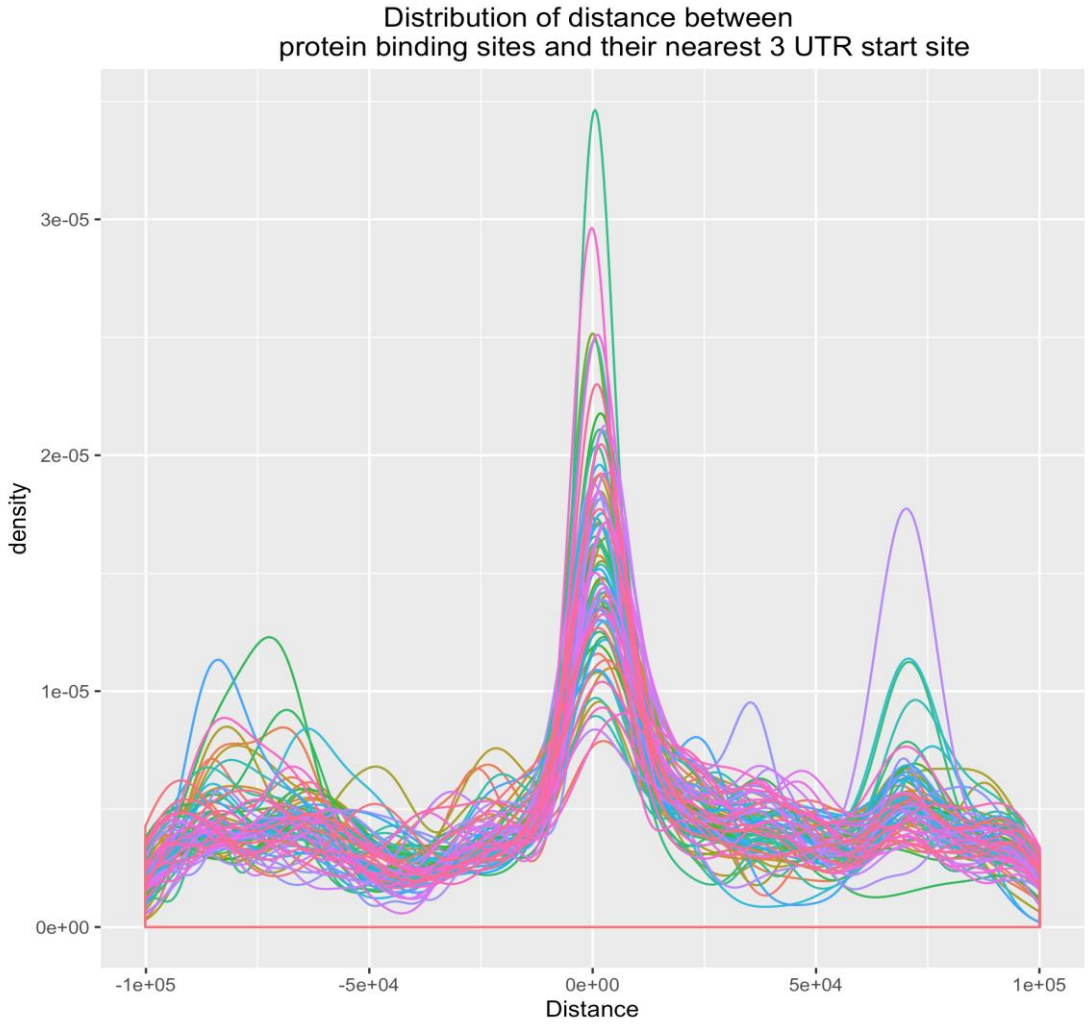

Figure 31, Distance distribution of protein binding sites to their nearest 3’ UTR start site. Each line is a protein, and x-axes represent the distance of all protein binding sites to their nearest 3’ UTR on the located gene, while the y-axis represents the relative density distribution.

We hypothesis that those RNAs that are functionally required for many RBPs will be more abundant overall. To address this question, we used a database of approximately 100 RBPs from eCLIP to study if the number of possible binding partners affected the average change in expression level of the protein binding RNAs. The expectation was that RNAs with greater RBPs binding capacities will be more stable during viral infection compared to RNAs that bind to fewer RBPs overall.

The data shows that the average log2 fold change from Mock to 18 hpi of the set of higher binding genes was significantly higher than that of genes that bind to fewer RBPs. As shown in Figure 32, while most RNAs are down-regulated at 18 hpi by more than one fold due to viral infection, there is reduced change in expression level for RNAs that can bind to more RBPs.

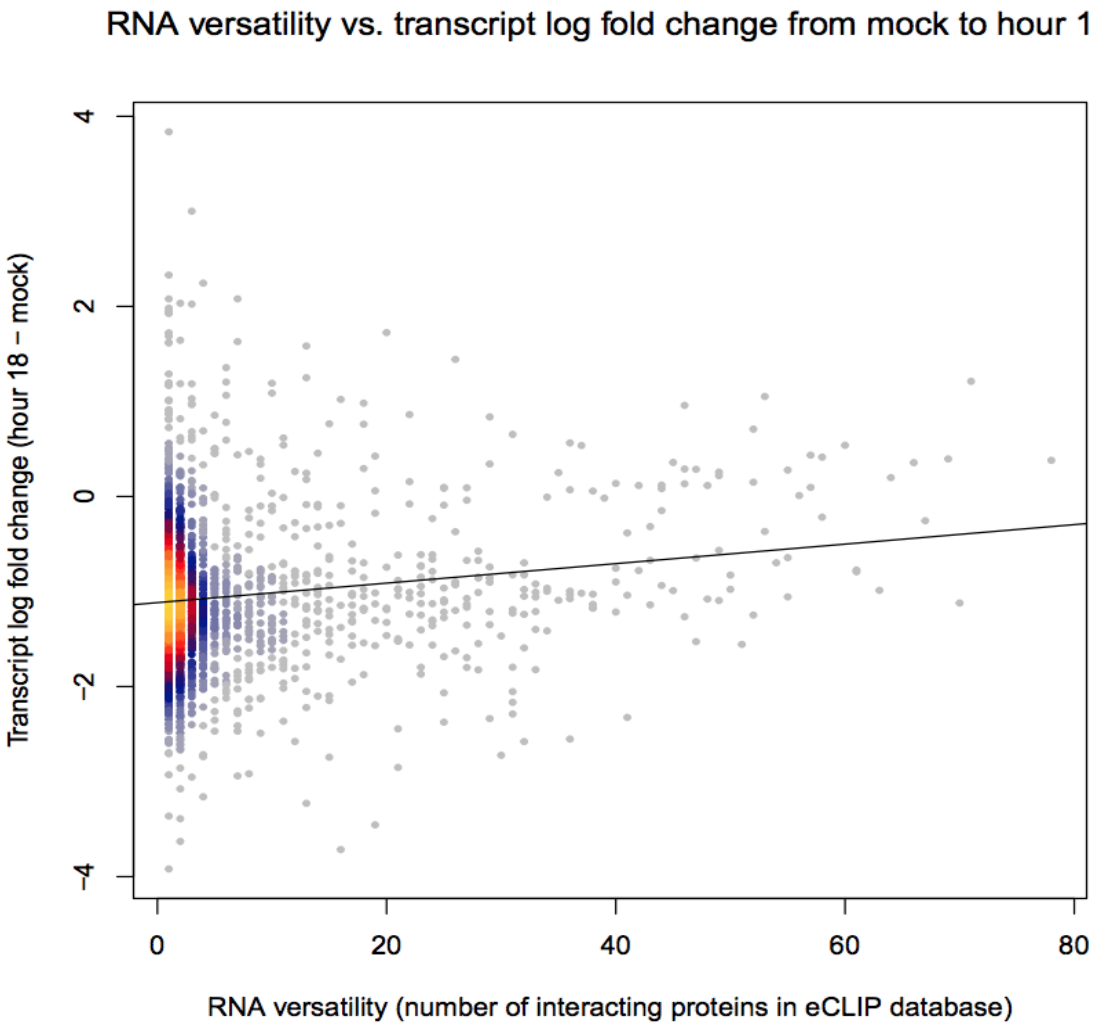

Figure 32. RNA versatility vs. transcript log fold change from Mock to 18 hpi. The x-axis gives number of possible interacting proteins, while log2 fold change of RNA abundance from uninfected to late infection stage is shown on the y-axis. The black line is a linear regression line between the average log2 fold change from Mock to 18 hpi and the possible number of proteins they may interact with

3.3.8) Predicting binding sites on viral RNA using DeepBind

DeepBind is a software that uses deep learning algorithms to analyze binding specificity of DNA and RNA sequences to which a set of proteins will bind. It is designed for the identification of genomic regions that are sensitive to deleterious mutations that can cause diseases. Trained with RNA-binding proteins and a set of their binding regions, DeepBind can analyze new sequences to and compute the likelihood of binding along the sequences for all possible RNA-binding

proteins. One of the critical objective of this study is to identify human RBPs that may bind to the RNA sequence of the Sindbis virus at 18hpi. Therefore, in the next analysis we used DeepBind to test the binding specificity for human originated RBPs on the RNA genome of the Sindbis virus.

First, to make sure that the binding sites suggested by DeepBind were more than just random hits, we compared the Deepbind scores for each RBP along with the virus with simulated scores obtained by 1000 random permutations of the Sindbis virus genome. To identify RBPs relevant to our study, we combined the information from Deepbind with the RBPs abundance change values from Mock to 18 hpi. This combined analysis successfully identified a list of RBPs that may bind to viral RNA and have a higher protein abundance at 18 hpi compared to Mock (Table 3). An example of predicted binding specificity for protein PABPC4 along the Sindbis virus genome is shown in Figure 33.

Table 3, Output of the DeepBind analysis showing the proteins that bound to viral RNA and had a higher protein abundance at 18 hours post infection compared to Mock (binding p-value <0.3)

	symbol	max	percentile	p.value	p.adj.deepbind	ENSGid	coefficient	p.adj.protein
13	PABPC4	5.7442	98.3	0.017	0.2183	ENSG00000090621	1.27898	0.007341
47	G3BP2	3.5803	91.3	0.087	0.4246	ENSG00000138757	0.40425	0.229163
48	HNRNPH2	13.2697	91.8	0.082	0.4246	ENSG00000126945	0.12824	0.654146
55	PCBP2	0.5832	90.1	0.099	0.4392	ENSG00000197111	0.08712	0.779228
73	PCBP1	5.6341	81.9	0.181	0.5957	ENSG00000169564	0.22001	0.421983
75	SNRNP70	2.1808	82.1	0.179	0.5957	ENSG00000104852	0.26753	0.481337
76	SRP54	2.9414	81.2	0.188	0.5957	ENSG00000100883	1.10059	0.081943
87	PABPC1	3.4635	78.9	0.211	0.6113	ENSG00000070756	1.06354	0.010322
92	TIA1	6.1766	76.3	0.237	0.6286	ENSG00000116001	0.40653	0.184288
98	RBM4	3.7620	74.4	0.256	0.6433	ENSG00000173933	0.42976	0.155194

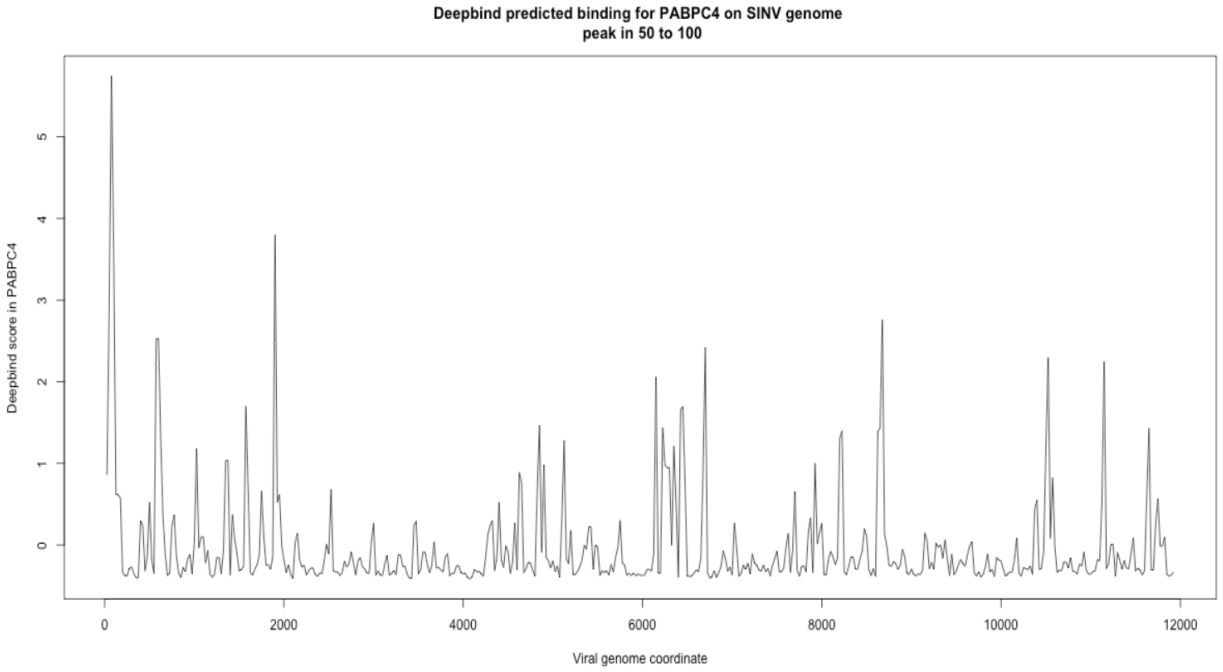

Figure 33, Predicted binding scores across the SINV genome using DeepBind for the PABPC4 protein. The x-axis shows the virus genome while the y-axis gives the binding score calculated by DeepBind.

3.4) Discussion

Our data reveals that virus infection causes a pervasive remodeling of the RBPome, modulating the activity of more than two hundred RBPs. These changes are likely driven by a combination of effects including global alterations of the transcriptome, and activation of cellular pathways related to RNA degradation and protein synthesis. We also revealed that RNA degradation plays a particularly critical role in regulating the RNA level in cell at 18hpi. Further, RNA regulation is likely modulated by the RNA binding protein XRN1. XRN1is a crucial molecular component of the 5’ to 3’ degradation pathway and is highly expressed at 18hpi. It is also demonstrated that the infection alters the codon composition at 18 hpi to favor the efficient synthesis of virus proteins.

Although we demonstrated that some of the human RBPs which increased their binding at 18 hpi have the potential to bind virus RNAs, it remains unclear which RBPs are exploited by virus. Our collaborative partner in the Castello group at Oxford University did a follow-up study with

i-CLIP using a stimulated ‘Gem-associated protein 5’ (GEMIN5) at 18 hpi (Garcia-Moreno, M., 2019). This analysis showed that GENIM5 binds to the cap and the 5’ and 3’ untranslated regions (UTR) in SINV RNAs at 18 hpi. Taking GEMIN5 as an example, it is likely that some other stimulated RBPs at 18 hpi may also interact with the Sindbis virus, and demonstrate either pro- or anti-viral activities.

3.4.1) Lack of annotation of rRNAs in public databases

The GENECODE annotation file only features four regions on chromosome 21 which are annotated as 8s rRNA and no 28s rRNA is annotated. The newest annotation file from ENSEMBL has fewer regions annotated as 8s rRNA on chromosome 21, and also has no annotated information about 28s rRNA. Unexpectedly, in this study we found a relatively comprehensive annotation of rRNAs in the human genome. For instance, in a sample of Mock experiment, which has the same experimental setup as other samples, more than 90% of the mapped reads had annotations as rRNA, as shown in Figure34.

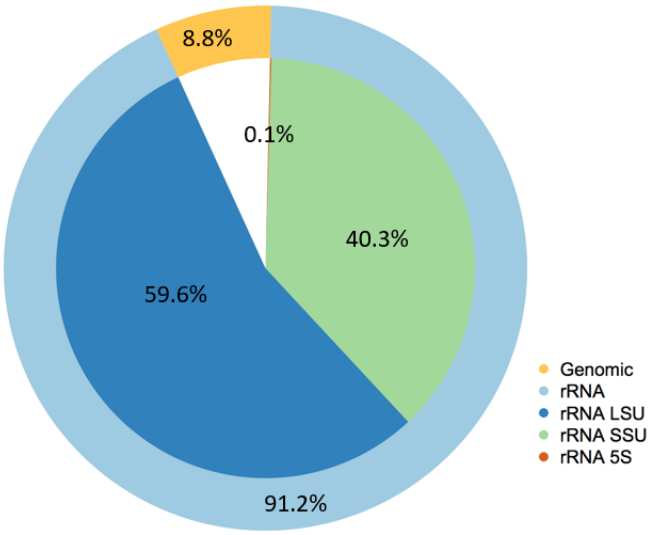

Figure 34, Distribution of mapped reads in different rRNAs in an example sample at Mock stage, the corresponding percentage are shown under each section.

3.4.2) Peak on reverse strand of virus genome at 18 hpi

Amongst our findings, the sharp peak on the reverse strand of the viral genome at 18 hpi drew special attention (Figure 21 in 3.3.1). This small peak at the 7800-8000bp region and has more than 10 fold increased expression compared to other peaks in its vicinity. Approximately 93.6% of the reads mapped to this region are paired reads, and only a small fraction of the reads were mismatches, suggesting the peak is not a false positive. After closer investigation, we found that this region maps to an adenovirus (Figure 35). In 1973, malignant transformation of the HEK293 cell line was achieved by integrating of human adenovirus type 5 into the healthy human embryonic kidney cells in Alex van der Eb's laboratory. It seems that this piece of DNA has been latent and carried by HEK293 cell line as a part of its genetic information ever since. Although our experiment has not been affected, this phenomenon does raise some concerns about studying virus infection models within virus-transformed cell-lines.

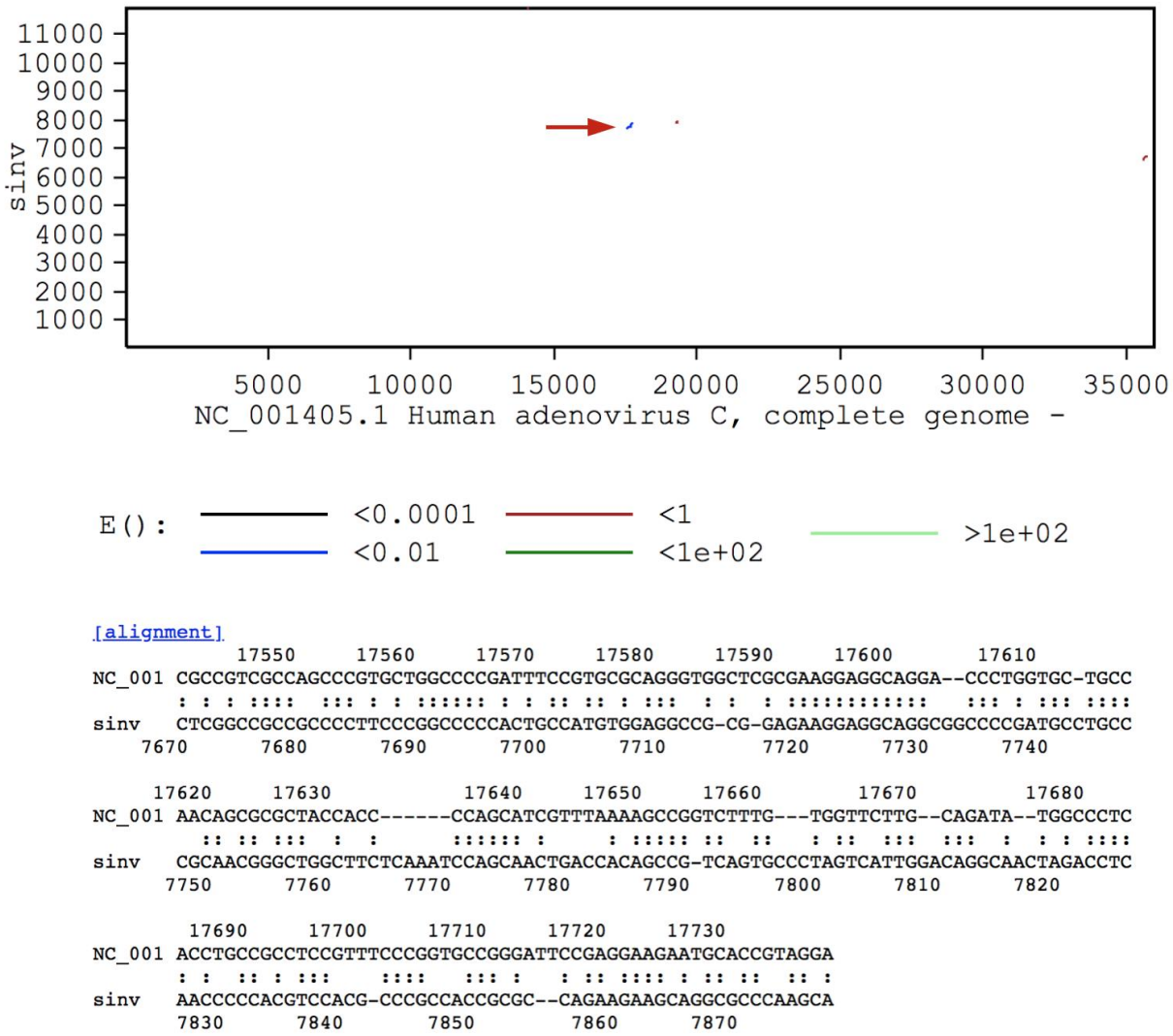

Figure 35, The local alignment of Sindbis virus and Adenovirus type 5, showing a considerable overlapping region between adenovirus type 5 (13548-13923) and Sindbis virus (7665-8038). The x axis and y axis represent the genome of adenovirus type 5 and Sindbis virus respectively. Regions with significant similarities are shown in color. A local sequence alignment is also shown in below.

This analysis uncovered 19 proteins that existed in both the DeepBind and ECLiP dataset. Given our hypothesis that binding to the viral genome will extend the functional life of proteins, we investigated the relationship of these proteins expression levels with the number of binding sites on their viral genome of these protein (DeepBInd value > 2). However, these proteins did not show correlation between the DeepBind binding score and their change in protein intensity. We speculate that this could be because this set of proteins could also bind to other RNAs, and the

RBPs that are exploited by the virus may not directly bind to the virus genome. Better conclusions can be made if there are more available RBPs in the eCLiP datasets.

4. The online interactive proteome differential analysis tool ‘Pepro’

4.1) Introduction

The identification and quantification of the proteome from biological samples has markedly improved with the advent of ‘next generation’ proteomics. However, proteomics data is unfortunately still somewhat ‘user unfriendly’ and is normally handled by trained experts only. Here we address the unmet need for a user friendly, accessible analysis tool with the development of our web-based application ‘Pepro’. Pepro uses a combination of well-established statistical methods for proteome analysis, offering an easy-handling and a friendly interface for differential analysis of quantified peptides. The software is running as online web application, which is available at https://nishuai.shinyapps.io/pepro/. Therefore, an up-to-date web browser is the only prerequisite for using the softare.

The Pepro application contains features which carry out peptides mapping to the proteome, quality checking, and removal of batch effects, interactive data visualization, and differential analysis. The core of Pepro lies in the implementation of a false discovery rate (FDR) controlled moderated t-test for peptide differential analysis. The input for Pepro is a table containing peptide sequence and their measured intensity in each experimental condition. This type of data can be found handy from popular next generation proteomics software such as MaxQuant (Tyanova, S., 2016; Cox, J., 2011). After feeding the data, Pepro guides users through initial peptide quantification to the visualization and download of results using simple step by step procedures. A comprehensive analysis pipeline of Pepro is shown in Figure 36.

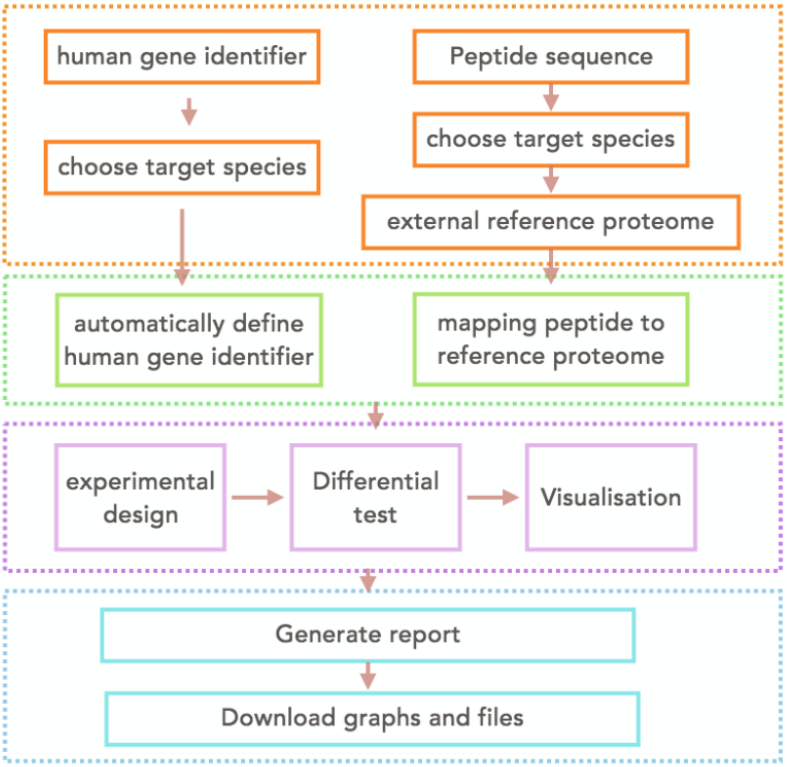

Figure 36, a schematic diagram of Pepro data analysis workflow

4.2) Methods

4.2.1) Shiny app platform

R is a programming environment that integrates statistical analysis and graphical display. It can run on both Windows and UNIX like operating systems. Compared with other statistical analysis software, R has the following features that contribute to its high popularity. First and foremost, it is free and open-source; all parts of R and R source code can be freely downloaded and distributed under GNU license. A second essential feature is its flexibility as a statistically orientated programming language, compared to popular statistical software like Stata or SAS. Packages and functions can be shared, and other users can edit or modify functions to meet their specific needs, to form a growing user ecosystem. So far, there are more than 8000 packages available for various types of analysis in different disciplines.

Shiny is an open-source R package that provides a framework for building interactive web applications. Taking advantage of R's strong capability in statistical computing and graphics, Shiny makes it easy to embed statistical analysis pipelines into a user-friendly web interface, without knowing other web-development languages. The web interface also very helpful in aiding people with no programming background to use statistical tools and visualize their data.

4.2.2) Quantitative proteome differential analysis

Proteins relevant to regulation may change expression under different biological conditions. To screen for these differentially expressed proteins, an accurate quantification of peptide/protein expression levels is critical. The differential protein screening process is achieved by statistical inference of the quantitative results and provides insight into biological significance. In this study, the proteome differential analysis was used to identify RNA binding proteins in human cell lines. The moderated t-test uses every single protein for differentiation. This produces a large number of independent tests which can elevate the FDR. We corrected the p-values for multiple testing by controlling the false discovery rate using the Benjamini-Hochberg method (Benjamini, Y., 1995)

Low or non-detectable protein abundance, technical issue with experimental design, miscleavage, imperfect ionization, and peptide misidentification can all lead to missing values that reduce the power to detect differentiated proteins. These challenging issues result in a high rate of missing data and limit the power to detect differentiated proteins (Tebbe, A., 2015; Wieczorek, S., 2017). Fortuntely, statistical approaches have been developed to overcome the challenge of missing values (Cho, H., 2007, Zhang, B., 2006, Karpievitch, Y., 2009, Ryu, S.Y., 2014). In Pepro, we choose a moderated t-test for protein differential analysis (Smyth, G.K., 2004). This test modifies the standard Student's t-test by replacing the ordinary standard deviation with posterior residual standard deviations for each comparison, providing a more robust test in a setting of high missingness (Smyth, G.K., 2004). We note that this t-test has only been found to work better than classical methods when the number of biological replications is larger or equal to 3 (Zhang, B., 2006). However, a minimum of 3 or more biological replicates is swiftly becoming a field standard (Unterlander, N., 2018; Gordon, A., 2018).

4.2.3) Semi-quantitative proteome differential analysis

For proteins for which the protein intensity was ‘zero’ in one of the two conditions compared, we applied a semi-quantitative approach that assumed proteins without quantitative information were below the detection limit, as described in Sysoev, V.O., et al. (Sysoev, V.O., 2016). This approach counts the number of replicates in each condition in which a given protein has an intensity value. When comparing two conditions and three biological replicates, this leads to a matrix with 16 different groups (detected 0, 1, 2 or 3 times in condition one versus detected 0, 1, 2 or 3 times in condition 2). A protein is classified as a ‘dynamic RBP’ by the semi-quantitative method if an intensity value is assigned to it in 2 or 3 of the replicates in one of the two conditions, with only 1 or 0 intensity values assigned to it in the other condition.

To validate the semi-quantitative method, the similarity of both quantitatively and semi-quantitatively identified RBPs in the A549 cell line to a known RBPome were measured. The known RBPome contains three sets of RBPs previously identified in 3 different human cell lines (Castello A., 2016). Using the A549 proteome, we classified the proteome into the following three categories: quantitatively identified RBP, semi-quantitatively identified RBP, and other proteins. The proteins in each category were further separated by counting instances where a protein matched with a known RBPome in the 3 cell lines. This could occur 0, 1, 2 and 3 times. The number 0 indicated that the identified A549 protein had not been previously reported as a RBP in any of the 3 cell lines, while 3 meant the identified proteins were reported as a RBPs in all 3 cell lines in the RBPome dataset and so forth. We termed this the “similarity distribution”. Next, we checked if the similarity distribution of semi-quantitatively identified RBPs resembled the similarity distribution of quantitatively identified RBPs, as shown in the Figure 37.

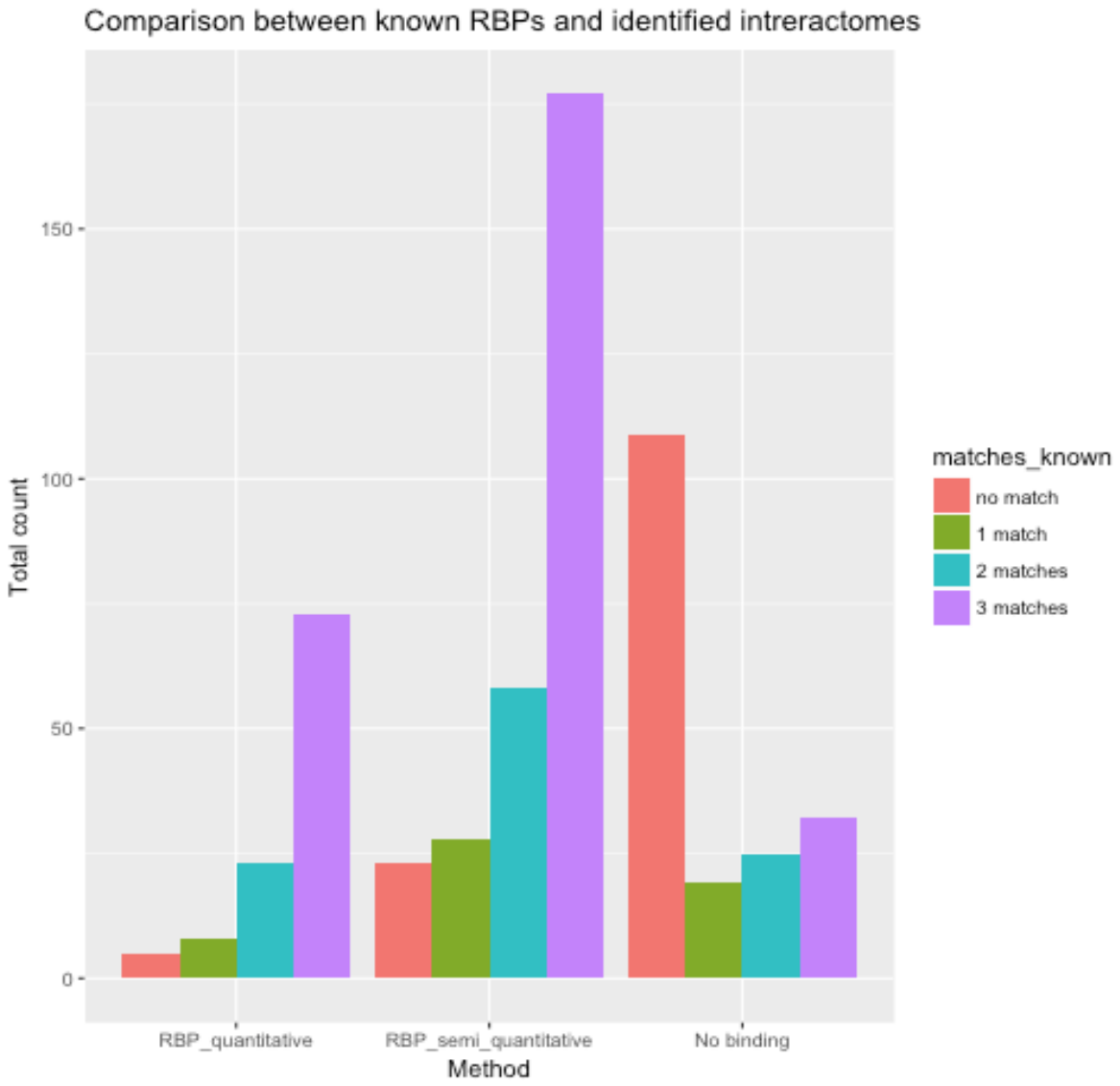

Figure 37: The similarity distribution in 3 categories to the known RBPome, identified RBPs by semi-quantification methods shows similar pattern to quantitatively identified RBPs

4.3) Discussion

4.3.1) Reactivity of Pepro

Reactivity is a deeply embedded part of Shiny and is one of Shiny's most essential features. With reactivity, web applications can instantly update in accordance with user input such as hovering, clicking, and zooming. Pepro takes advantage of reactivity in various ways to improve the user experience. In Pepro, the information provided by the user can simultaneously affect an

upcoming request from Pepro. For example, when defining pairs of samples for differential analysis, the user is provided with an input panel with two input areas where all sample names are available to choose from. Instantly after the user has filled the input panel with two sample names, the input panel is pushed down, and a new input panel will appear for more input. At the same time, Pepro will ensure that a sample can no longer be selected if it has has already been paired with all other samples. Besides the programmed reactivity, Pepro also makes use of recently developed packages to provide a sophisticated interactive display (Xie, Y., 2018, Dean Attali, D., 2016, Sievert, C., 2017). This means all plots and tables generated in Pepro are also interactive. To guide the user throughout the analysis pipeline, Pepro also gives users a global message that guides them through every step of the analysis and changes its content to inform the user about the next steps.

4.3.2) Robustness of Pepro

Pepro is designed to be an online app that serves a significant number of users. Therefore it is critical to have the ability to handle erroneous user input and exceptional operations. Pepro improves robustness in three main ways: via input control, modularization, and the use of wrappers for situations where the program might fail. Moreover, the input area in Prepro has been designed in different forms to ensure the input is restricted to a particular format. For example, when defining the number of conditions, a slider is used to expect an integer input; however when assigning sample names to conditions, the user is only allowed to select from a list of previously defined sample names. Controlling the input allows Pepro to ensure efficient and error-free analysis.

The Pepro backend is divided into four modules that perform cross checks before proceeding to the next module. To proceed, the user needs to click a button, which is only active when a defined condition is met. This condition validation feature ensures that all inputs and variables are valid in regard to the next module. The button becomes inactive again if the condition changes.

User-specified file uploading is an input area that is subject to an unpredictable input type where Pepro has less control. Even a robust file reading function will fail at some point given potentially

unlimited input file formats. To deal with this problem, Pepro uses a wrapper to the file reading function that encloses the potential error inside a smaller environment, limiting the possibility of more widespread ramifications. According to the wrapper test result, Prepro will continue to read the file or inform the user to check their input.

4.3.3) Protect user genetic privacy

Personal proteomics data contains a significant amount of sensitive information about an individual. Therefore it should be handled with special care to protect an individual's genetic privacy. It is advised that even sharing this type of data even anonymously should be restricted, especially if the data contains sufficient information to identify an individual, for example, in patient proteome sequencing data. While Pepro does not save or use any uploaded user-specific information, we cannot preclude the inherent risk that comes from uploading data to any server. Therefore, on the welcome page, Pepro users are reminded of data privacy concerns and asked to make sure they adhere to privacy regulations in their institute and country. The only information Pepro collects is the user IP region and total active time to compile basic usage statistics.

Although it is not advisable to share any individual identifiable proteomics data, two types of data are generally regarded as relatively safe. The first is proteomics data derived from well-known cell lines because it is publicly available for every research entity. The second is protein expression data, which does not contain an individual's genetic information. Although an individuals' SNP genotypes can potentially be predicted from RNA expression data (Schadt, E.E., 2012), the sheer size of public proteomics expression databases and lack of links between genotype to protein expression, make it is almost impossible to make that type of prediction from expression data. For this reason, Pepro is also offers differential analysis based on protein expression values and human gene identifiers, instead of peptide sequences.

5. Conclusions

This research is a combined effort of wet-lab experiments and statistical analysis workflows. Most of the experimental work in this study were carried at the Castello's lab at Oxford university, while Bernd and I focus on data analysis and development of the statistical methodologies. Some of the analytical tools we developed segued into independent side projects which will help others with similar research needs. Besides Pepro and RBDetect, we also created a web-based tool for comparing 2 groups of cell growth curves, named Growthcurves. Growthcurves can be used to discriminate differences between HEK293 cell growth under multiple culture conditions. This web application can be found at https://nishuai.shinyapps.io/growthcurves/.

The RIC shows great potential for identifying RNA-binding proteins and is experimentally robust and simple to implement. However, there are some experimental limitations that affect its specificity and sensitivity in RBP identification. It is reported that some DNA binding proteins and protein complexes also bind to RNAs (Conrad, T., 2016), making it possible that some proteins may be falsely classified as RBPs. Moreover, RNA interactome capture may underestimate the abundance of some RNA-binding proteins and their RNA counterparts. For instance, If the RNA-protein interaction site is spatially close to the mRNA poly(A) tail, it may obstruct the poly(A) tail exposure and inhibit oligo(dT) capture. This results in underestimation of the number of these RNA molecule and effects measurement precision of the corresponding RBP. However, these phenomena so rare it is extremely unlikely they will affect any global observations.

The standard application of mRNA sequencing usually involves the analysis of differential gene expression and alternative splicing events. However, the diversity of RNA species enabled us to uncover a lot of more interesting aspects with the examination of total cellular RNA. For example, by investigating the identified tRNA abundance and their associated anti codons, this study demonstrated that viral infection alters the codon composition of infected cells. We then show that these changes in codon composition favor translation of Sindbis viral proteins. This

finding is consistent previous research on other RNA viruses and underlines the utility of our methods (Zhou, J., 1999, Pavon-Eternod, M., 2012).

There is a strong focus in proteomics data analysis on peptide identification and quantification techniques, however, there is lack of user-friendly statistical methods for downstream differential protein analysis (Efstathiou, G., 2017). With its intuitive interface and straightforward design, Pepro makes the analysis of specialized proteomics data structures possible for end users with little knowledge in statistics. Users can set the experimental design, remove batch effects, carry out sophisticated quantitative and semi-quantitative differential expression analysis, interactively visualize, and save results in word or pdf format. Although some efforts have made to develop other user-friendly software (Efstathiou, G., 2017, Wieczorek, S., 2017), Pepro stands out from the crowd as one of the first truly comprehensive web-based toolkit that enables all the above-mentioned features. It fulfills an unmet need in a growing omics field that often includes many biologists who are non-specialists in this area.

In conclusion, this thesis reviewed the interplay dynamics between RNA and RBPs in human HEK293 cell line at three different viral infection stages. We observed a remodelling of binding activities of RBPs and the subsequent activation of the immune responses in the host cell. To our surprise most RBPs demonstrating altered RNA binding did not show protein-level changes. Besides using statistical methods to evaluate the relative effects of different RNA processes, we also demonstrated that RNA degradation pathways had the biggest contribution to changes in RNA abundance change in SINV infected cells. Similar machinery may also apply to other alphaviruses, such as Chikungunya and Mayaro viruses, and thus we hope this study may contribute for the development of drugs to help solving public health problems caused by similar viruses in some developing countries.

6. Reference

1. Alberts B, Johnson A, Lewis J, et al. 2002. Molecular Biology of the Cell. 4th edition. New York: Garland Science.
2. Alipanahi, B., A. Delong, M.T. Weirauch, B.J. Frey. 2015. Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. *Nat Biotechnol* 33(8) 831-838.
3. Aloni, Y., L.E. Hatlen, G. Attardi. 1971. Studies of fractionated HeLa cell metaphase chromosomes. II. chromosomal distribution of sites for transfer RNA and 5 s RNA. *J Mol Biol* 56(3) 555-563.
4. Anders, S., Huber, W., 2010. Differential expression analysis for sequence count data. *Genome Biology* 11, R106
5. Attali, D. 2016. Easily improve the user experience of your Shiny apps in seconds.
6. Aubert, M., M.F. O'Donohue, S. Lebaron, P.E. Gleizes. 2018. Pre-Ribosomal RNA Processing in Human Cells: From Mechanisms to Congenital Diseases. *Biomolecules* 8(4).
7. Barber, M., Robert S. Bordoli, R. Donald Sedgwick, Andrew N. Tyler, 1981, Fast atom bombardment of solids (F.A.B.): a new ion source for mass spectrometry, *Journal of the Chemical Society, Chemical Communications* 325-327
8. Baltz, A.G., M. Munschauer, B. Schwanhausser, A. Vasile, Y. Murakawa, M. Schueler, N. Youngs, D. Penfold-Brown, K. Drew, M. Milek, E. Wyler, R. Bonneau, M. Selbach, C. Dieterich, M. Landthaler. 2012. The mRNA-bound proteome and its global occupancy profile on protein-coding transcripts. *Mol Cell* 46(5) 674-690.
9. Benjamini, Y., Hochberg, Y., 1995. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. Journal of the Royal Statistical Society. Series B. 57-1, 289-300
10. Bjork, P., L. Wieslander. 2015. The Balbiani Ring Story: Synthesis, Assembly, Processing, and Transport of Specific Messenger RNA-Protein Complexes. *Annu Rev Biochem* 84 65-92.
11. Blakley, C.R., J.J. Carmody, M.L. Vestal. 1980. Liquid chromatograph-mass spectrometer for analysis of nonvolatile samples. *Analytical Chemistry* 52(11) 1636-1641.
12. Branton, D., D.W. Deamer, A. Marziali, H. Bayley, S.A. Benner, T. Butler, M. Di Ventra, S. Garaj, A. Hibbs, X. Huang, S.B. Jovanovich, P.S. Krstic, S. Lindsay, X.S. Ling, C.H. Mastrangelo, A. Meller, J.S. Oliver, Y.V. Pershin, J.M. Ramsey, R. Riehn, G.V. Soni, V. Tabard-Cossa, M. Wanunu, M. Wiggin, J.A. Schloss. 2008. The potential and challenges of nanopore sequencing. *Nat Biotechnol* 26(10) 1146-1153.
13. Brewer, G. 1991. An A + U-rich element RNA-binding factor regulates c-myc mRNA stability in vitro. *Mol Cell Biol* 11(5) 2460-2466.
14. Brooks, S.A., W.F. Rigby. 2000. Characterization of the mRNA ligands bound by the RNA binding protein hnRNP A2 utilizing a novel in vivo technique. *Nucleic Acids Res* 28(10) E49.
15. Buhler, J. 2001. Efficient large-scale sequence comparison by locality-sensitive hashing.

Bioinformatics 17(5) 419-428.
16. Burger, K., B. Muhl, M. Kellner, M. Rohrmoser, A. Gruber-Eber, L. Windhager, C.C. Friedel, L. Dolken, D. Eick. 2013. 4-thiouridine inhibits rRNA synthesis and causes a nucleolar stress response. *RNA Biol* 10(10) 1623-1630.
17. Burrows, M., D.J. Wheeler. 1994. A block sorting lossless data compression algorithm. *Technical Report*.
18. Butter, F., M. Scheibe, M. Morl, M. Mann. 2009. Unbiased RNA-protein interaction screen by quantitative proteomics. *Proc Natl Acad Sci U S A* 106(26) 10626-10631.
19. Cancer Genome Atlas Research, N., J.N. Weinstein, E.A. Collisson, G.B. Mills, K.R. Shaw, B.A. Ozenberger, K. Ellrott, I. Shmulevich, C. Sander, J.M. Stuart. 2013. The Cancer Genome Atlas Pan-Cancer analysis project. *Nat Genet* 45(10) 1113-1120.
20. Carrasco, L., M.A. Sanz, E. Gonzalez-Almela. 2018. The Regulation of Translation in Alphavirus-Infected Cells. *Viruses* 10(2).
21. Casneuf T., Van de Peer Y., Huber W., 2007, In situ analysis of cross-hybridisation on microarrays and the inference of expression correlation. *BMC Bioinformatics* 8:461
22. Castello, A., B. Fischer, K. Eichelbaum, R. Horos, Benedikt M. Beckmann, C. Strein, Norman E. Davey, David T. Humphreys, T. Preiss, Lars M. Steinmetz, J. Krijgsveld, Matthias W. Hentze. 2012. Insights into RNA Biology from an Atlas of Mammalian mRNA-Binding Proteins. *Cell* 149(6) 1393-1406.
23. Castello, A., B. Fischer, C.K. Frese, R. Horos, A.M. Alleaume, S. Foehr, T. Curk, J. Krijgsveld, M.W. Hentze. 2016. Comprehensive Identification of RNA-Binding Domains in Human Cells. *Mol Cell* 63(4) 696-710.
24. Castello, A., C.K. Frese, B. Fischer, A.I. Jarvelin, R. Horos, A.M. Alleaume, S. Foehr, T. Curk, J. Krijgsveld, M.W. Hentze. 2017. Identification of RNA-binding domains of RNA-binding proteins in cultured cells on a system-wide scale with RBDmap. *Nat Protoc* 12(12) 2447-2464.
25. Castello, A., R. Horos, C. Strein, B. Fischer, K. Eichelbaum, L.M. Steinmetz, J. Krijgsveld, M.W. Hentze. 2013. System-wide identification of RNA-binding proteins by interactome capture. *Nat Protoc* 8(3) 491-500.
26. Centers for Disease Control and Prevention. 2012. First Global Estimates of 2009 H1N1 Pandemic Mortality Released by CDC-Led Collaboration.
27. Cho, E.J., T. Takagi, C.R. Moore, S. Buratowski. 1997. mRNA capping enzyme is recruited to the transcription complex by phosphorylation of the RNA polymerase II carboxy-terminal domain. *Genes & development* 11(24) 3319-3326.
28. Cho, H., D.M. Smalley, D. Theodorescu, K. Ley, J.K. Lee. 2007. Statistical identification of differentially labeled peptides from liquid chromatography tandem mass spectrometry. *Proteomics* 7(20) 3681-3692.
29. Cole CN, Scarcelli JJ. 2006. Transport of messenger RNA from the nucleus to the cytoplasm. *Curr Opin Cell Biol*. 18(3):299-306
30. Conrad, T., A.S. Albrecht, V.R. de Melo Costa, S. Sauer, D. Meierhofer, U.A. Orom. 2016.

Serial interactome capture of the human cell nucleus. *Nat Commun* 7 11212.
31. Consortium, E.P. 2012. An integrated encyclopedia of DNA elements in the human genome. *Nature* 489(7414) 57-74.
32. Costa-Silva, J., D. Domingues, F.M. Lopes. 2017. RNA-Seq differential expression analysis: An extended review and a software tool. *PLoS One* 12(12) e0190152.
33. Cox, J., N. Neuhauser, A. Michalski, R.A. Scheltema, J.V. Olsen, M. Mann. 2011. Andromeda: A Peptide Search Engine Integrated into the MaxQuant Environment. *Journal of Proteome Research* 10(4) 1794-1805.
34. Crick, F.H. 1958. On protein synthesis. *Symp Soc Exp Biol* 12 138-163.
35. Cullen, B.R. 2004. Transcription and processing of human microRNA precursors. *Mol Cell* 16(6) 861-865.
36. Dean, J., S. Ghemawat. 2004. *MapReduce: Simplified Data Processing on Large Clusters*.
37. Deka, H., S. Chakraborty. 2014. Compositional Constraint Is the Key Force in Shaping Codon Usage Bias in Hemagglutinin Gene in H1N1 Subtype of Influenza A Virus. *International journal of genomics* 2014 349139.
38. Dempster, A. J., 1921. Positive Ray Analysis of Lithium and Magnesium, Phys. Rev. 18, 415
39. Deshler, J.O., M.I. Highett, T. Abramson, B.J. Schnapp. 1998. A highly conserved RNA-binding protein for cytoplasmic mRNA localization in vertebrates. *Current Biology* 8(9) 489-496.
40. Di Liegro, C.M., G. Schiera, I. Di Liegro. 2014. Regulation of mRNA transport, localization and translation in the nervous system of mammals (Review). *Int J Mol Med* 33(4) 747-762.
41. Ding, M., M.J. Schlesinger. 1989. Evidence that sindbis virus NSP2 is an autoprotease which processes the virus nonstructural polyprotein. *Virology* 171(1) 280-284.
42. Dobin, A., Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR. 2013, STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 29(1):15-21.
43. Domingo, E., C. Escarmis, N. Sevilla, A. Moya, S.F. Elena, J. Quer, I.S. Novella, J.J. Holland. 1996. Basic concepts in RNA virus evolution. *FASEB J* 10(8) 859-864.
44. Domingo, E., J.J. Holland. 1997. RNA virus mutations and fitness for survival. *Annu Rev Microbiol* 51 151-178.
45. Drake, J.W., J.J. Holland. 1999. Mutation rates among RNA viruses. *Proceedings of the National Academy of Sciences* 96(24) 13910.
46. Edery, I., N. Sonenberg. 1985. Cap-dependent RNA splicing in a HeLa nuclear extract. *Proc Natl Acad Sci U S A* 82(22) 7590-7594.
47. Edman, P., 1949. A method for the determination of amino acid sequence in peptides. *Arch Biochem* 22(3):475.
48. Efstathiou, G., A.N. Antonakis, G.A. Pavlopoulos, T. Theodosiou, P. Divanach, D.C. Trudgian, B. Thomas, N. Papanikolaou, M. Aivaliotis, O. Acuto, I. Iliopoulos. 2017. ProteoSign: an end-user online differential proteomics statistical analysis platform. *Nucleic Acids Res* 45(W1) W300-W306.
49. Foley, S.W., B.D. Gregory. 2016. Protein Interaction Profile Sequencing (PIP-seq). *Curr*

Protoc Mol Biol 116 27.25.21-27.25.15.
50. Freire, C., G. Palmisano, C. Braconi, F. Cugola, F. Russo, P. Beltrao Braga, A. Iamarino, D. Ferreira de Lima Neto, A. Sall, L. Rosa Fernandes, M. Larsen, P. Zanotto. 2018. NS1 codon usage adaptation to humans in pandemic Zika virus. *Memórias do Instituto Oswaldo Cruz* 113.
51. Garcia-Moreno, M., M. Noerenberg, S. Ni, A.I. Jarvelin, E. Gonzalez-Almela, C.E. Lenz, M. Bach-Pages, V. Cox, R. Avolio, T. Davis, S. Hester, T.J.M. Sohier, B. Li, G. Heikel, G. Michlewski, M.A. Sanz, L. Carrasco, E.P. Ricci, V. Pelechano, I. Davis, B. Fischer, S. Mohammed, A. Castello. 2019. System-wide Profiling of RNA-Binding Proteins Uncovers Key Regulators of Virus Infection. *Mol Cell* 74(1) 196-211 e111.
52. Gebauer, F., T. Preiss, M.W. Hentze. 2012. From cis-regulatory elements to complex RNPs and back. *Cold Spring Harb Perspect Biol* 4(7) a012245.
53. Geiduschek, E.P., T. Nakamoto, S.B. Weiss. 1961. The enzymatic synthesis of RNA: complementary interaction with DNA. *Proc Natl Acad Sci U S A* 47 1405-1415.
54. Gerstberger, S., M. Hafner, T. Tuschl. 2014. A census of human RNA-binding proteins. *Nat Rev Genet* 15(12) 829-845.
55. Glisovic, T., J.L. Bachorik, J. Yong, G. Dreyfuss. 2008. RNA-binding proteins and post-transcriptional gene regulation. *FEBS Lett* 582(14) 1977-1986.
56. Gorchakov, R., et al., 2005. Inhibition of Transcription and Translation in Sindbis Virus-Infected Cells. Journal of Virology. 79.15.9397-9409
57. Gordon, A., S.K. Kannan, K. Gousset. 2018. A Novel Cell Fixation Method that Greatly Enhances Protein Identification in Microproteomic Studies Using Laser Capture Microdissection and Mass Spectrometry. *Proteomics* 18(11) e1700294-e1700294.
58. Gualerzi, C.O., C.L. Pon. 1990. Initiation of mRNA translation in prokaryotes. *Biochemistry* 29(25) 5881-5889.
59. Gylfe, A., Ribers, Å., Forsman, O., Bucht, G., Alenius, G., Wållberg-Jonsson, S....Evander, M. (2018). Mosquitoborne Sindbis Virus Infection and Long-Term Illness. *Emerging Infectious Diseases* 24(6), 1141-1142
60. Hafner, M., M. Landthaler, L. Burger, M. Khorshid, J. Hausser, P. Berninger, A. Rothballer, M. Ascano, Jr., A.C. Jungkamp, M. Munschauer, A. Ulrich, G.S. Wardle, S. Dewell, M. Zavolan, T. Tuschl. 2010. Transcriptome-wide identification of RNA-binding protein and microRNA target sites by PAR-CLIP. *Cell* 141(1) 129-141.
61. Hahn, S. 2004. Structure and mechanism of the RNA polymerase II transcription machinery. *Nat Struct Mol Biol* 11(5) 394-403.
62. He, F., A. Jacobson. 2015. Nonsense-Mediated mRNA Decay: Degradation of Defective Transcripts Is Only Part of the Story. *Annual Review of Genetics* 49(1) 339-366.
63. Heller, D.N., C. Fenselau, R.J. Cotter, P. Demirev, J.K. Olthoff, J. Honovich, M. Uy, T. Tanaka, Y. Kishimoto. 1987. Mass spectral analysis of complex lipids desorbed directly from lyophilized membranes and cells. *Biochem Biophys Res Commun* 142(1) 194-199.
64. Hentze M.W., Castello A., Schwarzl T., Preiss T. 2018. A brave new world of RNA-binding proteins *Nat Rev Mol Cell Biol* 19(5):327-341.

65. Hershberg, R., D. Petrov. 2008. Selection on Codon Bias. *Annual review of genetics* 42 287-299.
66. Houseley, J., Tollervey, D., 2009. The Many Pathways of RNA Degradation. Cell 4-20, 763-776
67. Huynh, M. L., Russell, P., Walsh, B. 2009. Tryptic digestion of in-gel proteins for mass spectrometry analysis. *Methods in molecular biology* 519, 507-513
68. Jacob, A.G., Smith, C.W.J. 2017. Intron retention as a component of regulated gene expression programs. *Hum Genet* 136, 1043–1057
69. Jeanteur, P., G. Attardi. 1969. Relationship between HeLa cell ribosomal RNA and its precursors studied by high resolution RNA-DNA hybridization. *Journal of Molecular Biology* 45(2) 305-324.
70. Kammoun, M. 2016. RNA-binding site prediction using classification methods.
71. Kaner, J., S. Schaack. 2016. Understanding Ebola: the 2014 epidemic. *Global Health* 12(1) 53.
72. Karpievitch, Y., J. Stanley, T. Taverner, J. Huang, J.N. Adkins, C. Ansong, F. Heffron, T.O. Metz, W.J. Qian, H. Yoon, R.D. Smith, A.R. Dabney. 2009. A statistical framework for protein quantitation in bottom-up MS-based proteomics. *Bioinformatics* 25(16) 2028-2034.
73. Kim, M.S., 2014. A draft map of the human proteome. *Nature* 509(7502): 575–581.
74. König, J., K. Zarnack, G. Rot, T. Curk, M. Kayikci, B. Zupan, D.J. Turner, N.M. Luscombe, J. Ule. 2010. iCLIP reveals the function of hnRNP particles in splicing at individual nucleotide resolution. *Nature structural & molecular biology* 17(7) 909-915.
75. Kramer, K., T. Sachsenberg, B.M. Beckmann, S. Qamar, K.L. Boon, M.W. Hentze, O. Kohlbacher, H. Urlaub. 2014. Photo-cross-linking and high-resolution mass spectrometry for assignment of RNA-binding sites in RNA-binding proteins. *Nat Methods* 11(10) 1064-1070.
76. Kumar, N., B. Bera, B. Greenbaum, S. Bhatia, R. Sood, S. Pavulraj, T. Anand, B.N. Tripathi, N. Virmani. 2016. Revelation of Influencing Factors in Overall Codon Usage Bias of Equine Influenza Viruses. *PloS one* 11 e0154376.
77. Kung, J.T., D. Colognori, J.T. Lee. 2013. Long noncoding RNAs: past, present, and future. *Genetics* 193(3) 651-669.
78. Kurkela, S., Rätti, O., Huhtamo, E., Uzcátegui, N. Y., Nuorti, J. P., Laakkonen, J., Manni, T., Helle, P., Vaheri, A., & Vapalahti, O. 2008. Sindbis virus infection in resident birds, migratory birds, and humans, Finland. *Emerging infectious diseases* 14(1), 41–47
79. Labno, A., R. Tomecki, A. Dziembowski. 2016. Cytoplasmic RNA decay pathways - Enzymes and mechanisms. *Biochim Biophys Acta* 1863(12) 3125-3147.
80. Laine, M., Luukkainen, R., Toivanen, A., 2004. Sindbis viruses and other alphaviruses as cause of human arthritic disease. J Intern Med 256(6):457-71
81. Lee, W.P., M.P. Stromberg, A. Ward, C. Stewart, E.P. Garrison, G.T. Marth. 2014. MOSAIK: a hash-based algorithm for accurate next-generation sequencing short-read mapping. *PLoS One* 9(3) e90581.
82. Leung, J.Y.-S., M.M.-L. Ng, J.J.H. Chu. 2011. Replication of alphaviruses: a review on the

entry process of alphaviruses into cells. *Adv Virol* 2011 249640-249640.
83. Levy S., et al. 2007. The Diploid Genome Sequence of an Individual Human. *Plos Biology* 5(10): e254.
84. Li, H., J. Ruan, R. Durbin. 2008a. Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome research* 18(11) 1851-1858.
85. Li, H., Handsaker, B., Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R; 1000 Genome Project Data Processing Subgroup. 2009. The Sequence Alignment/Map format and SAMtools. Bioinformatics 25(16):2078-9.
86. Li, R., Y. Li, K. Kristiansen, J. Wang. 2008b. SOAP: short oligonucleotide alignment program. *Bioinformatics* 24(5) 713-714.
87. Li, R., C. Yu, Y. Li, T.W. Lam, S.M. Yiu, K. Kristiansen, J. Wang. 2009. SOAP2: an improved ultrafast tool for short read alignment. *Bioinformatics* 25(15) 1966-1967.
88. Li, Z., P.D. Nagy. 2011. Diverse roles of host RNA binding proteins in RNA virus replication. *RNA Biol* 8(2) 305-315.
89. Liao, Y., Smyth, G.K., Shi, W., 2014. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinforamtics 30(7):923-30
90. Licatalosi, D.D., A. Mele, J.J. Fak, J. Ule, M. Kayikci, S.W. Chi, T.A. Clark, A.C. Schweitzer, J.E. Blume, X. Wang, J.C. Darnell, R.B. Darnell. 2008. HITS-CLIP yields genome-wide insights into brain alternative RNA processing. *Nature* 456(7221) 464-469.
91. Long, Y.C., Wang, X.Y., Youmans, D.T., Cech, T.R., 2017, How do lncRNAs regulate transcription? *Sci Adv* 3(9): eaao2110.
92. MacBeath, G., S.L. Schreiber. 2000. Printing Proteins as Microarrays for High-Throughput Function Determination. *Science* 289(5485) 1760.
93. Macfarlane, R.D., D.F. Torgerson. 1976. Californium-252 plasma desorption mass spectroscopy. *Science* 191(4230) 920-925.
94. Maniatis, T., R. Reed. 2002. An extensive network of coupling among gene expression machines. *Nature* 416(6880) 499-506.
95. Marcais, G., C. Kingsford. 2011. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics* 27(6) 764-770.
96. Margulies, M., et al., 2005. Genome sequencing in microfabricated high-density picolitre reactors, Nature 437, 376–380
97. Marra, M.A., S.J. Jones, C.R. Astell, R.A. Holt, A. Brooks-Wilson, Y.S. Butterfield, J. Khattra, J.K. Asano, S.A. Barber, S.Y. Chan, A. Cloutier, S.M. Coughlin, D. Freeman, N. Girn, O.L. Griffith, S.R. Leach, M. Mayo, H. McDonald, S.B. Montgomery, P.K. Pandoh, A.S. Petrescu, A.G. Robertson, J.E. Schein, A. Siddiqui, D.E. Smailus, J.M. Stott, G.S. Yang, F. Plummer, A. Andonov, H. Artsob, N. Bastien, K. Bernard, T.F. Booth, D. Bowness, M. Czub, M. Drebot, L. Fernando, R. Flick, M. Garbutt, M. Gray, A. Grolla, S. Jones, H. Feldmann, A. Meyers, A. Kabani, Y. Li, S. Normand, U. Stroher, G.A. Tipples, S. Tyler, R. Vogrig, D. Ward, B. Watson, R.C. Brunham, M. Krajden, M. Petric, D.M. Skowronski, C. Upton, R.L. Roper. 2003. The Genome sequence of the SARS-associated coronavirus. *Science* 300(5624) 1399-1404.

98. Matthew N.B., 2006, Analysis of the prostate cancer cell line LNCaP transcriptome using a sequencing-by-synthesis approach. *BMC Genomics* 7, 246
99. McCracken, S., N. Fong, E. Rosonina, K. Yankulov, G. Brothers, D. Siderovski, A. Hessel, S. Foster, S. Shuman, D.L. Bentley. 1997. 5'-Capping enzymes are targeted to pre-mRNA by binding to the phosphorylated carboxy-terminal domain of RNA polymerase II. *Genes Dev* 11(24) 3306-3318.
100. McGeoch, D., P. Fellner, C. Newton. 1976. Influenza virus genome consists of eight distinct RNA species. *Proc Natl Acad Sci U S A* 73(9) 3045-3049.
101. McKenna, A., M. Hanna, E. Banks, A. Sivachenko, K. Cibulskis, A. Kernytsky, K. Garimella, D. Altshuler, S. Gabriel, M. Daly, M.A. DePristo. 2010. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* 20(9) 1297-1303.
102. Misra, S., A. Agrawal, W.K. Liao, A. Choudhary. 2011. Anatomy of a hash-based long read sequence mapping algorithm for next generation DNA sequencing. *Bioinformatics* 27(2) 189-195.
103. Morikawa, N., F. Imamoto. 1969. Degradation of Tryptophan Messenger: On the Degradation of Messenger RNA for the Tryptophan Operon in Escherichia coli. *Nature* 223(5201) 37-40.
104. Morris, H.R., et al., 1981. Fast atom bombardment: A new mass spectrometric method for peptide sequence analysis. *Biochemical and Biophysical Research Communications* 623-631
105. Moya, A., S.F. Elena, A. Bracho, R. Miralles, E. Barrio. 2000. The evolution of RNA viruses: A population genetics view. *Proc Natl Acad Sci U S A* 97(13) 6967-6973.
106. Mukherjee, N., Calviello, L., Hirsekorn, A., 2017. Integrative classification of human coding and noncoding genes through RNA metabolism profiles. *Nat Struct Mol Biol* 24, 86–96
107. Niwa, M., S.D. Rose, S.M. Berget. 1990. In vitro polyadenylation is stimulated by the presence of an upstream intron. *Genes & Development* 4(9) 1552-1559.
108. Ong, S.E., B. Blagoev, I. Kratchmarova, D.B. Kristensen, H. Steen, A. Pandey, M. Mann. 2002. Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics. *Mol Cell Proteomics* 1(5) 376-386.
109. Oshlack, A., M.D. Robinson, M.D. Young. 2010. From RNA-seq reads to differential expression results. *Genome Biol* 11(12) 220.
110. Pandit, A., S. Sinha. 2011. Differential trends in the codon usage patterns in HIV-1 genes. *PLoS One* 6(12) e28889.
111. Patel, R.K., M. Jain. 2012. NGS QC Toolkit: a toolkit for quality control of next generation sequencing data. *PLoS One* 7(2) e30619.
112. Pavon-Eternod, M., A. David, K. Dittmar, P. Berglund, T. Pan, J. Bennink, J. Yewdell. 2012. Vaccinia and influenza A viruses select rather than adjust tRNAs to optimize translation. *Nucleic acids research* 41.
113. Penalva, L.O., S.A. Tenenbaum, J.D. Keene. 2004. Gene expression analysis of messenger RNP complexes. *Methods Mol Biol* 257 125-134.
114. Perez-Cano, L., A. Solernou, C. Pons, J. Fernandez-Recio. 2010. Structural prediction of

protein-RNA interaction by computational docking with propensity-based statistical potentials. *Pac Symp Biocomput* 293-301.

115. Peter, G., H. Robert. 1969. Messenger RNA. *Annual Review of Biochemistry*.
116. Pretis, S.D., Theresia Kress, Marco J. Morelli, Giorgio E. M. Melloni, Laura Riva, Bruno Amati, Mattia Pelizzola. 2015. INSPEcT: a computational tool to infer mRNA synthesis, processing and degradation dynamics from RNA- and 4sU-seq time course experiments, *Bioinformatics* 31-17, 2829–2835
117. Quail, M.A., M. Smith, P. Coupland, T.D. Otto, S.R. Harris, T.R. Connor, A. Bertoni, H.P. Swerdlow, Y. Gu. 2012. A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. *BMC Genomics* 13 341.
118. Quinlan, A.R., Hall, I.M., 2010. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26(6):841-2
119. Rasmussen, E.B., J.T. Lis. 1993. In vivo transcriptional pausing and cap formation on three Drosophila heat shock genes. *Proc Natl Acad Sci U S A* 90(17) 7923-7927.
120. Regev, A., S.A. Teichmann, E.S. Lander, I. Amit, C. Benoist, E. Birney, B. Bodenmiller, P. Campbell, P. Carninci, M. Clatworthy, H. Clevers, B. Deplancke, I. Dunham, J. Eberwine, R. Eils, W. Enard, A. Farmer, L. Fugger, B. Gottgens, N. Hacohen, M. Haniffa, M. Hemberg, S. Kim, P. Klenerman, A. Kriegstein, E. Lein, S. Linnarsson, E. Lundberg, J. Lundeberg, P. Majumder, J.C. Marioni, M. Merad, M. Mhlanga, M. Nawijn, M. Netea, G. Nolan, D. Pe'er, A. Phillipakis, C.P. Ponting, S. Quake, W. Reik, O. Rozenblatt-Rosen, J. Sanes, R. Satija, T.N. Schumacher, A. Shalek, E. Shapiro, P. Sharma, J.W. Shin, O. Stegle, M. Stratton, M.J.T. Stubbington, F.J. Theis, M. Uhlen, A. van Oudenaarden, A. Wagner, F. Watt, J. Weissman, B. Wold, R. Xavier, N. Yosef, P. Human Cell Atlas Meeting. 2017. The Human Cell Atlas. *Elife* 6.
121. Rich, A., U.L. RajBhandary. 1976. Transfer RNA: molecular structure, sequence, and properties. *Annu Rev Biochem* 45 805-860.
122. Ross, A.F., Y. Oleynikov, E.H. Kislauskis, K.L. Taneja, R.H. Singer. 1997. Characterization of a beta-actin mRNA zipcode-binding protein. *Mol Cell Biol* 17(4) 2158-2165.
123. Ryu, S.Y., W.J. Qian, D.G. Camp, R.D. Smith, R.G. Tompkins, R.W. Davis, W. Xiao. 2014. Detecting differential protein expression in large-scale population proteomics. *Bioinformatics* 30(19) 2741-2746.
124. Sanger, F., Nicklen, S., Coulson, A.R., 1977. DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci U S A.* 74(12): 5463–5467.
125. Saldaña-Meyer, R., E. González-Buendía, G. Guerrero, V. Narendra, R. Bonasio, F. Recillas-Targa, D. Reinberg. 2014. CTCF regulates the human p53 gene through direct interaction with its natural antisense transcript, Wrap53. *Genes & Development* 28(7) 723-734.
126. Schadt, E.E., S. Woo, K. Hao. 2012. Bayesian method to predict individual SNP genotypes from gene expression data. *Nature Genetics* 44(5) 603-608.
127. Schena M., Shalon D., Davis R.W., Brown P.O., 1995, Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* 270(5235) 467-70

128. Scherrer, T., N. Mittal, S.C. Janga, A.P. Gerber. 2010. A screen for RNA-binding proteins in yeast indicates dual functions for many enzymes. *PLoS One* 5(11) e15499.
129. Schoenberg, D.R., L.E. Maquat. 2012. Regulation of cytoplasmic mRNA decay. *Nat Rev Genet* 13(4) 246-259.
130. Sievert, C., C. Parmer, T. Hocking, S. Chamberlai, K. Ram, M. Corvellec, P. Despouy. 2017. plotly: Create Interactive Web Graphics via 'plotly.js'.
131. Simon, M.D. 2013. Capture Hybridization Analysis of RNA Targets (CHART). *Curr Protoc Mol Biol* 101(1) 21.25.21-21.25.16.
132. Smyth, G.K. 2004. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Stat Appl Genet Mol Biol* 3 Article3.
133. Srimani, J.K., P.Y. Wu, J.H. Phan, M.D. Wang. 2010. A distributed system for fast alignment of next-generation sequencing data. *IEEE Int Conf Bioinform Biomed Workshops* 2010 579-584.
134. Stewart, D., S. Dhungana, R. Clark, W. Pathmasiri, S. McRitchie, S. Sumner. 2015. Chapter 4 - Omics Technologies Used in Systems Biology. R.C. Fry, ed. *Systems Biology in Toxicology and Environmental Health*. Academic Press, Boston, 57-83.
135. Strauss, E.G., C.M. Rice, J.H. Strauss. 1983. Sequence coding for the alphavirus nonstructural proteins is interrupted by an opal termination codon. *Proceedings of the National Academy of Sciences of the United States of America* 80(17) 5271-5275.
136. Strauss, E.G., C.M. Rice, J.H. Strauss. 1984. Complete nucleotide sequence of the genomic RNA of Sindbis virus. *Virology* 133(1) 92-110.
137. Strauss, J.H., E.G. Strauss. 1994. The alphaviruses: gene expression, replication, and evolution. *Microbiol Rev* 58(3) 491-562.
138. Sysoev, V.O., B. Fischer, C.K. Frese, I. Gupta, J. Krijgsveld, M.W. Hentze, A. Castello, A. Ephrussi. 2016. Global changes of the RNA-bound proteome during the maternal-to-zygotic transition in Drosophila. *Nat Commun* 7 12128.
139. Takamizawa, A., C. Mori, I. Fuke, S. Manabe, S. Murakami, J. Fujita, E. Onishi, T. Andoh, I. Yoshida, H. Okayama. 1991. Structure and organization of the hepatitis C virus genome isolated from human carriers. *J Virol* 65(3) 1105-1113.
140. Tebbe, A., M. Klammer, S. Sighart, C. Schaab, H. Daub. 2015. Systematic evaluation of label-free and super-SILAC quantification for proteome expression analysis. *Rapid Commun Mass Spectrom* 29(9) 795-801.
141. Tenenbaum, S.A., Carson, C.C., Lager, P.J., Keene, J.D., 2000. Identifying mRNA subsets in messenger ribonucleoprotein complexes by using cDNA arrays. Proc Natl Acad Sci U S A. 97(26):14085-90
142. Tenenbaum, S.A., P.J. Lager, C.C. Carson, J.D. Keene. 2002. Ribonomics: identifying mRNA subsets in mRNP complexes using antibodies to RNA-binding proteins and genomic arrays. *Methods* 26(2) 191-198.
143. Terribilini, M., J.D. Sander, J.-H. Lee, P. Zaback, R.L. Jernigan, V. Honavar, D. Dobbs. 2007. RNABindR: a server for analyzing and predicting RNA-binding sites in proteins. *Nucleic*

Acids Research 35(suppl_2) W578-W584.
144. Trapnell, C., L. Pachter, S.L. Salzberg. 2009. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* 25(9) 1105-1111.
145. Treiber, T., N. Treiber, U. Plessmann, S. Harlander, J.-L. Daiß, N. Eichner, G. Lehmann, K. Schall, H. Urlaub, G. Meister. 2017. A Compendium of RNA-Binding Proteins that Regulate MicroRNA Biogenesis. *Molecular Cell* 66(2) 270-284.e213.
146. Tsvetanova, N.G., D.M. Klass, J. Salzman, P.O. Brown. 2010. Proteome-wide search reveals unexpected RNA-binding proteins in Saccharomyces cerevisiae. *PLoS One* 5(9).
147. Tyanova, S., T. Temu, J. Cox. 2016. The MaxQuant computational platform for mass spectrometry-based shotgun proteomics. *Nat Protoc* 11(12) 2301-2319.
148. Ule, J., K.B. Jensen, M. Ruggiu, A. Mele, A. Ule, R.B. Darnell. 2003. CLIP identifies Nova-regulated RNA networks in the brain. *Science* 302(5648) 1212-1215.
149. Unterlander, N., A.A. Doucette. 2018. Membrane-Based SDS Depletion Ahead of Peptide and Protein Analysis by Mass Spectrometry. *Proteomics* 18(9) e1700025.
150. Urlaub, H., K. Hartmuth, R. Luhrmann. 2002. A two-tracked approach to analyze RNA-protein crosslinking sites in native, nonlabeled small nuclear ribonucleoprotein particles. *Methods* 26(2) 170-181.
151. Van Nostrand, E.L., P. Freese, G.A. Pratt, X. Wang, X. Wei, R. Xiao, S.M. Blue, J.-Y. Chen, N.A.L. Cody, D. Dominguez, S. Olson, B. Sundararaman, L. Zhan, C. Bazile, L.P.B. Bouvrette, J. Bergalet, M.O. Duff, K.E. Garcia, C. Gelboin-Burkhart, M. Hochman, N.J. Lambert, H. Li, T.B. Nguyen, T. Palden, I. Rabano, S. Sathe, R. Stanton, A. Su, R. Wang, B.A. Yee, B. Zhou, A.L. Louie, S. Aigner, X.-d. Fu, E. Lécuyer, C.B. Burge, B.R. Graveley, G.W. Yeo. 2018. A Large-Scale Binding and Functional Map of Human RNA Binding Proteins. *bioRxiv* 179648.
152. Van Nostrand, E.L., G.A. Pratt, A.A. Shishkin, C. Gelboin-Burkhart, M.Y. Fang, B. Sundararaman, S.M. Blue, T.B. Nguyen, C. Surka, K. Elkins, R. Stanton, F. Rigo, M. Guttman, G.W. Yeo. 2016. Robust transcriptome-wide discovery of RNA-binding protein binding sites with enhanced CLIP (eCLIP). *Nat Methods* 13(6) 508-514.
153. Wang, L., S.J. Brown. 2006. BindN: a web-based tool for efficient prediction of DNA and RNA binding sites in amino acid sequences. *Nucleic Acids Res* 34(Web Server issue) W243-248.
154. M. Gerstein, M. Snyder. 2009. RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics* 10(1) 57-63.
155. Wang, Z., M. Kayikci, M. Briese, K. Zarnack, N.M. Luscombe, G. Rot, B. Zupan, T. Curk, J. Ule. 2010. iCLIP predicts the dual splicing effects of TIA-RNA interactions. *PLoS Biol* 8(10) e1000530.
156. Welch, B. L., 1951. On the Comparison of Several Mean Values: An Alternative Approach. *Biometrika* 38 (3/4): 330–336.
157. Wickham, H., 2009. Ggplot2: *Elegant Graphics for Data Analysis. 2nd Edition*, *Springer, New York.*

158. Wieczorek, S., F. Combes, C. Lazar, Q. Giai Gianetto, L. Gatto, A. Dorffer, A.M. Hesse, Y. Coute, M. Ferro, C. Bruley, T. Burger. 2017. DAPAR & ProStaR: software to perform statistical analyses in quantitative discovery proteomics. *Bioinformatics* 33(1) 135-136.
159. Will, C.L., R. Luhrmann. 2011. Spliceosome structure and function. *Cold Spring Harb Perspect Biol* 3(7).
160. Wong, J., Gao, D., Nguyen, T. et al. 2017. Intron retention is regulated by altered MeCP2-mediated splicing factor recruitment. *Nat Commun* 8, 15134
161. Xie, Y. 2018. DT: An R interface to the DataTables library.
162. Yamashita, M., Fenn, J. B., 1984, Electrospray ion source. Another variation on the free-jet theme, *J. Phys. Chem*. 88, 20, 4451-4459
163. Yang, X., Di Liu, Fei Liu, Jun Wu, Jing Zou, Xue Xiao, Fangqing Zhao, Baoli Zhu. 2013. HTQC: a fast quality control toolkit for Illumina sequencing data, BMC Bioinforamtics, 14, 33
164. Zhang, B., N.C. VerBerkmoes, M.A. Langston, E. Uberbacher, R.L. Hettich, N.F. Samatova. 2006. Detecting differential and correlated protein expression in label-free shotgun proteomics. *J Proteome Res* 5(11) 2909-2918.
165. Zhang, J., J. Baran, A. Cros, J.M. Guberman, S. Haider, J. Hsu, Y. Liang, E. Rivkin, J. Wang, B. Whitty, M. Wong-Erasmus, L. Yao, A. Kasprzyk. 2011. International Cancer Genome Consortium Data Portal--a one-stop shop for cancer genomics data. *Database (Oxford)* 2011 bar026.
166. Zhang, W., B.J. Wagner, K. Ehrenman, A.W. Schaefer, C.T. DeMaria, D. Crater, K. DeHaven, L. Long, G. Brewer. 1993. Purification, characterization, and cDNA cloning of an AU-rich element RNA-binding protein, AUF1. *Mol Cell Biol* 13(12) 7652-7665.
167. Zhao, H., Y. Yang, Y. Zhou. 2011. Structure-based prediction of RNA-binding domains and RNA-binding sites and application to structural genomics targets. *Nucleic Acids Res* 39(8) 3017-3025.
168. Zhao, J., L. Hyman, C. Moore. 1999. Formation of mRNA 3' ends in eukaryotes: mechanism, regulation, and interrelationships with other steps in mRNA synthesis. *Microbiol Mol Biol Rev* 63(2) 405-445.
169. Zheng, B., S. Jeong, Y. Zhu, L. Chen, Q. Xia. 2017. miRNA and lncRNA as biomarkers in cholangiocarcinoma(CCA) *Oncotarget*, 100819-100830.
170. Zhou, J., W. Liu, S. Peng, X. Sun, I. Frazer. 1999. Papillomavirus Capsid Protein Expression Level Depends on the Match between Codon Usage and tRNA Availability. Journal of Virology 73.